IN PARTNERSHIP WITH:
CNRS

Université de Montpellier

Activity Report 2019

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
A9.2. - Machine learning
A9.3. - Signal analysis

Other Research Topics and Application Domains:

B1. - Life sciences
B1.1. - Biology
B1.1.7. - Bioinformatics
B1.1.11. - Plant Biology
B3.3. - Geosciences
B4. - Energy
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, HDR]
Alexis Joly [Inria, Researcher, HDR]
Antoine Liutkus [Inria, Researcher]
Florent Masseglia [Inria, Senior Researcher, HDR]
Didier Parigot [Inria, Researcher, HDR]
Christophe Pradal [CIRAD, Researcher]
Dennis Shasha [Inria, International Chair, Advanced Research Position]
Hervé Goëau [CIRAD, Researcher]

Faculty Members

Esther Pacitti [Univ de Montpellier, Associate Professor, HDR]
Michel Riveill [Univ de Nice - Sophia Antipolis, Professor, from Feb 2019 until Jul 2019]

PhD Students

Christophe Botella [INRA, until Sep 2019]
Benjamin Deneu [Inria, from Oct 2019]
Lamia Djebour [Averroes fellowship, Algeria, from Oct 2019]
Gaetan Heidsieck [Inria]
Titouan Lorieul [Univ de Montpellier]
Khadidja Meguelati [Averroes fellowship, Algeria, until Nov 2019]
Renan Souza [UFRJ, Brazil]
Heraldo Borges [CEFET/RJ, Brazil]
Alena Shilova [Inria]
Daniel Rosendoo [Inria]
Quentin Leroy [INA, from Oct 2019]
Mathieu Fontaine [Inria, until July 2019]

Technical staff

Antoine Affouard [Inria, Engineer]
Julien Champ [Inria, Engineer, from Apr 2019]
Benjamin Deneu [Inria, Engineer, until Sep 2019]
Alain Ibrahim [Inria, from Mar 2019 until May 2019]
Boyan Kolev [Inria, Engineer, until Nov 2019]
Quentin Leroy [Inria, Engineer, from Aug 2019 until Sep 2019]
Oleksandra Levchenko [Inria, Engineer]
Tanmoy Mondal [Inria, Engineer, from Feb 2019]
Fabian Robert Stoter [Inria, Engineer]
Jean-Christophe Lombardo [Inria, Engineer]

Intern and Apprentice

Delton de Andrade Vaz [Inria, from Jun 2019 until Aug 2019]

Administrative Assistant

Nathalie Brillouet [Inria]

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, as produced through empirical observation and simulation. Similarly, digital humanities are faced for decades with the problem of exploiting vast amounts of digitized cultural and historical data, such as broadcasted radio or TV content. Such data must be processed (cleaned, transformed, analyzed) in order to draw new conclusions, prove scientific theories and eventually produce knowledge. 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 has hundreds of exabytes.

Scientific data is very complex, in particular because of the heterogeneous methods, the uncertainty of the captured data, the inherently multiscale nature (spatial, temporal) of many sciences and the growing use of imaging (e.g. molecular imaging), resulting in data with hundreds of dimensions (attibutes, features, etc.). 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. Finally, a major difficulty is to interpret scientific data. Unlike web data, e.g. web page keywords or user recommendations, which regular users can understand, making sense out of scientific data requires high expertise in the scientific domain. And interpretation errors can have highly negative consequences, e.g. deploying an oil driller under water at a wrong position.

Despite the variety of scientific data, we can identify common features: big data; manipulated through workflows; typically complex, e.g. multidimensional; 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).

The three main challenges of scientific data management can be summarized by: (1) scale (big data, big applications); (2) complexity (uncertain, high-dimensional data), (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, we strive to develop architectures, models and algorithms that can be implemented as components or services in specific computing environments, e.g. the cloud. We design and validate our solutions by working closely with our scientific partners in Montpellier such as CIRAD, INRA and IRD, which provide the scientific expertise to interpret the data. 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; declarative languages to manipulate data and workflows; and highly distributed and parallel environments such as cluster and cloud for scalability and performance. We also exploit machine learning, probabilities and statistics for high-dimensional data processing, data analytics and data 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 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, and is applied to categorical and continuous data. In the Zenith team, we are interested in both of these data types. Categorical data designates a set of data that can be described as "check boxes". It can be names, products, items, towns, etc. A common illustration is the market basket data, where each item bought by a client is recorded and the set of items is the basket. The typical data mining problems with this kind of data are:

- **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. 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 $i + j + k$".
- **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.).

Continuous data are numeric records that can have an infinite number of values between any two values. A temperature value or a timestamp are examples of such data. They are involved in a widely used type of data known as time series: a series of values, ordered by time, and giving a measure, e.g. coming from a sensor. There is a large number of problems that can apply to this kind of data, including:

- **Indexing and retrieval.** The goal, here, is usually to find, given a query q and a time series dataset D, the records of D that are most similar to q. This may involve any transformation of D by means of an index or an alternative representation for faster execution.
- **Pattern and outlier detection.** The discovery of recurrent patterns or atypical sub-windows in a time series has applications in finance, industrial manufacture or seismology, to name a few. It calls for techniques that avoid pairwise comparisons of all the sub-windows, which would lead to prohibitive response times.
- **Clustering**. The goal is the same as categorical data clustering: group similar time series and separate dissimilar ones.

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. In this context, we use probability theory to quantify uncertainty and 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 our research is oriented towards computationally effective methods.
- **Deep neural networks**. A major breakthrough in machine learning performance has been the advent of deep neural networks, which are characterized by high numbers (millions) of parameters and scalable learning procedures. We are striving towards original architectures and methods that are theoretically grounded and offer state-of-the-art performance for data search and 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 main communities are targeted: botany, and digital humanities. In both cases, our observation 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. Thus, the team is actively involved in the organization of international evaluation campaigns that allow machine learning researchers to propose new methods while solving important application problems.

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.

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 has hundres of TB. 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. Awards

- Antoine Liutkus and Fabian Stoter won the second place at the Global Pytorch Summer Hackaton 2019 organized by FaceBook with the open-unmix software.
- Antoine Liutkus obtained the *Outstanding Reviewer Award* from IEEE.
- Vitor Silva obtained the *best PhD thesis award* from SBBD.

5.1.2. Software

The Pl@ntNet mobile application reached its ten million downloads.

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%), which is quite impressive considering the fact that a majority of users are beginners when they start playing.

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

6.3. 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

6.4. 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: Vitor 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

6.5. 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

6.6. 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

6.7. 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 Programming, ...) to analyse and simulate shapes and their development.

- Participants: Christian Fournier, Christophe Godin, Christophe Pradal, Frédéric Boudon, Patrick Valduriez, Esther Pacitti and Yann Guédon
- 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

6.8. 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

6.9. museval

KEYWORDS: Source Separation - Metric

SCIENTIFIC DESCRIPTION: museval is a Python package aimed at evaluating audio source separation algorithm on the musdb corpus.

It as a scientific tool of high impact, but of limited transfer importance, since it is only (but widely) used by the community to evaluate performance in scientific publications.

FUNCTIONAL DESCRIPTION: The BSSEval metrics, as implemented in the [MATLAB toolboxes](http://bass-db.gforge.inria.fr/bss_eval/) and their re-implementation in [mir_eval](http://craffel.github.io/mir_eval/#module-mir_eval.separation) are widely used in the audio separation literature. One particularity of BSSEval is to compute the metrics after optimally matching the estimates to the true sources through linear distortion filters. This allows the criteria to be robust to some linear mismatches. Apart from the optional evaluation for all possible permutations of the sources, this matching is the reason for most of the computation cost of BSSEval, especially considering it is done for each evaluation window when the metrics are computed on a framewise basis.

For this package, we enabled the option of having _time invariant_ distortion filters, instead of necessarily taking them as varying over time as done in the previous versions of BSS eval. First, enabling this option _significantly reduces_ the computational cost for evaluation because matching needs to be done only once for the whole signal. Second, it introduces much more dynamics in the evaluation, because time-varying matching filters turn out to over-estimate performance. Third, this makes matching more robust, because true sources are not silent throughout the whole recording, while they often were for short windows.

RELEASE FUNCTIONAL DESCRIPTION: This version makes museval compatible with the latest MUSDB package version

- Participant: Antoine Liutkus
- Contact: Antoine Liutkus
- Publication: The 2018 Signal Separation Evaluation Campaign

6.10. Imitates

Indexing and mining Massive Time Series

KEYWORDS: Time Series - Indexing - Nearest Neighbors

FUNCTIONAL DESCRIPTION: Time series indexing is at the center of many scientific works or business needs. The number and size of the series may well explode depending on the concerned domain. These data are still very difficult to handle and, often, a necessary step to handling them is in their indexing. Imitates is a Spark Library that implements two algorithms developed by Zenith. Both algorithms allow indexing massive amounts of time series (billions of series, several terabytes of data).

