RESEARCH CENTRE

Inria Center
at Université Côte d'Azur

IN PARTNERSHIP WITH:

CNRS, Université de Montpellier

2022 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

Inria

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

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

B1.1.11. – Plant Biology

B3.3. – Geosciences

B3.5. – Agronomy

B3.6. – Ecology

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, HDR]
- Diego Marcos Gonzalez [INRIA, ISFP, from Oct 2022]
- Florent Masseglia [INRIA, Senior Researcher, HDR]
- Christophe Pradal [CIRAD, Researcher]

Faculty Members

- François Munoz [UGA, Associate Professor]
- Esther Pacitti [Univ Montpellier, Professor, HDR]

Post-Doctoral Fellows

- Benjamin Bourel [CNRS, from Jun 2022]
- Ondrej Cifka [Univ Montpellier, from Feb 2022]
- Raphael De Freitas Saldanha [INRIA, from Dec 2022]
- Baldwin Dumortier [Univ Montpellier]

PhD Students

- Benjamin Deneu [INRIA]
- Lamia Djebour [Univ Montpellier]
- Joaquim Estopinan [INRIA]
- Camille Garcin [Univ Montpellier]
- Cesar Leblanc [INRIA, from Sep 2022]
- Tanguy Lefort [Univ Montpellier]

Technical Staff

- Antoine Affouard [INRIA, Engineer]
- Mathias Chouet [INRIA, Engineer]
- Hugo Gresse [INRIA, Engineer]
- Théo Larcher [INRIA, Engineer, from Oct 2022]
- Pierre Leroy [INRIA, Engineer]
- Jean-Christophe Lombardo [INRIA, Engineer]
- Shamprikta Mehreen [INRIA, Engineer, until Jul 2022]

Interns and Apprentices

- Cesar Leblanc [INRIA, from Apr 2022 until Sep 2022]

Administrative Assistant

- Cathy Desseaux [INRIA, from Apr 2022]

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 [11] 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 and Parallel Data Management

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 (MapReduce, Spark, Pregel), file systems (GFS, HDFS), NoSQL systems (BigTable, Hbase, MongoDB), NewSQL systems (Spanner, 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 or data lake. 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 workflow systems [10] are also useful for data integration and data analytics. 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 Machine Learning for High Dimensional Data Processing

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, 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 spurious 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 HIRROS 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, but also for plant modelling. We address this application in the context of the French initiative OpenAlea, with CIRAD and INRAE.

- **Large language models for genomics.** In the context of a collaboration with CNRS - INRAE, we are developing an activity on large language models applied to genomics. In particular , our work focuses on *inverse folding*, i.e., predicting a sequence of amineo acids that are able to generate a given protein structure, with applications in the drug design industry. These models involve training large deep models on several millions of structural data samples. We also investigated explanatory methods for large language models.

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.

- We 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 2022, we have 19 journal publications versus 14 conference publications.

6 Highlights of the year

6.1 Awards

The paper "A Data-Driven Model Selection Approach to Spatio-Temporal Prediction" by Rocío Zorrilla, Eduardo Ogasawara, Patrick Valduriez and Fabio Porto [46] was nominated for best paper and obtained the second prize at SBBD 2022 – Brazilian Symposium on Databases, Buzios, Brazil, 2022.

6.2 Data

In 2022, Pl@ntNet data has been used in 220 scientific publications in various domains and prestigious journals such as Nature, Plos One or Annals of Botany.

7 New software and platforms

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, 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 Distributed Data and Model Management

8.1.1 Scientific Workflows for Plant Phenotyping and Modeling

Participants: Christophe Pradal, Esther Pacitti, Patrick Valduriez.

With the development of new plant phenotyping platforms, biologists are faced with a deluge of data to be computationally analyzed. The OpenAlea software addresses this challenge, with the support of data-intensive scientific workflows [10] and cache-aware techniques to process huge datasets in the cloud [5].

In [18], we propose an automatic phenotyping method called PhenoTrack3D to extract to extract a $3D+t$ reconstruction of maize from images. It allows the study of plant architecture and individual organ development over time during the entire growth cycle. The method tracks the development of each organ from a time-series of plants whose organs have already been segmented in 3D using existing methods. In [22], we propose another automatic phenotyping method for root systems. This method combines an imaging device and an automatic analysis workflow based on registration and topological tracking, capable of accurately describing the topology and geometry of observed root systems in $2D+t$. In complement to these phenotyping methods, available in OpenAlea, we designed the mathematical model HydroRoot to study the role of root architecture and anatomy in water uptake. In [17], we investigate how the interplay between conductivities along radial (e.g. aquaporins) and axial (e.g. xylem vessels) pathways determines water transport properties of highly branched root system architectures. In [14], we model the concomitant transport of water and solutes in roots under water deficit, by considering hydrostatic and osmotic forces on maize roots. In [28], we assess to what extent the phenomics and modelling communities can address the issues of interoperability and data exchange, using a science mapping approach.

8.1.2 Elastic Scalable Transaction Processing

Participants: Patrick Valduriez.

Scaling ACID transactions in a cloud database is hard, and providing elastic scalability even harder. In [24], we present a solution for elastic scalable transaction processing. Unlike previous solutions, it does not require any hardware assistance. Yet, it does scales linearly to 100s of servers. We show the correctness of our solution. Finally, we provide a thorough performance evaluation of our solution using LeanXcale, an industrial-strength NewSQL database system on Amazon Web Services (AWS). The results show linear scalability, e.g., 5 million TPC-C NewOrder transactions per minute (TPM) with 200 nodes, which is greater than the TPC-C throughput obtained by the 9th highest result in all history using dedicated hardware used exclusively (not shared like in our evaluation) for the benchmark. Furthermore, the efficiency in terms of TPM per core is double that of the two top TPC-C results.

8.1.3 Distributed Intelligence on the Edge-to-Cloud Continuum

Participants: Daniel Rosendo, Patrick Valduriez.

The large scale and optimized deployment of learning-based workflows across the Edge-to-Cloud Continuum requires extensive and reproducible experimental analysis of application execution on representative testbeds. A thorough experimental analysis requires the assessment of the impact of multiple factors, such as: model accuracy, training time, network overhead, energy consumption, processing latency, among others. In [27], we propose a comprehensive review of the state-of-the-art in learning-based analytics for the Edge-to-Cloud Continuum. The main simulation, emulation, deployment systems, and testbeds for experimental research on the Edge-to-Cloud Continuum available today are also surveyed, with special attention on experiment reproducibility.

