RESEARCH CENTRE

Sophia Antipolis - Méditerranée

IN PARTNERSHIP WITH:

CNRS, Université de Montpellier

2020 ACTIVITY REPORT

Project-Team ZENITH

Scientific Data Management

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

DOMAIN

Perception, Cognition and Interaction

THEME

Data and Knowledge Representation and Processing

Contents

Project-Team ZENITH

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

Keywords

Computer sciences and digital sciences

A1.1. – Architectures

A3.1. – Data

A3.3. – Data and knowledge analysis

A4. – Security and privacy

A4.8. – Privacy-enhancing technologies

A5.4.3. – Content retrieval

A5.7. – Audio modeling and processing

A9.2. – Machine learning

A9.3. – Signal analysis

Other research topics and application domains

B1. – Life sciences

B1.1. – Biology

B1.1.7. – Bioinformatics

B1.1.11. – Plant Biology

B3.3. – Geosciences

B4. – Energy

B6. – IT and telecom

B6.5. – Information systems

1 Team members, visitors, external collaborators

Research Scientists

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

Faculty Member

- Esther Pacitti [Univ of Montpellier, Professor, HDR]

PhD Students

- Heraldo Borges [CEFET/RJ, Brazil]
- Benjamin Deneu [Inria]
- Lamia Djebour [Ministry of Higher Education, Algeria]
- Joaquim Estopinan [Inria, from Nov 2020]
- Camille Garcin [Univ of Montpellier, from Oct 2020]
- Gaetan Heidsieck [Inria]
- Quentin Leroy [INA]
- Titouan Lorieul [Univ of Montpellier, until Sep 2020]
- Khadidja Meguelati [INRAE, until May 2020]
- Daniel Rosendo [Inria, Rennes]
- Alena Shilova [Inria, Bordeaux]

Technical Staff

- Antoine Affouard [Inria, Engineer]
- Heraldo Borges [Inria, Engineer, from Dec 2020]
- Julien Champ [Inria, Engineer]
- Mathias Chouet [Inria, Engineer, from Oct 2020]
- Theo Delfieu [Inria, Engineer, from Jun 2020]
- Baldwin Dumortier [Inria, Engineer, from Feb 2020]

- Hugo Gresse [Inria, Engineer, from Jun 2020]
- Oleksandra Levchenko [Inria, Engineer]
- Tanmoy Mondal [Inria, Engineer, until May 2020]
- Fabian Robert Stoter [Inria, Engineer]

Interns and Apprentices

- Felix Pucheral [Université Gustave Eiffel, from May 2020 until Jun 2020]

Administrative Assistant

- Nathalie Brillouet [Inria]

2 Overall objectives

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

Scientific data is very complex, in particular because of the heterogeneous methods, the uncertainty of the captured data, the inherently multiscale nature (spatial, temporal) of many sciences and the growing use of imaging (e.g. molecular imaging), resulting in data with hundreds of dimensions (attibutes, features, etc.). Modern science research is also highly collaborative, involving scientists from different disciplines (e.g. biologists, soil scientists, and geologists working on an environmental project), in some cases from different organizations in different countries. Each discipline or organization tends to produce and manage its own data, in specific formats, with its own processes. Thus, integrating such distributed data gets difficult as the amounts of heterogeneous data grow. Finally, a major difficulty is to interpret scientific data. Unlike web data, e.g. web page keywords or user recommendations, which regular users can understand, making sense out of scientific data requires high expertise in the scientific domain. Furthermore, interpretation errors can have highly negative consequences, e.g. deploying an oil driller under water at a wrong position.

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

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

Our approach is to capitalize on the principles of distributed and parallel data management. In particular, we exploit: high-level languages as the basis for data independence and automatic optimization; declarative languages to manipulate data and workflows; and highly distributed and parallel environments such as cluster and cloud for scalability and performance. We also exploit machine learning, probabilities and statistics for high-dimensional data processing, data analytics and data search.

3 Research program

3.1 Distributed Data Management

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

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

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

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

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

3.2 Big Data

Big data (like its relative, data science) has become a buzz word, with different meanings depending on your perspective, e.g. 100 terabytes is big for a transaction processing system, but small for a web search engine.

Although big data has been around for a long time, it is now more important than ever. We can see overwhelming amounts of data generated by all kinds of devices, networks and programs, e.g. sensors,

mobile devices, connected objects (IoT), social networks, computer simulations, satellites, radiotelescopes, etc. Storage capacity has doubled every 3 years since 1980 with prices steadily going down, making it affordable to keep more data around. Furthermore, massive data can produce high-value information and knowledge, which is critical for data analysis, decision support, forecasting, business intelligence, research, (data-intensive) science, etc.

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

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

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

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

3.3 Data Integration

Scientists can rely on web tools to quickly share their data and/or knowledge. Therefore, when performing a given study, a scientist would typically need to access and integrate data from many data sources (including public databases). Data integration can be either physical or logical. In the former, the source data are integrated and materialized in a data warehouse. In logical integration, the integrated data are not materialized, but accessed indirectly through a global (or mediated) schema using a data integration system. These two approaches have different trade-offs, e.g. efficient analytics but only on historical data for data warehousing versus real-time access to data sources for data integration systems (e.g. web price comparators).

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

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

3.4 Data Analytics

Data analytics refers to a set of techniques to draw conclusions through data examination. It involves data mining, statistics, and data management, and is applied to categorical and continuous data. In the Zenith team, we are interested in both of these data types. Categorical data designates a set of data that can be described as "check boxes". It can be names, products, items, towns, etc. A common illustration is the market basket data, where each item bought by a client is recorded and the set of items is the basket. The typical data mining problems with this kind of data are:

- **Frequent itemsets and association rules**. In this case, the data is usually a table with a high number of rows and the data mining algorithm extracts correlations between column values. A typical example of frequent itemset from a sensor network in a smart building would say that "in 20% rooms, the door is closed, the room is empty, and lights are on."
- **Frequent sequential pattern extraction.** This problem is very similar to frequent itemset discovery but considering the order between. In the smart building example, a frequent sequence could say that "in 40% of rooms, lights are on at time i, the room is empty at time $i + j$ and the door is closed at time $i + j + k$".
- **Clustering**. The goal of clustering is to group together similar data while ensuring that dissimilar data will not be in the same cluster. In our example of smart buildings, we could find clusters of rooms, where offices will be in one category and copy machine rooms in another because of their differences (hours of people presence, number of times lights are turned on/off, etc.).

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

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

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

3.5 High Dimensional Data Processing and Search

High dimensionality is inherent in applications involving images, audio and text as well as in many scientific applications involving raster data or high-throughput data. Because of the *dimensionality curse*, technologies for processing and analyzing such data cannot rely on traditional relational DBMS or data mining methods. It rather requires machine learning methods such as dimensionality reduction, representation learning or random projection. The activity of Zenith in this domain focuses on methods for large-scale data processing and search, in particular in the presence of strong uncertainty and/or ambiguity. Actually, while small datasets are often characterized by a careful collection process, massive amounts of data often come with outliers and spurrious items, because it appears impossible to guarantee

faultless collection at massive bandwidth. Another source of noise is often the sensor itself, that may be of low quality but of high sampling rate, or even the actual content, e.g. in cultural heritage applications when historical content appears seriously damaged by time. To attack these difficult problems, we focus on the following research topics:

- **Uncertainty estimation**. Items in massive datasets may either be uncertain, e.g. for automatically annotated data as in image analysis, or be more or less severely corrupted by noise, e.g. in noisy audio recordings or in the presence of faulty sensors. In both cases, the concept of *uncertainty* is central for the end-user to exploit the content. In this context, we use probability theory to quantify uncertainty and propose machine learning algorithms that may operate robustly, or at least assess the quality of their output. This vast topic of research is guided by large-scale applications (both data search and data denoising), and our research is oriented towards computationally effective methods.

- **Deep neural networks**. A major breakthrough in machine learning performance has been the advent of deep neural networks, which are characterized by high numbers (millions) of parameters and scalable learning procedures. We are striving towards original architectures and methods that are theoretically grounded and offer state-of-the-art performance for data search and processing. The specific challenges we investigate are: very high dimensionality for static data and very long-term dependency for temporal data, both in the case of possibly strong uncertainty or ambiguity (e.g. hundreds of thousands of classes).

- **Community service**. Research in machine learning is guided by applications. In Zenith, two main communities are targeted: botany, and digital humanities. In both cases, our observation is that significant breakthroughs may be achieved by connecting these communities to machine learning researchers. This may be achieved through wording application-specific problems in classical machine learning parlance. Thus, the team is actively involved in the organization of international evaluation campaigns that allow machine learning researchers to propose new methods while solving important application problems.

4 Application domains

4.1 Data-intensive Scientific Applications

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

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

- **Management of astronomical catalogs**. An example of data-intensive scientific applications is the management of astronomical catalogs generated by the Dark Energy Survey (DES) project on which we are collaborating with researchers from Brazil. In this project, huge tables with billions of tuples and hundreds of attributes (corresponding to dimensions, mainly double precision real numbers) store the collected sky data. Data are appended to the catalog database as new observations are performed and the resulting database size has hundreds of TB. Scientists around the globe can query the database with queries that may contain a considerable number of attributes. The volume of data that this application holds poses important challenges for data management. In particular, efficient solutions are needed to partition and distribute the data in several servers. An efficient partitioning scheme should try to minimize the number of fragments accessed in the execution of a query, thus reducing the overhead associated to handle the distributed execution.

