RESEARCH CENTRE

Sophia Antipolis - Méditerranée

IN PARTNERSHIP WITH:

CNRS, Université de Montpellier

2021 ACTIVITY REPORT

Project-Team ZENITH

Scientific Data Management

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

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team ZENITH

Creation of the Project-Team: 2012 January 01

Keywords

Computer sciences and digital sciences

A1.1. – Architectures

A3.1. – Data

A3.3. – Data and knowledge analysis

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 members, visitors, external collaborators

Research Scientists

- Patrick Valduriez [Team leader, Inria, Senior Researcher, HDR]
- Reza Akbarinia [Inria, Researcher, HDR]
- Hervé Goëau [CIRAD, Researcher]
- Alexis Joly [Inria, Senior Researcher, HDR]
- Antoine Liutkus [Inria, Researcher]
- Florent Masseglia [Inria, Senior Researcher, HDR]
- Didier Parigot [Inria, Researcher, until Sep 2021, HDR]
- Christophe Pradal [CIRAD, Researcher]

Faculty Members

- Francois Munoz [Univ Grenoble Alpes, Associate Professor, from Feb 2021]
- Esther Pacitti [Univ de Montpellier, Professor, HDR]

Post-Doctoral Fellow

- Baldwin Dumortier [Univ de Montpellier, from Feb 2021]

PhD Students

- Benjamin Deneu [Inria]
- Lamia Djebour [Univ de Montpellier]
- Joaquim Estopinan [Inria]
- Camille Garcin [Univ de Montpellier]
- Tanguy Lefort [Univ de Montpellier, from Oct 2021]
- Quentin Leroy [INA, CIFRE]

Technical Staff

- Antoine Affouard [Inria, Engineer]
- Julien Champ [Inria, Engineer, until May 2021]
- Mathias Chouet [Inria, Engineer]
- Theo Delfieu [Inria, Engineer, until May 2021]
- Baldwin Dumortier [Inria, Engineer, Jan 2021]
- Hugo Gresse [Inria, Engineer, until Nov 2021]
- Pierre Leroy [Inria, Engineer, from Oct 2021]
- Oleksandra Levchenko [Inria, Engineer, until Mar 2021]
- Titouan Lorieul [Inria, Engineer, from Jun 2021]
- Heraldo Pimenta Borges Filho [Inria, Engineer, until Jul 2021]
- Fabian Robert Stoter [Inria, Engineer, Jan 2021]

Interns and Apprentices

- Aimi Okabayashi [Inria, from May 2021 until Aug 2021]

Administrative Assistant

- Nathalie Brillouet [Inria, until Mar 2021]

2 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 have been 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. Furthermore, 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, INRAE 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 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.

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, making it affordable to keep more data around. Furthermore, 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 SPARQL) 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. 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 machine learning methods such as dimensionality reduction, representation learning or random projection. The activity of Zenith in this domain focuses on methods for large-scale data processing and search, 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 INRAE, 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 hundreds 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 INRAE 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). 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. In the same vein, we have an ongoing collaboration with Radio France, that has large amounts of archives to restore, for repurposing applications.

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 Social and environmental responsibility

We do consider the ecological impact of our technology, especially large data management.

- In our work on cache-based scheduling of scientific workflows in multisite clouds, we can minimize the monetary cost of the cloud, which directly reflects the energy consumption.

- We have also started to address the (major) problem of energy consumption of our ML models, by introducing energy-based metrics to assess the energy consumption during the training on GPU of our ML models. Furthermore, we want to improve training pipelines that reduce the need for training models from scratch. At inference, network compression methods can reduce the memory footprint and the computational requirements when deploying models.

- In the design of the Pl@ntnet mobile application, we adopt an eco-responsible approach, taking care not to integrate addictive, energy-intensive or non-essential functionalities to uses that promote the preservation of biodiversity and environment.

- To reduce our carbon footprint, we reduce to the minimum the number of long-distance trips, and favor train as much as possible. We also trade conference publications for journal publications, to avoid traveling. For instance, in 2020, we have 27 journal publications versus 19 conference publications.

6 Highlights of the year

6.1 Awards

- Renan Souza's 2019 PhD thesis "Supporting User Steering in Large-Scale Workflows with Provenance Data", supervised by Marta Mattoso (COPPE-UFRJ) and Patrick Valduriez, received the Honorable Mention in the Biannual Contest of the 2021 SBBD (Brazilian Symposium of Databases) conference.
- According to the Inria Academy survey on the software distributed by Inria and its partners and the needs of companies, Pl@ntnet ranks second.
- The paper "Efficient Incremental Computation of Aggregations over Sliding Window" [51] by Chao Zhang, Reza Akbarinia and Farouk Toumani, obtained the best paper award from BDA 2021.

7 New software and platforms

Let us describe new/updated software.

7.1 New software

7.1.1 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 200 countries (30M downloads) and allows identifying about 35K plant species at present time.

Publications: hal-01629195, hal-02937618, hal-03343235, hal-01182775

Contact: Alexis Joly

Participants: Antoine Affouard, Jean-Christophe Lombardo, Pierre Bonnet, Hervé Goëau, Mathias Chouet, Hugo Gresse, Julien Champ, Alexis Joly

7.1.2 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.

Publication: hal-01629149

Contact: Alexis Joly

Participants: Maximilien Servajean, Alexis Joly

7.1.3 Savime

Name: 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.

Publication: lirmm-01620376

Contact: Patrick Valduriez

Participants: Hermano Lustosa, Fabio Porto, Patrick Valduriez

Partner: LNCC - Laboratório Nacional de Computação Científica

7.1.4 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 Contributions: 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.

Publications: hal-01166298, hal-00831811

Contact: Christophe Pradal

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

Partners: CIRAD, INRAE

7.1.5 Imitates

Name: 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).

Publication: lirmm-01886794

Contact: Florent Masseglia

Partners: New York University, Université Paris-Descartes

7.1.6 UMX

Name: open-unmix

Keywords: Source Separation, Audio

Scientific Description: UMX implements state of the art audio/music source separation with deep neural networks (DNNs). It is intended to serve as a reference in the domain. It has been presented in two major scientific communications: An Overview of Lead and Accompaniment Separation in Music (https://hal-lirmm.ccsd.cnrs.fr/lirmm-01766781) and Music Separation with DNNs (Making it work (ISMIR 2018 Tutorial) https://sigsep.github.io/ismir2018_tutorial/index.html#/cover).

Functional Description: UMX implements audio source separation with deep learning, using the 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.

Publication: lirmm-01766781

Authors: Antoine Liutkus, Fabian Robert Stoter, Emmanuel Vincent

Contact: Antoine Liutkus

7.1.7 TDB

Keywords: Data assimilation, Big data, Data extraction

Scientific Description: TDB comes as a building block for audio machine learning pipelines. It is a scraping tool that allows large scale data augmentation. Its different components allow building a large dataset of samples composed of related audio tracks, as well as the associated metadata. Each sample comprises a dynamic number of entries.

Functional Description: TDB is composed of two core submodules. First, a data extraction pipeline allows scraping a provider url so as to extract large amounts of audio data. The provider is assumed to offer audio content in a freely-accessible way through a hardcoded specific structure. The software automatically downloads the data locally under a raw data format. To aggregate the raw data set, a list of item ids is used. The item ids will be requested from the provider given a url in parallel fashion. Second, a data transformation pipeline allows transforming the raw data into a dataset that is compatible with machine learning purposes. Each produced subfolder contains a set of audio files corresponding to a predefined set of sources, along with the associated metadata. A working example is provided.

Each component has several submodules, in particular, network handling and audio transcoding. Thus, TDB can be viewed as an extract-transform-load (ETL) pipeline that enables applications such as deep learning on large amounts of audio data, assuming that an adequate data provider url is fed into the software.

