INRIA International Program
Associate Team Proposal
2019 – 2021

# REDAS– REPRODUCIBLE DATA SCIENCE
# Analysis Techniques and Workflow Methodologies for Reproducible Data Science

– An associated team proposal for INRIA/Polaris –

| | |
|---|---|
| Title | Analysis Techniques and Workflow Methodologies for Reproducible Data Science |
| Acronym | REDAS– REPRODUCIBLE DATA SCIENCE |
| Coordinator | Guillaume HUARD (INRIA-Polaris)<br>Jean-Marc VINCENT (INRIA-Polaris)<br>Arnaud LEGRAND (INRIA-Polaris) |
| Coordinator (Brazil) | Lucas MELLO SCHNORR (UFRGS)<br>João Luiz DIHL COMBA (UFRGS) |

September/2018

**Key Words**

- A – Research themes on digital science: (at most 5 keywords)
  - A1.1.4. HPC
  - A3.4 Apprentissage et statistiques
  - A5.2. Visualisation de données
  - A6.2.4. Méthodes statistiques
  - A8.6 Théorie de l'information
- B – Other research themes and application areas: (at most 5 keywords)
  - B8.3. Urbanisme et planification
  - B9.5.7. Géographie
  - B9.6. Recherche reproductible
  - B9.7.2. Données ouvertes

# Contents

## Presentation

The remaining of this proposal is organized as follows. Section 1 presents the list of participants and the history of the collaboration. Section 2 presents the scientific context of this project, its main and specific objectives and the work-program for the first year. Section 3 presents the budget for the project. Section 4 lists the added value of this project. The document ends with a list of references.

# 1 Partnership

### 1.1 Detailed list of participants

The Polaris project team is the only INRIA project team involved. From the brazilian side, the members are from the Informatics Institute of UFRGS. The Table 1 below shows the list of permanent participants. The non-permanent researchers are listed in Table 2. Please note that the list might evolve since new students might get involved in the next years. The area of expertise indicates on which domain the investigation of student mostly lies.

Table 1: Permanent members, their affiliation and status.

| **Name** | **Team** | **Status** | **Main expertise** |
|---|---|---|---|
| Guillaume Huard | INRIA-Polaris | Associate Professor<br>REDAS Coordinator | Performance evaluation<br>Software Engineering |
| Jean-Marc Vincent | INRIA-Polaris | Associate Professor | Statistics<br>Information Theory |
| Arnaud Legrand | INRIA-Polaris | Researcher<br>Polaris Leader | Reproducible Research<br>Performance Modeling |
| Lucas Mello Schnorr | UFRGS | Adjunct Professor<br>REDAS Coordinator | Performance Evaluation<br>Trace Analysis & Workflow |
| João Comba | UFRGS | Associate Professor | Information Visualization<br>Data Science |

Table 2: Non-permanent members, their affiliation, status and area of expertise.

| **Name** | **Origin** | **Status** | **Area of Expertise** |
|---|---|---|---|
| Tom Cornebize | Polaris | PhD Candidate | Data Analytics for HPC |
| Nils Defauw | Polaris | Master | Information Theory and Aggregation |
| Flora Gautheron | Polaris | Master | Data Analytics for Social Sciences |
| Cicero Pahins | UFRGS | Post-doc | Efficient Data Structures<br>for Big Data Visual Analysis |
| Vinicius G. Pinto | UFRGS | PhD Candidate | Task-based Performance Analysis |
| Lucas Nesi | UFRGS | Master | Task-based Performance Analysis |
| Guilherme Alles | UFRGS | Master | Phenology analysis |

### 1.2 Nature and history of the collaboration

The collaboration between UFRGS and Grenoble is active since 1992 in domains such as High-Performance Computing. More recently, we have seen the establishment of new areas of investigation such as trace visualization and data science. This project intends to support this

recent involvement and effectively represents a renewing of the cooperation between UFRGS and Grenoble. We gather complementary skills from the participants to make this project possible. The Polaris members have knowledge in information theory, data aggregation, statistical learning, and parallel systems. The UFRGS members have experience with information visualization, data manipulation, data science, and parallel programming. All members have a strong interest in reproducible research with an active involvement in open science and improvement of research methodologies. This complementarity has been recently translated in several co-advised Master and PhD students supported by CAPES/Cofecub projects. As of September 2018, there is one PhD candidate being co-advised between Grenoble-Alpes University and UFRGS.

LICIA[1] has been financed by CNRS between 2011 and 2018 as a *Laboratoire International Associé* (LIA) with the objective to support new scientific collaborations on new domains with other teams at UFRGS. While, at the beginning, the high performance computing area was at the heart of the LICIA activities, now, its action enabled the appearance of new collaborations in other domains, including the data science which is the main topic of REDAS. An Associate Team would be a excellent opportunity for this group of researchers to secure the collaboration in the data science topic.