- **Personal health data analysis and privacy**. Today, it is possible to acquire data on many domains related to personal data. For instance, one can collect data on her daily activities, habits or health. It is also possible to measure performance in sports. This can be done thanks to sensors, communicating devices or even connected glasses. Such data, once acquired, can lead to valuable knowledge for these domains. For people having a specific disease, it might be important to know if they belong to a specific category that needs particular care. For an individual, it can be interesting to find a category that corresponds to her performances in a specific sport and then adapt her training with an adequate program. Meanwhile, for privacy reasons, people will be reluctant to share their personal data and make them public. Therefore, it is important to provide them with solutions that can extract such knowledge from everybody's data, while guaranteeing that their private data won't be disclosed to anyone.

- **Botanical data sharing**. Botanical data is highly decentralized and heterogeneous. Each actor has its own expertise domain, hosts its own data, and describes them in a specific format. Furthermore, botanical data is complex. A single plant's observation might include many structured and unstructured tags, several images of different organs, some empirical measurements and a few other contextual data (time, location, author, etc.). A noticeable consequence is that simply identifying plant species is often a very difficult task; even for the botanists themselves (the so-called taxonomic gap). Botanical data sharing should thus speed up the integration of raw observation data, while providing users an easy and efficient access to integrated data. This requires to deal with social-based data integration and sharing, massive data analysis and scalable content-based information retrieval. We address this application in the context of the French initiative Pl@ntNet, with CIRAD and IRD.

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

- **Audio heritage preservation**. Since the end of the 19^{th} century, France has commissioned ethnologists to record the world's immaterial audio heritage. This results in datasets of dozens of thousands of audio recordings from all countries and more than 1200 ethnies. Today, this data is gathered under the name of 'Archives du CNRS — Musée de l'Homme' and is handled by the CREM (Centre de Recherche en Ethno-Musicologie). Profesional scientists in digital humanities are accessing this data daily for their investigations, and several important challenges arise to ease their work. The KAMoulox project, lead by A. Liutkus, targets at offering online processing tools for the scientists to automatically restore this old material on demand. In the same vein, we have an ongoing collaboration with Radio France, that has large amounts of archives to restore, for repurposing applications.

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

5 Social and environmental responsibility

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

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

- We have also started to address the (major) problem of energy consumption of our ML models, by introducing energy-based metrics to assess the energy consumption during the training on GPU of our ML models. Furthermore, we want to improve training pipelines that reduce the need for training models from scratch. At inference, network compression methods can reduce the memory footprint and the computational requirements when deploying models.
- In the design of the Pl@ntnet mobile application, we adopt an eco-responsible approach, taking care not to integrate addictive, energy-intensive or non-essential functionalities to uses that promote the preservation of biodiversity and environment.
- To reduce our carbon footprint, we reduce to the minimum the number of long-distance trips, and favor train as much as possible. We also trade conference publications for journal publications, to avoid traveling. For instance, in 2020, we have 27 journal publications versus 19 conference publications.

6 Highlights of the year

6.1 Awards

- The Pl@ntnet project (Alexis Joly, Pierre Bonnet, Hervé Goëau, Julien Champ, Jean-Christophe Lombardo and Antoine Affouard) won the innovation price from Inria, the French Academy of Science and Dassault Systems.
- The paper "Distributed Caching of Scientific Workflows in Multisite Cloud" [46] by Gaëtan Heidsieck, Daniel de Oliveira, Esther Pacitti, Christophe Pradal, François Tardieu, and Patrick Valduriez, obtained the best paper award from DEXA 2020.
- Fabian Stoter in collaboration with Inria Nancy won the first place at the Global Pytorch Summer Hackaton 2020, with the DeMask software that provides an end-to-end model for enhancing speech while wearing face masks.
- Antoine Liutkus received an IEEE outstanding reviewer award for his reviewing in IEEE transactions and conferences (notably TASLP and ICASSP).

6.2 International

- The Inria Brasil web site has been created to reflect the long-live collaboration between Inria and LNCC, the Brazilian National Scientific Computing Laboratory, and associated Brazilian universities in HPC, AI, Data Science and Scientific Computing. The collaboration is headed by Frédéric Valentin (LNCC, Inria International Chair) and Patrick Valduriez.
- Pl@ntNet has become a data provider to GBIF, the world's largest government-funded biodiversity data platform.

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 (20M downloads) and allows identifying about 30K plant species at present time.

Publication: hal-01629195

Contact: Alexis Joly

Participants: Antoine Affouard, Jean-Christophe Lombardo, Pierre Bonnet, Hervé Goëau, Mathias Chouet, 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 DfAnalyzer

Name: Dataflow Analysis

Keywords: Data management, Monitoring, Runtime Analysis

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

URL: https://github.com/vssousa/dfanalyzer-spark

Publication: lirmm-01867887

Contact: Patrick Valduriez

Participants: Vitor Sousa Silva, Daniel De Oliveira, Marta Mattoso, Patrick Valduriez

Partners: COPPE/UFRJ, Uff

7.1.4 CloudMdsQL Compiler

Keywords: Optimizing compiler, NoSQL, Data integration

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

Publication: lirmm-01184016

Authors: Boyan Kolev, Patrick Valduriez

Contact: Patrick Valduriez

Participants: Boyan Kolev, Oleksandra Levchenko, Patrick Valduriez

7.1.5 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

7.1.6 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

Authors: Samuel Dufour Kowalski, Christophe Pradal

Contact: Christophe Pradal

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

Partners: CIRAD, INRA

7.1.7 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

7.1.8 VersionClimber

Keywords: Software engineering, Deployment, Versionning

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

URL: https://versionclimber.readthedocs.io/

Publication: hal-02262591

Contact: Christophe Pradal

Participants: Christophe Pradal, Dennis Shasha, Sarah Cohen-Boulakia, Patrick Valduriez

Partners: CIRAD, New York University

7.1.9 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: This software 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

7.1.10 TDB

Keywords: Data assimilation, Big data, Data extraction

Scientific Description: The TDB software 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: The TDB software is composed of two core submodules: First, a data extraction pipeline permits to scrape 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 permits to transform 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 one of theses core components comprises several submodules, notably network handling and audio transcoding. The TDB software must hence be understood 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.

Authors: Fabian Robert Stoter, Antoine Liutkus

Contact: Antoine Liutkus

Participants: Antoine Liutkus, Fabian Robert Stoter

7.1.11 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-definite 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.

The software comes with full documentation, detailed comments and unit tests.

Functional Description: UMX-PRO is a TensorFlow v2 implementation for an end-to-end music separation system including network architecture, data pipeline, training code, inference code as well as pre-trained weights.

Authors: Antoine Liutkus, Fabian Robert Stoter

Contact: Antoine Liutkus

8 New results

8.1 Scientific Workflows

8.1.1 Runtime Dataflow Analysis with DfAnalyzer

Participants Patrick Valduriez.

DfAnalyzer is a tool for monitoring, debugging, and analyzing dataflows generated by Computational Science and Engineering (CSE) applications. It collects strategic raw data, registering provenance data, and enabling query processing, all asynchronously and at runtime. DfAnalyzer provides lightweight dataflow components to be invoked by CSE applications using HPC, in the same way computational scientists plug HPC (e.g., PETSc) and visualization (e.g., ParaView) libraries. In [37], we show DfAnalyzer's main functionalities and how to analyze dataflows in CSE applications at runtime. The performance evaluation of CSE executions for a complex multiphysics application shows that DfAnalyzer has negligible time overhead on the total elapsed time.

8.1.2 Data Reduction in Scientific Workflows

Participants Patrick Valduriez.

Scientific workflows need to be iteratively, and often interactively, executed for large input datasets. Reducing data from input datasets is a powerful way to reduce overall execution time in such workflows. In [38], we adopt the “human-in-the-loop” approach, which enables users to steer the running workflow and reduce subsets from datasets online. We propose an adaptive workflow monitoring approach that combines provenance data monitoring and computational steering to support users in analyzing the evolution of key parameters and determining the subset of data to remove. We extend a provenance data model to keep track of users' interactions when they reduce data at runtime. In our experimental validation, we develop a test case from the oil and gas domain, using a 936-cores cluster. The results on this test case show that the approach yields reductions of 32% of execution time and 14% of the data processed.

8.1.3 Caching of Scientific Workflows in Multisite Cloud

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

Many scientific experiments are performed using scientific workflows, which are becoming more and more data-intensive. We consider the efficient execution of such workflows in the cloud, leveraging

the heterogeneous resources available at multiple cloud sites (geo-distributed data centers). Since it is common for workflow users to reuse code or data from other workflows, a promising approach for efficient workflow execution is to cache intermediate data in order to avoid re-executing entire workflows. In [25], we propose an adaptive caching solution for data-intensive workflows in the cloud. Our solution is based on a new scientific workflow management architecture that automatically manages the storage and reuse of intermediate data and adapts to the variations in task execution times and output data size. In [46] , we propose a distributed solution for caching of scientific workflows in a multisite cloud. We implemented our solutions fro adaptive and distributed caching in the OpenAlea workflow system, together with cache-aware distributed scheduling algorithms. Our experimental evaluation on a three-site cloud with a data-intensive application in plant phenotyping shows that our solution can yield major performance gains.

8.2 Query Processing

8.2.1 Uncertainty Quantification Queries over Big Spatial Data

Participants Esther Pacitti, Patrick Valduriez.