Contact: Antoine Liutkus

Participants: Antoine Liutkus, Fabian Robert Stoter

7.1.8 UMX-PRO

Name: Unmixing Platform - PRO

Keywords: Audio signal processing, Source Separation, Deep learning

Scientific Description: UMX-PRO is written in Python using the TensorFlow 2 framework and provides an off-the-shelf solution for music source separation (MSS). MSS consists in extracting different instrumental sounds from a mixture signal. In the scenario considered by UMX-PRO, a mixture signal is decomposed into a pre-defined set of so called targets, such as: (scenario 1) {"vocals", "bass", "drums", "guitar", "other"} or (scenario 2) {"vocals", "accompaniment"}.

The following key design choices were made for UMX-PRO. The software revolves around the training and inference of a deep neural network (DNN), building upon the TensorFlow v2 framework. The DNN implemented in UMX-PRO is based on a BLSTM recurrent network. However, the software has been designed to be easily extended to other kinds of network architectures to allow for research and easy extensions. Given an appropriately formatted database (not part of UMX-PRO), the software trains the network. The database has to be split into train and valid subsets, each one being composed of folders called samples. All samples must contain the same set of audio files, having the same duration: one for each desired target. For instance: {vocals.wav, accompaniment.wav}. The software can handle any number of targets, provided they are all present in all samples. Since the model is trained jointly, a larger number of targets increases the GPU memory usage during training. Once the models have been trained, they can be used for separation of new mixtures through a dedicated end-to-end separation network. Interestingly, this end-to-end network comprises an optional refining step called expectation-maximization that usually improves separation quality.

Functional Description: UMX-PRO implements a full audio separation deep learning pipeline in Tensorflow v2. It provides everything needed to train and use a deep learning model for separating music signals, including network architecture, data pipeline, training code, inference code as well as pre-trained weights. The software comes with full documentation, detailed comments and unit tests.

Authors: Antoine Liutkus, Fabian Robert Stoter

Contact: Antoine Liutkus

8 New results

8.1 Scientific Workflows

8.1.1 Data Provenance and Recommendation in ML Workflows

Participants: Esther Pacitti, Patrick Valduriez.

Scientific ML is multidisciplinary, heterogeneous, and affected by the physical constraints of the domain, making analyses challenging. In [33], we leverage workflow provenance techniques to build a holistic view to support the entire lifecycle of scientific ML. The experiments show that the decisions enable queries that integrate domain semantics with ML models while keeping low overhead (<1%), high scalability and high performance.

Provenance can also contribute to the interpretation of models resulting from the life cycle in deep neural networks (DNN). In [45], we propose a provenance data-based approach for the collection and analysis of configuration data in DNN, with an experimental validation with Keras and a real application which provides evidence of the flexibility and efficiency of the approach.

Data provenance can also be exploited for recommendation. In [32],we present FReeP-Feature Recommender from Preferences, a parameter value recommendation method that is designed to suggest

values for workflow parameters, taking into account past user preferences. FReeP is based on ML techniques, particularly in Preference Learning. FReeP is composed of three algorithms, where two of them aim at recommending the value for one parameter at a time, and the third makes recommendations for n parameters at once.

8.1.2 Reproducible Performance Optimization of Complex wokflows on the Edge-to-Cloud Continuum

Participants: Daniel Rosendo, Patrick Valduriez.

Complex workflows typically combine computing, analytics and learning. They often require a hybrid execution infrastructure with IoT devices interconnected to cloud/HPC systems. Such workflows are subject to complex constraints and requirements in terms of performance, resource usage, energy consumption and financial costs. This makes it challenging to optimize their configuration and deployment. In [46, 50], we propose a methodology to support the optimization of such workflows on the Edge-to-Cloud Continuum. We implement it as an extension of E2Clab, a previously proposed framework supporting the complete experimental cycle across the Edge-to-Cloud Continuum (see our white paper in [57]). Our approach relies on a rigorous analysis of possible configurations in a controlled testbed environment to understand their behaviour and related performance trade-offs. We illustrate our methodology with our Pl@ntNet application.

8.2 Data Analytics

8.2.1 Efficient Similarity Search in Large Time Series Databases

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

Fast and accurate similarity search is critical to performing many data mining tasks like motif discovery, classification or clustering. In [25], we present our parallel solutions, developed based on two state-of-the-art approaches iSAX and sketch, for k nearest-neighbor (kNN) search in large databases of time series. We compare the two solutions based on various measures of quality and time performance, and propose a tool that uses the characteristics of application data to determine which solution to choose for that application and how to set the parameters for that solution. Our experiments show that: (i) iSAX and its derivatives perform best in both time and quality when the time series can be characterized by a few low frequency Fourier Coefficients, a regime where the iSAX pruning approach works well; (ii) iSAX performs significantly less well when high frequency Fourier Coefficients have much of the energy of the time series; (iii) A random projection approach based on sketches by contrast is more or less independent of the frequency power spectrum. The experiments show the close relationship between pruning ratio and time for exact iSAX as well as between pruning ratio and the quality of approximate iSAX. Our toolkit analyzes typical time series of an application (i) to determine optimal segment sizes for iSAX and (ii) when to use Parallel Sketches instead of iSAX. Our solutions have been implemented using Spark, evaluated over a cluster of nodes, and have been applied to both real and synthetic data.

8.2.2 Efficient Computation of Aggregations for Analyzing Large Streaming Data

Participants: Reza Akbarinia.

In stream processing systems, *aggregations* having the inherent property of summarizing information from data, constitute a fundamental operator to compute real-time statistics. They are typically computed over finite subsets of a stream, called sliding windows. One of the challenges faced by the

sliding window aggregation (SWAG) algorithms is to incrementally compute aggregations over moving data, i.e., without recomputing the aggregation from scratch after inserting new data items or evicting old data items to/from the window. High throughput and low latency are essential requirements as stream processing systems are typically designed for real-time applications.

In [14], we propose PBA (Parallel Boundary Aggregator), a novel algorithm that computes incremental aggregations in parallel. PBA groups continuous slices into chunks, and maintains two buffers for each chunk containing, respectively, the cumulative slice aggregations (denoted as csa) and the left cumulative slice aggregations (denoted as lcs) of the chunk's slices. Using PBA, SWAGs can be computed in constant time for both amortized and worst-case time. We also propose an approach to optimize the chunk size, which guarantees the minimum latency for PBA. We conducted extensive empirical experiments using both synthetic and real-world datasets. Our experiments show that PBA behaves very well for average and large sliding windows (e.g., with sizes higher than 1024 values) compared to the state-of-the-art algorithms. For small-size windows, the results show the superiority of the non-parallel version of PBA (denoted as SBA) that outperforms other algorithms in terms of throughput.

8.2.3 Anomaly Detection in Time Series

Participants: Heraldo Borges, Reza Akbarinia, Florent Masseglia.

In many real-world applications such as health monitoring, siesmology, aircraft engine test, etc, the data is collected in the form of time series. Anomaly detection is very important in this type of data. It refers to discovering any abnormal behavior within the data encountered in a specific time interval. For instance, cardiologists are interested in identifying anomalous parts of ECG signals to diagnose heart disorders.

In [18], we present a survey of existing techniques for anomaly detection in time series. The objective is to provide an understanding of the problem of detecting anomalies and how existing techniques are related. The paper is divided into three main parts. First, the main concepts are presented. Then, the anomaly detection task is defined. Afterward, the main approaches and strategies to solve the problem are presented.

8.2.4 Efficient kNN Search in Large Chemometrics Databases

Participants: Reza Akbarinia, Florent Masseglia.