- Partners: New York University - Université Paris-Descartes
- Contact: Florent Masseglia
- Publication: ParCorr: efficient parallel methods to identify similar time series pairs across sliding windows

6.11. VersionClimber

KEYWORDS: Software engineering - Deployment - Versionning

FUNCTIONAL DESCRIPTION: VersionClimber is an automated system to help update the package and data infrastructure of a software application based on priorities that the user has indicated (e.g. I care more about having a recent version of this package than that one). The system does a systematic and heuristically efficient exploration (using bounded upward compatibility) of a version search space in a sandbox environment (Virtual Env or conda env), finally delivering a lexicographically maximum configuration based on the user-specified priority order. It works for Linux and Mac OS on the cloud.

- Participants: Christophe Pradal, Dennis Shasha, Sarah Cohen-Boulakia and Patrick Valduriez
- Partners: CIRAD - New York University
- Contact: Christophe Pradal
- Publication: VersionClimber: version upgrades without tears
- URL: https://versionclimber.readthedocs.io/

6.12. UMX

open-unmix

KEYWORDS: Source Separation - Audio

SCIENTIFIC DESCRIPTION: Implements state of the art audio/music source separation with DNNs.

This software is intended to serve as a reference in the domain. It has notably been the object of several scientific communications: 1. An Overview of Lead and Accompaniment Separation in Music https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766781/ 2. Music separation with DNNs: making it work (ISMIR 2018 Tutorial) https://sigsep.github.io/ismir2018_tutorial/index.html#/cover

FUNCTIONAL DESCRIPTION: This software implements audio source separation with deep learning, using pytorch and tensorflow frameworks.

It comprises the code for both training and testing the separation networks, in a flexible manner.

Pre and post-processing around the actual deep neural nets include sophisticated specific multichannel filtering operations.

- Authors: Antoine Liutkus, Fabian Robert Stoter and Emmanuel Vincent
- Contact: Antoine Liutkus
- Publication: An Overview of Lead and Accompaniment Separation in Music

7. New Results

7.1. Scientific Workflows

7.1.1. User Steering in Dynamic Workflows

Participants: Renan Souza, Patrick Valduriez.

In long-lasting scientific workflow executions in HPC machines, computational scientists (users) often need to fine-tune several workflow parameters. These tunings are done through user steering actions that may significantly improve performance or improve the overall results. However, in executions that last for weeks, users can lose track of what has been adapted if the tunings are not properly registered. In [18], we address the problem of tracking online parameter fine-tuning in dynamic workflows steered by users. We propose a lightweight solution to capture and manage provenance of the steering actions online with negligible overhead. The resulting provenance database relates tuning data with data for domain, dataflow provenance, execution, and performance, and is available for analysis at runtime. We show how users may get a detailed view of execution, providing insights to determine when and how to tune. We discuss the applicability of our solution in different domains and validate it with a real workflow in Oil and Gas extraction. In this experiment, the user could determine which tuned parameters influence simulation accuracy and performance. The observed overhead for keeping track of user steering actions at runtime is negligible.

7.1.2. ProvLake: Efficient Runtime Capture of Multiworkflow Data

Participants: Renan Souza, Patrick Valduriez.

Computational Science and Engineering (CSE) projects are typically developed by multidisciplinary teams. Despite being part of the same project, each team manages its own workflows, using specific execution environments and data processing tools. Analyzing the data processed by all workflows globally is critical in a CSE project. However, this is hard because the data generated by these workflows are not integrated. In addition, since these workflows may take a long time to execute, data analysis needs to be done at runtime to reduce cost and time of the CSE project. A typical solution in scientific data analysis is to capture and relate workflow runtime data in a provenance database, thus allowing for runtime data analysis. However, such data capture competes with the running workflows, adding significant overhead to their execution. To solve this problem, we introduce a system called ProvLake [39]. While capturing the data, ProvLake logically integrates and ingests them into a provenance database ready for runtime analysis. We validate ProvLake in a real use case in Oil and Gas extraction with four workflows that process 5 TB datasets for a deep learning classifier. Compared with Komadu, the closest competing solution, our approach has much smaller overhead.

7.1.3. Adaptive Caching of Scientific Workflows in the Cloud

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

We consider the efficient execution of data-intensive scientific workflows in the cloud. Since it is common for workflow users to reuse other workflows or data generated by other workflows, a promising approach for efficient workflow execution is to cache intermediate data and exploit it to avoid task re-execution. In [27], we propose an adaptive caching solution for data-intensive workflows in the cloud. Our solution is based on a new scientific workflow management architecture that automatically manages the storage and reuse of intermediate data and adapts to the variations in task execution times and output data size. We evaluated our solution by implementing it in the OpenAlea system and performing extensive experiments on real data with a data-intensive application in plant phenotyping. The results show that adaptive caching can yield major performance gains.

7.2. Query Processing

7.2.1. 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 [12], we propose a system, called SD-TOPK (Secure Distributed TOPK), that encrypts and stores user data in a cloud across a set of nodes, and is able to evaluate top-k queries over the encrypted data. SD-TOPK comes with a novel top-k query processing algorithm that finds a set of encrypted data that is proven to contain the top-k data items. This is done without having to decrypt the data in the nodes where they are stored. In addition, we propose a powerful filtering algorithm that removes the false positives as much as possible without data decryption. We implemented and evaluated the performance of our system over synthetic and real databases. The results show excellent performance for SD-TOPK compared to TA-based approaches.

7.2.2. Parallel Query Rewriting in Key-Value Stores under Single-Key Constraints

Participant: Reza Akbarinia.

Semantic constraints bring important knowledge about the structure and the domain of data. They allow users to better exploit their data thanks to the possibility of formulating high-level queries, which use a vocabulary richer than that of the single sources. However, the constraint-based rewriting of a query may lead to a huge set of new queries, which has a consequent impact on the query answering time.

In [37], we propose a novel technique for parallelizing both the generation and the evaluation of the rewriting set of a query serving as the basis for distributed query evaluation under constraints. Our solution relies on a schema for encoding the possible rewritings of a query on an integer interval. This allows us to generate equi-size partitions of rewritings, and thus to balance the load of the parallel working units that are in charge of generating and evaluating the queries. The experimental evaluation of our technique shows a significant reduction of query rewriting and execution time by means of parallelization.

7.3. Data Analytics

7.3.1. SAVIME: Simulation Data Analysis and Visualization

Participant: Patrick Valduriez.

Limitations in current DBMSs prevent their wide adoption in scientific applications. In order to make scientific applications benefit from DBMS support, enabling declarative data analysis and visualization over scientific data, we present an in-memory array DBMS system called SAVIME. In [34], we describe the system SAVIME, along with its data model. Our preliminary evaluation show how SAVIME, by using a simple storage definition language (SDL) can outperform the state-of-the-art array database system, SciDB, during the process of data ingestion. We also show that it is possible to use SAVIME as a storage alternative for a numerical solver without affecting its scalability.

7.3.2. 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 [20], we propose a parallel solution to construct the state of the art iSAX-based index over billions of time series by making the most of the parallel environment by carefully distributing the work load. Our solution takes advantage of frameworks such as MapReduce or Spark. We provide dedicated strategies and algorithms for a deep combination of parallelism and indexing techniques. We also propose a parallel query processing algorithm that, given a query, exploits the available processing nodes to answer the query in parallel using the constructed parallel index. We implemented our index construction and query processing algorithms, and evaluated their performance over large volumes of data (up to 4 billion time series of length 256, for a total volume of 6 TB). Our experiments demonstrate high performance of our algorithm with an indexing time of less than 2 hours for more than 1 billion time series, while the state of the art centralized algorithm needs more than 5 days. They also illustrate that our approach is able to process 10M queries in less than 140 seconds, while the state of the art centralized algorithm need almost 2300 seconds.

We have implemented our solutions in the Imitates software. The demonstration of Imitates [32] is available at http://imitates.gforge.inria.fr/. The demo visitors are able to choose query time series, see how each algorithm approximates nearest neighbors and compare times in a parallel environment.