We consider big spatial data, which is typically produced in scientific areas such as geological or seismic interpretation. The spatial data can be produced by observation (e.g. using sensors or soil instruments) or numerical simulation programs and correspond to points that represent a 3D soil cube area. However, errors in signal processing and modeling create some uncertainties associated with model calculations of true, physical quantities of interest (QOIs), and thus a lack of accuracy in identifying geological or seismic phenomenons. Uncertainty Quantification (UQ) is the process of quantifying such uncertainties. In [29], we consider the problem of answering UQ queries over large spatio-temporal simulation results. We propose the SUQ2 method based on the Generalized Lambda Distribution (GLD) function. To further analyze uncertainty, the main solution is to compute a Probability Density Function (PDF) of each point in the spatial cube area. However, computing PDFs on big spatial data can be very time consuming (from several hours to even months on a computer cluster). In [32], we propose a new solution to efficiently compute such PDFs in parallel using Spark, with three methods: data grouping, machine learning prediction and sampling. We evaluate our solution by extensive experiments on different computer clusters using big data ranging from hundreds of GB to several TB. The experimental results show that our solution scales up very well and can reduce the execution time by a factor of 33 (in the order of seconds or minutes) compared with a baseline method.

8.3 Data Analytics

8.3.1 Massively Distributed Time Series Indexing

Participants Djamel Edine Yagoubi, Reza Akbarinia, Florent Masseglia.

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

In [40], we propose a parallel solution to construct the state of the art iSAX-based index over billions of time series by carefully distributing the workload. Our solution takes advantage of parallel data processing frameworks such as MapReduce or Spark. We provide dedicated strategies and algorithms for a deep combination of parallelism and indexing techniques. We also propose a parallel query processing algorithm that, given a query, exploits the available processing nodes to answer the query in parallel using the constructed parallel index. We implemented our algorithms, and evaluated their performance

over large volumes of data (up to 4 billion time series of length 256, for a total volume of 6 TB). Our experiments demonstrate high performance with an indexing time of less than 2 hours for more than 1 billion time series, while the state of the art centralized algorithm needs more than 5 days. They also illustrate that our approach is able to process 10M queries in less than 140 seconds, while the centralized algorithm needs almost 2300 seconds.

8.3.2 Efficient Similarity Search in Large Time Series Databases

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

Fast and accurate similarity search is critical to performing many data mining tasks like motif discovery, classification or clustering.

In [30], we present our parallel solutions, developed based on two state-of-the-art approaches iSAX and sketch, for k nearest-neighbor (kNN) search in large databases of time series. We compare the two solutions based on various measures of quality and time performance, and propose a tool that uses the characteristics of application data to determine which solution to choose for that application and how to set the parameters for that solution. Our experiments show that: (i) iSAX and its derivatives perform best in both time and quality when the time series can be characterized by a few low frequency Fourier Coefficients, a regime where the iSAX pruning approach works well. (ii) iSAX performs significantly less well when high frequency Fourier Coefficients have much of the energy of the time series. (iii) A random projection approach based on sketches by contrast is more or less independent of the frequency power spectrum. The experiments show the close relationship between pruning ratio and time for exact iSAX as well as between pruning ratio and the quality of approximate iSAX. Our toolkit analyzes typical time series of an application (i) to determine optimal segment sizes for iSAX and (ii) when to use Parallel Sketches instead of iSAX. Our solutions have been implemented using Spark, evaluated over a cluster of nodes, and have been applied to both real and synthetic data.

8.3.3 Efficient kNN Search in Large Chemometrics Databases

Participants Reza Akbarinia, Florent Masseglia.

Chemometrics scientists exploit a wide range of tools for the analysis and interpretation of spectroscopic data. One of the objectives of these tools is to associate spectral information with physico-chemical properties in order to predict their properties. Among them, a reference method is PLSR (Partial Least Squares Regression). It is composed of a dimension reduction step (PLS) followed by a regression on the scores produced. A well known issue regarding PLS lies in the difficulty to apprehend non linearities. As a solution, an extension of the method, called KNN-PLS, was developed. However, this solution is based on a neighborhood selection method whose execution time is highly dependent on the size of the database, leading to prohibitive response times.

In [34], we propose a new method, called parSketch-PLS, designed to perform kNN search in large spectral databases. It combines parSketch, a solution we developed for indexing and querying time series, and the PLS method. We compare the PLS and KNN-PLS methods with the parSketch-PLS method. The experiments illustrate that parSketch-PLS offers a good operational trade-off between prediction performance and computational cost. Furthermore, we propose a framework to interpret the neighborhoods returned by comparing their relative sizes with the evolution of performance and the input parameters of parSketch-PLS.

8.3.4 Time Series Clustering via Dirichlet Mixture Models

Participants Khadidja Meguelati, Florent Masseglia.

Dirichlet Process Mixture (DPM) is a model for clustering, with the advantage of automatic discovery of clusters and nice properties, such as the potential convergence to the actual clusters in the data. These advantages come at the price of prohibitive response times, which impairs its adoption and makes centralized DPM approaches inefficient. In [52], we gave a demonstration of DC-DPM (Distributed Computing DPM) and HD4C (High Dimensional Data Distributed Dirichlet Clustering). DC-DPM is a parallel clustering solution that gracefully scales to millions of data points while remaining DPM compliant, which is the challenge of distributing this process. HD4C (High Dimensional Data Distributed Dirichlet Clustering) is a parallel clustering solution that addresses the curse of dimensionality by distributed computing and performs clustering of high dimensional data such as time series (as a function of time), hyperspectral data (as a function of wavelength) etc. The demonstration site is available at: http://147.100.179.112:3838/team/kmeguelati/dpmclustering/

8.3.5 Spatial-Time Series Clustering

Participants Heraldo Borges, Florent Masseglia.

Discovering motifs in time series data and clustering such data have been widely explored. However, when it comes to spatial-time series, a clear gap can be observed according to the literature review. [12] presents a short overview of space–time series clustering, which can be generally grouped into three main categories such as: hierarchical, partitioning-based, and overlapping clustering. The first category is to identify hierarchies in space–time series data. The second category focuses on determining disjoint partitions among the space–time series data, whereas the third category explores fuzzy logic to determine the different correlations between the space–time series clusters. This work can provide guidance to practitioners for selecting the most suitable methods for their used cases, domains, and applications. [16] presents an approach to discover and rank motifs in spatial-time series, denominated Combined Series Approach (CSA). CSA is based on partitioning the spatial-time series into blocks. Inside each block, subsequences of spatial-time series are combined by means of a hash-based motif discovery algorithm. The approach was evaluated using both synthetic and seismic datasets. CSA outperforms traditional methods designed only for time series. CSA was also able to prioritize motifs that were meaningful both in the context of synthetic data and also according to seismic specialists.

8.4 Machine Learning for Biodiversity Informatics

8.4.1 Machine Learning using Digitized Herbarium Specimens in Phenology

Participants Julien Champ, Alexis Joly.

Phenology, i.e., the timing of life-history events, is a key trait for understanding responses of organisms to climate. The digitization and online mobilization of herbarium specimens is rapidly advancing our understanding of plant phenological response to climate and climatic change. The current practice of manually harvesting data from individual specimens, however, greatly restricts our ability to scale-up data collection. Our recent investigations have demonstrated that machine learning can facilitate this effort [36]. However, present attempts have focused largely on simplistic binary coding of reproductive phenology (e.g., presence/absence of flowers). In [21] (jointly with Harvard University, Boston University, UFBA and CIRAD), we use crowd-sourced phenological data of buds, flowers, and fruits from more than 3,000 specimens of six common wildflower species of the eastern United States to train models using Mask R-CNN to segment and count phenological features. A single global model was able to automate the binary coding of each of the three reproductive stages with more than 87% accuracy. We

also successfully estimated the relative abundance of each reproductive structure on a specimen with more than 90% accuracy. Precise counting of features was also successful, but accuracy varied with phenological stage and taxon. Specifically, counting flowers was significantly less accurate than buds or fruits likely due to their morphological variability on pressed specimens. Moreover, our Mask R-CNN model provided more reliable data than non-expert crowd-sourcers but not botanical experts, highlighting the importance of high-quality human training data. Finally, we also demonstrated the transferability of our model to automated phenophase detection and counting of the three Trillium species, which have large and conspicuously-shaped reproductive organs. These results highlight the promise of our two-phase crowd-sourcing and machine-learning pipeline to segment and count reproductive features of herbarium specimens, thus providing high-quality data with which to investigate plant responses to ongoing climatic change.

8.4.2 Analysis of the Use of Pl@ntNet Services for Biodiversity conservation

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

In [11] (jointly with the UK Centre for Ecology and Hydrology and CIRAD), we apply the Pl@ntNet identification engine to social media imagery (Flickr in particular) to generate new biodiversity observations. We find that this approach is able to generate new data on species occurrence but that there are biases in both the social media data and the AI image classifier that need to be considered in analyses. This approach could be applied outside the biodiversity domain, to any phenomena of interest that may be captured in social media imagery. The checklist we provide at the end of this paper should therefore be of interest to anyone considering this approach to generating new data. In [15], we present two Pl@ntNet-based citizen science initiatives piloted by conservation practitioners in Europe (France) and Africa (Kenya). We discuss various perspectives of AI-based plants identification, including benefits and limitations. Based on the experiences of field managers, we formulate several recommendations for future initiatives. The recommendations are aimed at a diverse group of conservation managers and citizen science practitioners.

8.4.3 New Methods and Perspectives on Plant Disease Characterization

Participants Herve Goeau, Alexis Joly.