In precision agriculture and plant breeding, the amount of data tends to increase, and is becoming more and more complex, leading to difficulties in managing and analyzing it. Optical instruments such as NIR Spectroscopy or hyperspectral imaging are gradually expanding directly in the field, increasing the amount of spectral database. Using these tools allows access to non-destructive and rapid measurements to classify new varieties according to breeding objectives. However, processing this massive amount of spectral data is challenging

In the context of genotype discrimination, we propose a method called parSketch-PLSDA [30] in order to analyze spectral data. ParSketch-PLSDA is a combination of our indexing technique parSketch and the reference method PLSDA for predicting classes from multivariate data. We evaluated ParSketch-PLSDA through extensive experimentation over a large spectra dataset generated from hyperspectral images of leaves of four different sunflower genotypes. We compared Sketch-PLSDA with the state-of-the-art PLSDA method. The prediction model obtained by PLSDA has a classification error close to 23% on average across all genotypes. Our ParSketch-PLSDA method outperforms PLSDA by significantly reducing the prediction error to 10%.

8.3 Machine Learning for Biodiversity

8.3.1 New Methods for Species Distribution Modeling at Large Scale

Participants: Benjamin Deneu, Christophe Botella, Alexis Joly, François Munoz.

Species Distribution Models (SDMs) are fundamental tools in ecology for predicting the geographic distribution of species based on environmental data. The generalizability and spatial accuracy of a SDM depend very strongly on the type of model used and the environmental data used as explanatory variables.

In [3], we applied Convolutional Neural Networks (CNN) to a very large dataset of plant occurrences in France (GBIF), on a large taxonomical scale. We found that the landscape structure around location crucially contributes to improve predictive performance in particular for rare species, which open promising perspectives for biodiversity monitoring and conservation strategies. These models were computed on Jean ZAY super-computer as one of the Grand Challenge of GENCI.

In [39], we study for the first time a country-wide species distribution model based on very high resolution (1m) remote sensing images processed by a CNN. We demonstrate that this model can capture habitat information at very fine spatial scales, while providing overall better predictive performance than conventional models.

In [1], we propose a new method to correct the biases of SDMs coming from the non-uniform spatial sampling effort. The proposed method is based on estimating the variation in sampling effort over units of a spatial mesh in parallel with the environmental density of multiple species using a marked Poisson process model. It is particularly suited to the analysis of massive but highly heterogeneous presence-only data such as citizen science data.

8.3.2 AI-based Herbarium Specimens Analysis

Participants: Hervé Goëau, Alexis Joly.

Imaging of biological collections has been progressing at pace with tens of millions of images becoming available online in the last two decades. As such, they are an irreplaceable asset for research of all kinds, including ecology, natural history and epidemiology. In 2021, we worked on two new studies involving the analysis of digitized herbarium specimens [26, 21].

In [26], herbarium specimens were scored both manually by human observers and by a mask R-CNN object detection model to (1) evaluate the concordance between ML and manually-derived phenological data and (2) determine whether ML-derived data can be used to reliably assess phenological patterns. The ML model generally underestimates the number of reproductive structures present on each specimen; however, when these counts are used to provide a quantitative estimate of the phenological stage of plants on a given sheet, the ML and manually-derived PI's were highly concordant, demonstrating that phenological data extracted using machine learning can be used reliably to estimate the phenological stage of herbarium specimens and to detect phenological patterns.

In [21], we investigated whether digitized herbarium specimens can be used to improve the identification performance of species for which we have very few (if any) photos. Therefore, we introduced a new open dataset and a new domain adaptation method that provides significant improvement; however, the task remains highly challenging and progress is still needed.

8.3.3 Deep Learning Models for Digital Agriculture

Participants: Julien Champ, Herve Goeau, Alexis Joly, Baldwin Dumortier, Antoine Liutkus.

The ubiquity, portability and mobility of digital technologies are transforming agriculture and food production. Within ZENITH, we develop new deep learning approaches supporting this transformation.

In [42], we propose a new deep learning architecture for plant disease recognition. Contrary to classical models based on the classification of host species-disease pairs, we propose a new conditional multi-task learning (CMTL) approach which allows the distribution of host species and disease characteristics learned simultaneously with a conditional link between them. We show that our approach can improve the performance of plant disease identification while improving the understanding of the prediction. Meanwhile, we also compose a new dataset that could serve as an important benchmark in this field.

In [37], we report our participation to the ROSE challenge evaluating new approaches for automatic weeding by agricultural robots. The solution we develop jointly with INRAE is based on the following tasks: weed detection and location based on a new instance segmentation deep learning framework, positioning of the weeding probe according to location results and electrical destruction of weeds. Discrimination results obtained between crops and weeds attained 76% accuracy and mortality rates of weeds ranged from 75% (Polygonaceae) to 13% (Brassicaceae). Survivors were nevertheless impacted with a severe limitation of their growth demonstrating the relevance of the overall approach.

Some current ongoing research in the context of a collaboration with INRAE (AI3P project) furthermore aims at predicting protein interaction and structure, in a way similar to how Transformer-based model predict text in natural language processing except that positions are not taken as integers, but as scalars in $\mathbb{R}^3$. The most notable application of this work would be to predict which protein should be synthesized to interact with some other target protein, which is a common problem in drug design. The study is risky but promising.

8.3.4 Evaluation of Species Identification and Prediction Algorithms

Participants: Alexis Joly, Herve Goeau, Benjamin Deneu, Titouan Lorieul, Camille Garcin.

We ran a new edition [53] of the LifeCLEF evaluation campaign with the involvement of hundreds of data scientists and research teams worldwide. It results in a new snapshot of the performance of state-of-the-art computer vision, bio-acoustic and machine learning techniques towards building real-world biodiversity monitoring systems. One of the main new outcomes was the arrival of Visual Transformers among the best models of the SnakeCLEF task. Even if their performance is still slightly inferior to that of convolutional neural networks, there is no doubt that they are now an alternative to be considered in the future. On the contrary, the 50 best methods of the BirdCLEF sound recognition task were solely based on convolutional neural networks ensembles. Interestingly, the choice of the CNN backbone does not seem to be the most determining factor of the better performance. The devil is in the detail, typically in the pre-processing and post-processing methodologies. The geolifeclef task also confirms the power of convolutional neural networks for this type of task, revealing their ability to recognise species habitats even when they are only trained on remote sensing images only (i.e. without any additional environmental data as input).

In addition to LifeCLEF, we published a new plant identification dataset and benchmark in the context of NeurIPS 2021 conference [41]. We highlight two particular features of this new dataset called Pl@ntNet-300K: (i) its strong class imbalance, and, (ii) the presence of many species visually similar. These two characteristics make Pl@ntNet-300K well suited for the evaluation of set-valued classification methods and algorithms. Therefore, we introduce two set-valued evaluation metrics (macro-average top-k and average-k accuracy) and we provide baseline results established by training deep neural networks using the cross-entropy loss.

8.3.5 Innovative services in Pl@ntNet platform

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

Pl@ntnet is a citizen observatory that relies on artificial intelligence (AI) technologies to help people identify plants with their smartphones. Over the past few years, Pl@ntNet has become one of the largest plant biodiversity observatories in the world with several million contributors. A set of new tools and innovative services following the FAIR (Findable, Accessible, Interoperable, Reusable) principles were developed in 2021.

We first enriched Pl@ntNet public API (my.plantnet.org) with new features such as similar image search, automatic organ recognition or geo-localized species prediction (based on the results of Benjamin Deneu's PhD work [3]). The API was published as a new service in the European Open Science Cloud. It currently has nearly 4900 user accounts including start up developers, IT services providers, researchers, citizen observatories, etc.

Besides, two new complementary concepts were developed to support the creation of customized and collaborative e-floras in the mobile application [15]: groups and monitoring work spaces. Groups allow any user to create a private or public space on the platform . These groups are used by people to structure their collaborative activity related to a given area or taxonomic group. Over 3300 groups have already been created by e.g., professional land managers, educators, and plant enthusiasts. Monitoring work spaces allow a given stakeholder to access all Pl@ntNet observations and identification requests of a given species list in a particular area. For example, this service has been mobilized to follow the recent development of an invasive species (i.e., Hakea sericea Schrad. & J.C. Wendl.) in and around a natural reserve on the Mediterranean coast.

8.4 Machine Learning for audio and long time series

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.

8.4.1 Setting the State of the Art in Music Demixing

Participants: Fabian-Robert Söter, Baldwin Dumortier, Antoine Liutkus.