7.3.3. Online Correlation Discovery in Sliding Windows of Time Series

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

In some important applications (such as finance, retail, etc.), we need to find correlated time series in a time window, and then continuously slide this window. Doing this efficiently in parallel could help gather important insights from the data in real time. In [30], we address the problem of continuously finding highly correlated pairs of time series over the most recent time window. Our solution, called ParCorr, uses the sketch principle for representing the time series. We implemented ParCorr on top of UPM-CEP, a Complex Event Processing streaming engine developed by our partner Universitat Politecnica de Madrid. Our solution improves the parallel processing of UPM-CEP, allowing higher throughput using less resources. An interesting aspect of our solution is the discovery of time series that are correlated to a certain subset of time series. The discovered correlations can be used to select features for training a regression model for prediction.

7.3.4. Time Series Clustering via Dirichlet Mixture Models

Participants: Khadidja Meguelati, Florent Masseglia.

Dirichlet Process Mixture (DPM) is a model used for clustering with the advantage of discovering the number of clusters automatically and offering nice properties like, *e.g.*, the potential convergence to the actual clusters in the data. These advantages come at the price of prohibitive response times, which impairs its adoption and makes centralized DPM approaches inefficient. In [35], we propose DC-DPM (Distributed Computing DPM), a parallel clustering solution that gracefully scales to millions of data points while remaining DPM compliant, which is the challenge of distributing this process. In [36], we propose HD4C (High Dimensional Data Distributed Dirichlet Clustering), a parallel clustering solution that addresses the curse of dimensionality by distributed computing and performs clustering of high dimensional data such as time series (as a function of time), hyperspectral data (as a function of wavelength) etc. For both methods, our experiments on synthetic and real world data show high performance.

7.4. Machine Learning for Biodiversity Informatics

7.4.1. Phenological Stage Annotation with Deep Convolutional Neural Networks

Participants: Titouan Lorieul, Herve Goeau, Alexis Joly.

Herbarium based phenological research offers the potential to provide novel insights into plant diversity and ecosystem processes under future climate change. The goal of this study [11], conducted in collaboration with US and French ecologists, is to automate the scoring of reproductive phenological stages within a huge amount of digitized herbaria and provide significant resources for the ecological and organismal scientific communities. Specifically, we address three questions: 1) Can fertility, i.e., the presence of reproductive structures, be automatically detected from digitized specimens using deep learning? 2) Are the detection models generalizable to different herbarium data sets? and 3) Is it possible to automatically record stages (i.e., phenophases) within longer phenological events on herbarium specimens? This is the first time that such an analysis has been conducted at this scale, on such a large number of herbarium specimens and species. The results obtained for 7782 species of plants representing angiosperms, gymnosperms, and ferns suggest that it is possible to consider large-scale phenological annotation across broad phylogenetic groups.

7.4.2. Deep Species Distribution Modelling

Participants: Benjamin Deneu, Christophe Botella, Alexis Joly.

Species distribution models (SDM) are widely used for ecological research and conservation purposes. Given a set of species occurrences and environmental data (such as climatic rasters, soil occupation, altitude, etc.), the aim is to infer the spatial distribution of the species over a given territory. In a previous work, we showed that using deep convolutional networks significantly improved predictive performance compared to conventional punctual approaches. We have deepened this methodology with two main contributions. The first one is to extend the model to explicitly take into account species co-occurrences [22]. This is achieved through a new multimodal architecture that allows the joint learning of biotic and abiotic patterns in a common representation space. The second contribution is to experiment deep SDMs at the scale of several tens of thousands of species and tens of millions of occurrences. These contributions were made possible thanks to the use of supercomputer supercomputer Jean Zay (more than 1000 GPUs) of the GENCI national infrastructure.

7.4.3. Evaluation of Species Identification and Prediction Algorithms

Participants: Alexis Joly, Herve Goeau, Christophe Botella, Benjamin Deneu, Fabian Robert Stoter.

We run a new edition of the LifeCLEF evaluation campaign [29] with the involvement of 16 research teams worldwide. The main outcomes of the 2019-th edition are:

- **GeoLifeCLEF.** The main result of the second edition of this challenge [24] is that deep convolutional models outperform the most efficient machine learning models used in ecology (such as random forests or boosted trees). In particular, they are able to transfer knowledge from animals distribution to plant distribution, which had never been shown before.
- **PlantCLEF.** The 2019-th edition of the plant identification challenge [26] was designed to evaluate automated identification on the flora of data deficient regions, tropical ones in particular. It is based on a dataset of 10K species mainly focused on the Guiana shield and the Northern Amazon rainforest, an area known to have one of the greatest diversity of plants and animals in the world. The results reveal that the identification performance in this context is considerably lower than the one obtained on temperate plants of Europe and North America. The performance of convolutional neural networks fall due to the very low number of training images for most species and the higher degree of noise that is occurring in such data.
- **Bird sounds identification.** The 2019-the edition of the BirdCLEF challenge [41] focuses on the difficult task of recognizing all birds vocalizing in omni-directional soundscape recordings. Therefore, the dataset of the previous year has been extended with more than 350 hours of manually annotated soundscapes that were recorded using 30 field recorders in Ithaca (NY, USA). The main outcome is that the recognition performance can be significantly improved thanks to sophisticated data augmentation methods adapted to the problem.

In addition to organizing these challenges, we published a synthesis of the LifeCLEF evaluation campaign since its inception in 2011. This synthesis [44] is part of a larger book published on the occasion of the 20th anniversary of the CLEF international research forum. It highlights the rapid progress that automatic identification has made over the past decade, and allows us to take a step back on the future challenges of this discipline.

7.4.4. Optimal Checkpointing for Heterogeneous Chains: How to Train Deep Neural Networks with Limited Memory

Participants: Alena Shilova, Alexis Joly.

In many deep learning tasks for biodiversity, limited GPU memory is a performance limiting factor. The use of larger image sizes, in particular, is often not possible because the back-propagation algorithm requires storing all network activation maps in memory during for the backward stage. A larger image size could improve the performance of many tasks such as the analysis of digitized herbarium beds, range modeling or early detection of crop weeds in precision agriculture.
In this work [47], done in collaboration with the REAL-OPT team, we introduce a new activation checkpointing method which allows to significantly decrease memory usage when training Deep Neural Networks with the back-propagation algorithm. Similarly to checkpointing techniques coming from the literature on Automatic Differentiation, it consists in dynamically selecting the forward activations that are saved during the training phase, and then automatically recomputing missing activations from those previously recorded. We propose an original computation model that combines two types of activation savings: either only storing the layer inputs, or recording the complete history of operations that produced the outputs (this uses more memory, but requires fewer recomputations in the backward phase), and we provide an algorithm to compute the optimal computation sequence for this model, when restricted to memory persistent sequences. We provide a PyTorch implementation that processes the entire chain, dealing with any sequential DNN whose internal layers may be arbitrarily complex and automatically executing it according to the optimal checkpointing strategy computed given a memory limit. Through extensive experiments, we show that our implementation consistently outperforms existing checkpointing approaches for a large class of networks, image sizes and batch sizes.

7.5. Machine Learning for Audio Heritage Data

Audio data is typically exploited through large repositories. For instance, music right holders face the challenge of exploiting back catalogues of significant sizes while ethnologists and ethnomusicologists need to browse daily through archives of heritage audio recordings that have been gathered across decades. The originality of our research on this aspect is to bring together our expertise in large volumes and probabilistic music signal processing to build tools and frameworks that are useful whenever audio data is to be processed in large batches. In particular, we leverage on the most recent advances in probabilistic and deep learning applied to signal processing from both academia (e.g. Telecom Paris, PANAMA & Multispeech Inria project-teams, Kyoto University) and industry (e.g. Mitsubishi, Sony), with a focus towards large scale community services.

7.5.1. Setting the State of the Art in Music Demixing

Participants: Fabian-Robert Söter, Antoine Liutkus.

We have been very active in the topic of music demixing, with a prominent role in defining the state of the art in this domain. This has been achieved through several means.

- In the previous years, we have been organizing the Signal Separation Evaluation Challenge (SiSEC), an international event in the signal processing community that is held since 2007. Its objective is to bring together researchers to evaluate their algorithms on music separation/demixing on the same data and with the same metrics. From 2016 to 2019, A. Liutkus was the lead chair of SiSEC.
- We have developed the *open-unmix* [19] software, which is a reference implementation for music source separation. For the first time, it makes it possible for any researcher to use and improve a state-of-the art implementation (MIT-licensed) in the domain. In terms of performance, open-unmix mathes the best results we observed over the years as the organizers of SiSEC. The open-unmix software won the second place at the Global Pytorch Summer Hackaton 2019 organized by FaceBook.

The *pro* private version of this software is currently under active development for transfer to industry.

- In [6], we present the field to the non-specialist researcher, in a wide-audience scientific magazine. We are also core contributors of the audio section for the position paper on the use of AI for the creation industry [48].

7.5.2. Generative Modelling for Audio

Participants: Antoine Liutkus, Fabian-Robert Söter, Mathieu Fontaine.