The control of plant diseases is a major challenge to ensure global food security and sustainable agriculture. Several recent studies have proposed to improve existing procedures for early detection of plant diseases through automatic image recognition systems based on deep learning. In [28], we study these methods in detail, especially those based on convolutional neural networks. We first examine whether it is more relevant to fine-tune a pre-trained model on a plant identification task rather than a general object recognition task. In particular, we show through visualization techniques, that the characteristics learned differ according to the approach adopted and that they do not necessarily focus on the part affected by the disease. Therefore, we introduce a more intuitive method that considers diseases independently of crops, and show that it is more effective than the classic crop-disease pair approach, especially when dealing with disease involving crops that are not illustrated in the training database. In [27], we develop a new technique based on a Recurrent Neural Network (RNN) to automatically locate infected regions and extract relevant features for disease classification. We show experimentally that our RNN-based approach is more robust and has greater ability to generalize to unseen infected crop species and different plant disease domain images compared to classical CNN approaches. We also show that our approach is capable of accurately locating infectious diseases in plants. Our approach, which has been tested on a large number of plant species, should thus contribute to the development of more effective means of detecting and classifying crop pathogens in the near future.

8.4.4 Evaluation of Species Identification and Prediction Algorithms

Participants Alexis Joly, Herve Goeau, Benjamin Deneu, Titouan Lorieul, Fabian Robert Stoter.

We run a new edition of the LifeCLEF evaluation campaign [48] with the involvement of 16 research teams worldwide. The main outcomes of the 2020 edition are:

- **Location-based Species prediction (GeoLifeCLEF).** Jointly with Caltech and Microsoft research, we released a new outstanding dataset of 1.9 million species observations paired with high-resolution remote sensing imagery, land cover data, and altitude, in addition to traditional low-resolution climate and soil variables [64]. It allowed to highlight for the first time the ability of remote sensing imagery and convolutional neural networks to improve predictive performance of Species Distribution Models (SDM), complementary to traditional approaches [43].
- **Plant identification (PlantCLEF)**. The PlantCLEF 2020 challenge was designed to evaluate to what extent automated identification on the flora of data deficient regions can be improved by the use of herbarium collections. It is based on a dataset of about 1,000 species mainly focused on the South America's Guiana Shield, an area known to have one of the greatest diversity of plants in the world. The challenge was evaluated as a cross-domain classification task where the training set consist of several hundred thousand herbarium sheets and few thousand of photos to enable learning a mapping between the two domains. The results revealed that the recent advances in domain adaptation enable the use of herbarium data to facilitate the identification of rare tropical species for which no or very few other training photos are available [59].
- **Bird sounds recognition (BirdCLEF)**. Passive acoustic monitoring is a cornerstone of the assessment of ecosystem health and the improvement of automated assessment systems has the potential to have a transformative impact on global biodiversity monitoring. The BirdCLEF challenge [49] focuses on the development of reliable detection systems for avian vocalizations in continuous soundscape data. It is the largest evaluation that specifically aims at developing state-of-the-art classifiers to help researchers to cope with conservation challenges of our time. Results obtained in 2020 show Deep neural networks provide good overall baselines but there is still large room for improvement.

8.4.5 Deep Learning Based Instance Segmentation for Precision Agriculture

Participants Julien Champ, Herve Goeau, Alexis Joly.

Weed removal in agriculture is typically achieved using herbicides. The use of autonomous robots to reduce weeds is a promising alternative solution, although their implementation requires the precise detection and identification of crops and weeds to allow an efficient action. In [20] we propose an instance segmentation approach to this problem making use of a Mask R-CNN model for weeds and crops detection on farmland. Therefore, we created a new data set comprising field images on which the outlines of 2489 specimens from two crop species and four weed species were manually drawn. The probability of detection using the model was quite good but varied significantly depending on the species and size of the plants. In practice, between 10% and 60% of weeds could be removed without too high of a risk of confusion with crop plants. Furthermore, we show that the segmentation of each plant enabled the determination of precise action points such as the barycenter of the plant surface.

8.4.6 Correcting bias in Species Distribution Models

Participants Christophe Botella, Alexis Joly.

Presence-only Species Distribution Models require background points, which should be consistent with sampling effort across the environmental space to avoid bias. A standard approach is to use uniformly distributed background points (UB). When multiple species are sampled, another approach is to use a set of occurrences from a Target-Group of species as background points (TGOB). In this work [17], we investigate estimation biases when applying TGOB and UB to opportunistic naturalist occurrences. We model species occurrences and observation process as a thinned Poisson point process, and express asymptotic likelihoods of UB and TGOB as a divergence between environmental densities, in order to characterize biases in species niche estimation. To illustrate our results, we simulate species occurrences with different types of niche (specialist/generalist, typical/marginal), sampling effort and TG species density. We conclude that none of the methods are immune to estimation bias, although the pitfalls are different.

8.5 Machine Learning for audio and long time series

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

8.5.1 Setting the State of the Art in Music Demixing

Participants Fabian-Robert Söter, Antoine Liutkus.

We have been very active for years in the topic of music demixing, with a prominent role in defining the state of the art in this domain. Our contributions this year in the domain are numerous. After years of leading SiSEC, the international separation evaluation campaign, we handled the lead to another team. This year, we continued handling our *MUSDB18* dataset, which takes some time, notably for granting access rights to all the interested teams and sending out links. It is the #11 dataset on Zenodo with 7500 downloads, making it the most popular music dataset worldwide.

We maintain the *open-unmix* software, which is an established reference implementation for music source separation. We also participated in the design and implementation of Asteroid [53], a research effort towards a unified software platform for audio separtion research, lead by the Multispeech Inria team. One of our contributions with Asteroid won the first place at the Global Pytorch Summer Hackaton 2020 organized by Facebook.

8.5.2 Deep models for audio and long-range data

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

Our strategy is to go beyond our current expertise on music demixing to address the new and very active topics of audio style transfer, enhancement, and generation, with large scale applications for the exploitation and repurposing of large audio corpora. This means leaving our comfort zone on source separation to address new exciting challenges, notably the use of Transformers in audio. For this purpose, our strategy is to develop new deep learning models, based on Transformers, that allow processing very long time series. On the engineering side, our contributions mostly concern data management and curating large corpora, as mentioned above.

An ongoing research effort concerns *long-term* interactions in time-series. We fully embraced the recently proposed Transformer architecture, that models inter-sample dependencies in a very flexible manner. However, it couldn't properly account for *relative* attention at scale. A significant research effort was done in this direction, and papers will be submitted soon. In preceding years, we proposed several models to leverage time-frequency dependencies for processing (Kernel Additive Models). Current trends make it possible to train such dependencies.

8.5.3 Robust Probabilistic Models for Time-series

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

Processing large amounts of data for denoising or analysis comes with the need to devise models that are robust to outliers and permit efficient inference. For this purpose, we advocate the use of non-Gaussian models for this purpose, which are less sensitive to data-uncertainty. We developed a new filtering paradigm that goes beyond least-squares estimation. In collaboration with researchers from Telecom Paris, we introduce several methods that generalize least-squares Wiener filtering to the case of α-stable processes. This very theoretically important contribution has been published as a journal paper [22].

9 Bilateral contracts and grants with industry

9.1 INA (2019-2022)

Participants Quentin Leroy, Alexis Joly.

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

9.2 Transfer of UMX-PRO

Participants Antoine Liutkus, Fabian-Robert Stöter.

A. Liutkus and F.-R. Stoter are the authors of the UMX-PRO software, which has been transferred to a north-american company for several hundred thousand euros.This software is a complete solution for audio source separation.All other details regarding this software transfer are confidential and subject to a non-disclosure agreement.

9.3 Transfer of TDB

Participants Antoine Liutkus, Fabian-Robert Stöter.

A. Liutkus and F.-R. Stoter are the authors of the TDB software, which is a solution for audio scraping. It allows gathering the largest audio separation dataset available today, and has been successfully transferred to a European company named AudioSourceRE.

10 Partnerships and cooperations

10.1 International initiatives

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

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

10.1.1 Inria associate team not involved in an IIL

HPDaSc

Title: *High Performance Data Science (HPDaSc)*

Site web: https://team.inria.fr/zenith/hpdasc/

Duration: 2020 - 2022

Coordinator: Patrick Valduriez

Partners:

- *LNCC, COPPE/UFRJ, UFF and CEFET, (Brazil)*

Inria contact: Patrick Valduriez

Summary: Data-intensive science refers to modern science, such as astronomy, geoscience or life science, where researchers need to manipulate and explore massive datasets produced by observation or simulation. It requires the integration of two fairly different paradigms: high-performance computing (HPC) and data science. We address the following requirements for high-performance data science (HPDaSc): support realtime analytics and visualization (in either in situ or in transit architectures) to help make high-impact online decisions; combine ML with analytics and simulation, which implies dealing with uncertain training data, autonomously built ML models and combine ML models and simulation models; support scientific workflows that combine analytics, modeling and simulation, and exploit provenance in realtime and HIL (Human in the Loop) for efficient workflow execution.

To address these requirements, we will exploit new distributed and parallel architectures and design new techniques for ML, realtime analytics and scientific workflow management. The architectures will be in the context of multisite cloud, with heterogeneous data centers with data nodes, compute nodes and GPUs. We will validate our techniques with major software systems on real applications with real data. The main systems will be OpenAlea and Pl@ntnet from Zenith and DfAnalyzer and SAVIME from the Brazilian side. The main applications will be in agronomy and plant phenotyping (with plant biologists from CIRAD and INRA), biodiversity informatics (with biodiversity scientists from LNCC and botanists from CIRAD), and oil & gas (with geoscientists from UFRJ and Petrobras).