8.1.4 Data and Model Management with Gypscie

Participants: Patrick Valduriez.

As predictive analytics using ML models (or models for short) become prevalent in different stages of scientific exploration, a new set of artifacts are produced during the models' life-cycle that need to be managed. In addition to the models' versions, ML life-cycle artifacts include the collected training data and pre-processing workflows, data labels and selected features, model training, tuning and monitoring statistics and provenance information. However, to realize the full potential of data science, these artifacts must be built and combined, which can be very complex as there can be many. Furthermore, they should be shared and reused, in particular, in different execution environments such as HPC or Spark clusters. In order to support the complete ML life-cycle process and produced artifacts, we propose the Gypscie framework [45], to develop, share, improve and publish ML artifacts. Gypscie is integrated with the SAVIME database system for querying tensor data. In particular, we proposed a data-driven approach for selecting pre-trained temporal models to be applied at each query point [46], which avoids training a different model for each domain point, thus saving model training time. We implemented this approach into the SAVIME database system [44].

8.2 Data Analytics

8.2.1 Time Series Prediction and Anomaly Detection

Participants: Esther Pacitti.

TSPred is a prediction process that seamlessly integrates nonstationary time series transformations with state-of-the-art statistical and machine learning methods. In [29], we describe a novel implementation of TSPred for time series prediction in association with data preprocessing. It is made available as an R-package, which provides functions for defining and conducting time series prediction, including data pre(post)processing, decomposition, modeling, prediction, and accuracy assessment. Furthermore, TSPred enables user-defined methods, which significantly expands the applicability of the framework.

Time series event detection is related to studying methods for detecting observations in a series with special meaning. These observations differ from the expected behavior of the data set. In [41], we propose a novel method for detecting events in nonstationary time series The method, entitled Forward and Backward Inertial Anomaly Detector (FBIAD), analyzes inconsistencies in observations concerning surrounding temporal inertia (forward and backward). FBIAD is shown to outperform other methods both in accuracy and elapsed time.

8.2.2 Entropy-Based Segmentation of Time Series

Participants: Lamia Djebour, Reza Akbarinia, Florent Masseglia.

Many applications today generate data at increasing rates. The data may concern personal activities (e.g., electricity or water consumption using smart-meters or smart plugs) or professional activities (e.g., heart activity through monitoring or farming using plants sensors). This results in the production of large and complex data, usually in the form of time series. Data mining techniques on massive sets of time series have drawn a lot of interest since their application may lead to improvements in a large number of these activities, relying on fast and accurate similarity search in time series for performing tasks like , e.g., classification, clustering, and motif discovery. Because of the big data volumes, these tasks can be slow on raw data. This is why approximate representation of time series is needed as a means to allow fast computation of similarity search. In [34, 20], we propose a new representation technique, called ASAX (Adaptive SAX), which yields a variable-size segmentation of time series with high precision in retrieval tasks thanks to its lower information loss. Our representation is based on entropy measurement for detecting the time intervals that should be split. We also propose a lower bounding method that allows approximating the distance between the original time series based on their representations in ASAX. The experimental results show that the more the data distribution in the time domain is unbalanced, the greater the precision gain of ASAX.

8.2.3 Parallel Techniques for Variable Size Segmentation of Time Series Datasets

Participants: Lamia Djebour, Reza Akbarinia, Florent Masseglia.

Given the high data volumes in time series applications, or simply the need for fast response times, it is usually necessary to rely on alternative, shorter representations of the series, usually with some precision loss. This incurs approximate comparisons of time series where precision is a major issue. In [33], we propose efficient parallel techniques for improving the execution time of our segmentation algorithm. In this work, we first propose a new representation technique, called ASAX-SSE, which allows obtaining a variable-size segmentation of time series based on SSE (sum of squared error). It can reduce significantly the error incurred by possible splittings at different steps of the representation calculation. Then, we propose efficient parallel algorithms for improving the execution time of our segmentation approach using GPUs. We implemented our approach and conducted empirical experiments using more than 120 real world datasets. The results show significant performance gains in terms of precision for

similarity search compared to SAX. They show the effectiveness of our parallel algorithms, e.g., up to ×45 faster than the sequential algorithm for 1M time series.

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 an SDM strongly depend on the type of model used and the environmental data used as explanatory variables. In [19], we introduce a country-wide species distribution model based on very high resolution (1m) remote sensing images processed by a convolutional neural network. We demonstrate that this model can capture landscape and habitat information at very fine spatial scales, while providing overall better predictive performance than conventional models. To demonstrate the ecological significance of the model, we propose an original analysis to visualise the relation between input data and species traits or environment learned by the model as well as some statistical tests verifying them.

In [21], we rather ask whether the temporal dimension of remote sensing images can also be exploited by deep-SDMs. We therefore built a substantial and original dataset (called DeepOrchidSeries) aimed at modelling the distribution of orchids on a global scale based on Sentinel-2 Image Time-series. Thanks to this ambitious dataset, we trained several deep-SDMs based on Convolutional Neural Networks (CNNs) whose input was extended to include the temporal dimension. To quantify the contribution of the temporal dimension, we designed a novel interpretability methodology based on temporal permutation tests, temporal sampling and temporal averaging.

In [16], we introduce a new Bayesian dynamic species distribution model to explore the roles of long-distance dispersal and age-structured fecundity in the transient invasion dynamics of Plectranthus barbatus, a woody plant invader in South Africa.

8.3.2 AI-based Herbarium Specimens Analysis

Participants: Hervé Goëau, Alexis Joly.

Imaging of biological collections has been progressing at increasing pace, with tens of millions of images becoming available online over the last two decades. As such, they are an irreplaceable asset for research of all kinds, including ecology, natural history and epidemiology. In 2022, we worked on a new study involving the analysis of digitized herbarium specimens [23]. We investigated the extent to which the use of deep learning can help detect and type-classify relatively rare vegetative structures in herbarium collections. Our results demonstrate the relevance approaches for growing shoot detection.