In the last years, we have been very active in the topic of music demixing, with a prominent role in defining the state of the art in the domain and organizing numerous related events. Our contributions this year in the domain are numerous. First, we manage the *MUSDB18* dataset, which is the most popular music separation/processing dataset worldwide, with 5000 downloads this year. Second, we continued maintaining the *open-unmix* software, which is an established reference implementation for music source separation. This year, we released a new set of weights for it, that incorporates months of work in making the system up to date in terms of performance by training on a much larger dataset, while keeping the same top-level standard in terms of software development practice. Third, we co-organized the Music Demixing Workshop, a satellite event from ISMIR 2021, that enjoyed a strong participation with approximately a hundred participants. Fourth, we developed a web-based interface to open-unmix, that allows knowledgeable sound engineers to use it in a more user-friendly way. This was done in collaboration with Radio France.

8.4.2 Deep models for audio and long-range data

Participants: Antoine Liutkus, Fabian-Robert Söter, Baldwin Dumortier.

Our strategy is to go beyond our current expertise on music demixing to address new interesting topics in audio. This means leaving our comfort zone on source separation to address new exciting challenges. This year, we pursued our theoretical efforts on Transformers, and were the first to propose a new way to combine relative positional encodings with linear complexity Transformers [7].

9 Bilateral contracts and grants with industry

9.1 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.

9.2 Pl@ntNet donations

Participants: Alexis Joly, Hugo Gresse, Mathias Chouet, David Margery.

A contract has been signed between Inria and Agropolis Foundation to allow the use of Pl@ntNet donations to pay the salaries of InriaSOFT engineers working on the development of the platform. In 2021, a volume of 195K euros of donations from approximately 20K donors was collected.

10 Partnerships and cooperations

10.1 International initiatives

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

HPDaSc

Title: High Performance Data Science

Project web site : HPDaSc

Duration: 2020 ->

Coordinator: Fabio Porto (fporto@lncc.br)

Partners:

- LNCC, UFRJ, UFF, CEFET/RJ (Brazil)

Inria contact: Patrick Valduriez

Summary: Data-intensive science requires the integration of two fairly different paradigms: high-performance computing (HPC) and data science. HPC is compute-centric and focuses on high-performance of simulation applications, typically using powerful, yet expensive supercomputers whereas data science is data-centric and focuses on scalability and fault-tolerance of web and cloud

applications using cost-effective clusters of commodity hardware. This project addresses the grand challenge of High Performance Data Science (HPDaSc), by developing architectures and methods to combine simulation, ML and data analytics.

10.2 European initiatives

10.2.1 FP7 & H2020 projects

RISC2

Title: RISC2: A network for supporting the coordination of High-Performance Computing research between Europe and Latin America

Project web site : RISC2

Duration: 2021 - 2022

Coordinator: Barcelona Supercomputing Center, Spain

Partners:

- BULL ATOS (France)
- CIEMAT (Spain)
- CINECA (Italy)
- Inria (HiePACS, Nachos, Zenith)
- JUELICH (Germany)
- UNIVERSIDADE DE COIMBRA (Portugal)

Inria contact: Stéphane Lanteri

Summary: The RISC2 project is a coordination network for High Performance Computing (HPC) between Europe and Latin America, funded by the European H2020 FETHPC program and the partner countries. It is managed by Barcelona Supercomputing Center and has eight main European HPC actors and the main HPC actors from Brazil, including LNCC, Mexico, Argentina, Colombia, Uruguay, Costa Rica and Chile. The objective is to encourage stronger cooperation between their research and industrial communities on HPC applications and infrastructure deployment. The main project deliverable will be a cooperation roadmap aimed at policymakers, the scientific community and industry, identifying key application areas, HPC infrastructure and policy requirements, and exploring ways for the activities established during the project to last beyond its lifetime. The activities and results will be disseminated widely through dedicated project communication tools and will take advantage of existing platforms such as Campus Iberoamerica. The training carried out in the project will help capacitate Latin American HPC, and the structured interaction between researchers and policymakers in both regions will reinforce links and help define a coordinated policy and a clear roadmap for the future.)

Cos4Cloud

Title: Cos4Cloud: Co-designed Citizen Observatories Services for the EOS-Cloud

Project web site : Cos4Cloud

Duration: 2019 - 2023

Coordinator: CSIC, Spain

Partners:

- The Open University (UK)

- CREAF (Spain)
- Bineo (Spain)
- EarthWatch (UK)
- NKUA (Greece)
- SVERIGES LANTBRUKSUNIVERSITET (Sweden)
- DynAikon (UK)
- Trobola (Columbia)
- 52°North Initiative for Geospatial Open Source Software (Germany)

Inria contact: Alexis Joly

Summary: Cos4Cloud integrates 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.

10.3 National initiatives

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

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.

ANR PerfAnalytics (2021-2024), 100 Keuro.

Participants: Reza Akbarinia, Florent Masseglia.

The objective of the PerfAnalytics project is to analyze sport videos in order to quantify the sport performance indicators and provide feedback to coaches and athletes, particularly to French sport federations in the perspective of the Paris 2024 Olympic games. A key aspect of the project is to couple the existing technical results on human pose estimation from video with scientific methodologies from biomechanics for advanced gesture objectivation. The motion analysis from video represents a great potential for any monitoring of physical activity. In that sense, it is expected that exploitation of results will be able to address not only sport, but also the medical field for orthopedics and rehabilitation.

PPR "Antibiorésistance": structuring tool "PROMISE" (2021-2024), 240 Keuro.

Participants: Reza Akbarinia, Florent Masseglia.

The objective of the PROMISE (PROfessional coMmunIty network on antimicrobial reSistancE) project is to build a large data warehouse for managing and anlyzing antimicrobial resitance (AMR) data. It gathers 21 existing professional networks and 42 academic partners from three sectors, human, animal, and environment. The project is based on the following transdisciplinary and cross-sectoral pillars: i) fostering synergies to improve the one-health surveillance of antibiotic consumption and AMR, ii) data sharing for improving the knowledge of professionals, iii) improving clinical research by analyzing the shared data.

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.

CASDAR CARPESO (2020-2022), 87 Keuro.

Participants: Julien Champ, Hervé Goëau, Alexis Joly.

In order to facilitate the agro-ecological transition of livestock systems, the main objective of the project is to enable the practical use of meslin (grains and forages) by demonstrating their interests and remove sticking points on the nutritional value of the meslin. Therefore, it develops AI-based tools allowing to automatically assess the nutritional value of meslin from images. The consortium includes 10 chambers of agriculture, 1 Technical Institute (IDELE) and 2 research organizations (Inria, CIRAD).

10.3.1 Others

Pl@ntNet InriaSOFT consortium (2019-20XX), 80 Keuro / year

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

This contract between four research organisms (Inria, INRAE, 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.

Ministry of Culture (2019-2021), 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).

Ministry of Culture (2020-2021): Audio separation, 75 Keuro

Participants: Baldwin Dumortier, Antoine Liutkus.

This project is a collaboration with the innovation department at Radio France. It is funded in the context of the *convention cadre* between Inria and the *Ministère de la culture*. Its objective is to provide expert sound engineers from Radio France with state of the art separation tools developped at Inria. It involves both research on source separation and software engineering.

DINUM, 80 Keuro

Participants: Reza Akbarinia, Florent Masseglia.

The objective of the contract is to analyze the evolution of the time series of coordinates provided by the IGN (National Institute of Geographic and Forest Information), and to detect the anomalies of different origins, for example, seismic or material movements.

CACTUS Inria exploratory action (2020-2022), 200 Keuro

Participants: Alexis Joly, Joaquim Estopinan.

CAcTUS is an Inria exploratory action led by Alexis Joly and focused on predictive approaches to determining the conservation status of species.

MUSE AI3P (2021-2024), 80 Keuro

Participants: Baldwin Dumortier, Antoine Liutkus.