10.1.2 Inria international partners

Informal international partners We have regular scientific relationships with research laboratories in:

- North America: Univ. of Waterloo (Tamer Özsu), UCSB Santa Barbara (Divy Agrawal and Amr El Abbadi), Northwestern Univ. (Chicago), university of Florida (Pamela Soltis, Cheryl Porter, Gil Nelson), Harvard (Charles Davis), UCSB (Susan Mazer).
- Asia: National Univ. of Singapore (Beng Chin Ooi, Stéphane Bressan), Wonkwang University, Korea (Kwangjin Park), Kyoto University (Japan), Tokyo University (Hiroyoshi Iwata), Academica Sinica, Taiwan (Y. Yang).

- Europe: Univ. of Madrid (Ricardo Jiménez-Periz), UPC Barcelona (Josep Lluis Larriba Pey), HES-SO (Henning Müller), University of Catania (Concetto Spampinatto), Cork School of Music (Ireland), RWTH (Aachen, Germany), Chemnitz technical university (Stefan Kahl), Berlin Museum für Naturkunde (Mario Lasseck), Stefanos Vrochidis (Greece, ITI), UK center for hydrology and ecology (Tom August)

- Africa: Univ. of Tunis (Sadok Ben-Yahia), IMSP, Bénin (Jules Deliga)

- Australia: Australian National University (Peter Christen)

- Central America: Technologico de Costa-Rica (Erick Mata, former director of the US initiative Encyclopedia of Life)

10.1.3 Participation in other international programs

Inria Brasil, 2020-2022 The Inria Brasil web site is now open.

Inria and LNCC, the Brazilian National Scientific Computing Laboratory, signed a Memory of Understanding to collaborate, with associated Brazilian universities, in HPC, AI, Data Science and Scientific Computing. This objective is to create an Inria International Lab., Inria Brasil. The collaboration is headed by Frédéric Valentin (LNCC, Inria International Chair) and Patrick Valduriez

10.2 International research visitors

10.2.1 Visits of international scientists

- Heraldo Borges (CEFET-RJ, Brazil), working on “Discovering Patterns in Restricted Space-Time Datasets” visited us until May.

10.3 European initiatives

10.3.1 FP7 & H2020 Projects

COS4CLOUD

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

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

Duration: 2019 - 2022

Coordinator: CSIC (Spain)

Partners: The Open University, CREAF, Bineo, EarthWatch, SLU, NKUA, CERT, Bineo, ECSA.

Inria contact: Alexis Joly

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

10.4 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.

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.

ANR WeedElec (2018-2021), 106 Keuro.

Participants Julien Champ, Hervé Goëau, Alexis Joly.

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

ANR KAMOuLOX (2016-2020), 290 Keuro.

Participants Antoine Liutkus, Fabian-Robert Stöter.

The KAMOuLOX project aimed at providing online unmixing tools for ethnologists, that are not specialists in audio engineering. It was the opportunity for cutting-edge signal processing research, a strong dissemination activity in terms of (open-source) software release, and important contributions in deep learning research for audio.

CASDAR CARPESO (2020-2022), 87 Keuro.

Participants Julien Champ, Hervé Goëau, Alexis Joly.

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

10.4.1 Others

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

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

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

Ministry of Culture (2019-2021), 130 Keuro

Participants Alexis Joly, Jean-Christophe Lombardo.

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

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

Participants Baldwin Dumortier, Antoine Liutkus.

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

DINUM, 80 Keuro

Participants Reza Akbarinia, Florent Masseglia.

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

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

Participants Alexis Joly, Joaquim Estopinan.

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

11 Dissemination

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

General chair, scientific chair

- E. Pacitti: Bases de Données Avancées (BDA), 2021, PC chair.
- C. Pradal: Crop Modeling for the Future, ICROPM Symposium, 3-5 February 2020, https://www.icropm2020.org

Member of the organizing committees

- F. Masseglia : Extraction et Gestion des Connaissances (EGC), 2021, https://egc2021.sciencesconf.org/
- F. Masseglia : the 1st VIVA European Summer School on Artificial Intelligence and Software Verification and Validation https://www.lirmm.fr/viva2020/
- R. Akbarinia: Extraction et Gestion des Connaissances (EGC), 2021, https://egc2021.sciencesconf.org/
- A. Joly: organizing committee of the international conference CLEF 2020 and the chair of the LifeCLEF track (http://clef2020.clef-initiative.eu/)

11.1.2 Scientific events: selection

Member of the conference program committees

- IEEE Artificial Intelligence & Knowledge Engineering (AIKE), 2020: F. Masseglia
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (PKDD), 2020: F. Masseglia
- Int. Conf. on Information Management and Big Data (SIMBig), 2020: F. Masseglia
- IEEE Int. Conf. on Data Mining (ICDM), 2020: F. Masseglia
- ACM Symposium on Applied Computing (ACM SAC), Data Mining Track (DM), 2020: F. Masseglia
- ACM Symposium on Applied Computing (ACM SAC), Data Stream Track (DS), 2020: F. Masseglia
- Extraction et Gestion des Connaissances (EGC), 2020: F. Masseglia
- Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2020 : F. Masseglia
- International Conference on Very Large Data Bases (VLDB), 2020: R. Akbarinia
- Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA), 2020: R. Akbarinia
- Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2020: A. Joly, A. Liutkus
- Neural Information Processing Systems (NeurIPS): A. Liutkus, A. Joly
- Int. Conf. on Machine Learning (ICML): A. Liutkus

- Int. Conf. on Learning Representations (ICLR): A. Liutkus
- Int. Conf. on Computer Vision (CVPR), 2020: A. Joly
- Int. Conf. and Labs of the Evaluation Forum (CLEF), 2020: A. Joly
- European. Conf. on Information Retrieval (ECIR), 2020: A. Joly
- European. Conf. on Computer Vision (ECCV), 2020: A. Joly
- IEEE/ACM Int. Symposium in Cluster, Cloud, and Grid Computing (CCGrid) 2019: Esther Pacitti
- Int. Conf. on functional-structural plant models (FSPM), 2020: C. Pradal

11.1.3 Journal

Member of the editorial boards

- VLDB Journal: P. Valduriez.
- Transactions on Large Scale Data and Knowledge Centered Systems: R. Akbarinia.
- Distributed and Parallel Databases: E. Pacitti, P. Valduriez.
- Plant Methods: C. Pradal.

Reviewer - reviewing activities

- Annals of Telecommunications (ANTE) : F. Masseglia
- Distributed and Parallel Databases (DAPD): E. Pacitti, P. Valduriez
- IEEE Transactions on Knowledge and Data Engineering (TKDE): R. Akbarinia, F. Masseglia
- Information Systems: R. Akbarinia
- IEEE access: R. Akbarinia
- Knowledge and Information Systems (KAIS): R. Akbarinia, F. Masseglia
- Ecosphere: A. Joly
- Methods in Ecology and Evolution: A. Joly
- Plant methods: A. Joly
- Science of the Total Environment: A. Joly
- Machine Learning: A. Joly
- Pattern Recognition Letters: A. Joly
- Transactions on Image Processing: A. Joly
- Multimedia Tools and Applications: A. Joly
- Environmental Research Letters: A. Joly
- Information Processing and Management: A. Joly
- ACM Trans. on Database Systems: E. Pacitti
- IEEE Transaction on Signal Processing (TSP): A. Liutkus
- IEEE Transactions on Audio Speech and Language Processing (TASLP): A. Liutkus
- IEEE Signal Processing Magazine: A. Liutkus
- IEEE Signal Processing Letters: A. Liutkus
- Frontiers in Plant Science: C. Pradal

11.1.4 Invited talks

- A. Joly: "L'IA au service de la biodiversité végétale", 24 Nov, Accadémie des sciences; "Deep learning and Pl@ntNet", 16 Nov, Imaginecology conference;
- A. Liutkus: tutorial on "music source separation" at Int. Symposium on Music Information Retrieval (ISMIR 2018).
- P. Valduriez: Lecture: Distributed Database Systems: the case for NewSQL, 19 November, CWI Lectures, Amsterdam (Virtual).
- P. Valduriez: Tutorial: Principles of Distributed Database Systems: spotlight on NewSQL, 29 september, Brazilian Symposium on Databases (SBBD).
- C. Pradal: "Multiscale plant modelling and Phenotyping" on 8 october at Tottori University and on 16 october at Nagoya University, Japan; workshop on plant modelling on 28 october at Tokyo University, Japan.

11.1.5 Leadership within the scientific community

- A. Joly: Scientific manager of the LifeCLEF research forum.
- A. Liutkus: Member of the IEEE Technical Committee on Audio and Acoustic Signal Processing.
- E. Pacitti: Member of the Steering Committee of the BDA conference.
- P. Valduriez: President of the Steering Committee of the BDA conference (until October).

11.1.6 Scientific expertise

- A. Joly: scientific advisory board of the ANR program "AI for biodiversity"
- A. Joly: expert for the National HPC grand equipment (GENCI) programs
- A. Joly: scientific advisory board of LepiNoc project (automated beetle tracking)
- E. Pacitti: reviewer for STIC AmSud international program.
- P. Valduriez: reviewer for STIC AmSud international program.
- P. Valduriez: reviewer for NSERC (Canada).
- C. Pradal: member of CSS EGBIP (Commissions Scientifiques Spécialisées) INRAE.
- A. Liutkus: reviewer for FONDECYT (Chile) competition 2020.

11.1.7 Research administration

- A. Joly: Technical director of the InriaSOFT consortium Pl@ntNet and representative of Inria in the steering committee.
- F. Masseglia: until september 2020, “Chargé de mission pour la médiation scientifique Inria” and head of Inria’s national network of colleagues involved in science popularization.
- F. Masseglia: since october 2020, “Deputy Scientific Director of Inria, in charge of the domain Perception, Cognition and Interaction”.
- 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 (DPEI).
- Antoine Liutkus is an elected member of the IEEE technical committee on Audio.