8.3.3 Evaluation of Species Identification and Prediction Algorithms

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

We ran a new edition [38] of the LifeCLEF evaluation campaign with the involvement of hundreds of data scientists and research teams worldwide. The edition provided 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 insights of this edition is that vision transformers perform better than CNNs in several tasks, in particular in the PlantCLEF task involving a

very large dataset of 4M images of 80K plant species. In the other challenges, what seems to best explain the highest performance is the model selection methodology employed, given the time constraints and the available computational resources. The challenge with the most methodological novelty was the GeoLifeCLEF challenge, which is quite unusual due to its multi-modal nature (mixing very different types) and the originality of the task itself (set-valued classification based on presence-only data). The way all the modalities were combined was clearly one of the main drivers of success. Moreover, the set-valued classification problem has encouraged the implementation of original strategies that have proven to be effective in the end.

8.3.4 New features in the Pl@ntNet platform

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

Pl@ntnet is a citizen observatory that relies on 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 features were developed in 2022.

First, we developed an offline mode of the Pl@ntNet mobile application, which makes it possible to identify plant images without any connexion by downloading a compressed version of the full Pl@ntNet model. This embedded version can be automatically updated when a wifi connexion is available and the newly observed plants are uploaded automatically to Pl@ntNet's server. Data such as images and common names can also be downloaded to improve user's experience (at the price of more storage on the smartphone). About 10K users already have been using this new feature since its release in October 2022.

Another development is GeoPl@ntNet, an original web application that enables to find out which plant species have already been observed or are potentially present in this area by selecting a point on a map and drawing a rectangle around it. In order to predict species in places where there is little or no data, the application uses Deep-SDM [19], a deep learning model trained to predict the species present from high-resolution satellite data (from the IGN) and environmental data (temperature, precipitation, soil type, etc.). This technology is based on the outcomes of the PhD thesis of Benjamin Deneu and has been implemented within the framework of the European project Cos4Cloud, which develops open technologies for citizen science. The service is currently adapted for predictions in Metropolitan France.

8.3.5 Migratory Data Modeling

Participants: Ondrej Cifka, Antoine Liutkus.

We developed new deep neural networks that are able to successfully model and predict migration of animals, in the context of a collaboration with ecologists from the CEFE CNRS laboratory in Montpellier. The training data for such models has been gathered from the international MoveBank dataset and consists of time series of GPS data of various length, one for each animal in an acquisition campaign. Such data offers many challenges on its own, such as irregular sampling frequency, different species being represented, missing data, etc.

Our work involved designing Transformer models that could account for past positions as well as complementary geographical context data that was extracted from thirdparty sources like temperature, altitude, soil type, etc. This involved a significant engineering effort to allow training on GPU at effective speed.

8.3.6 Large Language Models for Protein Design

Participants: Baldwin Dumortier, Antoine Liutkus.

AlphaFold is a recent breakthrough in computational biology that successfully predicts the *folding*, i.e. the spatial configuration, of a protein given its defining sequence of amino acids. This achievement unlocked many new research efforts from a large community of researchers in computational biology. In the context of an ongoing collaboration with INRAE and CNRS, we focused on the original idea of designing a model that could successfully achieve *inverse folding*, i.e., outputting a sequence of amino acids when its input is a sequence of spatial positions. Such a model would actually be very useful for protein design. Leveraging our recent work on positional encoding and the large AlphaFold database that was released publicly on 2022, we were able to train such a network that achieved a very good accuracy [51]. This involved a significant engineering effort to develop an efficient data pipeline.

Since the release of our technical report, several other teams have been interested in the same area and our perspectives on this topic are numerous. We notably intend to leverage our unique collaboration between a computer science and biology labs to organize the first large scale biological validation campaign for the several protein generative models that were proposed by the community in 2022, yet evaluated only through simulations.

8.3.7 Blackbox Explanatory Methods for Large Language Models

Participants: Ondrej Cifka, Antoine Liutkus.

While working on large language models for migratory modeling, we realized that an increasingly interesting research topic is the design of explanatory methods for such models. Indeed, although sheer performance is often fundamental, it is actually not always the case. For instance, ecologists are not interested so much in the capacity of a model in accutely predicting the future position of animals, but rather in successfully identifying which geographical variables are relevant in doing these predictions, or to which extent past context is exploited to make such predictions. The question of context length hence soon became central, because is can directly be related to the notion of *memory* for animal migrations.

This collaboration led us to develop *context length probing* [50], which is a new original way to study the behavior of a large language model by studying how it reacts to input contexts of different lengths. Doing so, its simplicity is remarkable, departing from several other explanatory methods that often require accessing the actual internals of a model or further specific training.

9 Partnerships and cooperations

9.1 International initiatives

9.1.1 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 - 2024

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.

9.2 International research visitors

9.2.1 Visits of international scientists

Other international visits to the team

Dennis Shasha

Status: researcher

Institution of origin: New York University

Country: USA

Dates: 1 april- 30 may

Context of the visit: AI3P project (MUSE iSite)

Mobility program/type of mobility: research stay, lecture

Fabio Porto

Status: researcher

Institution of origin: LNCC

Country: Brazil

Dates: 16 may - 29 may

Context of the visit: HPDaSc associated team

Mobility program/type of mobility: research stay, lecture

9.3 European initiatives

9.3.1 Horizon Europe

GUARDEN GUARDEN project on cordis.europa.eu

Title: safeGUARDing biodivErsity aNd critical ecosystem services across sectors and scales

Duration: From November 1, 2022 to October 31, 2025

Partners:

- Inria
- PARC NATIONAL DE PORT-CROS (CONSERVATOIRE BOTANIQUE NATIONAL MEDITERRANEEN DE PORQUEROLLES), France
- STICHTING NATURALIS BIODIVERSITY CENTER (NATURALIS), Netherlands