Discriminative training for audio signal processing is inherently limited in the sense that it boils down to assuming that the target signals are present in the input, and can be recovered through some kind of filtering, even if this involves sophisticated deep models. We move forward to a new paradigm for signal processing, in which the observed signals and time series are not assumed to comprise the totality of the target, but rather some arbitrarily degraded version of it. The objective then can be understood as *generating new content given this input*. For instance, bandwidth extension may be thought of as audio super-resolution.

Our research on generative modelling concerns both methodological/theoretical aspects and applied research. On the former, we introduce the Sliced Wasserstein Flow in our ICML paper [33], which enables the optimal transport of particles from two probability spaces in a principled way. On the latter, we study the combination of heavy-tailed probabilistic models with generative audio models for source separation in [31], [25].

Our strategy is to go beyond our current expertise on music demixing to address the new and very active topics of audio style transfer and enhancement, with large scale applications for the exploitation and repurposing of large audio corpora.

7.5.3. Robust Probabilistic Models for Time-series

Participants: Mathieu Fontaine, Antoine Liutkus, Fabian-Robert Söter.

Processing large amounts of data for denoising or analysis comes with the need to devise models that are robust to outliers and permit efficient inference. For this purpose, we advocate the use of non-Gaussian models for this purpose, which are less sensitive to data-uncertainty. Our contributions on this topic can be split in two parts. First, we develop new filtering methods that go beyond least-squares estimation. In collaboration with researchers from Telecom Paris, we introduce several methods that generalize least-squares Wiener filtering to the case of α-stable processes [2]. This work is currently also under review as a journal paper. Second, as mentioned in the previous section, we have been working on generative models for audio, with the particular twist that the deep models we consider are trained probabistically under α-stable assumptions. This has the remarkable effect of significantly augmenting robustness [31], [25].

8. Bilateral Contracts and Grants with Industry

8.1. SAFRAN (2018-2019)

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).

8.2. INA (2019-2022)

Participants: Quentin Leroy, Alexis Joly.

The PhD of Quentin Leroy is funded in the context of an industrial contract (CIFRE) with INA, the French company in charge of managing the French TV archives and audio-visual heritage. The goal of the PhD is to develop new methods and algorithms for the interactive learning of new classes in INA archives.

9. Partnerships and Cooperations

9.1. National Initiatives

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

Participants: Alexis Joly, 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 and biodiversity data sharing.

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

Participants: Julien Champ, 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.

9.1.3. Others

9.1.3.1. Pl@ntNet InriaSOFT consortium, 80 Keuro / year

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

This contract between four research organisms (Inria, INRA, IRD and CIRAD) aims at sustaining the Pl@ntNet platform in the long term. It has been signed in November 2019 in the context of the InriaSOFT national program of Inria. Each partner subscribes a subscription of 20K euros per year to cover engineering costs for maintenance and technological developments. In return, each partner has one vote in the steering committee and the technical committee. He can also use the platform in his own projects and benefit from a certain number of service days within the platform. The consortium is not fixed and is not intended to be extended to other members in the coming years.

9.1.3.2. Ministry of Culture, 130 Keuro

Participants: Alexis Joly, Jean-Christophe Lombardo.

Two contracts have been signed with the ministry of culture to adapt, extend and transfer the content-based image retrieval engine of Pl@ntNet ("Snoop") toward two major actors of the French cultural domain: the French National Library (BNF) and the French National institute of audio-visual (INA).

9.1.3.3. INRA/Inria PhD program, 100 Keuro

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).

9.2. European Initiatives

9.2.1. FP7 & H2020 Projects

9.2.1.1. 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
Partners: 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. In this project, Zenith is in charge of designing and implementing the components for analytics and parallel query processing.

9.2.1.2. Cos4Cloud

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

Project title: Cos4Cloud
Instrument: H2020
Duration: 2019 - 2022
Total funding: 5 Meuros (Zenith: 400Keuros)
Coordinator: CSIC (Spain)
Partners: The Open University, CREAF, Bineo, EarthWatch, SLU, NKUA, CERT, Bineo, ECSA.
Inria contact: Alexis Joly

Cos4Cloud will integrate citizen science in the European Open Science Cloud (EOSC) through the co-design of innovative services to solve challenges faced by citizen observatories, while bringing Citizen Science (CS) projects as a service for the scientific community and the society and providing new data sources. In this project, Zenith is in charge of developing innovative web services related to automated species identification, location-based species prediction and training data aggregation services.

9.3. International Initiatives

The team has two PhD students funded by an Algerian initiative ("Bourses d'excellence Algériennes "):

- Khadidja Meguelati, since 2016, "Massively Distributed Time Series Clustering via Dirichlet Mixture Models"
- Lamia Djebour, since 2019, "Parallel Time Series Indexing and Retrieval with GPU architectures"

9.3.1. Inria International Labs

In the context of LIRIMA, P. Valduriez gave a one week course in big data at IMSP, Bénin, in march, and an online seminar on Blockchain on 13 dec at Inria Rennes.

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

9.3.2.1. 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.

9.3.3. Inria International Partners

9.3.3.1. 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, Cheryl Porter, Gil Nelson), Harvard (Charles Davis), UCSB (Susan Mazer).
- Asia: National Univ. of Singapore (Beng Chin Ooi, Stéphane Bressan), Wonkwang University, Korea (Kwangjin Park), Kyoto University (Japan), Tokyo University (Hiroyoshi Iwata)
- 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), UK center for hydrology and ecology (Tom August)
- 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)

9.3.4. Participation in Other International Programs

9.3.4.1. Inria International Chairs

Dennis Shasha (NYU)

Title: Data Science in a Dynamic World

International Partner: New York University (NYU), USA

Duration: 2015 - 2019

Start year: 2015

Many fundamental problems in natural science from astronomy to microbiology require data from heterogeneous sources, hence giving rise to a new “data science”. The basic workflow is to collect that data, form some kind of similarity metric between objects based on each data source, and then weight those different similarity metrics for some data analysis task. The goal is to gain actionable insight such as the cause of some symptoms, the function of some protein, or the likely source of some epidemic. Most often this is conceived of as "do-it-once" exercise. However, as data acquisition techniques improve, data may evolve continuously. When that happens the question is whether new revised insights can be obtained in a close to real time manner. Whether this is possible depends on the qualities of the new data, the weighting of the data sources, and the machine learning algorithms used. This project addresses data science in a dynamic world, aiming to find fast and minimalist methods to update insights as new data appears. This will result in new data management algorithms that will be implemented in tools and validated in the context of real data, in particular biology data.

9.3.5. Visits of International Scientists

- Renan Souza (COPPE/UFRJ and IBM,Brazil): “Providing Online Data Analytical Support for Humans in the Loop of Computational Science and Engineering Applications” on Jan 15.
- Youcef Djenouri (Norwegian University of Science and Technology, Trondheim): “Urban traffic outlier detection” on Feb 14.
- Dennis Shasha (NYU) “Bounce Blockchain: a secure, energy-efficient permission less blockchain” on May 27.
- Alvaro Coutinho (COPPE/UFRJ, Brazil): “Some Reflections on Predictive Science in Geophysical Applications” on Nov 20.
- Marta Mattoso (COPPE/UFRJ, Brazil): “Adding Provenance Data to Experiments: From Computational Science to Deep Learning” on Nov 20.
- Eduardo Ogasawara, (CEFET-RJ, Brazil): “Event Detection in Time Series” on Nov 20.
- Heraldo Borges (CEFET-RJ, Brazil): “Discovering Patterns in Restricted Space-Time Datasets” on Nov 20.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

10.1.1.1. General Chair, Scientific Chair

- Florent Masseglia: co-organizer of the first edition of the ADITCA workshop, at CLOSER 2019: http://closer.scitevents.org/ADITCA.aspx?y=2019
- Florent Masseglia:

  of the second edition of the ADITCA workshop, at DATA 2019: http://www.dataconference.org/ADITCA.aspx

10.1.1.2. Member of the Organizing Committees

- A. Joly: organizing committee of the international conference CLEF 2019 and the chair of the LifeCLEF track, Lugano, sept. 2019 (http://clef2019.clef-initiative.eu/)