This project aims at leveraging recent AI breakthroughs in stimulating new applications in life science, notably for digital agronomy. Our contribution in this respect concerns the development of new deep learning models for genomics.

MUSE MULTINODE (2021-2024), 100 Keuro

Participants: Antoine Liutkus.

This project has been funded to promote collaboration between researchers in functional ecology (UMR CEFE) and ZENITH. Our contribution will lie in the development of new deep learning models for a better understanding of animal migrations.

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- A. Joly: General chair of LifeCLEF 2021 workshop.
- A. Liutkus: Program chair of MDX 2021 workshop.

Member of the organizing committees

- R. Akbarinia, F. Masseglia: Extraction et Gestion des Connaissances (EGC), 2021.
- A. Joly: CLEF 2021 international conference, Evaluation Lab Chair.

11.1.2 Scientific events: selection

Chair of conference program committees

- R. Akbarinia: Bases de Données Avancées (BDA), 2021, Demo PC chair.
- E. Pacitti: Bases de Données Avancées (BDA), 2021, PC chair.
- A. Liutkus: Source separation PC chair at ICASSP, 2021.

Member of the conference program committees

- R. Akbarinia: VLDB 2021, AIML Systems 2021, ADBIS 2021.
- P. Valduriez: Bases de Données Avancées (BDA), 2021.
- A. Joly: NeurIPS 2021, CVPR 2021, ICASSP 2021, ICLR 2021, ECIR 2021.
- F. Masseglia: PAKDD 2021, IJCAI 2021, PKDD 2021, DATA 2021, DSAA 2021, AIKE 2021, EGC 2021, SimBig 2021, SDM 2021, ACM SAC DM 2021, ACM SAC DS 2021, ICDM 2021.
- A. Liutkus: NeurIPS 2021, ICLR 2021, ICASSP 2021, WASPAA 2021.

11.1.3 Journal

Member of the editorial boards

- P. Valduriez: Distributed and Parallel Databases.
- R. Akbarinia: Transactions on Large Scale Data and Knowledge Centered Systems.
- C. Pradal: Plant Methods.
- A. Joly: Frontiers in Plant Science.

Reviewer - reviewing activities

- R. Akbarinia: IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Access, Journal of Information and Data Management.
- A. Joly: TPAMI, PLOS One, BioScience, Environmental Research Letters, Transactions on Multimedia, Transactions on Information Systems, Computers and Electronics in Agriculture, Ecological Indicators.
- A. Liutkus: IEEE TASLP, IEEE SPL, JOSS, arxive moderator for eess.AS.

11.1.4 Invited talks

- E. Pacitti: "Uma Perspectiva Evolutiva e Multidisciplinar do Tratamento de Dados", Seminários 2021, Instituto de Computação – UFF, Rio de Janeiro, 17 March 2021.
- P. Valduriez: "Innovation : startup strategies", 12th edition of the Marcus Evans "Innovation Strategies" conference, 20 May 2021.
- A. Joly: "Which scientific challenges beyond data collection?", keynote talk, CEMEB-NUMEV day, 30 March 2021.
- A. Joly: "Pl@ntNet, la science des données au service de la biodiversité végétale", EGC 2021 conference, 25 January 2021.

11.1.5 Leadership within the scientific community

- E. Pacitti: Member of the Steering Committee of the BDA conference.
- A. Joly: Foundator & scientific coordinator of LifeCLEF virtual: computer-assisted identification of living organisms (19 collaborators for the organization, 100s of registrants/participants, 100s publications, 1000s citations)
- A. Joly: Scientific and technical director of the Pl@ntNet platform: AI-based citizen science for plant biodiversity (3 permanent researchers, 4 engineers, 3 PhD students, 1 postdoc, tens of national and international partners, thousands of developers with an account on Pl@ntNet API, millions of end-users).
- A. Joly: Steering board of Cos4Cloud, H2020 research project (6M euros) aimed at integrating citizen science in the European Open Science Cloud (EOSC) through the co-design of innovative services.
- A. Liutkus: elected member of IEEE technical committee on audio.

11.1.6 Scientific expertise

- R. Akbarinia: Expert for STIC AmSud international program.
- A. Joly: Expert for the French National HPC grand equipment (GENCI)
- A. Liutkus: expert for ANR.

11.1.7 Research administration

- E. Pacitti: manager of Polytech' Montpellier's International Relationships for the computer science department (100 students).
- P. Valduriez: scientific manager for the Latin America zone at Inria's Direction of Foreign Relationships (DRI).
- F. Masseglia: deputy scientific director of Inria for the domain "Perception, Cognition And Interaction".

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

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

Esther Pacitti responsibilities on teaching (theoretical, home works, practical courses,exams) and supervision at Polytech' Montpellier UM, for engineer students:

- IG3: Database design, physical organization, 54h, level, L3, 50 students.

- IG4: Distributed Databases and NoSQL, 80h , level M1, 50 students.
- Large Scale Information Management (Iot, Recommendation Systems, Graph Databases), 27h, level M2, 20 students.
- Supervision of industrial projects
- Supervision of master internships.
- Supervision of computer science discovery projects.

Patrick Valduriez:

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

Alexis Joly:

- Univ. Montpellier: Machine Learning, 10h, level M2
- Polytech' Montpellier: Content-Based Image Retrieval, 6h, level M2.
- AgroParisTech: Deep Learning, 10h, level M1.

Antoine Liutkus

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

Christophe Pradal

- Univ. Montpellier: Root System Modelling, 15h, level M2.
- Univ. Cambridge: Functional-Structural Plant Modelling, 20h, level M2.

11.2.2 Supervision

PhD & HDR:

- Defended PhD: Alena Shilova, Memory Saving Strategies for Deep Neural Network Training, Defended December 2021, Univ. Bordeaux. Advisors: Olivier Beaumont, Lionel Eyraud-Dubois, Alexis Joly.
- PhD in progress: Daniel Rosendo, Enabling HPC-Big Data Convergence for Intelligent Extreme-Scale Analytics, started Oct 2019, Univ. Rennes. Advisors: Gabriel Antoniu, Alexandru Costan, Patrick Valduriez.
- PhD in progress: Benjamin Deneu, Large-scale and High-resolution Species Distribution Modelling using Deep Learning, Univ. Montpellier. Advisors: Alexis Joly, François Munoz, Maximilien Servajean, Pierre Bonnet.
- PhD in progress: Camille Garcin, Multi-class classification with high label ambiguity and a long-tailed distribution. Advisors: Joseph Salmon, Maximilien Servajean, Alexis Joly.
- PhD in progress: Joaquim Estopinan, Species Conservation Status Prediction. Advisors: Alexis Joly, François Munoz, Maximilien Servajean, Pierre Bonnet.
- PhD in progress: Quentin Leroy, Interactive retrieval of non-common visual entities in large archives. Advisors: Olivier Buisson, Alexis Joly.
- PhD in progress: Lamia Djebour, Parallel Time Series Indexing and Retrieval with GPU architectures. Advisors: Reza Akbarinia, Florent Masseglia.

11.2.3 Juries

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

- A. Joly: Etienne David, PhD, University of Avignon.
- A. Joly: Waleed RAGHEB, PhD, University of Montpellier.
- A. Joly: Alena Shilova, PhD, University of Bordeaux.
- F. Masseglia: Khaled Zaouk, PhD, Institut Polytechnique de Paris.
- F. Masseglia : Tao Peng, PhD, University of Aix-Marseille.
- E. Pacitti: Victorien Elvinger, PhD, Univ Lorraine, 2021.
- E. Pacitti: Lulian Sandu Popa, HDR, UVSQ - Université Paris-Saclay.
- P. Valduriez: Jorge Galicia Auyon, HDR, ENSMA, Univ Poitiers.
- A. Liutkus: Pritish Chandna (UPF Barcelona), Stylianos Mimilakis (Fraunhofer IDMT Ilmenau)

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

- Pl@ntNet Community management 2010-2021: A. Joly and H. Gresse spend several hours a week animating Pl@ntNet's user community. This includes: animating the community of developers using Pl@ntNet API (thousands of users, animating Pl@ntNet's social networks (twitter account, facebook account), managing the mailbox (contact@plantnet-project.org) and writing articles in the news section of the Pl@ntNet web site.
- F. Masseglia is co-head of the national project "1 Scientifique - 1 Classe, Chiche!"
- F. Masseglia is Member of the strategic committee of Fondation Blaise Pascal.

