IN PARTNERSHIP WITH:
CNRS

Université de Montpellier

Activity Report 2018

Project-Team ZENITH

Scientific Data Management

IN COLLABORATION WITH: Laboratoire d'informatique, de robotique et de microélectronique de Montpellier (LIRMM)

RESEARCH CENTER
Sophia Antipolis - Méditerranée

THEME
Data and Knowledge Representation and Processing

Table of contents

Project-Team ZENITH

Creation of the Team: 2011 January 01, updated into Project-Team: 2012 January 01

Keywords:

Computer Science and Digital Science:

A1. - Architectures, systems and networks
A1.1. - Architectures
A1.3. - Distributed Systems
A1.3.4. - Peer to peer
A1.3.5. - Cloud
A3.1. - Data
A3.3. - Data and knowledge analysis
A3.5. - Social networks
A3.5.2. - Recommendation systems
A4. - Security and privacy
A4.8. - Privacy-enhancing technologies
A5.4.3. - Content retrieval
A5.7. - Audio modeling and processing

Other Research Topics and Application Domains:

B1. - Life sciences
B1.1. - Biology
B1.1.7. - Bioinformatics
B6. - IT and telecom
B6.5. - Information systems

1. Team, Visitors, External Collaborators

Research Scientists

Patrick Valduriez [Team leader, Inria, Senior Researcher, HDR]
Reza Akbarinia [Inria, Researcher]
Alexis Joly [Inria, Researcher]
Antoine Liutkus [Inria, Researcher]
Florent Masseglia [Inria, Senior Researcher, HDR]
Didier Parigot [Inria, Researcher, HDR]
Christophe Pradal [CIRAD, Researcher]
Hervé Goëau [CIRAD, Researcher]

Faculty Members

Esther Pacitti [Univ of Montpellier, Professor, HDR]
Michel Riveill [Univ of Nice - Sophia Antipolis, Professor, HDR]
Dennis Shasha [NYU, Inria Int. Chair]

PhD Students

Christophe Botella [INRA]
Gaetan Heidsieck [Inria]
Titouan Lorieul [Univ of Montpellier]

Sakina Mahboubi [Inria, until Nov 2018]
Khadidja Meguelati [Averroes fellowship, Algeria]
Djamel-Edine Yagoubi [Inria, until Feb 2018]

Technical staff

Jean-Christophe Lombardo [Inria, Engineer]
Antoine Affouard [Inria, from Jul 2018]
Boyan Kolev [Inria, granted by H2020 ClouddbAppliance project]
Oleksandra Levchenko [Inria]
Valentin Leveau [Inria, until May 2018, granted by Agropolis Fondation]
Fabian-Robert Stoter [Inria]

Intern

Benjamin Deneu [Inria, from Mar 2018 until Sep 2018]

Administrative Assistant

Nathalie Brillouet [Inria, from Apr 2018]

Visiting Scientists

Vitor Silva [UFRJ, Brazil, until Jan 2018]
Mehdi Zitouni [Univ of Tunis, until Mar 2018]

2. Overall Objectives

2.1. Overall Objectives

Data-intensive science such as agronomy, astronomy, biology and environmental science must deal with overwhelming amounts of experimental data produced through empirical observation and simulation. Such data must be processed (cleaned, transformed, analyzed) in all kinds of ways in order to draw new conclusions, prove scientific theories and produce knowledge. Similarly, digital humanities are faced with the problem of exploiting vast amounts of digitized cultural and historical data, such as broadcasted radio or TV content over decades. However, constant progress in scientific observational instruments (e.g. satellites, sensors, large hadron collider), simulation tools (that foster in silico experimentation) or digitization of new content by archivists create a huge data overload. For example, climate modeling data are growing so fast that they will lead to collections of hundreds of exabytes by 2020.

Scientific data is very complex, in particular because of heterogeneous methods used for producing data, the uncertainty of captured data, the inherently multi-scale nature (spatial scale, temporal scale) of many sciences and the growing use of imaging (e.g. molecular imaging), resulting in data with hundreds of attributes, dimensions or descriptors. Modern science research is also highly collaborative, involving scientists from different disciplines (e.g. biologists, soil scientists, and geologists working on an environmental project), in some cases from different organizations in different countries. Each discipline or organization tends to produce and manage its own data, in specific formats, with its own processes. Thus, integrating such distributed data gets difficult as the amounts of heterogeneous data grow.

Despite their variety, we can identify common features of scientific data: big data; manipulated through complex, distributed workflows; typically complex, e.g. multidimensional or graph-based; with uncertainty in the data values, e.g., to reflect data capture or observation; important metadata about experiments and their provenance; and mostly append-only (with rare updates).

Relational DBMSs, which have proved effective in many application domains (e.g. business transactions, business intelligence), are not efficient at dealing with scientific data or big data, which is typically unstructured. In particular, they have been criticized for their “one size fits all” approach. As an alternative , more specialized solutions are being developped such as NoSQL/NewSQL DBMSs and data processing frameworks (e.g. Spark) on top of distributed file systems (e.g. HDFS).

The three main challenges of scientific data management can be summarized by: (1) scale (big data, big applications); (2) complexity (uncertain, multi-scale data with lots of dimensions), (3) heterogeneity (in particular, data semantics heterogeneity). These challenges are also those of data science, with the goal of making sense out of data by combining data management, machine learning, statistics and other disciplines. The overall goal of Zenith is to address these challenges, by proposing innovative solutions with significant advantages in terms of scalability, functionality, ease of use, and performance. To produce generic results, these solutions are in terms of architectures, models and algorithms that can be implemented in terms of components or services in specific computing environments, e.g. cloud. We design and validate our solutions by working closely with our scientific application partners such as CIRAD, INRA and IRD in France, or the National Research Institute on e-medicine (MACC) in Brazil. To further validate our solutions and extend the scope of our results, we also foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

Our approach is to capitalize on the principles of distributed and parallel data management. In particular, we exploit: high-level languages as the basis for data independence and automatic optimization; data semantics to improve information retrieval and automate data integration; declarative languages to manipulate data and workflows; and highly distributed and parallel environments such as P2P, cluster and cloud. We also exploit machine learning, probabilities and statistics for high-dimensional data processing, data analytics and data search. To reflect our approach, we organize our research program in five complementary themes:

- Data integration, including data capture and cleaning;
- Data management, in particular, indexing and privacy;
- Scientific workflows, in particular, in grid and cloud;
- Data analytics, including data mining and statistics;
- Machine learning for high-dimensional data processing and search.

3. Research Program

3.1. Distributed Data Management

Data management is concerned with the storage, organization, retrieval and manipulation of data of all kinds, from small and simple to very large and complex. It has become a major domain of computer science, with a large international research community and a strong industry. Continuous technology transfer from research to industry has led to the development of powerful DBMS, now at the heart of any information system, and of advanced data management capabilities in many kinds of software products (search engines, application servers, document systems, etc.).

To deal with the massive scale of scientific data, we exploit large-scale distributed systems, with the objective of making distribution transparent to the users and applications. Thus, we capitalize on the principles of large-scale distributed systems such as clusters, peer-to-peer (P2P) and cloud.

Data management in distributed systems has been traditionally achieved by distributed database systems which enable users to transparently access and update several databases in a network using a high-level query language (e.g. SQL). Transparency is achieved through a global schema which hides the local databases' heterogeneity. In its simplest form, a distributed database system is a centralized server that supports a global schema and implements distributed database techniques (query processing, transaction management, consistency management, etc.). This approach has proved to be effective for applications that can benefit from centralized control and full-fledge database capabilities, e.g. information systems. However, it cannot scale up to more than tens of databases.

Parallel database systems extend the distributed database approach to improve performance (transaction throughput or query response time) by exploiting database partitioning using a multiprocessor or cluster system. Although data integration systems and parallel database systems can scale up to hundreds of data sources or database partitions, they still rely on a centralized global schema and strong assumptions about the network.

In contrast, peer-to-peer (P2P) systems adopt a completely decentralized approach to data sharing. By distributing data storage and processing across autonomous peers in the network, they can scale without the need for powerful servers. P2P systems typically have millions of users sharing petabytes of data over the Internet. Although very useful, these systems are quite simple (e.g. file sharing), support limited functions (e.g. keyword search) and use simple techniques (e.g. resource location by flooding) which have performance problems. A P2P solution is well-suited to support the collaborative nature of scientific applications as it provides scalability, dynamicity, autonomy and decentralized control. Peers can be the participants or organizations involved in collaboration and may share data and applications while keeping full control over their (local) data sources. But for very-large scale scientific data analysis, we believe cloud computing (see next section), is the right approach as it can provide virtually infinite computing, storage and networking resources. However, current cloud architectures are proprietary, ad-hoc, and may deprive users of the control of their own data. Thus, we postulate that a hybrid P2P/cloud architecture is more appropriate for scientific data management, by combining the best of both approaches. In particular, it will enable the clean integration of the users' own computational resources with different clouds.

3.2. Big Data

Big data (like its relative, data science) has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine. It is also a moving target, as shown by two landmarks in DBMS products: the Teradata database machine in the 1980's and the Oracle Exadata database machine in 2010.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors, mobile devices, connected objects (IoT), social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down (e.g. 1 Gigabyte of Hard Disk Drive for: 1M$ in 1982, 1K$ in 1995, 0.02$ in 2015), making it affordable to keep more data around. And massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

The problem of big data has three main dimensions, quoted as the three big V's:

- Volume: refers to massive amounts of data, making it hard to store, manage, and analyze (big analytics);
- Velocity: refers to continuous data streams being produced, making it hard to perform online processing and analysis;
- Variety: refers to different data formats, different semantics, uncertain data, multiscale data, etc., making it hard to integrate and analyze.

There are also other V's such as: validity (is the data correct and accurate?); veracity (are the results meaningful?); volatility (how long do you need to store this data?).

Many different big data management solutions have been designed, primarily for the cloud, as cloud and big data are synergistic. They typically trade consistency for scalability, simplicity and flexibility, hence the new term Data-Intensive Scalable Computing (DISC). Examples of DISC systems include data processing frameworks (e.g. Hadoop MapReduce, Apache Spark, Pregel), file systems (e.g. Google GFS, HDFS), NoSQL systems (Google BigTable, Hbase, MongoDB), NewSQL systems (Google F1, CockroachDB, LeanXcale). In Zenith, we exploit or extend DISC technologies to fit our needs for scientific workflow management and scalable data analysis.

3.3. Data Integration

Scientists can rely on web tools to quickly share their data and/or knowledge. Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). Data integration can be either physical or logical. In the former, the source data are

integrated and materialized in a data warehouse. In logical integration, the integrated data are not materialized, but accessed indirectly through a global (or mediated) schema using a data integration system. These two approaches have different trade-offs, e.g. efficient analytics but only on historical data for data warehousing versus real-time access to data sources for data integration systems (e.g. web price comparators).

In both cases, to understand a data source content, metadata (data that describe the data) is crucial. Metadata can be initially provided by the data publisher to describe the data structure (e.g. schema), data semantics based on ontologies (that provide a formal representation of the domain knowledge) and other useful information about data provenance (publisher, tools, methods, etc.). Scientific metadata is very heterogeneous, in particular because of the autonomy of the underlying data sources, which leads to a large variety of models and formats. Thus, it is necessary to identify semantic correspondences between the metadata of the related data sources. This requires the matching of the heterogeneous metadata, by discovering semantic correspondences between ontologies, and the annotation of data sources using ontologies. In Zenith, we rely on semantic web techniques (e.g. RDF and SparkQL) to perform these tasks and deal with high numbers of data sources.

Scientific workflow management systems (SWfMS) are also useful for data integration. They allow scientists to describe and execute complex scientific activities, by automating data derivation processes, and supporting various functions such as provenance management, queries, reuse, etc. Some workflow activities may access or produce huge amounts of distributed data. This requires using distributed and parallel execution environments. However, existing workflow management systems have limited support for data parallelism. In Zenith, we use an algebraic approach to describe data-intensive workflows and exploit parallelism.

3.4. Data Analytics

Data analytics refers to a set of techniques to draw conclusions through data examination. It involves data mining, statistics, and data management. Data mining provides methods to discover new and useful patterns from very large datasets. These patterns may take different forms, depending on the end-user's request, such as:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the data mining algorithm extracts correlations between column values. This problem was first motivated by commercial and marketing purposes (*e.g.* discovering frequent correlations between items bought in a shop, which could help selling more). A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset discovery but considering the order between. In the smart building example, a frequent sequence could say that "in 40% of rooms, lights are on at time i, the room is empty at time $i + j$ and the door is closed at time $i + j + k$". Discovering frequent sequences has become critical in marketing, as well as in security (e.g. detecting network intrusions), in web usage analysis and any domain where data come in a specific order, typically given by timestamps.
- **Clustering**. The goal of clustering is to group together similar data while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we could find clusters of rooms, where offices will be in one category and copy machine rooms in another because of their differences (hours of people presence, number of times lights are turned on/off, etc.).