10.1.2. Scientific Events: Selection

10.1.2.1. Member of the Conference Program Committees

- Artificial Intelligence & Knowledge Engineering (AIKE), 2019: F. Masseglia

- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (PKDD), 2019: F. Masseglia
- Int. Conf. on Data Science, Technology and Applications (DATA), 2019: F. Masseglia
- Int. Conf. on Information Management and Big Data (SIMBig), 2019: F. Masseglia
- IEEE Int. Conf. on Data Mining (ICDM), 2019: F. Masseglia
- ACM Symposium on Applied Computing (ACM SAC), Data Mining Track (DM), 2019: F. Masseglia
- ACM Symposium on Applied Computing (ACM SAC), Data Stream Track (DS), 2019: F. Masseglia
- Extraction et Gestion des Connaissances (EGC), 2019: F. Masseglia
- Int. Conf. on Extending DataBase Technologies (EDBT), 2019: E. Pacitti
- Int. Conf. on Multimedia Retrieval (ICMR), 2019: A. Joly
- Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2019: A. Joly, A. Liutkus
- Int. Conf. on Computer Vision (CVPR), 2019: A. Joly
- Int. Conf. and Labs of the Evaluation Forum (CLEF), 2019: A. Joly
- European. Conf. on Information Retrieval (ECIR), 2019: A. Joly
- EAI Int. Conf. on e-Infrastructure and e-Services for Developing Countries (AFRICOMM 2019): P. Valduriez.
- Bases de Données Avancées (BDA), 2019: E. Pacitti, R. Akbarinia
- IEEE/ACM Int. Symposium in Cluster, Cloud, and Grid Computing (CCGrid) 2019: Esther Pacitti

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- VLDB Journal: P. Valduriez.
- Transactions on Large Scale Data and Knowledge Centered Systems: R. Akbarinia.
- Distributed and Parallel Databases: E. Pacitti, P. Valduriez.
- Book series “Data Centric Systems and Applications” (Springer): P. Valduriez.
- Plant Methods: C. Pradal.

10.1.3.2. Reviewer - Reviewing Activities

- EVISE Future Generation Computer Systems Journal (FGCS): F. Masseglia
- Information Systems (IS): T. Mondal, F. Masseglia
- Distributed and Parallel Databases (DAPD): E. Pacitti, P. Valduriez
- IEEE Transactions on Knowledge and Data Engineering (TKDE): R. Akbarinia, F. Masseglia
- IEEE Transactions on Industrial Informatics: R. Akbarinia
- Knowledge and Information Systems (KAIS): R. Akbarinia
- Plant methods: A. Joly
- Machine Learning: A. Joly
- Pattern Recognition Letters: 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

- IEEE Signal Processing Letters: A. Liutkus
- Frontiers in Plant Science: C. Pradal
- Neural Information Processing Systems (NeurIPS): A. Liutkus
- Int. Conf. on Machine Learning (ICML): A. Liutkus
- Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP): A. Liutkus

10.1.4. Invited Talks

- E.Pacitti: Inaugural lecture, “Data Processing: an evolutionary and multidisciplinary perspective”, CEFET/RJ, Rio de Janeiro on 12 August 2019.
- A. Joly: "AI for plant phenology" on 17 January at a workshop of the University of Florida (keynote); "AI for plant biodiversity monitoring" on 7 November at HPC-AI-BigData Convergence days (Conv’ 2019); "An end-to-end deep learning approach to biodiversity monitoring", on 12 December at British Ecological Society conference (BES 2019);
- A. Liutkus: tutorial on "music source separation" at Int. Symposium on Music Information Retrieval (ISMIR 2018).
- F. Masseglia: “Données, humanités et démarche scientifique”. Panel at the Inria Science Days (Journées Scientifiques Inria), on June 6, Lyon; “Analyse de données scientifiques”. IRD, on June 27, Montpellier. “Analyse de données à grande échelle”. TILECS Workshop. on July 4, Grenoble.
- P. Valduriez: "The Case for Hybrid Transaction Analytical Processing" on 25 April at IBM Research Brazil, Rio de Janeiro, Brazil; "Blockchain 2.0: opportunities and risks" on 14 Nov at Online Franco-African LIRIMA Seminar, Inria, Rennes; "Data-intensive Science" on 6 Nov. at HPC- AI-BigData Convergence Days (Conv’2019), EDF Lab Paris-Saclay; tutorial "NewSQl: principles, systems and trends" on 12 Dec at IEEE Bigdata 2019; participation in panel "Big Data Heterogeneity Challenges" on 11 Dec at IEEE Bigdata 2019; "Scalable transaction and polystore data management in LeanXcale" on 6 Dec at UC Berkeley and on 13 Dec at UCLA and UC Irvine.
- C. Pradal: "Multiscale plant modelling and Phenotyping" on 8 october at Tottori University and on 16 october at Nagoya University, Japan; workshop on plant modelling on 28 october at Tokyo University, Japan.

10.1.5. Leadership within the Scientific Community

- A. Joly: Scientific manager of the LifeCLEF research forum.
- A. Liutkus: elected member of the IEEE Technical Committee on Audio and Acoustic Signal Processing.
- P. Valduriez: President of the Steering Committee of the BDA conference.

10.1.6. Scientific Expertise

- F. Masseglia: expert for the HCERES evaluation of the DAVID Lab (UVSQ). January 2019.
- R. Akbarinia: expert for the French National Research Agency (ANR).
- A: Joly: scientific advisory board of the ANR program "AI for biodiversity", expert for the National HPC grand equipment (GENCI) "grand challenge" program, reviewer for STIC AmSud international program.
- E. Pacitti: reviewer for STIC AmSud international program.
- P. Valduriez: reviewer for STIC AmSud international program.
- P. Valduriez: reviewer for NSERC (Canada).
- C. Pradal: member of CSS EGBIP (Commissions Scientifiques Spécialisées) INRA.

10.1.7. Research Administration

- A. Joly: Technical director of the InriaSOFT consortium Pl@ntNet and representative of Inria in the steering committee.
- 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).

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, 50 students, UM2

IG4: Distributed Databases, Big Data, 80h , level M1, Polytech' Montpellier, 50 students UM2

IG5: Iot- Information Management, 27h, level M2, Polytech' Montpellier, 15 students, UM2

Industry internship supervision and committees, level M2, Polytech' Montpellier, 40h

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

InnObs technical school: Innovations in the observation of seasonal biological events and associated data management, 6h, professionals.

Antoine Liutkus

University Paul Valery (Montpellier): multidimensional data analysis, 15h, level M1

Polytech' Montpellier: Audio Machine Learning, 1.5h, level M1

10.2.2. Supervision

PhD & HDR:

HDR: Reza Akbarinia, Parallel Techniques for Big Data Analytics, Univ. Montpellier, 24 May.

PhD: Renan Souza, Massively Distributed Clustering, UFRJ, Brazil, 17 Dec. Advisors: Marta Mattoso (UFRJ), Patrick Valduriez.

PhD: Mathieu Fontaine, Alpha-stable models for signal processing, IAEM, Nancy, 18 July. Advisors: Roland Badeau (Telecom Paris), Antoine Liutkus.

PhD: Christophe Botella, Large-scale Species Distribution Modelling based on Citizen Science data, Montpellier, 8 October. started Oct 2016, Univ. Montpellier. Advisors: Alexis Joly, François Munoz (univ. of Grenoble), Pascal Monestiez (INRA).

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: Heraldo Borges, Discovering Tight Space-Time Sequences, started Oct 2018, Univ. Montpellier. Advisors: Esther Pacitti, Eduardo Ogaswara.

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: Lamia Djebour, Parallel Time Series Indexing and Retrieval with GPU architectures, started Oct 2019, Univ. Montpellier. Advisors: Reza Akbarinia, Florent Masseglia.

PhD in progress: Quentin Leroy, Active learning of unknown classes, started Oct 2019, Univ. Montpellier, Industrial contract with INA, Advisors: Alexis Joly

10.2.3. Juries

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

- R. Akbarinia: Chao Zhang (Univ. Clermont Auvergne, reviewer)
- A. Joly: Christophe Botella (Univ. of Montpellier, advisor)
- F. Masseglia: Rebecca Pontes Salles (Master thesis of 2.5 years, CEFET, Rio de Janeiro), Ricardo Sperandio (Univ. Rennes, reviewer), Thibault Desprez (Univ. Bordeaux, reviewer).
- E. Pacitti: Daniel de Oliveira Junior (Univ. Federal Fluminense, Brazil)
- E. Pacitti: Heraldo Borges Phd qualification (CEFET, Brazil)
- P. Valduriez: Alexandru Costan (HDR, ENS Rennes, reviewer), Patricio Cerda (Univ. Paris Saclay, reviewer), Renan Souza (UFRJ, Rio de Janeiro, advisor)
- E. Pacitti: Selection Committe for professor position (University of Montpellier)

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.

10.3.2. Articles and contents

- F. Masseglia is co-author of [21], [40] on the feedback of Class’Code after 3 years of project and on a vision about future work for computational thinking education and computer science popularisation.
- A. Joly has given several interviews to different media giving rise to web articles about Pl@ntNet (see e.g. Google news with keyword Pl@ntNet).