11.2 Teaching - Supervision - Juries

11.2.1 Teaching

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

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

- IG3: Database design, physical organization, 54h, level L3, 50 student
- 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 with defense: 1 group of 3 students, level M1 (3 mounths) and 1 group of 3 students level M2 (3 mounths)
- Supervision of master internships with defense: 1 group of 3 students, level M1 (3 mounths) and 3 students, level M2 (6 mounths each)
- Supervision of computer science discovery projects with defense: one group of 3 students level L2 (4 mounths)

Patrick Valduriez:

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

Alexis Joly:

- Univ. Montpellier: Machine Learning, 10h, level M2
- Polytech' Montpellier: Content-Based Image Retrieval, 4.5h, level M2
- AgroParisTech: Deep Learning, 18h, level M1
- Innobs technical school: Innovations in the observation of seasonal biological events and associated data management, 6 hours, for professionals.

Antoine Liutkus

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

Christophe Pradal

- Univ. Montpellier: Root System Modelling, 15h, level M2
- Univ. Montpellier: Functional-Structural Plant Modelling, 9h, level M2

11.2.2 Supervision

PhD & HDR:

- PhD: Gaetan Heidsieck, Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping, 9 December, Univ. Montpellier. Advisors: Esther Pacitti, Christophe Pradal, François Tardieu (INRAE).
- PhD: Titouan Lorieul, Pro-active Crowdsourcing, 2 December, Univ. Montpellier. Advisor: Alexis Joly.
- PhD : Khadidja Meguelati, Massively Distributed Clustering, 13 March, Univ. Montpellier. Advisors: Nadine Hilgert (INRAE), Florent Masseglia.

- PhD in progress: Heraldo Borges, Discovering Tight Space-Time Sequences, started Oct 2018, CEFET/Rio, Brazil. Advisors: Esther Pacitti, Eduardo Ogaswara (CEFET/Rio, Brazil).
- PhD in progress: Lamia Djebour, Parallel Time Series Indexing and Retrieval with GPU architectures, started Oct 2019, Univ. Montpellier. Advisors: Reza Akbarinia, Florent Masseglia.
- PhD in progress: Quentin Leroy, Active learning of unknown classes, started Oct 2019, Univ. Montpellier. Advisors: Alexis Joly
- PhD in progress: Alena Shilova, Scheduling Strategies for High Performance Deep Learning, started Oct 2019, Univ. Bordeaux. Advisors: Olivier Beaumont, Alexis Joly
- PhD in progress: Daniel Rosendo, Enabling HPC-Big Data Convergence for Intelligent Extreme-Scale Analytics, started Oct 2019, Univ. Rennes. Advisors: Gabriel Antoniu, Alexandru Costan, Patrick Valduriez
- PhD in progress: Joaquim Estopinan, Prediction of the conservation status of species, started Nov 2020, Univ. Montpellier. Advisors: Alexis Joly
- PhD in progress: Camille Garcin, Set-valued classification in the case of long-tail distributions, started Oct 2020, Univ. Montpellier. Advisors: Joseph Salmon, Alexis Joly
- PhD in progress: Rodrigo Alves Prado da Silva Data-centric Workflow Scheduling with Privacy Restrictions started Oct 2020, UFF, Brazil. Advisors: Daniel de Oliveira (UFF,Brazil), Esther Pacitti

11.2.3 Juries

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

- R. Akbarinia: Amine El Ouassouli (INSA Lyon, reviewer)
- A. Joly: Titouan Lorieul (Univ. Montpellier, advisor), Waleed RAGHEB (Univ. of Montpellier, jury president)
- F. Masseglia: Mehdi Kaytoue (HDR, Univ. Lyon 1, reviewer)
- E. Pacitti: Gaetan Heidscieck (Univ. Montpellier, advisor), Arnaud Grall (Univ. Nantes), Riad Mokadem (HDR. UPS, Toulouse, reviewer)
- P. Valduriez: Riad Mokadem (HDR. UPS, Toulouse)
- C. Pradal: Gaetan Heidscieck (Univ. Montpellier, advisor), Cyrille Midingoyi (Institut Agro. Montpellier, advisor), Kévin Dubois, Univ. Toulouse, reviewer)
- A. Liutkus: Gabriel Meseguer Brocal (Sorbonne Univ. Paris, reviewer).

Members of the team participated to the following hiring committees:

- F. Masseglia: Selection committee 4567 Polytech LIRIS (june 2020)
- F. Masseglia: Inria CRCN + ISFP n°4 - Nancy - Grand-Est (june 2020)
- R. Akbarinia: Associate professor position, INSA Lyon (May 2020)
- R. Akbarinia: ATER position, Univ. Montpellier (May 2020)
- A. Joly: Selection committee 3892, CIRAD Montpellier (Nov 2020)

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

- F. Masseglia was “Chargé de mission auprès de la DGD-S Inria pour la médiation scientifique” (50% of his time) until September 2020, and headed Inria’s national network of colleagues involved in science popularization.

- A. Joly spends several hours a week animating Pl@ntnet user community. This includes: (i) animating the community of developers using Pl@ntNet API (about 500 users), (ii) animating Pl@ntnet’s social networks (Twitter and Facebook accounts) and (iii) managing the mailbox contact@plantnet-project.org.

11.3.2 Articles and contents

- F. Masseglia is co-author of a guide for teacher and summer camp counselors on tracking apps in the context of covid-19, edited by the Académie des sciences: https://www.academie-sciences.fr/fr/Promouvoir-l-enseignement-des-sciences/cet-ete-avec-la-science.html
- A. Joly has given several interviews to different media giving rise to web articles about Pl@ntNet (see e.g. Google news with keyword Pl@ntNet).
- A. Joly has co-authored several popularization articles, e.g. for the Ministry of Culture magazine, the GENCI annual report, Inria national website, etc.
- A. Joly has co-produced a popularization video about Pl@ntNet’s French Academy of Science award.
- A. Joly actively participates to the design and development of all Pl@ntNet dissemination tools in particular Pl@ntNet web site that contains contents for the press, articles for the general public, tutorials of Pl@ntNet tools, guidelines for users of the API, etc.

11.3.3 Education

- F. Masseglia is member of the steering committee and the initiator, with Serge Abiteboul, of the program called "1 scientifique — 1 classe : Chiche !" with the goal of reaching *all* the students of a specific level. This massive plan should concern all scientists at Inria and our partners in France. It has been slowed down by the pandemic but should get back on rail by mid-2021.

- A. Joly gave a webinar for nearly 30 greek teachers about the use of Pl@ntNet in the context of a formal educative program organized in collaboration with the greek national education.

11.3.4 Interventions

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

11.3.5 Creation of media or tools for science outreach

- F. Masseglia participated to the covid-19 mission project "Parlons Maths" with the goal of making easier the organisation of online talks with high public interaction.
- F. Masseglia co-organised the first Inria online public event. It was for the science celebration national event, October 2-12, with 9 days of live talks streamed on the Inria Youtube channel. Everything had to be invented for this exceptional event where Inria had zero experience: https://www.inria.fr/fr/inria-fete-la-science.
- The softwares developed in the context of the Pl@ntNet project (Pl@ntNet mobile app, Pl@ntNet web, ThePlantGame) are used in a large number of formal educational programs as well as informal educational actions of individual teachers, associations, natural area managers, etc.

12 Scientific production

12.1 Major publications

[1] C. Botella, A. Joly, P. Monestiez, P. Bonnet and F. Munoz. 'Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection'. In: *PLoS ONE* 15.5 (May 2020), e0232078. DOI: 10.1371/journal.pone.0232078. URL: https://hal.archives-ouvertes.fr/hal-02639237.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

12.2 Publications of the year

International journals

[11] T. August, O. Pescott, A. Joly and P. Bonnet. 'AI Naturalists Might Hold the Key to Unlocking Biodiversity Data in Social Media Imagery'. In: *Patterns* 1.7 (Oct. 2020), p. 100116. DOI: 10.1016/j.patter.2020.100116. URL: https://hal.inria.fr/hal-02989043.

[12] A. Belhadi, Y. Djenouri, K. Nørvåg, H. Ramampiaro, F. Masseglia and J. C.-W. Lin. 'Space Time Series Clustering: Algorithms, Taxonomy, and Case Study on Urban Smart Cities'. In: *Engineering Applications of Artificial Intelligence* 95 (Oct. 2020), #103857. DOI: 10.1016/j.engappai.2020.103857. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03036868.

[13] F. Belhassine, D. Fumey, J. Chopard, C. Pradal, S. Martinez, E. Costes and B. Pallas. 'Modelling transport of inhibiting and activating signals and their combined effects on floral induction: application to apple tree'. In: *Scientific Reports* 10 (2020), p. 13085. DOI: 10.1038/s41598-020-69861-8. URL: https://hal.archives-ouvertes.fr/hal-02919190.

[14] P. Bonnet, J. Champ, H. Goëau, F.-R. Stöter, B. Deneu, M. Servajean, A. Affouard, J.-C. Lombardo, O. Levchenko, H. Gresse and A. Joly. 'Biodiversity Information Science and Standards 4: e58933 Pl@ntNet Services, a Contribution to the Monitoring and Sharing of Information on the World Flora'. In: *Biodiversity Information Science and Standards* 4 (2020). DOI: 10.3897/biss.4.58933. URL: https://hal.inrae.fr/hal-02973673.