- MINISTRY OF AGRICULTURE, RURAL DEVELOPMENT AND ENVIRONMENT OF CYPRUS, Cyprus
- PLYMOUTH MARINE LABORATORY LIMITED (PML), United Kingdom
- UNIVERSITY OF ANTANANARIVO, Madagascar
- CHAROKOPEIO PANEPISTIMIO (HAROKOPIO UNIVERSITY OF ATHENS (HUA)), Greece
- AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS (CSIC), Spain
- DRAXIS ENVIRONMENTAL SA (DRAXIS), Greece
- EBOS TECHNOLOGIES LIMITED (eBOS), Cyprus
- CENTRE DE COOPERATION INTERNATIONALE EN RECHERCHE AGRONOMIQUE POUR LEDEVELOPPEMENT - C.I.R.A.D. EPIC (CIRAD), France
- AGENTSCHAP PLANTENTUIN MEISE (AGENCE JARDIN BOTANIQUE DE MEISE), Belgium
- ENVECO ANONYMI ETAIRIA PROSTASIAS KAI DIAHIRISIS PERIVALLONTOS A.E. (ENVECO S.A. ENVIRONMENTAL PROTECTION AND MANAGEMENT), Greece
- AREA METROPOLITANA DE BARCELONA (AMB), Spain
- FREDERICK UNIVERSITY FU (FREDERICK UNIVERSITY FU), Cyprus
- EREVNITIKO PANEPISTILIAKO INSTITOUTO SYSTIMATION EPIKOINONIAS KAI YPOLOGISTON-EMP (RESEARCH UNIVERSITY INSTITUTE COMMUNICATION AND COMPUTER SYSTEMS), Greece

Inria contact: Alexis Joly

Coordinator:

Summary: GUARDEN's main mission is to safeguard biodiversity and its impact on people by bringing them at the forefront of policy and decision-making. This will be achieved through the development of user-oriented Decision Support Applications (DSAs), and leveraging on Multi-Stakeholder Partnerships (MSPs). They will take into account policy and management objectives and priorities across sectors and scales, build consensus to tackle data gaps, analytical uncertainties or conflicting objectives, and assess options to implement adaptive transformative change. To do so, GUARDEN will make use of a suite of methods and tools using Deep Learning, Earth Observation, and hybrid modelling to augment the amount of standardized and geo-localized biodiversity data, build-up a new generation of predictive models of biodiversity and ecosystem status indicators under multiple pressures (human and climate), and propose a set of complementary ecological indicators likely to be incorporated into local management and policy. The GUARDEN approach will be applied at sectoral case studies involving end users and stakeholders through Multi-Stakeholder Partnerships, and addressing critical cross-sectoral challenges (at the nexus of biodiversity and deployment of energy/transport infrastructure, agriculture, and coastal urban development). Thus, the GUARDEN DSAs shall help stakeholders engaged in the challenge to improve their holistic understanding of ecosystem functioning, biodiversity loss and its drivers and explore the potential ecological and societal impacts of alternative decisions. Upon the acquisition of this new knowledge and evidence, the DSAs will help end-users not only navigate but also shape the policy landscape to make informed all-encompassing decisions through cross-sectoral integration.

MAMBO MAMBO project on cordis.europa.eu

Title: Modern Approaches to the Monitoring of BiOdiversity

Duration: From September 1, 2022 to August 31, 2026

Partners:

- Inria
- AARHUS UNIVERSITET (AU), Denmark

- STICHTING NATURALIS BIODIVERSITY CENTER (NATURALIS), Netherlands
- THE UNIVERSITY OF READING, United Kingdom
- HELMHOLTZ-ZENTRUM FUR UMWELTFORSCHUNG GMBH - UFZ, Germany
- ECOSTACK INNOVATIONS LIMITED (EcoINN), Malta
- UK CENTRE FOR ECOLOGY & HYDROLOGY, United Kingdom
- CENTRE DE COOPERATION INTERNATIONALE EN RECHERCHE AGRONOMIQUE POUR LEDEVELOPPEMENT - C.I.R.A.D. EPIC (CIRAD), France
- PENSOFT PUBLISHERS (PENSOFT), Bulgaria
- UNIVERSITEIT VAN AMSTERDAM (UvA), Netherlands

Inria contact: Alexis Joly

Coordinator:

Summary: EU policies, such as the EU biodiversity strategy 2030 and the Birds and Habitats Directives, demand unbiased, integrated and regularly updated biodiversity and ecosystem service data. However, efforts to monitor wildlife and other species groups are spatially and temporally fragmented, taxonomically biased, and lack integration in Europe. To bridge this gap, the MAMBO project will develop, test and implement enabling tools for monitoring conservation status and ecological requirements of species and habitats for which knowledge gaps still exist. MAMBO brings together the technical expertise of computer science, remote sensing, social science expertise on human-technology interactions, environmental economy, and citizen science, with the biological expertise on species, ecology, and conservation biology. MAMBO is built around stakeholder engagement and knowledge exchange (WP1) and the integration of new technology with existing research infrastructures (WP2). MAMBO will develop, test, and demonstrate new tools for monitoring species (WP3) and habitats (WP4) in a co-design process to create novel standards for species and habitat monitoring across the EU and beyond. MAMBO will work with stakeholders to identify user and policy needs for biodiversity monitoring and investigate the requirements for setting up a virtual lab to automate workflow deployment and efficient computing of the vast data streams (from on the ground sensors, and remote sensing) required to improve monitoring activities across Europe (WP4). Together with stakeholders, MAMBO will assess these new tools at demonstration sites distributed across Europe (WP5) to identify bottlenecks, analyze the cost-effectiveness of different tools, integrate data streams and upscale results (WP6). This will feed into the co-design of future, improved and more cost-effective monitoring schemes for species and habitats using novel technologies (WP7), and thus lead to a better management of protected sites and species.

9.3.2 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 COS4CLOUD project on cordis.europa.eu

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

Duration: From November 1, 2019 to February 28, 2023

Partners:

- Inria
- TREBOLA ORGANIZACION ECOLOGICA (TREBOLA ORGANIZACION ECOLOGICA), Colombia
- ETHNIKO KAI KAPODISTRIAKO PANEPISTIMIO ATHINON (NKUA), Greece
- BINEO CONSULTING S.L., Spain
- SCHMIDT NORBERT CARL (DDQ), Netherlands
- CONSERVATION EDUCATION AND RESEARCH TRUST (EARTHWATCH), United Kingdom
- AGENCIA ESTATAL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS (CSIC), Spain
- SCIENCE FOR CHANGE, SL (SCIENCE FOR CHANGE), Spain
- SVERIGES LANTBRUKSUNIVERSITET (SWEDISH UNIVERSITY OF AGRICULTURAL SCIENCES), Sweden
- 52 NORTH SPATIAL INFORMATION RESEARCH GMBH (52°North GmbH), Germany
- DYNAIKON LTD, United Kingdom
- VEREIN DER EUROPAEISCHEN BURGERWISSENSCHAFTEN - ECSA E.V. (EUROPEAN CITIZEN SCIENCE ASSOCIATION), Germany
- CENTRO DE INVESTIGACION ECOLOGICA Y APLICACIONES FORESTALES (CREAF - CERCA), Spain
- SECURE DIMENSIONS GMBH (SECURE DIMENSIONS), Germany
- THE OPEN UNIVERSITY (OU), United Kingdom