10.3.3. Education

Computer science is, for the first time in France, an official discipline teached in high school (Lycée) with the common course about "Sciences Numériques et Technologie". As written by Inira’s CEO Bruno Sportisse: *"For over a decade, the institute has carried out actions that have paved the way with the firm conviction that "training in and through digital technology" is strategic: I am thinking here, in particular, of the Class’Code project."*

F. Masseglia is the initiator, with Serge Abiteboul, of the program called "1 scientifique — 1 classe : Chiche !" with the goal of reaching *all* the students of a specific level. This massive plan should concern all scientists at Inria and our partners in France.

F. Masseglia was ambassador of Inria for the Science Celebration Day (Fête de la science):

- https://www.youtube.com/watch?v=13957C9FxVg
- https://www.youtube.com/watch?v=ypnQe91Pztc

F. Masseglia gave a one day training on the Thymio robot for education in the media library networkprofessionals. February 1. Montpellier. 18 attendees.

F. Masseglia gave a one day training to teachers of French National Education on computational thinking. March 12. Montpellier. 20 attendees.

F. Masseglia gave a one day training to teachers in pediatric hospitals (CHU Arnaud de Villeneuve). February 2. Montpellier. 10 attendees.

F. Masseglia gave two days of training to computer science, robotics and computational thinking to reference teachers ("conseillers pédagogiques"). April 15 & 16. Montpellier. 20 attendees.

F. Masseglia gave two days of training to the media library network professionals on Poppy Ergo Jr. June 28 & September 6. Montpellier. 11 attendees.

F. Masseglia is a member of the scientific committee of a conference cycle on "Science and Society" organised with MSHSud.

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 Dec 11.

A. Joly was a member of the scientific advisory board of the AI Family Challenge organized by the NGO Iridescent that "supports girls, children, and their families to identify problems in their communities and find technology based solutions".

He organized an educational outing at the Montpellier Botanical Garden as part of the summer school organized each year by Polytech Montpellier (June 11). He participated in the organization and animation of the "wild salad" training organized by the association "Les écologistes de l'Euzière" 13-15 March.

10.3.4. Interventions

F. Masseglia participated in the LIRMM events on science popularisation ("Fête de la science" on Feb 19, "accueil de stagiaires" on Oct 11), Montpellier.

A. Joly participated in: the animation of the Inria stand at the SIDO exhibition (April 10-11, Lyon); the Pl@ntNet launch ceremony in Costa Rica in the presence of the Minister of Research and decision-makers (April 20).

E. Pacitti participated in Polytech'Montpellier International Summer School (Flow) on the subject of Data Science - Plant Phenotyping.

A. Joly participated in the day "Ramène ta science" co-organized by TelaBotanica and "Sciences Avenir" associations at the Halle Tropisme in Montpellier (as part of the Labbota initiative). He co-facilitated a participatory workshop and a debate with citizens on the theme of participatory science.

10.3.5. 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 provides support for education to computer science from the history point of view.

A. Joly actively participates to the design and development of all Pl@ntNet dissemination tools in particular Pl@ntNet web site that contains contents for the press, articles for the general public, tutorials of Pl@ntNet tools, guidelines for users of the API, etc.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] R. AKBARINIA. *Parallel Techniques for Big Data Analytics*, Université de Montpellier, May 2019, Habilitation à diriger des recherches, https://hal-lirmm.ccsd.cnrs.fr/tel-02169414

[2] M. FONTAINE. *Alpha-stable processes for signal processing*, Université de Lorraine, June 2019, https://tel.archives-ouvertes.fr/tel-02188304

[3] C. PRADAL. *Dataflow architecture for modular and generic plant simulation systems*, Université de Montpellier, July 2019, https://tel.archives-ouvertes.fr/tel-02193606

[4] R. SOUZA. *Supporting User Steering In Large-Scale Workflows With Provenance Data*, UFRJ, Rio de Janeiro, December 2019, https://hal-lirmm.ccsd.cnrs.fr/tel-02418022

Articles in International Peer-Reviewed Journals

[5] R. ALBASHA, C. FOURNIER, C. PRADAL, M. CHELLE, J. ALEJANDO PRIETO, G. LOUARN, T. SIMONNEAU, E. LEBON. *HydroShoot: a functional-structural plant model for simulating hydraulic structure, gas and energy exchange dynamics of complex plant canopies under water deficit - application to grapevine (Vitisvinifera L.)*, in "in silico Plants", June 2019 [*DOI :* 10.1093/INSILICOPLANTS/DIZ007], https://hal.inria.fr/hal-02253260

[6] E. CANO, D. FITZGERALD, A. LIUTKUS, M. D. PLUMBLEY, F. ROBERT-STÖTER. *Musical Source Separation: An Introduction*, in "IEEE Signal Processing Magazine", January 2019, vol. 36, n° 1, pp. 31-40 [*DOI :* 10.1109/MSP.2018.2874719], https://hal.inria.fr/hal-01945345

[7] T.-W. CHEN, L. 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", April 2019, vol. 70, n° 9, pp. 2523-2534 [*DOI :* 10.1093/JXB/ERY309], https://hal.inria.fr/hal-01895279

[8] M. FONTAINE, R. BADEAU, A. LIUTKUS. *Separation of Alpha-Stable Random Vectors*, in "Signal Processing", January 2020, 107465 p. [*DOI :* 10.1016/J.SIGPRO.2020.107465], https://hal.inria.fr/hal-02433213

[9] N. N. GAUDIO, A. E. G. ESCOBAR-GUTIÉRREZ, P. CASADEBAIG, J. EVERS, F. F. GERARD, G. LOUARN, N. COLBACH, S. MUNZ, M. LAUNAY, H. MARROU, R. BARILLOT, P. HINSINGER, J.-E. BERGEZ, D. COMBES, J.-L. DURAND, E. FRAK, L. PAGÈS, C. PRADAL, S. SAINT-JEAN, W. W. VAN DER WERF, E. JUSTES. *Current knowledge and future research opportunities for modeling annual crop mixtures. A review*, in "Agronomy for Sustainable Development", April 2019, vol. 39, n° 2 [*DOI :* 10.1007/s13593-019-0562-6], https://hal.inria.fr/hal-02228974

[10] J. LIU, N. MORENO LEMUS, E. PACITTI, F. PORTO, P. VALDURIEZ. *Parallel Computation of PDFs on Big Spatial Data Using Spark*, in "Distributed and Parallel Databases", 2019, pp. 1-38, forthcoming [*DOI :* 10.1007/s10619-019-07260-3], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02045144

[11] T. LORIEUL, K. D. PEARSON, E. R. ELLWOOD, H. GOËAU, J. MOLINO, P. W. SWEENEY, J. M. YOST, J. SACHS, G. NELSON, P. S. SOLTIS, P. BONNET, A. JOLY, E. MATA-MONTERO. *Toward a large-scale and deep phenological stage annotation of herbarium specimens: Case studies from temperate, tropical, and equatorial floras*, in "Applications in Plant Sciences", March 2019, vol. 7, n° 3, e01233 [*DOI :* 10.1002/aps3.1233], https://hal.umontpellier.fr/hal-02137748

[12] S. MAHBOUBI, R. AKBARINIA, P. VALDURIEZ. *Privacy-Preserving Top-k Query Processing in Distributed Systems*, in "Transactions on Large-Scale Data- and Knowledge-Centered Systems", 2019 [*DOI :* 10.1007/978-3-662-60531-8_1], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265730

[13] K. PARK, A. JOLY, P. VALDURIEZ. *DZI: An Air Index for Spatial Queries in One-dimensional Channels*, in "Data and Knowledge Engineering", November 2019, vol. 124, 101748 p. [*DOI :* 10.1016/j.datak.2019.101748], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02386429

[14] R. PEREZ, C. FOURNIER, L. CABRERA-BOSQUET, S. ARTZET, C. PRADAL, N. BRICHET, T.-W. CHEN, R. CHAPUIS, C. WELCKER, F. TARDIEU. *Changes in the vertical distribution of leaf area enhanced light interception efficiency in maize over generations of selection*, in "Plant, Cell and Environment", June 2019, vol. 42, n° 7, pp. 2105-2119 [*DOI :* 10.1111/pce.13539], https://hal.inria.fr/hal-02228393

[15] C. PRADAL, S. COHEN-BOULAKIA, P. VALDURIEZ, D. SHASHA. *VersionClimber: version upgrades without tears*, in "Computing in Science & Engineering", 2019, vol. 21, n° 5, pp. 87-93, forthcoming [*DOI :* 10.1109/MCSE.2019.2921898], https://hal.inria.fr/hal-02262591