One main problem in data analytics is to deal with data streams. Existing methods have been designed for very large data sets where complex algorithms from artificial intelligence were not efficient because of data size. However, we now must deal with data streams, sequences of data events arriving at high rate, where traditional data analytics techniques cannot complete in real-time, given the infinite data size. In order to extract knowledge from data streams, the data mining community has investigated approximation methods that could yield good result quality.

3.5. High dimensional data processing and search

High dimensionality is inherent in applications involving images, audio and text as well as in many scientific applications involving raster data or high-throughput data. Because of the *dimensionality curse*, technologies for processing and analyzing such data cannot rely on traditional relational DBMS or data mining methods. It rather requires to employ machine learning methods such as dimensionality reduction, representation learning or random projection. The activity of Zenith in this domain focuses on methods that permit data processing and search at scale, in particular in the presence of strong uncertainty and/or ambiguity. Actually, while small datasets are often characterized by a careful collection process, massive amounts of data often come with outliers and spurrious items, because it appears impossible to guarantee faultless collection at massive bandwidth. Another source of noise is often the sensor itself, that may be of low quality but of high sampling rate, or even the actual content, e.g. in cultural heritage applications when historical content appears seriously damaged by time. To attack these difficult problems, we focus on the following research topics:

- **Uncertainty estimation**. Items in massive datasets may either be uncertain, e.g. for automatically annotated data as in image analysis, or be more or less severely corrupted by noise, e.g. in noisy audio recordings or in the presence of faulty sensors. In both cases, the concept of *uncertainty* is central for the end-user to exploit the content and one core activity of Zenith in this context is the use of probability theory to quantify uncertainty and to propose machine learning algorithms that may operate robustly, or at least assess the quality of their output. This vast topic of research is guided by large-scale applications (both data search and data denoising), and research is hence oriented towards computationally effective methods.
- **Deep neural networks**. A major breakthrough in machine learning performance has been witnessed in the last ten years through the advent of deep neural nets. These models are characterized by a huge amount of parameters, that routinely reach dozens of millions, and by scalable learning procedures. Researchers in Zenith are striving towards proposing original architectures and methods that are theoretically grounded and offer state-of-the-art performance for data search and data processing. The specific challenges we investigate are: very high dimensionality for static data and very long-term dependency for temporal data, both in the case of possibly strong uncertainty or ambiguity (e.g. hundreds of thousands of classes).
- **Community service**. Research in machine learning is guided by applications. In Zenith, two core communities are targetted, which are: botany, and digital humanities. In both cases, the key observation done by Zenith is that significant breakthroughs may be achieved by connecting these communities to machine learning researchers. This may be achieved through wording application-specific problems in classical machine learning parlance. The team is active at the international level in organizing popular evaluation campaigns that allow machine learning researchers to propose new methods while solving important applicative problems. This activity has two distinct aspects: managing datasets, and offering tools to ease interoperability.

4. Application Domains

4.1. Data-intensive Scientific Applications

The application domains covered by Zenith are very wide and diverse, as they concern data-intensive scientific applications, i.e., most scientific applications. Since the interaction with scientists is crucial to identify and tackle data management problems, we are dealing primarily with application domains for which Montpellier has an excellent track record, i.e., agronomy, environmental science, life science, with scientific partners like INRA, IRD and CIRAD. However, we are also addressing other scientific domains (e.g. astronomy, oil extraction, music processing) through our international collaborations (e.g. in Brazil or the USA).

Let us briefly illustrate some representative examples of scientific applications on which we have been working on.

- **Management of astronomical catalogs**. An example of data-intensive scientific applications is the management of astronomical catalogs generated by the Dark Energy Survey (DES) project on which we are collaborating with researchers from Brazil. In this project, huge tables with billions of tuples and hundreds of attributes (corresponding to dimensions, mainly double precision real numbers) store the collected sky data. Data are appended to the catalog database as new observations are performed and the resulting database size is estimated to reach 100TB very soon. Scientists around the globe can query the database with queries that may contain a considerable number of attributes. The volume of data that this application holds poses important challenges for data management. In particular, efficient solutions are needed to partition and distribute the data in several servers. An efficient partitioning scheme should try to minimize the number of fragments accessed in the execution of a query, thus reducing the overhead associated to handle the distributed execution.
- **Personal health data analysis and privacy** Today, it is possible to acquire data on many domains related to personal data. For instance, one can collect data on her daily activities, habits or health. It is also possible to measure performance in sports. This can be done thanks to sensors, communicating devices or even connected glasses. Such data, once acquired, can lead to valuable knowledge for these domains. For people having a specific disease, it might be important to know if they belong to a specific category that needs particular care. For an individual, it can be interesting to find a category that corresponds to her performances in a specific sport and then adapt her training with an adequate program. Meanwhile, for privacy reasons, people will be reluctant to share their personal data and make them public. Therefore, it is important to provide them with solutions that can extract such knowledge from everybody's data, while guaranteeing that their private data won't be disclosed to anyone.
- **Botanical data sharing**. Botanical data is highly decentralized and heterogeneous. Each actor has its own expertise domain, hosts its own data, and describes them in a specific format. Furthermore, botanical data is complex. A single plant's observation might include many structured and unstructured tags, several images of different organs, some empirical measurements and a few other contextual data (time, location, author, etc.). A noticeable consequence is that simply identifying plant species is often a very difficult task; even for the botanists themselves (the so-called taxonomic gap). Botanical data sharing should thus speed up the integration of raw observation data, while providing users an easy and efficient access to integrated data. This requires to deal with social-based data integration and sharing, massive data analysis and scalable content-based information retrieval. We address this application in the context of the French initiative Pl@ntNet, with CIRAD and IRD.
- **Biological data integration and analysis**.

  Biology and its applications, from medicine to agronomy and ecology, are now producing massive data, which is revolutionizing the way life scientists work. For instance, using plant phenotyping platforms such as PhenoDyn and PhenoArch at INRA Montpellier, quantitative genetic methods allow to identify genes involved in phenotypic variation in response to environmental conditions. These methods produce large amounts of data at different time intervals (minutes to months), at different sites and at different scales ranging from small tissue samples to the entire plant until whole plant population. Analyzing such big data creates new challenges for data management and data integration.
- **Audio heritage preservation**.

  Since the end of the 19^{th} century, France has commissioned ethnologists to record the world's immaterial audio heritage. This results in datasets of dozens of thousands of audio recordings from all countries and more than 1200 ethnies. Today, this data is gathered under the name of **Archives du CNRS - Musée de l'Homme** and is handled by the CREM (Centre de Recherche en Ethno-Musicologie). Profesional scientists in digital humanities are accessing this data daily for their investigations, and several important challenges arise to ease their work. The KAMoulox project,

lead by A. Liutkus, targets at offering online processing tools for the scientists to automatically restore this old material on demand.

These application examples illustrate the diversity of requirements and issues which we are addressing with our scientific application partners. To further validate our solutions and extend the scope of our results, we also want to foster industrial collaborations, even in non scientific applications, provided that they exhibit similar challenges.

5. Highlights of the Year

5.1. Highlights of the Year

5.1.1. VLDB Conference

The VLDB conference (http://vldb2018.lncc.br) was in Rio de Janeiro. Its organization is a major outcome of the SciDISC associate team, with key positions held by members of the project: F. Porto: general chair, P. Valduriez: sponsor chair and many SciDISC members in the local organization. E. Ogasawara and P. Valduriez were chairs of the LADaS VLDB workshop. E. Pacitti was chair of the VLDB workshop on Big Social Data and Urban Computing (BiDU). The VLDB conference was a great success with about 700 participants.

5.1.2. New Book

A. Joly co-authored the book "Multimedia Tools and Applications for Environmental & Biodiversity Informatics" [69], which demonstrates how the latest advancements in data science impact the wide range of environmental and biodiversity studies.

6. New Software and Platforms

6.1. Pl@ntNet

KEYWORDS: Plant identification - Deep learning - Citizen science
FUNCTIONAL DESCRIPTION: Pl@ntNet is a participatory platform and information system dedicated to the production of botanical data through deep learning-based plant identification. It includes 3 main front-ends, an Android app (the most advanced and the most used one), an iOs app (being currently re-developed) and a web version. The main feature of the application is to return the ranked list of the most likely species providing an image or an image set of an individual plant. In addition, Pl@ntNet's search engine returns the images of the dataset that are the most similar to the queried observation allowing interactive validation by the users. The back-office running on the server side of the platform is based on Snoop visual search engine (a software developed by ZENITH) and on NewSQL technologies for the data management. The application is distributed in more than 180 countries (10M downloads) and allows identifying about 20K plant species at present time.

- Participants: Antoine Affouard, Jean-Christophe Lombardo, Pierre Bonnet, Hervé Goëau, Mathias Chouet and Julien Champ
- Contact: Alexis Joly
- Publication: Pl@ntNet app in the era of deep learning

6.2. ThePlantGame

KEYWORD: Crowd-sourcing

FUNCTIONAL DESCRIPTION: ThePlantGame is a participatory game whose purpose is the production of big taxonomic data to improve our knowledge of biodiversity. One major contribution is the active training of the users based on innovative sub-task creation and assignment processes that are adaptive to the increasing skills of the user. Thousands of players are registered and produce on average about tens new validated plant observations per day. The accuracy of the produced taxonnomic tags is very high (about 95

- Participants: Maximilien Servajean and Alexis Joly
- Contact: Alexis Joly
- Publication: Crowdsourcing Thousands of Specialized Labels: A Bayesian Active Training Approach

6.3. Snoop

KEYWORDS: Content-based Image Retrieval - Deep learning
FUNCTIONAL DESCRIPTION: Snoop is a C++ framework dedicated to large-scale content-based image retrieval. Its main features are (i) the extraction and efficient indexing of visual features (hand-crafted or learned through deep learning), (ii) the search of similar images through approximate k-nearest neighbors and (iii), the supervised recognition of trained visual concepts. The framework can be used either as a set of C++ libraries or as a set of web services through a RESTFUL API. Snoop is the visual search engine used by the Pl@ntNet applications (very large audience).

- Participants: Alexis Joly, Jean-Christophe Lombardo and Olivier Buisson
- Partner: INA (Institut National de l'Audiovisuel)
- Contact: Alexis Joly
- Publication: Random Maximum Margin Hashing

6.4. Chiaroscuro

KEYWORDS: Privacy - P2P - Data mining
FUNCTIONAL DESCRIPTION: Chiaroscuro is a complete solution for clustering personal data with strong privacy guarantees. The execution sequence produced by Chiaroscuro is massively distributed on personal devices, coping with arbitrary connections and disconnections. Chiaroscuro builds on our novel data structure, called Diptych, which allows the participating devices to collaborate privately by combining encryption with differential privacy. Our solution yields a high clustering quality while minimizing the impact of the differentially private perturbation.

- Participants: Tristan Allard, Georges Hebrail, Florent Masseglia and Esther Pacitti
- Contact: Florent Masseglia
- Publication: Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering

6.5. DfAnalyzer

Dataflow Analysis
KEYWORDS: Data management - Monitoring - Runtime Analysis
FUNCTIONAL DESCRIPTION: DfAnalyzer is a tool for monitoring, debugging, steering, and analysis of dataflows while being generated by scientific applications. It works by capturing strategic domain data, registering provenance and execution data to enable queries at runtime. DfAnalyzer provides lightweight dataflow monitoring components to be invoked by high performance applications. It can be plugged in scripts, or Spark applications, in the same way users already plug visualization library components.

- Participants: Vítor Sousa Silva, Daniel De Oliveira, Marta Mattoso and Patrick Valduriez
- Partners: COPPE/UFRJ - Uff
- Contact: Patrick Valduriez
- Publication: DfAnalyzer: Runtime Dataflow Analysis of Scientific Applications using Provenance
- URL: https://github.com/vssousa/dfanalyzer-spark

6.6. CloudMdsQL Compiler

KEYWORDS: Optimizing compiler - NoSQL - Data integration
FUNCTIONAL DESCRIPTION: The CloudMdsQL (Cloud Multi-datastore Query Language) polystore transforms queries expressed in a common SQL-like query language into an optimized query execution plan to be executed over multiple cloud data stores (SQL, NoSQL, HDFS, etc.) through a query engine. The compiler/optimizer is implemented in C++ and uses the Boost.Spirit framework for parsing context-free grammars. CloudMdsQL has been validated on relational, document and graph data stores in the context of the CoherentPaaS European project.