11.3.2 Articles and contents

- G. Heidsieck, E. Pacitti and F. Tardieu (INRAe) have co-produced a popularization video with Polytech' Montpellier on "The design of digital agriculture", May 2021.

11.3.3 Interventions

- A. Joly participated in a webinar organized by the European Citizen Science Association (ECSA) about the usage of Cos4Cloud services by citizen observatories.

12 Scientific production

12.1 Major publications

[1] C. Botella, A. Joly, P. Bonnet, F. Munoz and P. Monestiez. 'Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data'. In: *Methods in Ecology and Evolution* 12.5 (1st Feb. 2021), pp. 933–945. DOI: 10.1111/2041-210X.13565. URL: https://hal.umontpellier.fr/hal-03150701.

[2] J. Carranza-Rojas, H. Goëau, P. Bonnet, E. Mata-Montero and A. Joly. 'Going deeper in the automated identification of Herbarium specimens'. In: *BMC Evolutionary Biology* 17.1 (Dec. 2017), p. 181. DOI: 10.1186/s12862-017-1014-z. URL: https://hal.inria.fr/hal-01580070.

[3] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz and A. Joly. 'Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment'. In: *PLoS Computational Biology* 17.4 (19th Apr. 2021), e1008856. DOI: 10.1371/journal.pcbi.1008856. URL: https://hal.inrae.fr/hal-03220977.

[4] M. Fontaine, R. Badeau and A. Liutkus. 'Separation of Alpha-Stable Random Vectors'. In: *Signal Processing* (Jan. 2020), p. 107465. DOI: 10.1016/j.sigpro.2020.107465. URL: https://hal.inria.fr/hal-02433213.

[5] J. Liu, N. Moreno Lemus, E. Pacitti, F. Porto and P. Valduriez. 'Parallel Computation of PDFs on Big Spatial Data Using Spark'. In: *Distributed and Parallel Databases* 38 (2020), pp. 63–100. DOI: 10.1007/s10619-019-07260-3. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02045144.

[6] J. Liu, L. Pineda, E. Pacitti, A. Costan, P. Valduriez, G. Antoniu and M. Mattoso. 'Efficient Scheduling of Scientific Workflows using Hot Metadata in a Multisite Cloud'. In: *IEEE Transactions on Knowledge and Data Engineering* (2018). DOI: 10.1109/TKDE.2018.2867857. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01867717.

[7] A. Liutkus, O. Cífka, S.-L. Wu, U. Şimşekli, Y.-H. Yang and G. Richard. 'Relative Positional Encoding for Transformers with Linear Complexity'. In: ICML 2021 - 38th International Conference on Machine Learning. Proceedings of the 38th International Conference on Machine Learning. Virtual Only, United States, 18th July 2021. URL: https://hal.telecom-paris.fr/hal-03256451.

[8] A. Liutkus, U. Ş. Imşekli, S. Majewski, A. Durmus and F.-R. Stöter. 'Sliced-Wasserstein Flows: Nonparametric Generative Modeling via Optimal Transport and Diffusions'. In: *36th International Conference on Machine Learning (ICML)*. Long Beach, United States, June 2019. URL: https://hal.inria.fr/hal-02191302.

[9] M. Metz, M. Lesnoff, F. Abdelghafour, R. Akbarinia, F. Masseglia and J.-M. Roger. 'A "big-data" algorithm for KNN-PLS'. In: *Chemometrics and Intelligent Laboratory Systems* 203 (Aug. 2020), p. 104076. DOI: 10.1016/j.chemolab.2020.104076. URL: https://hal.inrae.fr/hal-02899789.

[10] D. Oliveira, J. Liu and E. Pacitti. *Data-Intensive Workflow Management: For Clouds and Data-Intensive and Scalable Computing Environments*. Vol. 14. Synthesis Lectures on Data Management 4. Morgan&Claypool Publishers, May 2019, pp. 1–179. DOI: 10.2200/S00915ED1V01Y201904DTM060. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02128444.

[11] T. Özsu and P. Valduriez. *Principles of Distributed Database Systems - Fourth Edition*. Télécharger la 3ieme et 4ieme édition : lien dans " voir aussi ". Springer, 2020, pp. 1–674. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265930.

[12] C. Pradal, S. Artzet, J. Chopard, D. Dupuis, C. Fournier, M. Mielewczik, V. Negre, P. Neveu, D. Parigot, P. Valduriez and S. Cohen-Boulakia. 'InfraPhenoGrid: A scientific workflow infrastructure for Plant Phenomics on the Grid'. In: *Future Generation Computer Systems* 67 (Feb. 2017), pp. 341–353. DOI: 10.1016/j.future.2016.06.002. URL: https://hal.inria.fr/hal-01336655.

[13] D.-E. Yagoubi, R. Akbarinia, F. Masseglia and T. Palpanas. 'Massively Distributed Time Series Indexing and Querying'. In: *IEEE Transactions on Knowledge and Data Engineering* 32.1 (2020), pp. 108–120. DOI: 10.1109/TKDE.2018.2880215. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02197618.

[14] C. Zhang, R. Akbarinia and F. Toumani. 'Efficient Incremental Computation of Aggregations over Sliding Windows'. In: 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2021). Singapore, Singapore, 2021, pp. 2136–2144. DOI: 10.1145/3447548.3467360. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03359490.

12.2 Publications of the year

International journals

[15] A. Affouard, M. Chouet, J.-C. Lombardo, H. Gresse, H. Goëau, T. Lorieul, P. Bonnet and A. Joly. 'Customized e-floras: How to develop your own project on the Pl@ntNet platform'. In: *Biodiversity Information Science and Standards* 5 (3rd Sept. 2021), pp. 1–4. DOI: 10.3897/biss.5.73857. URL: https://hal.inrae.fr/hal-03338811.

[16] M.-A. Benderra, A. Aparicio, J. Leblanc, D. Wassermann, E. Kempf, G. Galula, M. Bernaux, A. Canellas, T. Moreau, A. Bellamine, J.-P. Spano, C. Daniel, J. Champ, F. Canouï-Poitrine and J. Gligorov. 'Clinical Characteristics, Care Trajectories and Mortality Rate of SARS-CoV-2 Infected Cancer Patients: A Multicenter Cohort Study'. In: *Cancers* 13.19 (23rd Sept. 2021), p. 4749. DOI: 10.3390/cancers13194749. URL: https://hal.sorbonne-universite.fr/hal-03379679.

[17] E. Blanc, P. M. Barbillon, C. Fournier, C. Lecarpentier, C. Pradal and J. Enjalbert. 'Functional–Structural Plant Modeling Highlights How Diversity in Leaf Dimensions and Tillering Capability Could Promote the Efficiency of Wheat Cultivar Mixtures'. In: *Frontiers in Plant Science* 12 (29th Sept. 2021), p. 734056. DOI: 10.3389/fpls.2021.734056. URL: https://hal.inria.fr/hal-03358998.

[18] H. Borges, R. Akbarinia and F. Masseglia. 'Anomaly Detection in Time Series'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems* (2021). URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03359500.

[19] C. Botella, A. Joly, P. Bonnet, F. Munoz and P. Monestiez. 'Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data'. In: *Methods in Ecology and Evolution* 12.5 (1st Feb. 2021), pp. 933–945. DOI: 10.1111/2041-210X.13565. URL: https://hal.umontpellier.fr/hal-03150701.