Inria contact: Alexis Joly

Coordinator: CSIC

Summary: COS4CLOUD (Co-designed citizen observatories for the EOS-Cloud) aims at developing services that address the Open Science challenges shared by Citizen observatories of biodiversity, based on the experience of platforms such as Artportalen, Natusfera, iSpot, as well as other environmental quality monitoring platforms such as FreshWater Watch, KdUINO, OdourCollect, iSpex and CanAir.io. The innovative services will be designed, prototyped and implemented for improving the data and information quality using deep machine learning, automatic video recognition, advanced mobile app interfaces, and other cutting-edge technologies, based on data models and data protocols validated by traditional science. The new services will provide mechanisms to ensure the visibility and recognition of data contributors and the tools to improve networking between various stakeholders. Novel innovative digital services will be developed through the integration of CS products, generated by different providers, following open standards to ensure their interoperability, and offered in agile, fit-for-purpose and sustainable site available through EOSC hub, including a discovery service, to both traditional and citizen scientists. The design of new services will be user oriented, engaging a wide range of stakeholders in society, government, industry, academia, agencies, and research to co-design service requirements. As a result, COS4CLOUD will integrate citizen science in the European Open Science Cloud, bringing Citizen Science (CS) projects as a service for the scientific community and society at large.

9.4 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.

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).

9.4.1 Others

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

Participants: Alexis Joly, Jean-Christophe Lombardo, Hervé Goëau, Hugo Gresse, Mathias Chouet, Antoine Affouard, David Margery.

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 intended to be extended to other members in the coming years.

9.5 Pl@ntNet donations

Participants: Alexis Joly, Jean-Christophe Lombardo, Hervé Goëau, Hugo Gresse, Mathias Chouet, Antoine Affouard, 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 2022, a volume of 245K euros of donations from approximately 24K donors was collected.

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).

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. It is funding the PhD of Joaquim Estopinan and the development of a Python framework for the training of Deep Species Distribution Models.

9.6 Regional initiatives

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. In this context, we developed an *inverse folding* method (see Section 8.3.6) whose aim is to predict the amino acid sequence that is able to generate a protein with a prescribed spatial structure. This has strong applications in drug design.

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 lied in the development of new deep learning models for a better understanding and prediction of animal migrations.

10 Dissemination

Participants: Reza Akbarinia, Alexis Joly, Antoine Liutkus, Florent Masseglia, Esther Pacitti, Christophe Pradal, Patrick Valduriez.

10.1 Promoting scientific activities

10.1.1 Scientific events: organisation

General chair, scientific chair

- A. Joly: chair of LifeCLEF 2022 workshop.

Member of the organizing committees

- P. Valduriez: Workshop on HPC and Data Sciences meet Scientific Computing, CARLA Latin America HPC conference, Porto Alegre, Brazil, September 2022.
- R. Akbarinia: Discovery science conference, Montpellier, October 2022.
- F. Masseglia: Discovery science conference, Montpellier, October 2022.

10.1.2 Scientific events: selection

Member of the conference program committees

- R. Akbarinia: EDBT 2022, EURO-PAR 2022, AIMLSystems 2022, Discovery Science 2022, BDA 2022.
- E. Pacitti: Bases de Données Avancées (BDA), 2022.
- F. Masseglia: Discovery Science 2022, PAKDD 2022, PKDD 2022, DSAA 2022, AIKE 2022, SimBig 2022, ACM SAC DM 2022, ACM SAC DS 2022, ICDM 2022, KDIR 2022, IOTStream 2022, EGC 2021.
- A. Liutkus: NeurIPS 2022, ICML 2022, ICASSP 2022, ICLR 2022

Reviewer

- A. Joly: ICML 2022, ICASSP 2022, NEURIPS 2022

10.1.3 Journal

Member of the editorial boards

- P. Valduriez: Distributed and Parallel Databases.
- R. Akbarinia: Transactions on Large Scale Data and Knowledge Centered Systems (TLDKS).
- E. Pacitti: co-editor of the special issue of the TLDKS journal featuring the best papers of the BDA 2021 conference.

Reviewer - reviewing activities

- R. Akbarinia: IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Transactions on Fuzzy Systems, Journal of super computing.
- A. Joly: TPAMI, Methods in Ecology and Evolution, Ecology Letters, Computers and Electronics in Agriculture, Environmental Research Letters, Multimedia Systems.
- A. Liutkus: IEEE TASLP, IEEE SPL, JOSS, arxiv moderator for eess.AS
- F. Masseglia: Journal of Machine Learning Research (JMLR).

10.1.4 Invited talks

- P. Valduriez: "Innovation : startup strategies", CEFET, Rio de Janeiro, 5 August 2022.
- A. Joly: "Plant disease characterization based on deep learning", CAPTE conference, Avignon, 2022.
- A. Joly: "Environmental ethics and fairness issues in the design of AI algorithms for biodiversity", PIT colloquium, online, 2022.
- A. Joly: "Pl@ntNet overview", German federal workshop on Geotagged photos in the context of the CAP, online 2022.
- A. Joly: "Tutorial on Vision Transformers", Imaginecology, Villeurbane, 2022.

- C. Pradal: "Plant & Crop Modeling", One Planet Fellowship Seminar, Montpellier, May 2022
- C. Pradal: "Towards Digital Twins for Horticulture", keynote talk, IHC 2022, Angers, August 2022
- C. Pradal: "Multiscale plant modeling and phenotyping in OpenAlea", keynote talk, OMICS Symposium 2022, Cali, Colombia, November 2022.

10.1.5 Leadership within the scientific community

- E. Pacitti: Member of the Steering Committee of the BDA conference.
- A. Joly: Foundator & scientific coordinator of LifeCLEF virtual lab: 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 the Acoustic, Speech and Music signal Processing Technical Area Committee of the EURASIP (European Signal Processing Association).
- C. Pradal: Scientific and technical director of the OpenAlea platform: Multiscale platform for plant modelling (2 permanent researchers, 3 PhD students, 2 postdoc, tens of national partners, thousands of end-users).