[16] F. REYES, B. PALLAS, C. PRADAL, F. VAGGI, D. ZANOTELLI, T. MARCO, D. GIANELLE, E. COSTES. *MuSCA: a multi-scale source-sink carbon allocation model to 2 explore carbon allocation in plants. An application on static apple-tree*, in "Annals of Botany", October 2019 [*DOI :* 10.1093/aob/mcz122], https://hal.inria.fr/hal-02262908

[17] F. ROBERT-STÖTER, S. CHAKRABARTY, B. EDLER, E. A. P. HABETS. *CountNet: Estimating the Number of Concurrent Speakers Using Supervised Learning*, in "IEEE/ACM Transactions on Audio, Speech and Language Processing", February 2019, vol. 27, n° 2, pp. 268-282 [*DOI :* 10.1109/TASLP.2018.2877892], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02010805

[18] R. SOUZA, V. SILVA, J. J. CAMATA, A. L. G. A. COUTINHO, P. VALDURIEZ, M. MATTOSO. *Keeping Track of User Steering Actions in Dynamic Workflows*, in "Future Generation Computer Systems", October 2019, vol. 99, pp. 624-643, https://arxiv.org/abs/1905.07167 [*DOI :* 10.1016/j.future.2019.05.011], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02127456

[19] F.-R. STÖTER, S. UHLICH, A. LIUTKUS, Y. MITSUFUJI. *Open-Unmix - A Reference Implementation for Music Source Separation*, in "Journal of Open Source Software", September 2019, vol. 4, n° 41, 1667 p. [*DOI :* 10.21105/joss.01667], https://hal.inria.fr/hal-02293689

[20] D.-E. YAGOUBI, R. AKBARINIA, F. MASSEGLIA, T. PALPANAS. *Massively Distributed Time Series Indexing and Querying*, in "IEEE Transactions on Knowledge and Data Engineering", 2019, pp. 1-14 [*DOI :* 10.1109/TKDE.2018.2880215], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02197618

Articles in National Peer-Reviewed Journals

[21] C. ATLAN, J.-P. ARCHAMBAULT, O. BANUS, F. BARDEAU, A. BLANDEAU, A. COIS, M. COURBIN-COULAUD, G. GIRAUDON, S.-C. LEFÈVRE, V. LETARD, B. MASSE, F. MASSEGLIA, B. NINASSI, S. DE QUATREBARBES, M. ROMERO, D. ROY, T. VIÉVILLE. *Apprentissage de la pensée informatique : de la formation des enseignant·e·s à la formation de tou·te·s les citoyen·ne·s*, in "Revue de l'EPI (Enseignement Public et Informatique)", June 2019, https://arxiv.org/abs/1906.00647 , https://hal.inria.fr/hal-02145478

Invited Conferences

[22] B. DENEU, M. SERVAJEAN, C. BOTELLA, A. JOLY. *Evaluation of Deep Species Distribution Models using Environment and Co-occurrences*, in "CLEF 2019 - Conference and Labs of the Evaluation Forum", Lugano, Switzerland, September 2019, https://arxiv.org/abs/1909.08825 , https://hal.inria.fr/hal-02290310

[23] M. M. ZEKENG NDADJI, M. T. TCHENDJI, D. PARIGOT. *A Projection-Stable Grammatical Model to Specify Workflows for their P2P and Artifact-Centric Execution*, in "CRI'2019 - Conférence de Recherche en Informatique", Yaoundé, Cameroon, December 2019, https://hal.inria.fr/hal-02375958

International Conferences with Proceedings

[24] C. BOTELLA, M. SERVAJEAN, P. BONNET, A. JOLY. *Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences*, in "CLEF 2019 - Conference and Labs of the Evaluation Forum", Lugano, Switzerland, 2019, vol. CEUR Workshop Proceedings, n^{o} 2380, https://hal.archives-ouvertes.fr/hal-02190170

[25] M. FONTAINE, A. A. NUGRAHA, R. BADEAU, K. YOSHII, A. LIUTKUS. *Cauchy Multichannel Speech Enhancement with a Deep Speech Prior*, in "EUSIPCO 2019 - 27th European Signal Processing Conference", Coruña, Spain, September 2019, https://hal.telecom-paristech.fr/hal-02288063

[26] H. GOËAU, P. BONNET, A. JOLY. *Overview of LifeCLEF Plant Identification task 2019: diving into data deficient tropical countries*, in "CLEF 2019 - Conference and Labs of the Evaluation Forum", Lugano, Switzerland, Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, CEUR, September 2019, vol. 2380, pp. 1-13, https://hal.umontpellier.fr/hal-02283184

[27] G. HEIDSIECK, D. DE OLIVEIRA, E. PACITTI, C. PRADAL, F. TARDIEU, P. VALDURIEZ. *Adaptive Caching for Data-Intensive Scientific Workflows in the Cloud*, in "DEXA 2019 - 30th International Conference on Database and Expert Systems Applications", Linz, Austria, August 2019, https://hal.inria.fr/hal-02174445

[28] A. JOLY, H. GOËAU, C. BOTELLA, S. KAHL, M. POUPARD, M. SERVAJEAN, H. GLOTIN, P. BONNET, W.-P. VELLINGA, R. PLANQUÉ, J. SCHLÜTER, F.-R. STÖTER, H. MÜLLER. *LifeCLEF 2019: Biodiversity Identification and Prediction Challenges*, in "ECIR 2019 - 41st European Conference on IR Research", Cologne, Germany, L. AZZOPARDI, B. STEIN, N. FUHR, P. MAYR, C. HAUFF (editors), April 2019, vol. LNCS, n^{o} 11438, pp. 275-282 [*DOI :* 10.1007/978-3-030-15719-7_37], https://hal.umontpellier.fr/hal-02273257

[29] A. JOLY, H. GOËAU, C. BOTELLA, S. KAHL, M. SERVAJEAN, H. GLOTIN, P. BONNET, R. PLANQUÉ, F. ROBERT-STÖTER, W.-P. VELLINGA, H. MÜLLER. *Overview of LifeCLEF 2019: Identification of Amazonian Plants, South & North American Birds, and Niche Prediction*, in "CLEF 2019 - Conference and Labs of the Evaluation Forum", Lugano, Switzerland, F. CRESTANI, M. BRASCHER, J. SAVOY, A. RAUBER, H. MÜLLER, D. E. LOSADA, G. H. BÜRKI, G. H. BÜRKI, L. CAPPELLATO, N. FERRO (editors), Experimental

IR Meets Multilinguality, Multimodality, and Interaction, August 2019, vol. LNCS, n^{o} 11696, pp. 387-401 [*DOI :* 10.1007/978-3-030-28577-7_29], https://hal.umontpellier.fr/hal-02281455

[30] B. KOLEV, R. AKBARINIA, R. JIMENEZ-PERIS, O. LEVCHENKO, F. MASSEGLIA, M. PATINO, P. VALDURIEZ. *Parallel Streaming Implementation of Online Time Series Correlation Discovery on Sliding Windows with Regression Capabilities*, in "CLOSER 2019 - 9th International Conference on Cloud Computing and Services Science", Heraklion, Greece, 2019, pp. 681-687, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265729

[31] S. LEGLAIVE, U. SIMSEKLI, A. LIUTKUS, L. GIRIN, R. HORAUD. *Speech enhancement with variational autoencoders and alpha-stable distributions*, in "ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing", Brighton, United Kingdom, IEEE, 2019, pp. 541-545, https://arxiv.org/abs/1902.03926 [*DOI :* 10.1109/ICASSP.2019.8682546], https://hal.inria.fr/hal-02005106

[32] O. LEVCHENKO, B. KOLEV, D.-E. YAGOUBI, D. SHASHA, T. PALPANAS, P. VALDURIEZ, R. AKBARINIA, F. MASSEGLIA. *Distributed Algorithms to Find Similar Time Series*, in "ECML-PKDD : European Conference on Machine Learning and Knowledge Discovery in Databases", Wurtzbourg, Germany, 2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265726

[33] A. LIUTKUS, U. Ş. IMŞEKLI, S. MAJEWSKI, A. DURMUS, F. ROBERT-STÖTER. *Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions*, in "ICML 2019 - 36th International Conference on Machine Learning", Long Beach, United States, June 2019, https://hal.inria.fr/hal-02191302

[34] H. LUSTOSA, F. PORTO, P. VALDURIEZ. *SAVIME: A Database Management System for Simulation Data Analysis and Visualization*, in "SBBD 2019 - Simpósio Brasileiro de Banco de Dados", Fortaleza, Brazil, SBC, October 2019, pp. 1-12, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02266483