[20] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz and A. Joly. 'Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment'. In: *PLoS Computational Biology* 17.4 (19th Apr. 2021), e1008856. DOI: 10.1371/journal.pcbi.1008856. URL: https://hal.inrae.fr/hal-03220977.

[21] H. Goëau, P. Bonnet and A. Joly. 'AI-based Identification of Plant Photographs from Herbarium Specimens'. In: *Biodiversity Information Science and Standards* 5 (31st Aug. 2021), pp. 1–4. DOI: 10.3897/biss.5.73751. URL: https://hal.inrae.fr/hal-03338823.

[22] G. Heidsieck, D. De Oliveira, E. Pacitti, C. Pradal, F. Tardieu and P. Valduriez. 'Cache-aware scheduling of scientific workflows in a multisite cloud'. In: *Future Generation Computer Systems* 122 (2021), pp. 172–186. DOI: 10.1016/j.future.2021.03.012. URL: https://hal.archives-ouvertes.fr/hal-03189130.

[23] P. Kranas, B. Kolev, O. Levchenko, E. Pacitti, P. Valduriez, R. Jiménez-Peris and M. Patiño-Martinez. 'Parallel Query Processing in a Polystore'. In: *Distributed and Parallel Databases* (3rd Feb. 2021), p. 39. DOI: 10.1007/s10619-021-07322-5. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03148271.

[24] L. Kunstmann, D. Pina, F. Silva, A. Paes, P. Valduriez, D. De Oliveira and M. Mattoso. 'Online Deep Learning Hyperparameter Tuning based on Provenance Analysis'. In: *Journal of Information and Data Management* 12.5 (19th Nov. 2021), pp. 396–414. DOI: 10.5753/jidm.2021.1924. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03443660.

[25] O. Levchenko, B. Kolev, D.-E. E. Yagoubi, R. Akbarinia, F. Masseglia, T. Palpanas, P. Valduriez and D. Shasha. 'BestNeighbor: Efficient Evaluation of kNN Queries on Large Time Series Databases'. In: *Knowledge and Information Systems (KAIS)* 63.2 (2021), pp. 349–378. DOI: 10.1007/s10115-020-01518-4. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02973633.

[26] N. Love, P. Bonnet, H. Goëau, A. Joly and S. Mazer. 'Machine Learning Undercounts Reproductive Organs on Herbarium Specimens but Accurately Derives Their Quantitative Phenological Status: A Case Study of Streptanthus tortuosus'. In: *Plants* 10.11 (Nov. 2021), p. 2471. DOI: 10.3390/plants10112471. URL: https://hal.inria.fr/hal-03454183.

[27] C. A. Midingoyi, C. Pradal, A. Enders, D. Fumagalli, H. Raynal, M. Donatelli, I. N. Athanasiadis, C. Porter, G. Hoogenboomi, D. Holzworth, F. Garcia, P. Thorburn and P. M. Martre. 'Crop2ML: An open-source multi-language modeling framework for the exchange and reuse of crop model components'. In: *Environmental Modelling and Software* 142 (Aug. 2021). DOI: 10.1016/j.envsoft.2021.105055. URL: https://hal.inria.fr/hal-03231805.

[28] B. Pitchers, F. Do, C. Pradal, L. Dufour and P.-É. Lauri. 'Apple tree adaptation to shade in agroforestry: an architectural approach'. In: *American Journal of Botany* 108.5 (2nd May 2021), pp. 732–743. DOI: 10.1002/ajb2.1652. URL: https://hal-auf.archives-ouvertes.fr/hal-03214878.

[29] N. C. A. Pitman, T. Suwa, C. Ulloa Ulloa, J. Miller, J. Solomon, J. Philipp, C. Vriesendorp, A. Derby Lewis, S. Perk, P. Bonnet, A. Joly, M. Tobler, J. H. Best, J. P. Janovec, K. C. Nixon, B. M. Thiers, M. Tulig, E. E. Gilbert, R. Campostrini Forzza, G. Zimbrão, F. L. Ranzato Filardi, R. Turner, F. Zuloaga, M. Belgrano, C. Zanotti, J. M. de Vos, E. Hettwer Giehl, T. C. E. Paine, R. Texeira de Queiroz, K. Romoleroux and E. Hilo de Souza. 'Identifying gaps in the photographic record of the vascular plant flora of the Americas'. In: *Nature Plants* 7 (2021), pp. 1010–1014. DOI: 10.1038/s41477-021-00974-2. URL: https://hal.inrae.fr/hal-03312029.

[30] M. Ryckewaert, M. Metz, D. Héran, P. George, B. Grèzes-Besset, R. Akbarinia, J. M. Roger and R. Bendoula. 'Massive spectral data analysis for plant breeding using parSketch-PLSDA method: Discrimination of sunflower genotypes'. In: *Biosystems Engineering* 210 (Oct. 2021), pp. 69–77. DOI: 10.1016/j.biosystemseng.2021.08.005. URL: https://hal.inrae.fr/hal-03329674.

[31] R. Salles, E. Pacitti, E. Bezerra, F. Porto and E. Ogasawara. 'TSPred: A framework for nonstationary time series prediction'. In: *Neurocomputing* 467 (Jan. 2022), pp. 197–202. DOI: 10.1016/j.neucom.2021.09.067. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03452170.

[32] D. J. Silva, E. Pacitti, A. Paes and D. De Oliveira. 'Provenance-and machine learning-based recommendation of parameter values in scientific workflows'. In: *PeerJ Computer Science* 7 (5th July 2021), e606. DOI: 10.7717/peerj-cs.606. URL: https://hal.archives-ouvertes.fr/hal-03418836.

[33] R. Souza, L. G. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E. V. Brazil, M. Moreno, P. Valduriez, M. Mattoso, R. Cerqueira and M. A. S. Netto. 'Workflow Provenance in the Lifecycle of Scientific Machine Learning'. In: *Concurrency and Computation: Practice and Experience* (Aug. 2021). DOI: 10.1002/cpe.6544. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03324881.

[34] R. Souza, V. Silva, A. Lima, D. De Oliveira, P. Valduriez and M. Mattoso. 'Distributed in-memory data management for workflow executions'. In: *PeerJ Computer Science* 7 (2021), e527. DOI: 10.7717/peerj-cs.527. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03227618.

[35] H. Takahashi and C. Pradal. 'Root phenotyping: important and minimum information required for root modeling in crop plants'. In: *Breeding Science* (10th Feb. 2021). DOI: 10.1270/jsbbs.20126. URL: https://hal.inria.fr/hal-03139460.

[36] P. Valduriez, R. Jiménez-Peris and M. T. Özsu. 'Distributed Database Systems: The Case for NewSQL'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems*. Lecture Notes in Computer Science 12670 (18th May 2021), pp. 1–15. DOI: 10.1007/978-3-662-63519-3_1. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03228968.

International peer-reviewed conferences

[37] F. Abdelghafour, F. Rançon, S. Liu, J. Champ, V. De Rudnicki, C. Guizard, H. Goëau, C. Doussan, A. Joly, P. Bonnet and G. Rabatel. 'WeedElec: a robotic research platform for individual weed detection and selective electrical weeding'. In: EPCA 2021 - 13th European Conference on Precision Agriculture. Precision agriculture '21. Budapest, Hungary: Wageningen Academic Publishers, 2021, pp. 695–702. DOI: 10.3920/978-90-8686-916-9_83. URL: https://hal.inrae.fr/hal-03325689.

[38] A. Castro, H. Borges, R. Campisano, E. Pacitti, F. Porto, R. Coutinho and E. Ogasawara. 'Generalização de Mineração de Sequências Restritas no Espaço e no Tempo'. In: SBBD: Simpósio Brasileiro de Banco de Dados. Online, Brazil, 2021, pp. 313–318. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03452154.