- Participants: Boyan Kolev, Oleksandra Levchenko and Patrick Valduriez
- Contact: Patrick Valduriez
- Publication: CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language

6.7. Savime

Simulation And Visualization IN-Memory
KEYWORDS: Data management. - Distributed Data Management
FUNCTIONAL DESCRIPTION: SAVIME is a multi-dimensional array DBMS for scientific applications. It supports a novel data model called TARS (Typed ARray Schema), which extends the basic array data model with typed arrays. In TARS, the support of application dependent data characteristics is provided through the definition of TAR objects, ready to be manipulated by TAR operators. This approach provides much flexibility for capturing internal data layouts through mapping functions, which makes data ingestion independent of how simulation data has been produced, thus minimizing ingestion time.

- Participants: Hermano Lustosa, Fabio Porto and Patrick Valduriez
- Partner: LNCC - Laboratório Nacional de Computação Científica
- Contact: Patrick Valduriez
- Publication: TARS: An Array Model with Rich Semantics for Multidimensional Data

6.8. OpenAlea

KEYWORDS: Bioinformatics - Biology
FUNCTIONAL DESCRIPTION: OpenAlea is an open source project primarily aimed at the plant research community. It is a distributed collaborative effort to develop Python libraries and tools that address the needs of current and future works in Plant Architecture modeling. It includes modules to analyze, visualize and model the functioning and growth of plant architecture. It was formally developed in the Inria VirtualPlants team.

RELEASE FUNCTIONAL DESCRIPTION: OpenAlea 2.0 adds to OpenAlea 1.0 a high-level formalism dedicated to the modeling of morphogenesis that makes it possible to use several modeling paradigms (Blackboard, L-systems, Agents, Branching processes, Cellular Automata) expressed with different languages (Python, L-Py, R, Visual Porgramming, ...) to analyse and simulate shapes and their development.

- Participants: Christian Fournier, Christophe Godin, Christophe Pradal, Frédéric Boudon, Patrick Valduriez, Esther Pacitti and Yann Guedon
- Partners: CIRAD - INRA
- Contact: Christophe Pradal
- Publications: OpenAlea: Scientific Workflows Combining Data Analysis and Simulation - OpenAlea: A visual programming and component-based software platform for plant modeling

6.9. FP-Hadoop

Fast Parallel Hadoop

KEYWORDS: Hadoop - Data parallelism
FUNCTIONAL DESCRIPTION: FP-Hadoop makes the reduce side of Hadoop MapReduce more parallel and efficiently deals with the problem of data skew in the reduce side. In FP-Hadoop, there is a new phase, called intermediate reduce (IR), in which blocks of intermediate values, constructed dynamically, are processed by intermediate reduce workers in parallel. Our experiments using FP-Hadoop using synthetic and real benchmarks have shown excellent performance gains compared to native Hadoop, e.g. more than 10 times in reduce time and 5 times in total execution time.

- Participants: Reza Akbarinia, Miguel Liroz-Gistau and Patrick Valduriez
- Contact: Reza Akbarinia
- Publication: FP-Hadoop: Efficient Execution of Parallel Jobs Over Skewed Data

6.10. Hadoop_g5k

KEYWORD: Cluster
FUNCTIONAL DESCRIPTION: Hadoop_g5k is a tool that makes it easier to manage Hadoop and Spark clusters and prepare reproducible experiments in the Grid 5000 platform. Hadoop_g5k offers a set of scripts to be used in command-line interfaces and a Python API to interact with the clusters. It is currently active within the G5k community, facilitating the preparation and execution of experiments in the platform.

- Participants: Reza Akbarinia, Miguel Liroz-Gistau and Patrick Valduriez
- Contact: Reza Akbarinia
- URL: https://www.grid5000.fr/mediawiki/index.php/Hadoop_On_Execo

6.11. Triton Server

End-to-end Graph Mapper
KEYWORD: Web Application
FUNCTIONAL DESCRIPTION: A server for managing graph data and applications for mobile social networks. The server is built on top of the OrientDB graph database system and a distributed middleware. It provides an End-to-end Graph Mapper (EGM) for modeling the whole application as (i) a set of graphs representing the business data, the in-memory data structure maintained by the application and the user interface (tree of graphical components), and (ii) a set of standardized mapping operators that maps these graphs with each other.

- Participants: Didier Parigot, Patrick Valduriez and Benjamin Billet
- Contact: Didier Parigot
- Publication: End-to-end Graph Mapper

6.12. SON

Shared-data Overlay Network
KEYWORDS: Sharing - Ibuted exchange - Peer-to-peer.
FUNCTIONAL DESCRIPTION: SON is a development platform for P2P networks using web services, JXTA and OSGi. The development of a SON application is done through the design and implementation of a set of components. Each component includes a technical code that provides the component services and a code component that provides the component logic (in Java). The complex aspects of asynchronous distributed programming are separated from code components and automatically generated from an abstract description of services for each component by the component generator.

- Participants: Didier Parigot, Esther Pacitti and Patrick Valduriez
- Contact: Didier Parigot
- Publication: A Lightweight Middleware for developing P2P Applications with Component and Service-Based Principles
- URL: http://www-sop.inria.fr/members/Didier.Parigot/pmwiki/SON/index.php

6.13. SciFloware

Scientific Workflow Middleware
KEYWORDS: Bioinformatics - Distributed Data Management
FUNCTIONAL DESCRIPTION: SciFloware is a middleware for the execution of scientific workflows in a distributed and parallel way. It capitalizes on our experience with the Shared-Data Overlay Network and an innovative algebraic approach to the management of scientific workflows. SciFloware provides a development environment and a runtime environment for scientific workflows, interoperable with existing systems. We validate SciFloware with workflows for analyzing biological data provided by our partners CIRAD, INRA and IRD.

- Participants: Didier Parigot, Dimitri Dupuis and Patrick Valduriez
- Contact: Didier Parigot
- Publication: InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid
- URL: http://www-sop.inria.fr/members/Didier.Parigot/pmwiki/Scifloware

6.14. WebSmatch

Web Schema Matching
KEYWORD: Data integration
FUNCTIONAL DESCRIPTION: WebSmatch is a flexible, open environment for discovering and matching complex schemas from heterogeneous Web data sources. It provides three basic functions: (1) metadata extraction from data sources, (2) schema matching, (3) schema clustering to group similar schemas together. WebSmatch is delivered through Web services, to be used directly by data integrators or other tools with RIA clients. It is implemented in Java, delivered as Open Source Software (under LGPL). WebSmatch has been used by Data Publica and CIRAD to integrate public and private data sources.

- Participants: Emmanuel Castanier, Patrick Valduriez and Rémi Coletta
- Contact: Patrick Valduriez
- Publication: WebSmatch: a tool for Open Data
- URL: http://websmatch.gforge.inria.fr/

7. New Results

7.1. Query Processing

7.1.1. Top-k Query Processing Over Encrypted Data in the Cloud

Participants: Sakina Mahboubi, Reza Akbarinia, Patrick Valduriez.

Cloud computing provides users and companies with powerful capabilities to store and process their data in third-party data centers. However, the privacy of the outsourced data is not guaranteed by the cloud providers. One solution for protecting the user data against security attacks is to encrypt the data before being sent to the cloud servers. Then, the main problem is to evaluate user queries over the encrypted data.

In this work, we address the problem of top-k query processing over encrypted data. Top-k queries are important for many applications such as information retrieval, spatial data analysis, temporal databases, graph databases, etc. We consider two cases for top-k query processing over encrypted data: 1) centralized: the encrypted data are stored at a single node of a data center, which is useful if the database can fit at one node; 2) distributed: the encrypted data are partitioned and the partitions are encrypted and distributed across multiple nodes, which is useful if the database is very big.

In [52], we address the distributed case, and propose a system, called SD-TOPK, for top-k query processing over encrypted data distributed across several nodes of the cloud. SD-TOPK comes with a distributed top-k query processing algorithm that is executed in the nodes, and finds a set including the encrypted top-k data items. It also has an efficient filtering algorithm that removes most of the false positives included in the set returned by the top-k query processing algorithm. This filtering is done without needing to decrypt the data in the cloud.

In [51], we propose a complete system, called *BuckTop*, for the centralized case. BuckTop is able to efficiently evaluate top-k queries over encrypted data outsourced to a singe node, without having to decrypt it in that node. It includes a top-k query processing algorithm that works on the encrypted data stored in the cloud node, and returns a set that is proved to contain the encrypted data corresponding to the top-k results. We implemented BuckTop and compared its performance for processing top-k queries over encrypted data with that of the popular threshold algorithm (TA) over original (plaintext) data. The results show the effectiveness of BuckTop for outsourcing sensitive data in the cloud and answering top-k queries.

7.1.2. Privacy Preserving Index for Range Query Processing in the Clouds

Participants: Reza Akbarinia, Esther Pacitti.

During the last decade, a large body of academic work has tackled the problem of outsourcing databases to an untrusted cloud while maintaining both privacy and SQL-like querying functionality (at least partially). Range query is an important kind of query that expresses a bounded restriction over the retrieved records. In the database management systems, these queries are usually answered by using efficient indexes. However, developing privacy preserving indexes for untrusted environments is very challenging.

In [55], we propose a differentially private index to an outsourced encrypted dataset. Efficiency is enabled by using a plaintext index structure to perform range queries. Security relies on both differential privacy (of the index) and semantic security (of the encrypted dataset). Our solution, called PINED-RQ, develops algorithms for building and updating the differentially private index. Compared to state-of-the-art secure index based range query processing approaches, PINED-RQ executes queries in the order of at least one magnitude faster. The security of PINED-RQ is proved and its efficiency is assessed by an extensive experimental validation.

7.1.3. Constellation Queries to Analyze Geometrical Patterns

Participants: Dennis Shasha, Patrick Valduriez.

Constellation queries are useful to analyze geometrical patterns. A geometrical pattern is a set of points with all pairwise distances (or, more generally, relative distances) specified. Finding matches to such patterns, i.e. constellations, has applications to spatial data in seismic, astronomical, and transportation contexts. Finding geometric patterns is a challenging problem as the potential number of sets of elements that compose shapes is exponentially large in the size of the dataset and the pattern. In [53], we propose algorithms to find patterns in large data applications using constellation queries. Our methods combine quadtrees, matrix multiplication, and bucket join processing. Our distributed experiments show that the choice of the composition algorithm (matrix multiplication or nested loops) depends on the freedom introduced in the query geometry through the distance additive factor. Three clearly identified blocks of threshold values guide the choice of the best composition algorithm. Answering complex constellation queries, i.e. isotropic and non-isotropic queries, is challenging because scale factors and stretch factors may take any of an infinite number of values. In [53], we propose practically efficient sequential and distributed algorithms for pure, isotropic, and non-isotropic constellation queries. As far as we know, this is the first work to address isotropic and non-isotropic queries.

7.1.4. Parallel Polyglot Query Processing

Participants: Boyan Kolev, Oleksandra Levchenko, Esther Pacitti, Patrick Valduriez.

The blooming of different cloud data stores has turned polystore systems to a major topic in the nowadays cloud landscape. Especially, as the amount of processed data grows rapidly each year, much attention is being paid on taking advantage of the parallel processing capabilities of the underlying data stores. To provide data federation, a typical polystore solution defines a common data model and query language with translations to API calls or queries to each data store. However, this may lead to losing important querying capabilities. The polyglot approach of the CloudMdsQL query language allows data store native queries to be expressed as inline scripts and combined with regular SQL statements in ad-hoc integration queries. Moreover, efficient optimization techniques, such as bind join, can still take place to improve the performance of selective joins. In [47], we introduce the distributed architecture of the LeanXcale query engine that processes polyglot queries in the CloudMdsQL query language, yet allowing native scripts to be handled in parallel at data store shards, so that efficient and scalable parallel joins take place at the query engine level. The experimental evaluation of the LeanXcale parallel query engine on various join queries illustrates well the performance benefits of exploiting the parallelism of the underlying data management technologies in combination with the high expressivity provided by their scripting/querying frameworks

7.2. Scientific Workflows

7.2.1. In Situ Analysis of Simulation Data

Participants: Vitor Silva, Patrick Valduriez.

In situ analysis and visualization have been used successfully in large-scale computational simulations to visualize scientific data of interest, while data is in memory. Such data are obtained from intermediate (or final) simulation results, and once analyzed are typically stored in raw data files. However, existing in situ data analysis and visualization solutions (e.g. ParaView/Catalyst, VisIt) have limited online query processing and no support for dataflow analysis. The latter is a challenge for exploratory raw data analysis. In the context of the SciDISC associate team with Brazil [38], we propose a solution that integrates dataflow analysis with ParaView Catalyst for performing in-situ data analysis and monitoring dataflow from simulation runs [25].