10.1.6 Scientific expertise

- R. Akbarinia: Expert for STIC AmSud international program.
- A. Joly: expert for the French National HPC grand equipment (GENCI)
- A. Joly: member of the jury of the participatory research award of INRAE
- A. Joly: member of the jury of Inria researchers recruitment (CRCN and ISFP)
- A. Liutkus: expert for ANR.
- C. Pradal: member of the INRAE evaluation comitee CSS (Scientific Specialist Commission) in Plant Integrated Biology

10.1.7 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) and scientific director of the Inria-Brasil strategic partnership.
- F. Masseglia: deputy scientific director of Inria for the domain "Perception, Cognition And Interaction".
- R. Akbarinia: Scientific referent for research data at Inria Antenna of Montpellier; Member of Inria national commission for research data.
- C. Pradal: Team leader with C. Granier of the PhenoMEn team of the AGAP Institut

10.2 Teaching - Supervision - Juries

10.2.1 Teaching

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

Esther Pacitti 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: Big Data Architectures, 24h, level M2, Capgemini Institut.

Alexis Joly:

- Polytech' Montpellier: Content-Based Image Retrieval, 3h, level M2.

Christophe Pradal

- Univ. Montpellier: Root System Modelling, 15h, level M2.

10.2.2 Supervision

PhD & HDR:

- Defended PhD: 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: 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: 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: Cesar Leblanc, Predicting biodiversity future trajectories through deep learning. Advisors: Alexis Joly, Maximilien Servajean, Pierre Bonnet.
- PhD in progress: Tanguy Lefort, Ambiguity of classification labels and expert feedback. Advisors: Joseph Salmon, Benjamin Charlier, Alexis Joly.
- Defended PhD: Lamia Djebour, Adaptive Segmentation Techniques for Efficient Representation of Time Series Datasets. Defended September 2022. Advisors: Reza Akbarinia, Florent Masseglia.
- HDR defended: A. Liutkus on principled methods for signal processing [49].

10.2.3 Juries

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

- R. Akbarinia: Robin CARPENTIER, PhD, University of Paris-Saclay.
- R. Akbarinia: Hussein EL KHANSA, PhD, University of Montpellier.
- A. Joly: Paul Guelorget, PhD, Institut Polytechnique de Paris.
- A. Joly: Yang Loong, PhD, University of Malaya.
- A. Joly: Alexis Delaforge, University of Montpellier.
- A. Joly: Benjamin Deneu, PhD, University of Montpellier.
- A. Liutkus: Pritish Chandna (UPF Barcelona), Stylianos Mimilakis (Fraunhofer IDMT Ilmenau).
- F. Masseglia: Hubert Naacke, HDR defense, Sorbonne Université, reviewer.
- F. Masseglia: Walid Zeghdaoui, PhD, Université Lumière Lyon 2.

10.3 Popularization

10.3.1 Internal or external Inria responsibilities

- F. Masseglia is local contact for Montpellier of the national project "1 Scientifique - 1 Classe, Chiche!".
- F. Masseglia is Member of the strategic committee of Fondation Blaise Pascal.
- Pl@ntNet Community management: 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.

10.3.2 Articles and contents

- A. Joly is co-author of a popularization article in "The Conversation" magazine entitled "PlantNet, eBird, Spipoll, iNaturalist... ces applis au service de l'i-écologie".

11 Scientific production

11.1 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] 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.

[3] 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.

[4] C. Garcin, M. Servajean, A. Joly and J. Salmon. 'Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification'. In: ICML 2022 - 39th International Conference on Machine Learning. Vol. 162. Baltimore, United States: PMLR, 2022, pp. 7208–7222. URL: https://hal.inria.fr/hal-03828747.

[5] 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.

[6] 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.

[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] 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.

[13] 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.

11.2 Publications of the year

International journals

[14] F. Bauget, V. Protto, C. Pradal, Y. Boursiac and C. Maurel. 'A root functional-structural model allows to assess effects of water deficit on water and solute transport parameters'. In: *Journal of Experimental Botany* (14th Dec. 2022). DOI: 10.1093/jxb/erac471. URL: https://hal.inria.fr/hal-03915413.

[15] Q. Bertrand, Q. Klopfenstein, M. Massias, M. Blondel, S. Vaiter, A. Gramfort and J. Salmon. 'Implicit differentiation for fast hyperparameter selection in non-smooth convex learning'. In: *Journal of Machine Learning Research* 23.149 (Apr. 2022), pp. 1–43. URL: https://hal.archives-ouvertes.fr/hal-03228663.

[16] C. Botella, P. Bonnet, C. Hui, A. Joly and D. M. Richardson. 'Dynamic Species Distribution Modeling Reveals the Pivotal Role of Human-Mediated Long-Distance Dispersal in Plant Invasion'. In: *Biology* 11.9 (2022), p. 1293. DOI: 10.3390/biology11091293. URL: https://hal.archives-ouvertes.fr/hal-03771760.

[17] Y. Boursiac, C. Pradal, F. Bauget, M. Lucas, S. Delivorias, C. Godin and C. Maurel. 'Phenotyping and modeling of root hydraulic architecture reveal critical determinants of axial water transport'. In: *Plant Physiology* 190.2 (16th June 2022), pp. 1289–1306. DOI: 10.1093/plphys/kiac281. URL: https://hal.inrae.fr/hal-03701955.

[18] B. Daviet, R. Fernandez, L. Cabrera-Bosquet, C. Pradal and C. Fournier. 'PhenoTrack3D: an automatic high-throughput phenotyping pipeline to track maize organs over time'. In: *Plant Methods* 18.130 (8th Dec. 2022). DOI: 10.1186/s13007-022-00961-4. URL: https://hal.inria.fr/hal-03890913.

[19] B. Deneu, A. Joly, P. Bonnet, M. Servajean and F. Munoz. 'Very High Resolution Species Distribution Modeling Based on Remote Sensing Imagery: How to Capture Fine-Grained and Large-Scale Vegetation Ecology With Convolutional Neural Networks?' In: *Frontiers in Plant Science* 13 (6th May 2022). DOI: 10.3389/fpls.2022.839279. URL: https://hal.inrae.fr/hal-03695760.