[35] K. MEGUELATI, B. FONTEZ, N. HILGERT, F. MASSEGLIA. *Dirichlet Process Mixture Models made Scalable and Effective by means of Massive Distribution*, in "SAC 2019 - 34th Symposium On Applied Computing", Limassol, Cyprus, ACM/SIGAPP, 2019, pp. 502-509 [*DOI :* 10.1145/3297280.3297327], https://hal.archives-ouvertes.fr/hal-01999453

[36] K. MEGUELATI, B. FONTEZ, N. HILGERT, F. MASSEGLIA. *High Dimensional Data Clustering by means of Distributed Dirichlet Process Mixture Models*, in "IEEE International Conference on Big Data (IEEE BigData)", Los-Angeles, United States, December 2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02364411

[37] O. RODRIGUEZ, R. AKBARINIA, F. ULLIANA. *Querying Key-Value Stores under Single-Key Constraints: Rewriting and Parallelization*, in "RuleML+RR 2019 - the 3rd International Joint Conference on Rules and Reasoning", Bolzano, Italy, 2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02195593

[38] R. SOUZA, L. AZEVEDO, V. LOURENÇO, E. SOARES, R. THIAGO, R. BRANDÃO, D. CIVITARESE, E. VITAL BRAZIL, M. MORENO, P. VALDURIEZ, M. MATTOSO, R. CERQUEIRA, M. NETTO. *Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering*, in "Workshop on Workflows in Support of Large-scale Science (WORKS), co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)", Denver, United States, ACM, November 2019, 10 p. , https://hal-lirmm.ccsd.cnrs.fr/lirmm-02335500

[39] R. SOUZA, L. AZEVEDO, R. THIAGO, E. SOARES, M. NERY, M. NETTO, E. VITAL BRAZIL, R. CERQUEIRA, P. VALDURIEZ, M. MATTOSO. *Efficient Runtime Capture of Multiworkflow Data Using Prove-*

nance, in "eScience 2019 : 15th International eScience Conference", San Diego, United States, September 2019, pp. 1-10, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265932

Conferences without Proceedings

[40] C. ATLAN, J.-P. ARCHAMBAULT, O. BANUS, F. BARDEAU, A. BLANDEAU, A. COIS, M. COURBIN-COULAUD, G. GIRAUDON, S.-C. LEFÈVRE, V. LETARD, B. MASSE, F. MASSEGLIA, B. NINASSI, S. DE QUATREBARBES, M. ROMERO, D. ROY, T. VIÉVILLE. *Apprentissage de la pensée informatique : de la formation des enseignant·e·s à la formation de tou·te·s les citoyen·ne·s*, in "EIAH'19 Wokshop - Apprentissage de la pensée informatique de la maternelle à l'Université : retours d'expériences et passage à l'échelle", Paris, France, June 2019, https://hal.inria.fr/hal-02145480

[41] S. KAHL, F.-R. STÖTER, H. GOËAU, H. GLOTIN, R. PLANQUÉ, W.-P. VELLINGA, A. JOLY. *Overview of BirdCLEF 2019: Large-Scale Bird Recognition in Soundscapes*, in "CLEF 2019 : Conference and Labs of the Evaluation Forum", Lugano, Switzerland, CEUR Workshop Proceedings, CEUR, September 2019, vol. 2380, pp. 1-9, https://hal.umontpellier.fr/hal-02345644

[42] T. LORIEUL, A. JOLY. *Vers un désenchevêtrement de l'ambiguïté de la tâche et de l'incertitude du modèle pour la classification avec option de rejet à l'aide de réseaux neuronaux*, in "Conférence sur l'Apprentissage automatique (CAp)", Toulouse, France, July 2019, https://hal.archives-ouvertes.fr/hal-02421210

[43] M. NEGRI, M. SERVAJEAN, B. DENEU, A. JOLY. *Location-Based Plant Species Prediction Using A CNN Model Trained On Several Kingdoms - Best Method Of GeoLifeCLEF 2019 Challenge*, in "CLEF 2019 - Conference and Labs of the Evaluation Forum - Information Access Evaluation meets Multilinguality, Multimodality, and Visualization", Lugano, Switzerland, September 2019, https://hal.archives-ouvertes.fr/hal-02392637

Scientific Books (or Scientific Book chapters)

[44] A. JOLY, H. GOËAU, H. GLOTIN, C. SPAMPINATO, P. BONNET, W.-P. VELLINGA, J.-C. LOMBARDO, R. PLANQUÉ, S. PALAZZO, H. MÜLLER. *Biodiversity Information Retrieval Through Large Scale Content-Based Identification: A Long-Term Evaluation*, in "Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF", N. FERRO, C. PETERS (editors), The Information Retrieval Series, 2019, vol. 41, pp. 389-413, AcknowledgementsThe organization of the PlantCLEF task is supported by the French project Floris'Tic (Tela Botanica, Inria, CIRAD, INRA, IRD) funded in the context of the national investment program PIA. The organization of the BirdCLEF task is supported by the Xeno-Canto foundation for nature sounds as well as the French CNRS project SABIOD.ORG and EADM MADICS, and Floris'Tic. The annotations of some soundscapes were prepared with the late wonderful Lucio Pando at Explorama Lodges, with the support of Pam Bucur, Marie Trone and H. Glotin. The organization of the SeaCLEF task is supported by the Ceta-mada NGO and the French project Floris'Tic. [*DOI :* 10.1007/978-3-030-22948-1_16], https://hal.umontpellier.fr/hal-02273280

[45] D. OLIVEIRA, J. LIU, E. PACITTI. *Data-Intensive Workflow Management: For Clouds and Data-Intensive and Scalable Computing Environments*, Synthesis Lectures on Data Management, Morgan&Claypool Publishers, May 2019, vol. 14, n° 4, pp. 1-179 [*DOI :* 10.2200/S00915ED1V01Y201904DTM060], https://hal-lirmm.ccsd.cnrs.fr/lirmm-02128444

[46] T. ÖZSU, P. VALDURIEZ. *Principles of Distributed Database Systems - Fourth Edition*, Springer, 2019, pp. 1-700, forthcoming, https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265930

Research Reports

[47] O. BEAUMONT, L. EYRAUD-DUBOIS, J. HERRMANN, A. JOLY, A. SHILOVA. *Optimal checkpointing for heterogeneous chains: how to train deep neural networks with limited memory*, Inria Bordeaux Sud-Ouest, November 2019, n° RR-9302, https://arxiv.org/abs/1911.13214 , https://hal.inria.fr/hal-02352969

[48] B. CARAMIAUX, F. LOTTE, J. GEURTS, G. AMATO, M. BEHRMANN, F. BIMBOT, F. FALCHI, A. GARCIA, J. GIBERT, G. GRAVIER, H. HOLKEN, H. KOENITZ, S. LEFEBVRE, A. LIUTKUS, A. PERKIS, R. REDONDO, E. TURRIN, T. VIÉVILLE, E. VINCENT. *AI in the media and creative industries*, New European Media (NEM), April 2019, pp. 1-35, https://arxiv.org/abs/1905.04175 , https://hal.inria.fr/hal-02125504

Software

[49] A. AFFOUARD, J.-C. LOMBARDO, H. GOËAU, P. BONNET, A. JOLY. *Pl@ntNet*, April 2019, Software, https://hal.archives-ouvertes.fr/hal-02096020

[50] O. BUISSON, J.-C. LOMBARDO, A. JOLY. *Snoop*, April 2019, Software, https://hal.archives-ouvertes.fr/hal-02096036

[51] O. LEVCHENKO, D.-E. YAGOUBI, R. AKBARINIA, F. MASSEGLIA, B. KOLEV, D. SHASHA, T. PALPANAS, P. VALDURIEZ. *Imitates*, April 2019, Software, https://hal.inria.fr/hal-02095640

[52] M. LIROZ-GISTAU, R. AKBARINIA, P. VALDURIEZ. *FP-Hadoop*, April 2019, Software, https://hal.inria.fr/hal-02093002

[53] F. MASSEGLIA, J. DIENER. *LogMagnet*, April 2019, Software [*DOI :* 10.1145/2480362.2480419], https://hal.inria.fr/hal-02098365

[54] C. PRADAL, C. FOURNIER, F. BOUDON, P. VALDURIEZ, E. PACITTI, Y. Y. GUÉDON, C. GODIN. *OpenAlea*, April 2019, 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 Programming, ...) to analyse and simulate shapes and their development., https://hal.inria.fr/hal-02100181

[55] M. SERVAJEAN, A. JOLY. *ThePlantGame*, April 2019, Software, https://hal.archives-ouvertes.fr/hal-02096028