In [21], we propose a solution (architecture and algorithms), called Armful, to combine the advantages of a dataflow-aware SWMS and raw data file analysis techniques to allow for queries on raw data file elements that are related but reside in separate files. Its main components are a raw data extractor, a provenance gatherer and a query processing interface, which are all dataflow-aware.

An instantiation of Armful is DfAnalyzer [34], a library of components to support online in-situ and in-transit data analysis. DfAnalyzer components are plugged directly in the simulation code of highly optimized parallel applications with negligible overhead. With support of sophisticated online data analysis, scientists get a detailed view of the execution, providing insights to determine when and how to tune parameters or reduce data that does not need to be processed [35]. The source code of the DfAnalyzer implementation for Spark is available on github (github.com/hpcdb/RFA-Spark).

7.2.2. Scheduling of Scientific Workflows in Multisite Cloud

Participants: Esther Pacitti, Patrick Valduriez.

In [30], we consider the problem of efficient scheduling of a large SWf in a multisite cloud, i.e. a cloud with geo-distributed cloud data centers (sites). The reasons for using multiple cloud sites to run a SWf are that data is already distributed , the necessary resources exceed the limits at a single site, or the monetary cost is lower. In a multisite cloud, metadata management has a critical impact on the efficiency of SWf scheduling as it provides a global view of data location and enables task tracking during execution. Thus, it should be readily available to the system at any given time. While it has been shown that efficient metadata handling plays a key role in performance, little research has targeted this issue in multisite cloud. Then we propose to identify and exploit hot metadata (frequently accessed metadata) for efficient SWf scheduling in a multisite cloud, using a distributed approach. We implemented our approach within a scientific workflow management system, which shows that our approach reduces the execution time of highly parallel jobs up to 64% and that of the whole SWfs up to 55%.

7.2.3. Distributed Management of Scientific Workflows for Plant Phenotyping

Participants: Gaetan Heidsieck, Christophe Pradal, Esther Pacitti, Patrick Valduriez.

In the last decade, high-throughput phenotyping platforms have allowed acquisition of quantitative data on thousands of plants required for genetic analyses in well-controlled environmental conditions.The seven facilities of Phenome produce 200 terabytes of data annually, which are heterogeneous (images, time courses), multiscale (from the organ to the field) and originate from different sites. Hence, the major problem becomes the automatic analysis of these massive datasets and the ability to reproduce large and complex in-silico experiments.

In [31], we propose a solution (infrastructure) to distribute the computation of scientific workflows on very large grid computing facilities (EGI/France Grilles) to the 3D reconstruction, segmentation and tracking of plant organs. This infrastructure, InfraPhenoGrid, is based on OpenAlea, SciFloware and SON, a set of software and technology developed in the team. We have used this solution in [27] to dissect the genetic and environmental influence of biomass accumulation in complex multi-genotype maize canopies.

7.3. Data Analytics

7.3.1. Massively Distributed Indexing of Time Series

Participants: Djamel-Edine Yagoubi, Reza Akbarinia, Boyan Kolev, Oleksandra Levchenko, Florent Masseglia, Patrick Valduriez, Dennis Shasha.

Indexing is crucial for many data mining tasks that rely on efficient and effective similarity query processing. Consequently, indexing large volumes of time series, along with high performance similarity query processing, have became topics of high interest. For many applications across diverse domains though, the amount of data to be processed might be intractable for a single machine, making existing centralized indexing solutions inefficient.

In [36], we consider the problem of finding highly correlated pairs of time series across multiple sliding windows. Doing this efficiently and in parallel could help in applications such as sensor fusion, financial trading, or communications network monitoring, to name a few. We have developed a parallel incremental random vector/sketching approach, called ParCorr, to this problem and compared it with the state-of-the-art nearest neighbor method iSAX. Whereas iSAX achieves 100% recall and precision for Euclidean distance, the sketching approach is, empirically, at least 10 times faster and achieves 95% recall and 100% precision on real and simulated data. For many applications this speedup is worth the minor reduction in recall. Our method scales up to 100 million time series and scales linearly in its expensive steps (but quadratic in the less expensive ones).

In [48], we propose a demonstration of our sketch-based solution to efficiently perform both the parallel indexing of large sets of time series and a similarity search on them. Because our method is approximate, we explore the tradeoff between time and precision. A video showing the dynamics of the demonstration can be found at http://parsketch.gforge.inria.fr/video/parSketchdemo_720p.mov.

7.3.2. Parallel Mining of Maximally Informative k-Itemsets in Data Streams

Participants: Mehdi Zitouni, Reza Akbarinia, Florent Masseglia.

The discovery of informative itemsets is a fundamental building block in data analytics and information retrieval. While the problem has been widely studied, only few solutions scale. This is particularly the case when the dataset is massive, or the length k of the informative itemset to be discovered is high.

In [63], we address the problem of mining maximally informative k-itemsets (miki) in data streams based on joint entropy. We propose PentroS, a highly scalable parallel miki mining algorithm. PentroS renders the mining process of large volumes of incoming data very efficient. It is designed to take into account the continuous aspect of data streams, particularly by reducing the computations of need for updating the miki results after arrival/departure of transactions to/from the sliding window. PentroS has been extensively evaluated using massive real-world data streams. Our experimental results confirm the effectiveness of our proposal which allows excellent throughput with high itemset length.

7.3.3. Spatio-Temporal Data Mining

Participants: Esther Pacitti, Florent Masseglia.

The problem of discovering spatiotemporal sequential patterns affects a broad range of applications. Many initiatives find sequences constrained by space and time. We address in [40] an appealing new challenge for this domain: find tight space-time sequences, i.e., find within the same process: i) frequent sequences constrained in space and time that may not be frequent in the entire dataset and ii) the time interval and space range where these sequences are frequent. The discovery of such patterns along with their constraints may lead to extract valuable knowledge that can remain hidden using traditional methods since their support is extremely low over the entire dataset. Our contribution is a new Spatio-Temporal Sequence Miner (STSM) algorithm to discover tight space-time sequences.

7.4. Machine Learning for High-dimensional Data

7.4.1. Uncertainty in Fine-grained Classification

Participants: Titouan Lorieul, Alexis Joly.

Uncertainty is critical when considering classification problems that involve thousands of domain specific labels. A picture of a plant, for instance, contains only a partial information that is usually not sufficient to determine its scientific name with certainty. We first work on the modelling of such uncertainty in the context of crowdsourcing systems involving experts as well as non expert annotators. We rely on Bayesian inference to learn the annotators' confusion and to optimally assign them new items to be validated. In particular, we work on a non-parametric version of this model allowing to combine annotators' suggestions even when the number of possible labels is undetermined and might change over time [33]. In mirror to this research, we also work on the uncertainty of automatic classifiers, in particular deep convolutional neural networks trained on massive amounts of plant images. We conduct an experimental study aimed at evaluating quantitatively the intrinsic data ambiguity of image-based plant observations [64], and we started working on new methods for estimating the uncertainty of ensembles of deep neural networks by fitting a Dirichlet distribution on the set of their predictions. Besides, we study the use of different taxonomic levels as a source of potential reduction in prediction uncertainties [66].

7.4.2. Species Distribution Modelling based on Citizen Science Data

Participants: Christophe Botella, Alexis Joly.

Species distribution models (SDM) are widely used for ecological research and conservation purposes. Given a set of species occurrence, the aim is to infer its spatial distribution over a given territory. Because of the limited number of occurrences of specimens, this is usually achieved through environmental niche modeling approaches, i.e. by predicting the distribution in the geographic space on the basis of a mathematical representation of their known distribution in environmental space (= realized ecological niche). The environment is in most cases represented by climate data (such as temperature, and precipitation), but other variables such as soil type or land cover can also be used. In [24], we study for the first time the relevance of a species distribution model computed from automatically identified plant observations made by citizens rather than from classical inventories made by experts. The results show that the resulting models have a great potential for the early detection of new invasions. In [65] and [60], we propose a deep learning approach to species distribution modelling in order to improve the predictive effectiveness in the context of massive amount of occurrence data. Non-linear prediction models have been of interest for SDM for more than a decade but our study is the first one bringing empirical evidence that deep, convolutional and multilabel models might participate to resolve the limitations of SDM.

7.4.3. Evaluation of Species Identification and Prediction Algorithms

Participants: Alexis Joly, Hervé Goëau, Christophe Botella, Jean-Christophe Lombardo.

We ran a new edition of the LifeCLEF evaluation campaign [45] with the involvement of 13 research teams worldwide. The main novelties and outcomes of the 2018-th edition are the following:

- **GeoLifeCLEF**: a new challenge [71] dedicated to the location-based prediction of species based on spatial occurrences and environmental data tensors. The evaluation concludes that deep environmental convolutional neural networks perform better than spatial models or ponctual environmental models.
- **Man vs. Machine plant identification**: To evaluate how far automated identification systems are from the best possible performance, we organize a challenge involving 19 deep-learning systems implemented by 4 different research teams and 9 of the best expert botanists of the French flora. The main outcome of this work is that the performance of state-of-the-art deep learning models is now very close to the most advanced human expertise.
- **Bird sounds identification**: the 2018-the edition of the BirdCLEF challenge reveals impressive identification performance when considering bird sounds recorded by the Xeno-Canto community. Identifying birds in raw, multi-directional soundscapes, however, remains a very challenging task.

7.4.4. Towards the Recognition of The World's Flora: When HPC Meets Deep Learning

Participants: Hervé Goëau, Jean-Christophe Lombardo, Alexis Joly.

Automated identification of plants and animals have improved considerably in the last few years, in particular thanks to the recent advances in deep learning. In 2017, a challenge on 10,000 plant species (PlantCLEF) resulted in impressive performances with accuracy values reaching 90%. One of the most popular plant identification application, Pl@ntNet, nowadays works on 18K plant species. It accounts for million of users all over the world and already has a strong societal impact in several domains including education, landscape management and agriculture. Now, the big challenge is to train such systems at the scale of the world's biodiversity. Therefore, we built a training set of about 12M images illustrating 300K species of plants. Training a convolutional neural network on such a large dataset can take up to several months on a single node equipped with four recent GPUs. Moreover, to select the best performing architecture and optimize the hyper-parameters, it is often necessary to train several of such networks. Overall, this becomes a highly intensive computational task that has to be distributed on large HPC infrastructures. Therefore, we experiment two french national supercomputers through an access offered by GENCI (Occigen@CINES, a 3.5 Pflop/s Tier-1 cluster based on Broadwell-14cores@2.6Ghz nodes and Joliot-Curie»@TGCC, a BULL-Sequana-X1000 cluster integrating 1656 nodes Intel Skylake8168-24cores@2.7GHz). To implement the synchronized stochastic gradient descent on the CPU cluster Joliot-Curie, we are using the deep learning framework Intel CAFFE coupled with Intel MLSL library (in the context of a collaboration with Intel).

7.4.5. Evaluation of Music Separation Techniques

Participants: Antoine Liutkus, Fabian-Robert Stöter.

After the groundbreaking advent of deep learning, we feel the music processing community needs to step back and think about what had been accomplished and what remains challenging in the problems of musical signal processing and filtering. Therefore, we give a complete overview of the state of the art in music demixing in [32] comprising more than 350 references, as well as two chapters in dedicated books [68], [67]. These references may be considered as complete overviews of the state of the art in music demixing. Furthermore, we introduce the topic to non-expert researchers and engineers in [26].

Apart from this effort in presenting the most recent advances in music processing to the community, we organize yearly a systematic evaluation of state of the art. We report the results of the 2018 Signal Separation Evaluation Campaign in [58], gathering a record number of participants. A perceptual evaluation of the results obtained through this campaign is presented in [59], in collaboration with researchers from the Surrey University.

7.4.6. Robust Probabilistic Models for Time-series

Participants: Antoine Liutkus, Fabian-Robert Stöter.

Processing large amounts of data for denoising or analysis comes with the need to devise models that are robust to outliers and that permit efficient inference. For this purpose, we advocate the use of non-Gaussian models for this purpose, which are less sensitive to data-uncertainty. Most of our effort on this topic is split in two subtasks.

First, we develop new filtering methods that go beyond least-squares estimation. In collaboration with researchers from RWTH, Aachen, Germany, we introduce a new model based on mixtures of Gaussians for filtering in [50]. It combines tractability with a better account of phase consistency for complex data. Along with researchers from IRISA, Rennes and Telecom ParisTech, we also work on filtering α-stable processes [44], [46], [57], which enjoy important applications in robust signal processing.