[39] B. Deneu, A. Joly, P. Bonnet, M. Servajean and F. Munoz. 'How Do Deep Convolutional SDM Trained on Satellite Images Unravel Vegetation Ecology ?' In: *Lecture Notes in Computer Science*. ICPR 2020 - 25th International Conference on Pattern Recognition. Vol. 12666. Pattern Recognition. ICPR International Workshops and Challenges Virtual Event, January 10–15, 2021, Proceedings, Part VI. Milan / Virtual, Italy: Springer, 2021, pp. 148–158. DOI: 10.1007/978-3-030-68780-9_15. URL: https://hal.inrae.fr/hal-03167637.

[40] L. Djebour, R. Akbarinia and F. Masseglia. 'ASAX : Segmentation adaptative basée sur la quantité d'information pour SAX'. In: BDA 2021 - 37e Conférence sur la Gestion de Données - Principes, Technologies et Applications. Paris, France, 25th Oct. 2021. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03468535.

[41] C. Garcin, A. Joly, P. Bonnet, J.-C. Lombardo, A. Affouard, M. Chouet, M. Servajean, T. Lorieul and J. Salmon. 'Pl@ntNet-300K: a plant image dataset with high label ambiguity and a long-tailed distribution'. In: NeurIPS 2021 - 35th Conference on Neural Information Processing Systems. Virtual Conference, France, 6th Dec. 2021. URL: https://hal.inria.fr/hal-03474556.

[42] S. H. Lee, H. Goëau, P. Bonnet and A. Joly. 'Conditional Multi-Task learning for Plant Disease Identification'. In: ICPR 2020 - 25th International Conference on Pattern Recognition. Proceedings of ICPR 2020, 25th International Conference on Pattern Recognition. Milan, Italy: IEEE, 2021, pp. 3320–3327. DOI: 10.1109/ICPR48806.2021.9412643. URL: https://hal.inria.fr/hal-03415972.

[43] A. Liutkus, O. Cífka, S.-L. Wu, U. Şimşekli, Y.-H. Yang and G. Richard. 'Relative Positional Encoding for Transformers with Linear Complexity'. In: ICML 2021 - 38th International Conference on Machine Learning. Proceedings of the 38th International Conference on Machine Learning. Virtual Only, United States, 18th July 2021. URL: https://hal.telecom-paris.fr/hal-03256451.

[44] T. Lorieul, E. Cole, B. Deneu, M. Servajean, P. Bonnet and A. Joly. 'Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images'. In: *CEUR workshop Proceedings*. Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. Vol. 2936. Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum. Bucarest, Romania, 2021, pp. 1451–1462. URL: https://hal.inrae.fr/hal-03353487.

[45] D. Pina, L. Kunstmann, D. De Oliveira, P. Valduriez and M. Mattoso. 'Provenance Supporting Hyperparameter Analysis in Deep Neural Networks'. In: IPAW 2020-2021 - 8th and 9th International Provenance and Annotation Workshop. Vol. Lecture Notes in Computer Science. Provenance and Annotation of Data and Processes 12839. London, United Kingdom: Springer, 19th July 2021, pp. 20–38. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03324873.

[46] D. Rosendo, A. Costan, G. Antoniu, M. Simonin, J.-C. Lombardo, A. Joly and P. Valduriez. 'Reproducible Performance Optimization of Complex Applications on the Edge-to-Cloud Continuum'. In: Cluster 2021 - IEEE International Conference on Cluster Computing. Portland, OR, United States, 7th Sept. 2021. URL: https://hal.archives-ouvertes.fr/hal-03310540.

[47] R. A. P. Silva, E. Pacitti, Y. Y. Frota and D. De Oliveira. 'Análise de Desempenho da Distribuição de Workflows Científicos em Nuvens com Restrições de Confidencialidade'. In: Workshop on Computer and Communication Systems Performance (WPerformance 2021). Online, Brazil, 22nd July 2021, p. 12. DOI: 10.5753/wperformance.2021.15721. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03452298.

[48] R. M. Silva, D. Pina, L. Kunstmann, D. De Oliveira, P. Valduriez, A. L. G. A. Coutinho and M. Mattoso. 'Capturing Provenance to Improve the Model Training of PINNs: first handon experiences with Grid5000'. In: 42nd Ibero-Latin-American Congress on Computational Methods in Engineering and 3rd Pan American Congress on Computational Mechanics. Rio de Janeiro, Brazil, 9th Nov. 2021, pp. 1–7. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03443548.

[49] C. Zhang, R. Akbarinia and F. Toumani. 'Efficient Incremental Computation of Aggregations over Sliding Windows'. In: KDD 2021 - 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Singapore (Virtual), Singapore, 2021, pp. 2136–2144. DOI: 10.1145/3447548.3467360. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03359490.

National peer-reviewed Conferences

[50] D. Rosendo, A. Costan, G. Antoniu and P. Valduriez. 'Enabling Reproducible Analysis of Complex Workflows on the Edge-to-Cloud Continuum'. In: BDA 2021 - 37e Conférence sur la Gestion de Données - Principes, Technologies et Applications. Proceedings of the BDA 2021 conference. Paris, France, 28th Oct. 2021. URL: https://hal.archives-ouvertes.fr/hal-03332524.

[51] C. Zhang, R. Akbarinia and F. Toumani. 'Efficient Incremental Computation of Aggregations over Sliding Windows'. In: BDA 2021 - 37e Conférence sur la Gestion de Données - Principes, Technologies et Applications. Virtual, France: Colegio de Médicos y Cirujanos de Costa Rica, 25th Oct. 2021. DOI: 10.1145/3447548.3467360. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03468587.

Conferences without proceedings

[52] A. Joly, A. Affouard, M. Chouet, B. Deneu, J. Estopinan, H. Goëau, H. Gresse, J.-C. Lombardo, T. Lorieul, F. Munoz, M. Servajean and P. Bonnet. 'Pl@ntNet, ten years of automatic plant identification and monitoring'. In: IUCN - Congrès mondial de la nature. Marseille, France, 3rd Sept. 2021. URL: https://hal.inrae.fr/hal-03343235.

Scientific book chapters

[53] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, H. Glotin, R. Planqué, R. R. de Castañeda, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet and H. Müller. 'Overview of LifeCLEF 2021: an evaluation of Machine-Learning based Species Identification and Species Distribution Prediction'. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Vol. 12880. Lecture Notes in Computer Science. Springer International Publishing, 14th Sept. 2021, pp. 371–393. DOI: 10.1007/978-3-030-85251-1_24. URL: https://hal.inria.fr/hal-03415990.

Doctoral dissertations and habilitation theses

[54] A. Liutkus. 'Principled methods for mixtures processing'. Université de Montpellier, 11th Feb. 2022. URL: https://hal.inria.fr/tel-03578077.

Reports & preprints

[55] E. Chzhen, C. Denis, M. Hebiri and T. Lorieul. *Set-valued classification – overview via a unified framework*. 1st Mar. 2021. URL: https://hal.archives-ouvertes.fr/hal-03154625.

[56] C. Garcin, M. Servajean, A. Joly and J. Salmon. *Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification*. 8th Feb. 2022. URL: https://hal.archives-ouvertes.fr/hal-03562414.

Other scientific publications

[57] G. Antoniu, P. Valduriez, H.-C. Hoppe and J. Krüger. *Towards Integrated Hardware/Software Ecosystems for the Edge-Cloud-HPC Continuum*. 2021. DOI: 10.5281/zenodo.5534464. URL: https://hal.archives-ouvertes.fr/hal-03358930.

[58] D. Rosendo, A. Costan, G. Antoniu and P. Valduriez. 'E2Clab: Reproducible Analysis of Complex Workflows on the Edge-to-Cloud Continuum'. In: IPDPS 2021 - 35th IEEE International Parallel and Distributed Processing Symposium. Virtual, France, 17th May 2021. URL: https://hal.archives-ouvertes.fr/hal-03269852.

12.3 Other

Scientific popularization

[59] P. Valduriez. 'Making the Right Move to Senior Researcher: some challenges and hints'. In: *SIGMOD record* 50.2 (June 2021), pp. 30–32. DOI: 10.1145/3484622.3484628. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03240377.