[20] L. Djebour, R. Akbarinia and F. Masseglia. 'Variable-Size Segmentation for Time Series Representation'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems* (2022). URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03882927.

[21] J. Estopinan, M. Servajean, P. Bonnet, F. Munoz and A. Joly. 'Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family'. In: *Frontiers in Plant Science* 13 (Apr. 2022), p. 839327. DOI: 10.3389/fpls.2022.839327. URL: https://hal.inrae.fr/hal-03693593.

[22] R. Fernandez, A. Crabos, M. Maillard, P. Nacry and C. Pradal. 'High-throughput and automatic structural and developmental root phenotyping on Arabidopsis seedlings'. In: *Plant Methods* (1st Dec. 2022). DOI: 10.1186/s13007-022-00960-5. URL: https://hal.inria.fr/hal-03881548.

[23] H. Goëau, T. Lorieul, P. Heuret, A. Joly and P. Bonnet. 'Can Artificial Intelligence Help in the Study of Vegetative Growth Patterns from Herbarium Collections? An Evaluation of the Tropical Flora of the French Guiana Forest'. In: *Plants* 11.4 (Feb. 2022), pp. 530–552. DOI: 10.3390/plants11040530. URL: https://hal.inrae.fr/hal-03601464.

[24] R. Jimenez-Peris, D. Burgos-Sancho, F. Ballesteros, M. Patiño-Martinez and P. Valduriez. 'Elastic scalable transaction processing in LeanXcale'. In: *Information Systems* 108 (Sept. 2022), p. 102043. DOI: 10.1016/j.is.2022.102043. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03639381.

[25] J. Liu, C. Bondiombouy, L. Mo and P. Valduriez. 'Two-Phase Scheduling for Efficient Vehicle Sharing'. In: *IEEE Transactions on Intelligent Transportation Systems* 23.1 (2022), pp. 457–470. DOI: 10.1109/TITS.2020.3011952. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02913503.

[26] J. Liu, D. Dong, X. Wang, A. Qin, X. Li, P. Valduriez, D. Dou and D. Yu. 'Large-scale Knowledge Distillation with Elastic Heterogeneous Computing Resources'. In: *Concurrency and Computation: Practice and Experience* Special issue (2022), e7272. DOI: 10.1002/cpe.7272. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03740277.

[27] D. Rosendo, A. Costan, P. Valduriez and G. Antoniu. 'Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review'. In: *Journal of Parallel and Distributed Computing* 166 (Aug. 2022), pp. 71–94. DOI: 10.1016/j.jpdc.2022.04.004. URL: https://hal.archives-ouvertes.fr/hal-03654722.

[28] C. Saint Cast, G. Lobet, L. Cabrera-Bosquet, V. Couvreur, C. Pradal, F. Tardieu and X. Draye. 'Connecting plant phenotyping and modelling communities: lessons from science mapping and operational perspectives'. In: *in silico Plants* 4.1 (1st Jan. 2022), pp. 1–13. DOI: 10.1093/insilicoplants/diac005. URL: https://hal.inrae.fr/hal-03686060.

[29] 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.

[30] E. Senger, S. Osorio, K. Olbricht, P. Shaw, B. Denoyes, J. Davik, S. Predieri, S. Karhu, S. Raubach, N. Lippi, M. Höfer, H. Cockerton, C. Pradal, E. Kafkas, S. Litthauer, B. Usadel and B. Mezzetti. 'Towards smart and sustainable development of modern berry cultivars in Europe'. In: *Plant Journal* 111.5 (2022), pp. 1238–1251. DOI: 10.1111/tpj.15876. URL: https://hal.inrae.fr/hal-03711320.

[31] 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* 34.14 (25th June 2022), e6544. DOI: 10.1002/cpe.6544. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03324881.

[32] S. Woods, M. Daskolia, A. Joly, P. Bonnet, K. Soacha, S. Liñan, T. Woods, J. Piera and L. Ceccaroni. 'How Networks of Citizen Observatories Can Increase the Quality and Quantity of Citizen-Science-Generated Data Used to Monitor SDG Indicators'. In: *Sustainability* 14.7 (30th Mar. 2022). DOI: 10.3390/su14074078. URL: https://hal.inrae.fr/hal-03658842.

International peer-reviewed conferences

[33] L. Djebour, R. Akbarinia and F. Masseglia. 'Parallel Techniques for Variable Size Segmentation of Time Series Datasets'. In: ADBIS 2022 - 26th European Conference on Advances in Databases and Information Systems. Vol. 13389. Lecture Notes in Computer Science. Turin, Italy, 2022, pp. 148–162. DOI: 10.1007/978-3-031-15740-0_12. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03805997.

[34] L. Djebour, R. Akbarinia and F. Masseglia. 'Variable size segmentation for efficient representation and querying of non-uniform time series datasets'. In: SAC 2022 - 37th ACM/SIGAPP Symposium on Applied Computing. Virtual Event, United States, 25th Apr. 2022, pp. 395–402. DOI: 10.1145/3477314.3507000. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03806053.

[35] C. Garcin, M. Servajean, A. Joly and J. Salmon. 'Stochastic smoothing of the top-K calibrated hinge loss for deep imbalanced classification'. In: ICML 2022 - 39th International Conference on Machine Learning. Vol. 162. Baltimore, United States: PMLR, 2022, pp. 7208–7222. URL: https://hal.inria.fr/hal-03828747.

[36] H. Goëau, P. Bonnet and A. Joly. 'Overview of PlantCLEF 2022: Image-based plant identification at global scale'. In: CLEF 2022 - Conference and Labs of the Evaluation Forum. Vol. 3180. CEUR Workshop Proceedings 153. Bologne, Italy, 2022, pp. 1916–1928. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03793591.

[37] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, I. Bolon, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, I. Eggel, P. Bonnet, H. Müller and M. Šulc. 'LifeCLEF 2022 Teaser: An Evaluation of Machine-Learning Based Species Identification and Species Distribution Prediction'. In: *Lecture Notes in Computer Science*. ECIR 2022 - 44th European Conference on Information Retrieval Research. Vol. LNCS-13186. Advances in Information Retrieval : proceedings, part II. Stavanger, Norway: Springer International Publishing, 5th Apr. 2022, pp. 390–399. DOI: 10.1007/978-3-030-99739-7_49. URL: https://hal.inrae.fr/hal-03641389.