Second, we work on large amounts of musical archives. This includes an original way to scale up interference reduction in live musical recordings in collaboration with the managers of the Montreux Jazz Festival data at EPFL (Switzerland).

8. Bilateral Contracts and Grants with Industry

8.1. SAFRAN (2018)

Participants: Reza Akbarinia, Florent Masseglia.

SAFRAN and Inria are involved in the DESIR frame-agreement (Florent Masseglia is the scientific contact on "Data Analytics and System Monitoring" topic). In this context, SAFRAN dedicates 80K€ for a joint study of one year on time series indexing. The specific time series to be exploited are those of engine benchmarking with novel characteristics for the team (multiscale and multidimensional).

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. Labex NUMEV, Montpellier

URL: http://www.lirmm.fr/numev

We participate in the Laboratory of Excellence (labex) NUMEV (Digital and Hardware Solutions, Modelling for the Environment and Life Sciences) headed by University of Montpellier in partnership with CNRS, and Inria. NUMEV seeks to harmonize the approaches of hard sciences and life and environmental sciences in order to pave the way for an emerging interdisciplinary group with an international profile. The project is decomposed in four complementary research themes: Modeling, Algorithms and computation, Scientific data (processing, integration, security), Model-Systems and measurements.

9.1.2. Institute of Computational Biology (IBC), Montpellier

URL: http://www.ibc-montpellier.fr

IBC is a 6 year project (2012-2018) with a funding of 2Meuros by the MENRT (PIA program) to develop innovative methods and software to integrate and analyze biological data at large scale in health, agronomy and environment. Patrick Valduriez heads the workpackage on integration of biological data and knowledge.

9.2. National Initiatives

9.2.1. Institut de Convergence Agriculture numérique #DigitAg, (2017-2023), 275Keuro.

Participants: Florent Masseglia, Esther Pacitti, Christophe Pradal, Patrick Valduriez.

#DigitAg brings together in a partnership of seventeen actors (public research and teaching organizations, transfer actors and companies) with the objective of accelerating and supporting the development of agriculture companies in France and in southern countries based on new tools, services and uses. Based in Montpellier with an office in Toulouse and Rennes and led by Irstea, #DigitAg's ambition is to become a world reference for digital agriculture. In this project, Zenith is involved in the analysis of big data from agronomy, in particular, plant phenotyping.

9.2.2. PIA (Projets Investissements d'Avenir) Floris'Tic (2015-2018), 430Keuro.

Participants: Antoine Affouard, Jean-Christophe Lombardo, Hervé Goëau, Alexis Joly.

Floris'tic aims at promoting the scientific and technical culture of plant sciences through innovative pedagogic methods, including participatory initiatives and the use of IT tools such as the one built within the Pl@ntNet project. A. Joly heads the work package on the development of the IT tools. This is a joint project with the AMAP laboratory, the TelaBotanica social network and the Agropolis foundation.

9.2.3. ANR WeedElec (2018-2021), 106 Keuro.

Participants: Jean-Christophe Lombardo, Hervé Goëau, Alexis Joly.

The WeedElec project offers an alternative to global chemical weed control. It combines an aerial means of weed detection by drone coupled to an ECOROBOTIX delta arm robot equipped with a high voltage electrical weeding tool. WeedElec's objective is to remove the major related scientific obstacles, in particular the weed detection/identification, using hyperspectral and colour imaging, and associated chemometric and deep learning techniques.

9.2.4. Others

9.2.4.1. INRA/Inria PhD program, 100Keuros

Participant: Alexis Joly.

This contract between INRA and Inria allows funding a 3-years PhD student (Christophe Botella). The addressed challenge is the large-scale analysis of Pl@ntNet data with the objective to model species distribution (a big data approach to species distribution modeling). The PhD student is supervised by Alexis Joly with François Munoz (ecologist, IRD) and Pascal Monestiez (statistician, INRA).

9.3. European Initiatives

9.3.1. FP7 & H2020 Projects

9.3.1.1. CloudDBAppliance

Participants: Reza Akbarinia, Boyan Kolev, Florent Masseglia, Esther Pacitti, Patrick Valduriez.

Project title: CloudDBAppliance
Instrument: H2020
Duration: 2016 - 2019
Total funding: 5 Meuros (Zenith: 500Keuros)
Coordinator: Bull/Atos, France
Partner: Europe: Inria Zenith, U. Madrid, INESC and the companies LeanXcale, QuartetFS, Nordea, BTO, H3G, IKEA, CloudBiz, and Singular Logic.
Inria contact: Florent Masseglia, Patrick Valduriez

The project aims at producing a European Cloud Database Appliance for providing a Database as a Service able to match the predictable performance, robustness and trustworthiness of on premise architectures such as those based on mainframes. The cloud database appliance features: (i) a scalable operational database able to process high update workloads such as the ones processed by banks or telcos, combined with a fast analytical engine able to answer analytical queries in an online manner; (ii) an operational Hadoop data lake that integrates an operational database with Hadoop, so operational data is stored in Hadoop that will cover the needs from companies on big data; (iii) a cloud hardware appliance leveraging the next generation of hardware to be produced by Bull, the main European hardware provider. This hardware is a scale-up hardware similar to the one of mainframes but with a more modern architecture. Both the operational database and the in-memory analytics engine will be optimized to fully exploit this hardware and deliver predictable performance. Additionally, CloudDBAppliance will tolerate catastrophic cloud data centres failures (e.g. a fire or natural disaster) providing data redundancy across cloud data centres. In this project, Zenith is in charge of designing and implementing the components for analytics and parallel query processing.

9.4. International Initiatives

9.4.1. Inria Associate Teams Not Involved in an Inria International Labs

9.4.1.1. SciDISC

Title: Scientific data analysis using Data-Intensive Scalable Computing

International Partner (Institution - Laboratory - Researcher):

Universidade Federal do Rio de Janeiro (Brazil) - Computer Laboratory - Marta Mattoso

Start year: 2017

See also: https://team.inria.fr/zenith/scidisc/

Data-intensive science requires the integration of two fairly different paradigms: high-performance computing (HPC) and data-intensive scalable computing (DISC). Spurred by the growing need to analyze big scientific data, the convergence between HPC and DISC has been a recent topic of interest [[Coutinho 2014, Valduriez 2015]. This project will address the grand challenge of scientific data analysis using DISC (SciDISC), by developing architectures and methods to combine simulation and data analysis. The expected results of the project are: new data analysis methods for SciDISC systems; the integration of these methods as software libraries in popular DISC systems, such as Apache Spark; and extensive validation on real scientific applications, by working with our scientific partners such as INRA and IRD in France and Petrobras and the National Research Institute (INCT) on e-medicine (MACC) in Brazil.

9.4.2. Inria International Partners

9.4.2.1. Informal International Partners

We have regular scientific relationships with research laboratories in

- North America: Univ. of Waterloo (Tamer Özsu), UCSB Santa Barbara (Divy Agrawal and Amr El Abbadi), Northwestern Univ. (Chicago), university of Florida (Pamela Soltis), Vikram Salatore (Manager of Artificial Intelligence Products Group at Intel Corporation).
- Asia: National Univ. of Singapore (Beng Chin Ooi, Stéphane Bressan), Wonkwang University, Korea (Kwangjin Park), Kyoto University (Japan)
- Europe: Univ. of Madrid (Ricardo Jiménez-Periz), UPC Barcelona (Josep Lluis Larriba Pey), HES-SO (Henning Müller), University of Catania (Concetto Spampinatto), Cork School of Music (Ireland), RWTH (Aachen, Germany), Chemnitz technical university (Stefan Kahl), Berlin Museum für Naturkunde (Mario Lasseck), Stefanos Vrochidis (Greece, ITI)
- Africa: Univ. of Tunis (Sadok Ben-Yahia), IMSP, Bénin (Jules Deliga)
- Australia: Australian National University (Peter Christen)
- Central America: Technologico de Costa-Rica (Erick Mata, former director of the US initiative Encyclopedia of Life)

9.4.3. Participation in Other International Programs

BD-FARM

Title: Big Data Management and Analytics for Agriculture and Farming

International Partner (Institution - Laboratory - Researcher):

Chubu University - International Digital Earth Applied Science Research Center (IDEAS), Kiyoshi Honda

Duration: 2016 - 2018

Start year: 2016

See also: https://team.inria.fr/zenith/bdfarm-2016-2018-stic-asia/

World population is still growing and people are living longer and older. World demand for food rises sharply and current growth rates in agriculture are clearly not sufficient. But extreme flood, drought, typhoon etc, caused by climate change, give severe damages on traditional agriculture. Today, an urgent and deep redesign of agriculture is crucial in order to increase production and to reduce environmental impact. In this context, collecting, managing and analyzing dedicated, large, complex, and various datasets (Big Data) will allow improving the understanding of complex mechanisms behind adaptive, yield and crop improvement. Moreover, sustainability will require detailed studies such as the relationships between genotype, phenotype and environment. In other words, data science and ICT for agriculture must help improving production. Moreover, it has to be done while getting properly adapted to soil, climatic and agronomic constraints as well as taking into account the genetic specificities of plants.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

Several international scientists visited the team and gave seminars

- Vitor Silva (COPPE/UFRJ,Brazil): “A methodology for capturing and analyzing dataflow paths in computational simulations” on January 31.
- Dennis Shasha (NYU): “Reducing Errors by Refusing to Guess (Occasionally)” on June 1.
- Daniel de Oliveira (UFF, Brazil): “Parameter and Data Recommendation in Scientific Workflows based on Provenance” on June 5.
- Eduardo Ogasawara, (CEFET-RJ, Brazil): “Comparing Motif Discovery Techniques with Sequence Mining in the Context of Space-Time Series” on November 26.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. General Chair, Scientific Chair

- P. Valduriez: general chair of the VLDB Latin America Data Science Workshop (LaDAS@VLDB 2018)
- P. Valduriez: scientific chair, First Data Science School, IMSP, Django, Bénin

10.1.1.2. Member of the Organizing Committees

- A. Joly: organizing committee of the international conference CLEF 2018 and the chair of the LifeCLEF track, Avignon, sept. 2018 (http://clef2018.clef-initiative.eu/)

- A. Joly: organizing committee of the Floris'tic national workshop held in Montpellier, nov. 2018 (http://floristic.org/journeefloristic/)
- A. Liutkus: organizer of the 2018 Signal Separation Evaluation Campaign (https://sisec18.unmix.app/)
- F. Masseglia: finance chair of IEEE ICDE 2018 (https://icde2018.org)
- F. Masseglia: organization commitee of the Inria Science Days 2018 (https://www.inria.fr/en/news/news-from-inria/inria-science-days-2018)
- P. Valduriez: sponsor co-chair of IEEE ICDE 2018 (https://icde2018.org)
- P. Valduriez: sponsor co-chair of VLDB 2018 (https://vldb2018.lncc.br)

10.1.2. Scientific Events Selection

10.1.2.1. Chair of Conference Program Committees

- A. Joly: area chair of ACM Multimedia 2018 (http://www.acmmm.org/2018)
- E. Pacitti: PC chair of the VLDB workshop on Big Social Data and Urban Computing (BiDU@VLDB 2018)

10.1.2.2. Member of the Conference Program Committees

- ACM/SIGAPP Symposium On Applied Computing (ACM SAC) Data Mining track, 2018: F. Masseglia
- IEEE International Conference on Data Mining (IEEE ICDM), 2018: F. Masseglia
- International Joint Conference on Artificial Intelligence (IJCAI), "Sister Conference Best Paper Track", 2018: F. Masseglia
- International Symposium on Methodologies for Intelligent Systems (ISMIS), 2018: F. Masseglia
- Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD), 2018: F. Masseglia
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (PKDD), 2018: F. Masseglia
- IEEE Artificial Intelligence and Knowledge Engineering (IEEE AIKE), 2018: F. Masseglia
- International Conference on Information Management and Big Data (SIMBig), 2018: F. Masseglia
- International Conference on Data Science, Technology and Applications (DATA), 2018: F. Masseglia
- International Conference on Very Large Data Bases (VLDB), 2018: R. Akbarinia
- International Workshop on Big Data Management in Cloud Systems, 2018: R. Akbarinia
- Int. Conf. on Extending DataBase Technologies (EDBT), 2019: E. Pacitti
- Int. Conf. on Multimedia Retrieval (ICMR), 2018: A. Joly
- Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2018: A. Joly
- Int. Conf. on Computer Vision (CVPR), 2018: A. Joly
- Int. Conf. and Labs of the Evaluation Forum (CLEF), 2018: A. Joly
- European. Conf. on Information Retrieval (ECIR), 2019: A. Joly
- Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA 2018): F. Masseglia, E. Pacitti

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- VLDB Journal: P. Valduriez.
- Journal of Transactions on Large Scale Data and Knowledge Centered Systems: R. Akbarinia.