[15] P. Bonnet, A. Joly, J.-M. Faton, S. Brown, D. Kimiti, B. Deneu, M. Servajean, A. Affouard, J.-C. Lombardo, L. Mary, C. Vignau and F. Munoz. 'How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools'. In: *Ecological Solutions and Evidence* 1.2 (2020). DOI: 10.1002/2688-8319.12023. URL: https://hal.inrae.fr/hal-02937618.

[16] H. Borges, M. Dutra, A. Bazaz, R. Coutinho, F. Perosi, F. Porto, F. Masseglia, E. Pacitti and E. Ogasawara. 'Spatial-Time Motifs Discovery'. In: *Intelligent Data Analysis* (1st Oct. 2020). URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02984969.

[17] C. Botella, A. Joly, P. Monestiez, P. Bonnet and F. Munoz. 'Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection'. In: *PLoS ONE* 15.5 (20th May 2020), e0232078. DOI: 10.1371/journal.pone.0232078. URL: https://hal.archives-ouvertes.fr/hal-02639237.

[18] R. K. Braghiere, F. Gerard, J. Evers, C. Pradal and L. Pagès. 'Simulating the effects of water limitation on plant biomass using a 3D functional–structural plant model of shoot and root driven by soil hydraulics'. In: *Annals of Botany* 126.4 (6th Apr. 2020), pp. 713–728. DOI: 10.1093/aob/mcaa059. URL: https://hal.inrae.fr/hal-02912597.

[19] G. A. Brat, G. M. Weber, N. Gehlenborg, P. Avillach, N. P. Palmer, L. Chiovato, J. Cimino, B. K. Beaulieu-Jones, S. L'Yi, M. S. Keller et al. 'International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium'. In: *npj Digital Medicine* 3.1 (Dec. 2020), #109. DOI: 10.1038/s41746-020-00308-0. URL: https://hal.archives-ouvertes.fr/hal-02918344.

[20] J. Champ, A. Mora-Fallas, H. Goëau, E. Mata-Montero, P. Bonnet and A. Joly. 'Instance segmentation for the fine detection of crop and weed plants by precision agricultural robots'. In: *Applications in Plant Sciences* 8.7 (2020). DOI: 10.1002/aps3.11373. URL: https://hal.inrae.fr/hal-02910844.

[21] C. C. Davis, J. Champ, D. Park, I. Breckheimer, G. Lyra, J. Xie, A. Joly, D. Tarapore, A. M. Ellison and P. Bonnet. 'A New Method for Counting Reproductive Structures in Digitized Herbarium Specimens Using Mask R-CNN'. In: *Frontiers in Plant Science* 11 (31st July 2020). DOI: 10.3389/fpls.2020.01129. URL: https://hal.inrae.fr/hal-02909794.

[22] 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.

[23] M. Gauthier, R. Barillot, A. Schneider, C. Chambon, C. Fournier, C. Pradal, C. Robert and B. Andrieu. 'A functional structural model of grass development based on metabolic regulation and coordination rules'. In: *Journal of Experimental Botany* 71.18 (Sept. 2020), pp. 5454–5468. DOI: 10.1093/jxb/eraa276. URL: https://hal.inrae.fr/hal-02903070.

[24] H. Goëau, A. Mora-Fallas, J. Champ, N. L. Rossington Love, S. Mazer, E. Mata-Montero, A. Joly and P. Bonnet. 'A new fine-grained method for automated visual analysis of herbarium specimens: A case study for phenological data extraction'. In: *Applications in Plant Sciences* 8.6 (June 2020), #e11368. DOI: 10.1002/aps3.11368. URL: https://hal.inrae.fr/hal-02894994.

[25] G. Heidsieck, D. De Oliveira, E. Pacitti, C. Pradal, F. Tardieu and P. Valduriez. 'Efficient Execution of Scientific Workflows in the Cloud Through Adaptive Caching'. In: *Transactions on Large-Scale Data- and Knowledge-Centered Systems* (10th Sept. 2020), pp. 41–66. DOI: 10.1007/978-3-662-62271-1_2. URL: https://hal.archives-ouvertes.fr/hal-02969510.

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

[27] S. H. Lee, H. Goëau, P. Bonnet and A. Joly. 'Attention-Based Recurrent Neural Network for Plant Disease Classification'. In: *Frontiers in Plant Science* 11 (14th Dec. 2020). DOI: 10.3389/fpls.2020.601250. URL: https://hal.inria.fr/hal-03064464.

[28] S. H. Lee, H. Goëau, P. Bonnet and A. Joly. 'New perspectives on plant disease characterization based on deep learning'. In: *Computers and Electronics in Agriculture* 170 (Mar. 2020), p. 105220. DOI: 10.1016/j.compag.2020.105220. URL: https://hal.umontpellier.fr/hal-02470280.

[29] N. Lemus, F. Porto, Y. M. Souto, R. Pereira, J. Liu, E. Pacitti and P. Valduriez. 'SUQ2: Uncertainty Quantification Queries over Large Spatio-temporal Simulations'. In: *Bulletin of the Technical Committee on Data Engineering* 43.1 (Mar. 2020), pp. 47–59. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02531748.

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

[31] J. Liu, C. Bondiombouy, L. Mo and P. Valduriez. 'Two-Phase Scheduling for Efficient Vehicle Sharing'. In: *IEEE Transactions on Intelligent Transportation Systems* (2020), pp. 1–14. DOI: 10.1109/TITS.2020.3011952. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02913503.

[32] 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.

[33] H. L. S. Lustosa, A. C. da Silva, D. N. R. Da Silva, P. Valduriez and F. A. M. Porto. 'SAVIME: An Array DBMS for Simulation Analysis and ML Models Predictions'. In: *Journal of Information and Data Management* 11.3 (30th Dec. 2020), pp. 247–264. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03144324.

[34] 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 (15th Aug. 2020), p. 104076. DOI: 10.1016/j.chemolab.2020.104076. URL: https://hal.inrae.fr/hal-02899789.

[35] C. A. Midingoyi, C. Pradal, I. N. Athanasiadis, M. Donatelli, A. Enders, D. Fumagalli, F. Garcia, D. Holzworth, G. Hoogenboom, C. Porter, H. Raynal, P. Thorburn and P. M. Martre. 'Reuse of process-based models: automatic transformation into many programming languages and simulation platforms'. In: *in silico Plants* (2020), diaa007. DOI: 10.1093/insilicoplants/diaa007. URL: https://hal.inrae.fr/hal-02962262.

[36] K. D. Pearson, G. Nelson, M. Aronson, P. Bonnet, L. Brenskelle, C. C. Davis, E. Denny, E. R. Ellwood, H. Goëau, J. M. Heberling, A. Joly, T. Lorieul, S. Mazer, E. Meineke, B. Stucky, P. W. Sweeney, A. White and P. S. Soltis. 'Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research'. In: *Bioscience* 70.7 (2020), pp. 610–620. DOI: 10.1093/biosci/biaa044. URL: https://hal.umontpellier.fr/hal-02573627.

[37] V. Silva, V. Campos, T. Guedes, J. Camata, D. De Oliveira, A. L. Coutinho, P. Valduriez and M. Mattoso. 'DfAnalyzer: Runtime dataflow analysis tool for Computational Science and Engineering applications'. In: *SoftwareX* 12 (July 2020), p. 100592. DOI: 10.1016/j.softx.2020.100592. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02949807.

[38] R. Souza, V. Silva, A. L. G. d. A. Coutinho, P. Valduriez and M. Mattoso. 'Data reduction in scientific workflows using provenance monitoring and user steering'. In: *Future Generation Computer Systems* 110 (Sept. 2020), pp. 481–501. DOI: 10.1016/j.future.2017.11.028. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01679967.

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

[40] 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.

International peer-reviewed conferences

[41] F. Belhassine, D. Fumey, C. Pradal, J. Chopard, E. Costes and B. Pallas. 'A modelling framework for the simulation of signal transport within 3D structure: application for the simulation of within-tree variability in floral induction in apple trees'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Vol. 9. Hannover / Virtual, Germany, 4th Aug. 2020, pp. 16–17. DOI: 10.1038/s41598-020-69861-8. URL: https://hal.inria.fr/hal-03059482.

[42] F. Boudon, J. Vaillant and C. Pradal. 'Toward Virtual Modelling Environments using Notebooks for Phenotyping and Simulation of Plant Development'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Hanovre / Virtua, Germany, 2020, pp. 99–100. URL: https://hal.inria.fr/hal-03059535.

[43] B. Deneu, T. Lorieul, E. Cole, M. Servajean, C. Botella, P. Bonnet and A. Joly. 'Overview of LifeCLEF location-based species prediction task 2020 (GeoLifeCLEF)'. In: CLEF 2020 - 11th International Conference of the Cross-Language Evaluation Forum for European Languages. thessaloniki, Greece, 22nd Sept. 2020. URL: https://hal.inria.fr/hal-02989077.

[44] C. Fournier, S. Artzet, F. Tardieu and C. Pradal. 'What structural plant modelling and image-based phenotyping can learn from each other?' In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Hanovre / Virtua, Germany, 2020, pp. 55–56. URL: https://hal.inria.fr/hal-03059496.

[45] M. Gauthier, R. Barillot, A. Schneider, C. Chambon, C. Fournier, C. Pradal and B. Andrieu. 'A 3D architectural model of grass shoot morphogenesis and plasticity, driven by organ metabolite concentrations and coordination rules'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Hanovre / Virtua, Germany, 2020. URL: https://hal.inria.fr/hal-03059493.