[38] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet, M. Šulc and M. Hrúz. 'Overview of LifeCLEF 2022: An Evaluation of Machine-Learning Based Species Identification and Species Distribution Prediction'. In: *Lecture Notes in Computer Science*. CLEF 2022 - 13th International Conference of the CLEF Association. Vol. LNCS-13390. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Bologna, Italy: Springer International Publishing, 25th Aug. 2022, pp. 257–285. DOI: 10.1007/978-3-031-13643-6_19. URL: https://hal.inrae.fr/hal-03807746.

[39] S. Kahl, A. Navine, T. Denton, H. Klinck, P. Hart, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué and A. Joly. 'Overview of BirdCLEF 2022: Endangered bird species recognition in soundscape recordings'. In: CLEF 2022 - Conference and Labs of the Evaluation Forum. Vol. 3180. CEUR Workshop Proceedings 154. Bologne, Italy, 2022, pp. 1929–1939. URL: https://hal.inrae.fr/hal-03791428.

[40] C. Leblanc, A. Joly, T. Lorieul, M. Servajean and P. Bonnet. 'Species Distribution Modeling based on aerial images and environmental features with Convolutional Neural Networks'. In: *CEUR Workshop Proceedings 3180*. CLEF 2022 - Conference and Labs of the Evaluation Forum. Bologne, Italy, 5th Sept. 2022, pp. 2123–2150. URL: https://hal.inrae.fr/hal-03817452.

[41] J. Lima, P. Alpis, R. Salles, L. Escobar, F. Porto, E. Pacitti, R. Coutinho and E. Ogasawara. 'Forward and Backward Inertial Anomaly Detector: A Novel Time Series Event Detection Method'. In: IJCNN 2022 - IEEE International Joint Conference on Neural Networks. Padova, Italy: IEEE, 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892088. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03852429.

[42] T. Lorieul, E. Cole, B. Deneu, M. Servajean, P. Bonnet and A. Joly. 'Overview of GeoLifeCLEF 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data'. In: CLEF 2022 - Conference and Labs of the Evaluation Forum. Vol. 3180. CEUR Workshop Proceedings 155. Bologne, Italy, 2022, pp. 1940–1956. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03793595.

[43] P. Mangold, A. Bellet, J. Salmon and M. Tommasi. 'Differentially Private Coordinate Descent for Composite Empirical Risk Minimization'. In: ICML 2022 - 39th International Conference on Machine Learning. Vol. 162. Baltimore, United States: PMLR, 2022, pp. 14948–14978. URL: https://hal.inria.fr/hal-03424974.

[44] A. C. Silva, P. Valduriez and F. A. Machado Porto. 'Integrating Machine Learning Model Ensembles to the SAVIME Database System'. In: SBBD 2022 - Brazilian Symposium on Databases. Buzios, Brazil, Sept. 2022, pp. 1–8. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03850420.

[45] P. Valduriez and F. Porto. 'Data and Machine Learning Model Management with Gypscie'. In: CARLA 2022 - Workshop on HPC and Data Sciences meet Scientific Computing. Porto Alegre, Brazil, Sept. 2022, pp. 1–2. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03799097.

[46] R. Zorrilla, E. Ogasawara, P. Valduriez and F. Porto. 'A Data-Driven Model Selection Approach to Spatio-Temporal Prediction'. In: SBBD 2022 - Brazilian Symposium on Databases. Buzios, Brazil, 19th Sept. 2022, pp. 1–12. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03798483.

Edition (books, proceedings, special issue of a journal)

[47] *Transactions on Large-Scale Data- and Knowledge-Centered Systems LI: Special Issue on Data Management - Principles, Technologies and Applications*. Lecture Notes in Computer Science (LNCS) 13410 (2022). DOI: 10.1007/978-3-662-66111-6. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03883729.

Doctoral dissertations and habilitation theses

[48] L. Djebour. 'Adaptive Segmentation Techniques for Efficient Representation of Time Series Datasets'. Université de Montpellier, 13th Sept. 2022. URL: https://theses.hal.science/tel-03904591.

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

Reports & preprints

[50] O. Cífka and A. Liutkus. *Black-box language model explanation by context length probing*. 30th Dec. 2022. URL: https://hal.umontpellier.fr/hal-03917930.

[51] B. Dumortier, A. Liutkus, C. Carré and G. Krouk. *PeTriBERT : Augmenting BERT with tridimensional encoding for inverse protein folding and design*. 13th Aug. 2022. DOI: 10.1101/2022.08.10.503344. URL: https://hal.inrae.fr/hal-03759515.

[52] Q. Groom, M. Dillen, W. Addink, A. Ariño, C. Bölling, P. Bonnet, L. Cecchi, E. Ellwood, R. Figueira, P.-Y. Gagnier, O. Grace, A. Güntsch, H. Hardy, P. Huybrechts, R. Hyam, A. Joly, I. Larridon, V. K. Kommineni, L. Livermore, R. J. Lopes, J. Miller, S. Meeus, K. Milleville, M. Pignal, R. Panda, J. Poelen, B. Ristevski, T. Robertson, C. Rufino, J. Santos, M. Schermer, K. Seltmann, B. Scott, H. Teixeira, M. Trekels and J. Gaikwad. *Envisaging a global infrastructure to exploit the potential of digitised collections*. 2022. DOI: 10.22541/au.166678848.82362633/v2. URL: https://hal.inria.fr/hal-03871553.

[53] T. Lefort, B. Charlier, A. Joly and J. Salmon. *Improve learning combining crowdsourced labels by weighting Areas Under the Margin*. 12th Oct. 2022. URL: https://hal.archives-ouvertes.fr/hal-03812716.

[54] Q. Leroy, O. Buisson and A. Joly. *How does the degree of novelty impacts semi-supervised representation learning for novel class retrieval?* 25th Nov. 2022. URL: https://hal.inria.fr/hal-03871584.

[55] P. Mangold, A. Bellet, J. Salmon and M. Tommasi. *High-Dimensional Private Empirical Risk Minimization by Greedy Coordinate Descent*. 2022. URL: https://hal.inria.fr/hal-03714465.