- Distributed and Parallel Databases, Kluwer Academic Publishers: E. Pacitti, P. Valduriez.
- Book series “Data Centric Systems and Applications” (Springer): P. Valduriez.
- Multimedia Tools and Applications: A. Joly.
- Plant Methods: C. Pradal.

10.1.3.2. Reviewer - Reviewing Activities

Reviewing in international journals :

- Distributed and Parallel Databases (DAPD): R. Akbarinia, E. Pacitti, P. Valduriez
- IEEE Transactions on Knowledge and Data Engineering (TKDE): R. Akbarinia, F. Masseglia
- VLDB Journal: R. Akbarinia
- ACM Transactions on Database Systems (TODS): A. Joly
- IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI): A. Joly
- Information Sciences: A. Joly
- Ecological Informatics: A. Joly
- Multimedia Tools and Applications Journal (MTAP): A. Joly
- Multimedia Systems: A. Joly
- Transactions on Information Forensics & Security: A. Joly
- International Journal of Computer Vision: A. Joly
- Transactions on Image Processing: A. Joly
- ACM Trans. on Database Systems: E. Pacitti
- Knowledge and Information Systems (KAIS): F. Masseglia
- IEEE Transaction on Signal Processing (TSP): A. Liutkus
- IEEE Transactions on Audio Speech and Language Processing (TASLP): A. Liutkus
- IEEE Signal Processing Magazine: A. Liutkus
- Frontiers in Plant Science: C. Pradal

10.1.4. Invited Talks

- A. Joly: keynote talk on "Towards The Recognition of the World’s Flora: When HPC Meets Deep Learning" at Digital Infrastructures 2018 on Oct. 10
- A. Joly: keynote talk on "The Recognition of the World’s Flora" at Terratec 2018
- A. Liutkus: tutorial on music source separation at the International Symposium on Music Information Retrieval (ISMIR 2018).
- F. Masseglia: talk on "Massively Distributed Data Analytics", IRISA (Lacodam team), April 2018
- F. Masseglia: talk on "Massively Distributed Time Series Indexing and Querying", LIMOS, December 2018
- P. Valduriez: keynote talk on "Blockchain 2.0: opportunities and risks" on 29 may at Africatek 2018, Cotonou, Bénin, on 25 october at BDA 2018, Bucharest, Romania, and on 19 december at Colloquim COPPE/UFRJ, Rio de Janeiro
- C. Pradal: keynote talk on "OpenAlea : an open source project for plant modelling at different scales", August 2018, Crops in Silico Symposium, NCSA, Univ. Illinois, USA.
- C. Pradal: keynote talk on "OpenAlea : a modular platform for multiscale plant modelling", April 2018, EGU 2018, Vienna, Austria.

10.1.5. Leadership within the Scientific Community

- A. Joly: scientific manager of the LifeCLEF and Pl@ntNet research platforms

- A. Liutkus: elected member of the IEEE Technical Committee on Audio and Acoustic Signal Processing
- F. Masseglia: “Chargé de mission pour la médiation scientifique Inria” and head of Inria’s national network of colleagues involved in science popularization
- E. Pacitti: head of Polytech’ Montpellier’s Direction of Foreign Relationships
- P. Valduriez: scientific manager for the Latin America zone at Inria’s Direction of Foreign Relationships (DPEI)
- P. Valduriez: President of the Steering Committee of the BDA conference

10.1.6. Scientific Expertise

- R. Akbarinia, F. Masseglia: reviewer for international programs (STIC AmSud, ECOS SUD).
- R. Akbarinia: expert for the French National Research Agency (ANR).
- A: Joly: reviewer for STIC AmSud international program
- F. Masseglia: scientific referent for Inria on the frame agreement with SAFRAN about "System Monitoring and Data Analytics"
- E. Pacitti: reviewer for STIC AmSud international program
- P. Valduriez: reviewer for STIC AmSud international program
- P. Valduriez: reviewer for NSERC (Canada)
- C. Pradal: reviewer for STIC AmSud international program
- C. Pradal: member of CSS EGBIP (Commissions Scientifiques Spécialisées) INRA

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Most permanent members of Zenith teach at the Licence and Master degree levels at UM2.

Esther Pacitti:

IG3: Database design, physical organization, 54h, level L3, Polytech’Montpellier, UM2

IG4: Networks, 42h, level M1, Polytech’ Montpellier, UM2

IG4: Object-relational databases, 32h, level M1, Polytech’ Montpellier, UM2

IG5: Distributed systems, virtualization, 27h, level M2, Polytech’ Montpellier, UM2

Industry internship committee, 50h, level M2, Polytech’ Montpellier

Patrick Valduriez:

Professional: Distributed Information Systems, Big Data Architectures, 75h, level M2, Capgemini Institut

Alexis Joly:

University of Montpellier: Machine Learning, 15h, level M2

Polytech’ Montpellier: Content-Based Image Retrieval, 4.5h, level M1

AgroParisTech: Convolutional Neural Networks in Ecology and Agronomy, 2h, level M1

10.2.2. Supervision

PhD & HdR:

PhD : Vitor Silva, Analysis of raw data from multiple data sources during the execution of computational simulations, started 2014, UFRJ, Brazil, June 2018. Advisors: Marta Mattoso (UFRJ), Daniel Oliveira (UFF), Patrick Valduriez

PhD : Sakina Mahboubi, Privacy Preserving Top-k Query Processing over Outsourced Data, Univ. Montpellier, Nov. 21, 2018. Advisors: Reza Akbarinia, Patrick Valduriez.

PhD : Djamel-Edine Yagoubi, Massive distribution for indexing and mining time series, Univ. Montpellier, March 12, 2018. Advisors: Reza Akbarinia, Florent Masseglia, Themis Palpanas (Univ Paris Descartes).

PhD : Mehdi Zitouni, Parallel Itemsets Mining in Massively Distributed Environments, Univ. Tunis, Dec. 5, 2018. Advisors: Reza Akbarinia, Florent Masseglia, Sadok Ben Yahia (Univ Tunis).

PhD in progress: Gaetan Heidsieck, Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping, started Oct 2017, Univ. Montpellier. Advisors: Esther Pacitti, Christophe Pradal, François Tardieu (INRA).

PhD in progress: Christophe Botella, Large-scale Species Distribution Modelling based on crowdsrouced image streams, started Oct 2016, Univ. Montpellier. Advisors: Alexis Joly, François Munoz (IRD), Pascal Monestiez (INRA).

PhD in progress: Titouan Lorieul, Pro-active Crowdsourcing, started Oct 2016, Univ. Montpellier. Advisor: Alexis Joly.

PhD in progress: Khadidja Meguelati, Massively Distributed Clustering, started Oct 2016, Univ. Montpellier. Advisors: Nadine Hilgert (INRA), Florent Masseglia.

PhD in progress: Renan Souza, Massively Distributed Clustering, started 2015, UFRJ, Brazil. Advisors: Marta Mattoso (UFRJ), Daniel Oliveira (UFF), Patrick Valduriez.

PhD in progress: Mathieu Fontaine, Alpha-stable models for signal processing, started 2016, IAEM, Nancy, France. Advisors: Roland Badeau (Telecom ParisTech), Antoine Liutkus.

10.2.3. Juries

Members of the team participated to the following PhD or HDR committees:

- R. Akbarinia: Sakina Mahboubi (Univ. Montpellier, advisor), Djamel-Edine Yagoubi (Univ. Montpellier, advisor)
- A. Joly: Lee Sue Han (Univ. of Malaya)
- F. Masseglia: Yann Dauxais (Univ. Rennes), Steeve Vanel-Siyou (Univ. Clermont-Ferrand, reviewer), Marc Plantevit (HDR, Univ. Lyon, reviewer), Djamel-Edine Yagoubi (Univ. Montpellier, advisor)
- E. Pacitti: Abdoul Macine (Univ. Nice, reviewer)
- P. Valduriez: Louis Jachiet (Univ. Grenoble), Ovidiu-Cristian Marcu (Univ. Rennes 1, reviewer), Vitor Silva (UFRJ, Rio de Janeiro, advisor), Sakina Mahboubi (Univ. Montpellier, advisor), Yania Molina Souto (LNCC, Rio de Janeiro, reviewer)

Members of the team participated to the following hiring committees:

- A. Joly: associate professor position, Univ. Toulon
- F. Masseglia: Inria ARP/SRP; full professor position, INSA, Lyon

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

F. Masseglia is “Chargé de mission auprès de la DGD-S Inria pour la médiation scientifique” (50% of his time) and heads Inria’s national network of colleagues involved in science popularization (https://www.inria.fr/recherches/mediation-scientifique/actions-de-mediation-scientifique/presentation)

10.3.2. Articles and contents

Alexis Joly participated to the realization of a report on "Deep Learning and Agriculture" edited by the AgroTIC chair (https://www.agrotic.org/). He co-authored on article on data collection in citizen science projects [37].

10.3.3. Education

Teaching code is now officially in the school programs in France. Class'Code is a PIA project that aims at training the needed 300,000 teachers and professionals of education France. The project is a hybrid MOOC (both online courses and physical meetings). Florent Masseglia is co-author of the first course and scientific referent of the other courses.

Along with Class'Code, the association "La main à la pâte" has coordinated the writing of a school book on the teaching of computer science teaching, with Inria (Gilles Dowek, Pierre-Yves Oudeyer, Florent Masseglia and Didier Roy), France-IOI and the University of Lorraine. The book has been requested by and distributed to 15,000 readers in less than one month. The extension of this book for the French "Collège" has been released in 2017 with new activities and new scientific content.

F. Masseglia is giving a doctoral training at different doctoral schools in Montpellier, in order to train facilitators for helping teachers and people of the education world to better understand the "computational thinking". So far, 14 people have been trained.

P. Valduriez gave an invited talk on "Succeed in your Ph.D. Thesis: good practices and return of experience" at the Ph.D. meeting at LIRIS, Lyon, on December 11.

F. Masseglia is member of the pedagogic committee of "Edu'up", a project from France-IOI on learning code and computational thinking.

F. Masseglia gave a one day training session to school teachers in Créteil, on October 3.

Alexis Joly gave about 15 hours of professional training in the use of digital tools for environmental education (Pl@ntNet, ThePlantGame and Smart'Flore).

10.3.4. Interventions

Zenith participated to the following events:

- F. Masseglia co-organized the regional Code-Week events with the local network of media-library ("réseau des médiathèques de Montpellier Méditerranée Métropole").
- F. Masseglia is member of the project selection committee for "La fête de la science" in Montpellier.
- F. Masseglia animated a stand at the "semaine de la mémoire" event organized by Genopolys (September 20&21).
- F. Masseglia participated in a class visit, at Saussan, with Charles Torossian (co-author of the "Vallani-Torossian" report) and the rectrice, about code teaching.

10.3.5. Internal action

F. Masseglia organized, and participated to, a 2 days training session on the Poppy Ergo Jr robot (June 25&26).

10.3.6. Creation of media or tools for science outreach

In the context of the Floris'tic project, A. Joly participates regularly to popularization, educational and citizen science actions in France (with schools, cities, parks, associations, etc.). The softwares developed within the project (Pl@ntNet, Smart'Flore and ThePlantGame) are used in a growing number of formal educational programs and informal educational actions of individual teachers. For instance, Smart'Flore is used by the French National Education in a program for reducing early school leaving. Pl@ntNet app is used in the Reunion island in an educational action called Vegetal riddle organized by the Center for cooperation at school. It is also used in a large-scale program in Czech republic and Slovakia (with a total of 100 classrooms involved in the program). An impact study of the Pl@ntNet application did show that 6% of the respondents use it for educational purposes in the context of their professional activity.

F. Masseglia participated in the work group on "Jeu des 7 familles de l'informatique". This card game, to be announced officially in January 2019, provides support for education to computer science from the history point of view.