[46] G. Heidsieck, D. De Oliveira, E. Pacitti, C. Pradal, F. Tardieu and P. Valduriez. 'Distributed Caching of Scientific Workflows in Multisite Cloud'. In: *DEXA 2020: Database and Expert Systems Applications; DEXA: Database and Expert Systems Applications, Sep 2020, Bratislava, Slovakia. pp.51-65.* DEXA 2020 - 31st International Conference on Database and Expert Systems Applications. Vol. 12392. Lecture Notes in Computer Science. Bratislava, Slovakia: Springer Science and Business Media Deutschland GmbH, 13th Sept. 2020, pp. 51–65. DOI: 10.1007/978-3-030-59051-2_4. URL: https://hal.inrae.fr/hal-02962579.

[47] A. Joly, H. Goëau, C. Botella, R. Ruiz De Castaneda, H. Glotin, E. Cole, J. Champ, B. Deneu, M. Servajean, T. Lorieul, W.-P. Vellinga, F.-R. Stöter, A. Durso, P. Bonnet and H. Müller. 'LifeCLEF 2020 Teaser: Biodiversity Identification and Prediction Challenges'. In: ECIR 2020 - 42nd European Conference on IR Research on Advances in Information Retrieval. Vol. Lecture Notes in Computer Science. Advances in Information Retrieval. Proceedings, Part II 12036. Lisbon, Portugal, 8th Apr. 2020, pp. 542–549. DOI: 10.1007/978-3-030-45442-5_70. URL: https://hal.inrae.fr/hal-02873670.

[48] A. Joly, H. Goëau, S. Kahl, B. Deneu, W.-P. Vellinga, M. Servajean, E. Cole, L. Picek, R. Ruiz de Castañeda, I. Bolon, A. Durso, T. Lorieul, C. Botella, H. Glotin, J. Champ, I. Eggel, P. Bonnet and H. Müller. 'Overview of LifeCLEF 2020: A System-Oriented Evaluation of Automated Species Identification and Species Distribution Prediction'. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction
11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings.* CLEF 2020 - 11th International Conference of the Cross-Language Evaluation Forum for European Languages. Vol. 12260. Lecture Notes in Computer

Science. Thessaloniki, Greece: Springer, 15th Sept. 2020, pp. 342–363. DOI: 10.1007/978-3-030-58219-7_23. URL: https://hal.inrae.fr/hal-02945382.

[49] S. Kahl, M. Clapp, W. A. Hopping, H. Goëau, H. Glotin, R. Planqué, W.-P. Vellinga and A. Joly. 'Overview of BirdCLEF 2020: Bird Sound Recognition in Complex Acoustic Environments'. In: CLEF 2020 - 11th International Conference of the Cross-Language Evaluation Forum for European Languages. Thessaloniki, Greece, 22nd Sept. 2020. URL: https://hal.inria.fr/hal-02989101.

[50] S. Kirié, C. Pradal, H. Iwasaki, K. Noshita and H. Iwata. 'Three-dimensional morphological model of water lilies Nymphaea spp. for breeding historical study'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Hanovre / Virtua, Germany, 2020, pp. 61–62. URL: https://hal.inria.fr/hal-03059518.

[51] Q. Long, C. Pradal and W. Kurth. 'Co-simulation with OpenAlea and GroIMP for cross-platform functional-structural plant modelling'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Hanovre / Virtual, Germany, 2020, pp. 97–98. URL: https://hal.inria.fr/hal-03059527.

[52] K. Meguelati, B. Fontez, N. Hilgert, F. Masseglia and I. Sanchez. 'Massively Distributed Clustering via Dirichlet Process Mixture'. In: ECML PKDD 2020 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Ghent (virtual), Belgium, 14th Sept. 2020. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-03036910.

[53] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge and E. Vincent. 'Asteroid: the PyTorch-based audio source separation toolkit for researchers'. In: Interspeech 2020. Shanghai, China, 25th Oct. 2020. URL: https://hal.inria.fr/hal-02962964.

[54] D. Pina, L. Kunstmann, D. De Oliveira, P. Valduriez and M. Mattoso. 'An approach for the collection and analysis of configuration data in deep neural networks'. In: SBBD 2020 - 35ª Simpósio Brasileiro de Banco de Dados. Virtual, Brazil, Oct. 2020, pp. 1–6. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02969506.

[55] C. Pradal and C. Godin. 'MTG as a standard representation of plants in FSPMs'. In: FSPM 2020 -9th International Conference on Functional-Structural Plant Models. Hanovre / Virtua, Germany, 2020, pp. 86–87. URL: https://hal.inria.fr/hal-03059523.

[56] F. REES, R. Barillot, M. Gauthier, L. Pagès, C. Pradal and B. Andrieu. 'Simulating rhizodeposition as a function of shoot and root interactions within a new 3D Functional-Structural Plant Model'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Hanovre / Virtua, Germany, 2020, pp. 22–23. URL: https://hal.inrae.fr/hal-02964060.

[57] C. Saint Cast, G. Lobet, L. Cabrera-Bosquet, V. Couvreur, C. Pradal, B. Muller, F. Tardieu and X. Draye. 'Improving interoperability between phenomics and modelling communities by designing a Plant Modelling Ontology (PMO)'. In: FSPM 2020 - 9th International Conference on Functional-Structural Plant Models. Towards Computable Plants. Hanovre / Virtua, Germany, 2020, pp. 57–58. URL: https://hal.inria.fr/hal-03059507.

Conferences without proceedings

[58] Y. Boursiac, C. Pradal, F. Bauget, S. Delivorias, M. Lucas, C. Godin and C. Maurel. 'Phenotyping and modeling of water transport in roots'. In: iCROPM 2020 - Satellite workshop : Phenotyping and modeling of plant anchorage and physiology. Montpellier, France, 3rd Feb. 2020. URL: https://hal.inrae.fr/hal-02935069.

[59] H. Goëau, P. Bonnet and A. Joly. 'Overview of LifeCLEF Plant Identification task 2020'. In: CLEF 2020 - Conference and labs of the Evaluation Forum. CLEF 2020 - Conference and labs of the Evaluation Forum. Thessalonique, Greece, 22nd Sept. 2020. URL: https://hal.inrae.fr/hal-02980085.

Scientific books

[60] T. Özsu and P. Valduriez. *Principles of Distributed Database Systems - Fourth Edition*. Springer, 2020, pp. 1–674. URL: https://hal-lirmm.ccsd.cnrs.fr/lirmm-02265930.

Doctoral dissertations and habilitation theses

[61] G. Heidsieck. 'Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping'. Université Montpellier, 9th Dec. 2020. URL: https://hal.archives-ouvertes.fr/tel-03089552.

[62] T. Lorieul. 'Uncertainty in predictions of deep learning models for fine-grained classification'. Université de Montpellier (UM), FRA., 2nd Dec. 2020. URL: https://tel.archives-ouvertes.fr/tel-03040683.

Reports & preprints

[63] R. Akbarinia and B. Cloez. *Efficient Matrix Profile Computation Using Different Distance Functions*. 17th Jan. 2019. URL: https://hal.inrae.fr/hal-02788459.

[64] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet and A. Joly. *The GeoLifeCLEF 2020 Dataset*. 5th Nov. 2020. URL: https://hal.inria.fr/hal-02989062.

[65] B. Deneu, M. Servajean, P. Bonnet, F. Munoz and A. Joly. *Participation of LIRMM / Inria to the GeoLifeCLEF 2020 challenge*. 5th Nov. 2020. URL: https://hal.inria.fr/hal-02989084.

12.3 Other

Scientific popularization

[66] P. Bonnet, H. Goëau, F. Hopkins, E. Véla, A. Sahl, A. Affouard, J. Champ, H. Gresse and A. Joly. 'IUCN redlisting of some Irano-Anatolian plant species View project Flora Gallica View project'. In: *Carnets Botaniques* (2020), pp. 1–9. DOI: 10.34971/zaz0-n247. URL: https://hal.inrae.fr/hal-02981760.

[67] M. M. Zekeng Ndadji, M. Tchoupé Tchendji, C. Tayou Djamegni and D. Parigot. 'A Grammatical Model for the Specification of Administrative Workflow Using Scenario as Modelling Unit'. In: *Applied Informatics; Applied Informatics
Third International Conference, ICAI 2020, Ota, Nigeria, October 29–31, 2020, Proceedings*. ICAI 2020 - 3rd International Conference on Applied Informatics. Vol. 1277. Communications in Computer and Information Science. Ota, Nigeria: https://icai.itiud.org/, 19th Oct. 2020, pp. 131–145. DOI: 10.1007/978-3-030-61702-8_10. URL: https://hal.inria.fr/hal-02970761.

[68] M. M. Zekeng Ndadji, M. Tchoupé Tchendji, C. Tayou Djamegni and D. Parigot. 'A Language and Methodology based on Scenarios, Grammars and Views, for Administrative Business Processes Modelling'. In: *Paradigm Plus* 1.3 (25th Oct. 2020), pp. 1–22. URL: https://hal.inria.fr/hal-02977704.

[69] M. M. Zekeng Ndadji, M. Tchoupé Tchendji, C. Tayou Djamegni and D. Parigot. 'A Language for the Specification of Administrative Workflow Processes with Emphasis on Actors' Views'. In: *Computational Science and Its Applications – ICCSA 2020 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part VI; Computational Science and Its Applications – ICCSA 2020
20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part I*. ICCSA 2020 - 20th International Conference on Computational Science and Its Applications. Vol. 12254. Lecture Notes in Computer Science. Cagliari, Italy: https://iccsa.org/, 30th Sept. 2020, pp. 231–245. DOI: 10.1007/978-3-030-58817-5_18. URL: https://hal.archives-ouvertes.fr/hal-02968427.