11. Bibliography

Major publications by the team in recent years

[1] A. AFFOUARD, H. GOËAU, P. BONNET, J.-C. LOMBARDO, A. JOLY. *Pl@ntNet app in the era of deep learning*, in "ICLR: International Conference on Learning Representations", Toulon, France, April 2017, pp. 1-6, https://hal.archives-ouvertes.fr/hal-01629195

[2] T. ALLARD, G. HÉBRAIL, F. MASSEGLIA, E. PACITTI. *Chiaroscuro: Transparency and Privacy for Massive Personal Time-Series Clustering*, in "34th International ACM Conference on Management of Data (ACM SIGMOD)", Melbourne, Australia, ACM SIGMOD, May 2015 [*DOI :* 10.1145/2723372.2749453], https://hal.inria.fr/hal-01136686

[3] A. JOLY, P. BONNET, H. GOËAU, J. BARBE, S. SELMI, J. CHAMP, S. DUFOUR-KOWALSKI, A. AFFOUARD, J. CARRÉ, J.-F. MOLINO, N. BOUJEMAA, D. BARTHÉLÉMY. *A look inside the Pl@ntNet experience*, in "Multimedia Systems", 2015, 16 p. [*DOI :* 10.1007/s00530-015-0462-9], https://hal.inria.fr/hal-01182775

[4] A. JOLY, O. BUISSON. *Random Maximum Margin Hashing*, in "CVPR'11 - IEEE Computer Vision and Pattern Recognition", Colorado springs, United States, IEEE, June 2011, pp. 873-880 [*DOI :* 10.1109/CVPR.2011.5995709], https://hal.inria.fr/hal-00642178

[5] A. JOLY, H. GOEAU, P. BONNET, V. BAKIC, J. BARBE, S. SELMI, I. YAHIAOUI, J. CARRÉ, E. MOUYSSET, J.-F. MOLINO, N. BOUJEMAA, D. BARTHÉLÉMY. *Interactive plant identification based on social image data*, in "Ecological Informatics", 2013 [*DOI :* 10.1016/J.ECOINF.2013.07.006], http://www.sciencedirect.com/science/article/pii/S157495411300071X

[6] B. KOLEV, P. VALDURIEZ, C. BONDIOMBOUY, R. JIMÉNEZ-PERIS, R. PAU, J. O. PEREIRA. *CloudMdsQL: Querying Heterogeneous Cloud Data Stores with a Common Language*, in "Distributed and Parallel Databases", December 2016, vol. 34, n° 4, pp. 463-503 [*DOI :* 10.1007/s10619-015-7185-y], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01184016

[7] M. LIROZ-GISTAU, R. AKBARINIA, D. AGRAWAL, P. VALDURIEZ. *FP-Hadoop: Efficient Processing of Skewed MapReduce Jobs*, in "Information Systems", 2016, vol. 60, pp. 69-84 [*DOI :* 10.1016/J.IS.2016.03.008], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01377715

[8] J. LIU, E. PACITTI, P. VALDURIEZ, D. DE OLIVEIRA, M. MATTOSO. *Multi-Objective Scheduling of Scientific Workflows in Multisite Clouds*, in "Future Generation Computer Systems", 2016, vol. 63, pp. 76–95 [*DOI :* 10.1016/J.FUTURE.2016.04.014], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01342203

[9] H. LUSTOSA, F. PORTO, P. BLANCO, P. VALDURIEZ. *Database System Support of Simulation Data*, in "Proceedings of the VLDB Endowment (PVLDB)", September 2016, vol. 9, n° 13, pp. 1329-1340, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01363738

[10] S. MAHBOUBI, R. AKBARINIA, P. VALDURIEZ. *Privacy-Preserving Top-k Query Processing in Distributed Systems*, in "Euro-Par: European Conference on Parallel and Distributed Computing", Turin, Italy, August 2018, vol. LNCS, n° 11014 [*DOI :* 10.1007/978-3-319-96983-1_20], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886160

[11] E. PACITTI, R. AKBARINIA, M. EL DICK. *P2P Techniques for Decentralized Applications*, Morgan & Claypool Publishers, 2012, 104 p. , http://hal.inria.fr/lirmm-00748635

[12] C. SAHIN, T. ALLARD, R. AKBARINIA, A. ABBADI, E. PACITTI. *A Differentially Private Index for Range Query Processing in Clouds*, in "ICDE: International Conference on Data Engineering", Paris, France, April 2018, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886725

[13] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Fast Parallel Mining of Maximally Informative k-Itemsets in Big Data*, in "IEEE International Conference on Data Mining (ICDM)", Atlantic city, United States, August 2015, http://hal-lirmm.ccsd.cnrs.fr/lirmm-01187275

[14] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *A Highly Scalable Parallel Algorithm for Maximally Informative k-Itemset Mining*, in "Knowledge and Information Systems (KAIS)", January 2017, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01288571

[15] S. SALAH, R. AKBARINIA, F. MASSEGLIA. *Data placement in massively distributed environments for fast parallel mining of frequent itemsets*, in "Knowledge and Information Systems (KAIS)", 2017, vol. 53, n° 1, pp. 207-237 [*DOI :* 10.1007/s10115-017-1041-5], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620383

[16] M. SERVAJEAN, R. AKBARINIA, E. PACITTI, S. AMER-YAHIA. *Profile Diversity for Query Processing using User Recommendations*, in "Information Systems", March 2015, vol. 48, pp. 44-63 [*DOI :* 10.1016/j.is.2014.09.001], http://hal-lirmm.ccsd.cnrs.fr/lirmm-01079523

[17] M. SERVAJEAN, A. JOLY, D. SHASHA, J. CHAMP, E. PACITTI. *Crowdsourcing Thousands of Specialized Labels: A Bayesian Active Training Approach*, in "IEEE Transactions on Multimedia", June 2017, vol. 19, n° 6, pp. 1376 - 1391 [*DOI :* 10.1109/TMM.2017.2653763], https://hal.archives-ouvertes.fr/hal-01629149

[18] D. E. YAGOUBI, R. AKBARINIA, B. KOLEV, O. LEVCHENKO, F. MASSEGLIA, P. VALDURIEZ, D. SHASHA. *ParCorr: efficient parallel methods to identify similar time series pairs across sliding windows*, in "Data Mining and Knowledge Discovery", September 2018, vol. 32, n° 5, pp. 1481-1507 [*DOI :* 10.1007/s10618-018-0580-z], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886794

[19] D.-E. YAGOUBI, R. AKBARINIA, F. MASSEGLIA, T. PALPANAS. *DPiSAX: Massively Distributed Partitioned iSAX*, in "ICDM 2017: IEEE International Conference on Data Mining", New Orleans, United States, November 2017, pp. 1-6, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620125

[20] T. M. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems, third edition*, Springer, 2011, 845 p. , http://hal.inria.fr/hal-00640392/en

Publications of the year

Doctoral Dissertations and Habilitation Theses

[21] V. SILVA. *Analysis of raw data from multiple data sources during the execution of computational simulations*, Universidade Federal de Rio de Janeiro, June 2018, https://hal-lirmm.ccsd.cnrs.fr/tel-01830211

[22] D.-E. YAGOUBI. *Massive distribution for indexing and mining time series*, Université de Montpellier, March 2018, https://tel.archives-ouvertes.fr/tel-01945348

[23] M. ZITOUNI. *Parallel Itemset Mining in Massively Distributed Environments*, Université de Tunis El Manar ; Inria, December 2018, https://tel.archives-ouvertes.fr/tel-01953619

Articles in International Peer-Reviewed Journals

[24] C. BOTELLA, A. JOLY, P. BONNET, P. P. MONESTIEZ, F. MUNOZ. *Species distribution modeling based on the automated identification of citizen observations*, in "Applications in Plant Sciences", March 2018, vol. 6, n^{o} 2, pp. 1-11 [*DOI :* 10.1002/APS3.1029], https://hal.umontpellier.fr/hal-01739481

[25] J. CAMATA, V. SILVA, P. VALDURIEZ, M. MATTOSO, A. L. G. A. COUTINHO. *In situ visualization and data analysis for turbidity currents simulation*, in "Computers & Geosciences", January 2018, vol. 110, pp. 23-31 [*DOI :* 10.1016/J.CAGEO.2017.09.013], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01620127

[26] E. CANO, D. FITZGERALD, A. LIUTKUS, M. D. PLUMBLEY, F.-R. STÖTER. *Musical Source Separation: An Introduction*, in "IEEE Signal Processing Magazine", 2018, https://hal.inria.fr/hal-01945345

[27] T.-W. CHEN, L. C. CABRERA-BOSQUET, S. ALVAREZ PRADO, R. PEREZ, S. ARTZET, C. PRADAL, A. COUPEL-LEDRU, C. FOURNIER, F. TARDIEU. *Genetic and environmental dissection of biomass accumulation in multi-genotype maize canopies*, in "Journal of Experimental Botany", August 2018 [*DOI :* 10.1093/JXB/ERY309], https://hal.inria.fr/hal-01895279

[28] P. FERNIQUE, C. PRADAL. *AutoWIG: automatic generation of python bindings for C++ libraries*, in "PeerJ Computer Science", 2018, vol. 4 [*DOI :* 10.7717/PEERJ-CS.149], https://hal.inria.fr/hal-01756458

[29] J. LIU, E. PACITTI, P. VALDURIEZ. *A Survey of Scheduling Frameworks in Big Data Systems*, in "International Journal of Cloud Computing", 2018, vol. 7, n^{o} 2, pp. 103-128, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01692229

[30] J. LIU, L. PINEDA, E. PACITTI, A. COSTAN, P. VALDURIEZ, G. ANTONIU, M. MATTOSO. *Efficient Scheduling of Scientific Workflows using Hot Metadata in a Multisite Cloud*, in "IEEE Transactions on Knowledge and Data Engineering", 2018, pp. 1-20 [*DOI :* 10.1109/TKDE.2018.2867857], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01867717

[31] C. PRADAL, S. COHEN-BOULAKIA, G. HEIDSIECK, E. PACITTI, F. TARDIEU, P. VALDURIEZ. *Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping*, in "ERCIM News", 2018, pp. 36-37, https://hal.inria.fr/hal-01948568

[32] Z. RAFII, A. LIUTKUS, F.-R. STÖTER, S. IOANNIS MIMILAKIS, D. FITZGERALD, B. PARDO. *An Overview of Lead and Accompaniment Separation in Music*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", 2018 [*DOI :* 10.1109/TASLP.2018.2825440], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766781

[33] M. SERVAJEAN, R. CHAILAN, A. JOLY. *Non-parametric Bayesian annotator combination*, in "Information Sciences", April 2018, vol. 436-437, pp. 131-145 [*DOI :* 10.1016/J.INS.2018.01.020], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01703020

[34] V. SILVA, D. DE OLIVEIRA, P. VALDURIEZ, M. MATTOSO. *DfAnalyzer: Runtime Dataflow Analysis of Scientific Applications using Provenance*, in "Proceedings of the VLDB Endowment (PVLDB)", August 2018, vol. 11, n° 12, pp. 2082-2085, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01867887

[35] R. SOUZA, V. SILVA, A. L. COUTINHO, P. VALDURIEZ, M. MATTOSO. *Data reduction in scientific workflows using provenance monitoring and user steering*, in "Future Generation Computer Systems", 2018, pp. 1-21 [*DOI :* 10.1016/J.FUTURE.2017.11.028], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01679967

[36] D. E. YAGOUBI, R. AKBARINIA, B. KOLEV, O. LEVCHENKO, F. MASSEGLIA, P. VALDURIEZ, D. SHASHA. *ParCorr: efficient parallel methods to identify similar time series pairs across sliding windows*, in "Data Mining and Knowledge Discovery", September 2018, vol. 32, n° 5, pp. 1481-1507 [*DOI :* 10.1007/S10618-018-0580-Z], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886794

Articles in National Peer-Reviewed Journals

[37] S. BLANGY, V. LHOSTE, C. ARNAL, J. CARRÉ, A. CHAPOT, I. CHUINE, G. DARMON, A. JOLY, P. MONESTIEZ, P. BONNET. *Au-delà de la collecte des données dans les projets de sciences citoyennes : ouvrir le champ de l'analyse et de l'interprétation des données aux citoyens*, in "Technologie et innovation", 2018, https://hal.archives-ouvertes.fr/hal-01824900

Invited Conferences

[38] P. VALDURIEZ, M. MATTOSO, R. AKBARINIA, H. BORGES, J. CAMATA, A. L. G. A. COUTINHO, D. GASPAR, N. LEMUS, J. LIU, H. LUSTOSA, F. MASSEGLIA, F. NOGUEIRA DA SILVA, V. SILVA, R. SOUZA, K. OCAÑA, E. OGASAWARA, D. OLIVEIRA, E. PACITTI, F. PORTO, D. SHASHA. *Scientific Data Analysis Using Data-Intensive Scalable Computing: the SciDISC Project*, in "LADaS: Latin America Data Science Workshop", Rio de Janeiro, Brazil, CEUR-WS.org, August 2018, vol. CEUR Workshop Proceedings, n° 2170, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01867804

International Conferences with Proceedings

[39] M. R. BOUADJENEK, E. PACITTI, M. SERVAJEAN, F. MASSEGLIA, A. ABBADI. *A Distributed Collaborative Filtering Algorithm Using Multiple Data Sources*, in "DBKDA: Advances in Databases, Knowledge, and Data Applications", Nice, France, May 2018, https://arxiv.org/abs/1807.05853 , https://hal.archives-ouvertes.fr/hal-01911684

[40] R. CAMPISANO, H. BORGES, F. PORTO, F. PEROSI, E. PACITTI, F. MASSEGLIA, E. OGASAWARA. *Discovering Tight Space-Time Sequences*, in "DaWaK: Data Warehousing and Knowledge Discovery", Regensburg, Germany, September 2018, vol. LNCS, n° 11031, pp. 247-257 [*DOI :* 10.1007/978-3-319-98539-8_19], https://hal.archives-ouvertes.fr/hal-01925965

[41] A. B. CRUZ, J. FERREIRA, D. CARVALHO, E. MENDES, E. PACITTI, R. COUTINHO, F. PORTO, E. OGASAWARA. *Detecçao de Anomalias Frequentes no Transporte Rodoviario Urbano*, in "SBBD: Simpósio Brasileiro de Banco de Dados", Rio de Janeiro, Brazil, SBC, August 2018, pp. 271-276, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01868597

[42] D. DI CARLO, A. LIUTKUS, K. DÉGUERNEL. *Interference reduction on full-length live recordings*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech, and Signal Processing", Calgary, Canada, IEEE, April 2018, pp. 736-740 [*DOI :* 10.1109/ICASSP.2018.8462621], https://hal.inria.fr/hal-01713889

[43] J. FERREIRA, J. SOARES, F. PORTO, E. PACITTI, R. COUTINHO, E. OGASAWARA. *Rumo à Integração da Álgebra de Workflows com o Processamento de Consulta Relacional*, in "SBBD: Simpósio Brasileiro de Banco de Dados", Rio de Janeiro, Brazil, SBC, August 2018, pp. 205-210, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01868556

[44] M. FONTAINE, F.-R. STÖTER, A. LIUTKUS, U. SIMSEKLI, R. SERIZEL, R. BADEAU. *Multichannel Audio Modeling with Elliptically Stable Tensor Decomposition*, in "LVA ICA 2018 - 14th International Conference on Latent Variable Analysis and Signal Separation", Surrey, United Kingdom, July 2018, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766795

[45] A. JOLY, H. GOËAU, C. BOTELLA, H. GLOTIN, P. BONNET, W.-P. VELLINGA, R. PLANQUÉ, H. MÜLLER. *Overview of LifeCLEF 2018: A Large-Scale Evaluation of Species Identification and Recommendation Algorithms in the Era of AI*, in "CLEF: Cross-Language Evaluation Forum", Avignon, France, Experimental IR Meets Multilinguality, Multimodality, and Interaction, September 2018, vol. LNCS, n^{o} 11018, pp. 247-266 [*DOI :* 10.1007/978-3-319-98932-7_24], https://hal.archives-ouvertes.fr/hal-01913231

[46] N. KERIVEN, A. DELEFORGE, A. LIUTKUS. *Blind Source Separation Using Mixtures of Alpha-Stable Distributions*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, IEEE, April 2018, pp. 771-775, https://arxiv.org/abs/1711.04460 [*DOI :* 10.1109/ICASSP.2018.8462095], https://hal.inria.fr/hal-01633215

[47] B. KOLEV, O. LEVCHENKO, E. PACITTI, P. VALDURIEZ, R. VILAÇA, R. C. GONÇALVES, R. JIMÉNEZ-PERIS, P. KRANAS. *Parallel Polyglot Query Processing on Heterogeneous Cloud Data Stores with LeanXcale*, in "IEEE BigData", Seattle, United States, IEEE, December 2018, 10 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01921718

[48] O. LEVCHENKO, D.-E. YAGOUBI, R. AKBARINIA, F. MASSEGLIA, B. KOLEV, D. SHASHA. *Spark-parSketch: A Massively Distributed Indexing of Time Series Datasets*, in "CIKM: Conference on Information and Knowledge Management", Turin, Italy, October 2018, pp. 1951-1954 [*DOI :* 10.1145/3269206.3269226], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886760

[49] J. LIU, N. LEMUS, E. PACITTI, F. PORTO, P. VALDURIEZ. *Computation of PDFs on Big Spatial Data: Problem & Architecture*, in "LADaS: Latin America Data Science Workshop", Rio de Janeiro, Brazil, CEUR-WS.org, August 2018, vol. 2170, 6 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-01867758

[50] A. LIUTKUS, C. ROHLFING, A. DELEFORGE. *Audio source separation with magnitude priors: the BEADS model*, in "ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Canada, Signal Processing and Artificial Intelligence: Changing the World, April 2018, pp. 1-5 [*DOI :* 10.1109/ICASSP.2018.8462515], https://hal.inria.fr/hal-01713886

[51] S. MAHBOUBI, R. AKBARINIA, P. VALDURIEZ. *Answering Top-k Queries over Outsourced Sensitive Data in the Cloud*, in "DEXA: Database and Expert Systems Applications", Regensburg, Germany, September 2018, vol. LNCS, n^{o} 11029, pp. 218-231 [*DOI :* 10.1007/978-3-319-98809-2_14], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886164

[52] S. MAHBOUBI, R. AKBARINIA, P. VALDURIEZ. *Privacy-Preserving Top-k Query Processing in Distributed Systems*, in "Euro-Par: European Conference on Parallel and Distributed Computing", Turin, Italy, August 2018, pp. 281-292 [*DOI :* 10.1007/978-3-319-96983-1_20], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886160

[53] F. PORTO, A. KHATIBI, J. G. RITTMEYER, E. OGASAWARA, P. VALDURIEZ, D. SHASHA. *Constellation Queries over Big Data*, in "SBBD: Simpósio Brasileiro de Banco de Dados", Rio de Janeiro, Brazil, SBC, August 2018, pp. 85-96, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01867833

[54] F. PORTO, J. G. RITTMEYER, E. OGASAWARA, A. KRONE-MARTINS, P. VALDURIEZ, D. SHASHA. *Point Pattern Search in Big Data*, in "SSDBM: Scientific and Statistical Database Management", Bozen-Bolzano, Italy, ACM, July 2018 [*DOI :* 10.1145/3221269.3221294], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01819290

[55] C. SAHIN, T. ALLARD, R. AKBARINIA, A. ABBADI, E. PACITTI. *A Differentially Private Index for Range Query Processing in Clouds*, in "ICDE: International Conference on Data Engineering", Paris, France, April 2018, pp. 857-868, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01886725

[56] D. SILVA, A. PAES, E. PACITTI, D. DE OLIVEIRA. *F ReeP: towards parameter recommendation in scientific workflows using preference learning*, in "SBBD: Simpósio Brasileiro de Banco de Dados", Rio de Janeiro, Brazil, SBC, August 2018, n^{o} 211-216, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01868574

[57] U. SIMSEKLI, H. ERDOGAN, S. LEGLAIVE, A. LIUTKUS, R. BADEAU, G. RICHARD. *Alpha-stable low-rank plus residual decomposition for speech enhancement*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech, and Signal Processing", Calgary, Canada, April 2018, https://hal.inria.fr/hal-01714909

[58] F.-R. STÖTER, A. LIUTKUS, N. ITO. *The 2018 Signal Separation Evaluation Campaign*, in "LVA ICA: Latent Variable Analysis and Signal Separation", Surrey, United Kingdom, July 2018, https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766791

[59] D. WARD, R. D. MASON, C. KIM, F.-R. STÖTER, A. LIUTKUS, M. D. PLUMBLEY. *SiSEC 2018: State of the art in musical audio source separation - subjective selection of the best algorithm*, in "WIMP: Workshop on Intelligent Music Production", Huddersfield, United Kingdom, September 2018, https://hal.inria.fr/hal-01945362

Conferences without Proceedings

[60] B. DENEU, M. SERVAJEAN, C. BOTELLA, A. JOLY. *Location-based species recommendation using co-occurrences and environment-GeoLifeCLEF 2018 challenge*, in "CLEF: Conference and Labs of the Evaluation Forum", Avignon, France, September 2018, vol. CEUR Workshop Proceedings, n^{o} 2125, https://hal.archives-ouvertes.fr/hal-01913241

[61] H. GOËAU, P. BONNET, A. JOLY. *Overview of ExpertLifeCLEF 2018: how far automated identification systems are from the best experts?*, in "CLEF: Conference and Labs of the Evaluation Forum", Avignon, France, September 2018, https://hal.archives-ouvertes.fr/hal-01913244

[62] B. YUN. *How Can You Mend a Broken Inconsistent KBs in Existential Rules Using Argumentation*, in "SSA: Summer School on Argumentation", Varsovie, Poland, September 2018, https://hal.archives-ouvertes.fr/hal-01940651

[63] M. ZITOUNI, R. AKBARINIA, S. BEN YAHIA, F. MASSEGLIA. *Maximally Informative k-Itemset Mining from Massively Distributed Data Streams*, in "SAC: Symposium on Applied Computing", Pau, France, April 2018, pp. 1-10, https://hal.archives-ouvertes.fr/hal-01711990

Scientific Books (or Scientific Book chapters)

[64] P. BONNET, H. GOËAU, S. T. HANG, M. LASSECK, M. SULC, V. V. MALÉCOT, P. JAUZEIN, J.-C. MELET, C. YOU, A. JOLY. *Plant Identification: Experts vs. Machines in the Era of Deep Learning: Deep learning techniques challenge flora experts*, in "Multimedia Tools and Applications for Environmental & Biodiversity Informatics", June 2018, vol. Chapter 8, pp. 131-149 [*DOI :* 10.1007/978-3-319-76445-0_8], https://hal.archives-ouvertes.fr/hal-01913277

[65] C. BOTELLA, A. JOLY, P. BONNET, P. MONESTIEZ, F. MUNOZ. *A deep learning approach to Species Distribution Modelling*, in "Multimedia Tools and Applications for Environmental & Biodiversity Informatics", A. JOLY, S. VROCHIDIS, K. KARATZAS, A. KARPPINE, P. BONNE (editors), Springer, 2018, pp. 169-199 [*DOI :* 10.1007/978-3-319-76445-0_10], https://hal.archives-ouvertes.fr/hal-01834227

[66] J. CARRANZA-ROJAS, A. JOLY, H. GOËAU, E. MATA-MONTERO, P. BONNET. *Automated identification of herbarium specimens at different taxonomic levels*, in "Multimedia Tools and Applications for Environmental & Biodiversity Informatics", June 2018, vol. Multimedia Systems and Applications, pp. 151-167 [*DOI :* 10.1007/978-3-319-76445-0_9], https://hal.archives-ouvertes.fr/hal-01913272

[67] A. A. NUGRAHA, A. LIUTKUS, E. VINCENT. *Deep neural network based multichannel audio source separation*, in "Audio Source Separation", Springer, March 2018, https://hal.inria.fr/hal-01633858

[68] B. PARDO, A. LIUTKUS, Z. DUAN, G. RICHARD. *Applying source separation to music*, in "Audio Source Separation and Speech Enhancement", Wiley, August 2018, vol. Chapter 16 [*DOI :* 10.1002/9781119279860.CH16], https://hal.inria.fr/hal-01945320

Books or Proceedings Editing

[69] A. JOLY, S. VROCHIDIS, K. KARATZAS, A. KARPPINEN, P. BONNET (editors). *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, Springer International Publishing, 2018 [*DOI :* 10.1007/978-3-319-76445-0], https://hal-lirmm.ccsd.cnrs.fr/lirmm-01959343

Research Reports

[70] M. CONTRACTOR, C. PRADAL, D. SHASHA. *Platform Migrator*, New York University, May 2018, n^{o} TR2018-990, 43 p. , https://hal.inria.fr/hal-01948552

Other Publications

[71] C. BOTELLA, P. BONNET, F. MUNOZ, P. P. MONESTIEZ, A. JOLY. *Overview of GeoLifeCLEF 2018: location-based species recommendation*, CEUR Workshops Proceedings, September 2018, vol. CEUR-WS, n^{o} 2125, CLEF: Cross-Language Evaluation Forum, Poster, https://hal.archives-ouvertes.fr/hal-01913238

[72] F. REYES, B. PALLAS, D. GIANELLE, C. PRADAL, F. VAGGI, D. ZANOTELLI, M. TAGLIAVINI, D. GIANELLE, E. COSTES. *MuSCA: a multi-scale model to explore carbon allocation in plants*, October 2018, working paper or preprint [*DOI :* 10.1101/370189], https://hal.archives-ouvertes.fr/hal-01844390