# 2 Scientific Program

## 2.1 Context and Objective

We present a proposal for the Associated Team 2019–2021 call issued by INRIA, together with our partners from the Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil. The main scientific context of this project is to develop novel analysis techniques and workflow methodologies to support reproducible data science. We aim at leveraging the proposed methods to tackle analysis scenarios coming from computer science (performance analysis of parallel computer programs), biology (long-term phenology behavior analysis and correlation with climate change), and general public datasets from government transparency reports (public transportation, weather data, geographical correlation). We organize this general goal along three axes: (1) analysis techniques that will enable the construction of relevant information from large scale and noisy data, (2) workflow methodologies to combine the previous analysis techniques into a reproducible computation and (3) case studies originating from different domains to evaluate the relevance of our proposals. These axes become the work packages of the REDAS (REPRODUCIBLE DATA SCIENCE) project, detailed as follows:

1. **Analysis Techniques**: large volumes of data are hard to summarize using simple statistics that hides important behavior in the data. Therefore, raw information visualization plays a key role to explore such data, in particular when curating data and trying to develop intuition about the mathematical models underlying data. Yet, such visualizations require data aggregation, which may lead to significant information loss. It is thus essential to investigate adaptive data aggregation schemes that enable the reduction of the data while controlling the information loss.

   Once sound hypothesis about the data have been made, it is common to use statistical learning techniques (such as K-means, PCA, Ridge/Lasso, GAMs, CART, etc) [26] to find the model (and its parameters) that maximize the likelihood. These key characteristics of the phenomenon under study can be interpreted only if uncertainty associated to

[1] http://licia-lab.org

these parameters [25] can be computed (e.g., estimating their posterior distribution using Monte Carlo Markov Chain techniques [24]) and properly visualized. The main goal of this work package is thus to study how these different techniques (data aggregation, statistical learning, and information visualization) can be combined in a sound way and possibly to propose new ones.

2. **Workflow methodologies**: the analysis process often involves a mix of tools to produce the end result. The data has to be filtered before it can be passed to some standard statistical tool to, eventually, produce some projection of the transformed data that can be visualized and studied by the analyst. Furthermore, the process is interactive: when the analyst is unsatisfied with the end result, a part of the analysis has to be changed to produce a new visualization. These adaptations of the whole analysis typically start from intermediate data and only a part of the analysis has to be rerun. The issue comes with the increasing size of these analysis, the disparity of the analysis tools and the large space of analysis parameters. Indeed standard analysis tools are written in various languages (R, Python, C, a mix of them, ...) that make their combination difficult. As they all require several parameters, the analyst is tempted to keep many variants of the intermediate data in order to interactively adapt the analysis. Soon, the question of data provenance (how this dataset has been produced) become of utter importance in order to ensure reproducibility. As the computing cost of large scale analysis keeps on increasing, the problem only gets worse: large analysis should be distributed without losing track of all the steps in the process. The objective of this work package is to leverage workflow and labbook techniques to ensure that a whole analysis process is interactive, efficient and reproducible.

3. **Evaluation**: In the previous work packages we will propose both a theoretical and practical methodology whose relevance should be evaluated with real case studies. We will build our evaluation on well identified and quite different datasets originating from the following three areas, on which we already have some past experience:

   **Performance analysis of HPC applications** These applications and their underlying runtimes tend to be increasingly complex and dynamic. As a consequence, their execution traces become too large and impossible to analyze with classical tools. We have started investigating the use of data science tools to analyze such traces. Initial results are extremely encouraging but still suffer from many scalability issues. We believe the workflow organization would allow us to overcome this limitation, and to conduct deeper performance analysis.

   **Long-term phenology behavior analysis and correlation with climate change** The phenology is the study of plant grow through the use of digital cameras attached to towers installed in the middle of the natural environments. These cameras take photos every a certain number of minutes and enable the researcher to verify how certain species grow, including their relation with the climate. We plan to use the PEN (Phenology Eyes Network) datasets, a Japanese effort to congregate data from several observation stations around the world. For this dataset, specific analysis blocks have to be designed to eventually shape the analysis into the form of a reproducible workflow.

   **General public datasets from governement transparency reports** All public Brazilian institutions are obliged by law to provide datasets about any publicly-financed data measurements. The city of Porto Alegre has long-term weather datasets that contain temperature, pressure and other indicators from different parts of the city. The

goal in this case study is very exploratory, for example to envision a way to represent such data in a geographical manner to verify if certain parts of the city may suffer from flash flood more than others. This may effectively serve to plan new rain drainage systems for the city. In this case the challenge comes from the dimensionality of the dataset. We claim that an interactive approach is especially suited for the guided exploration of such a dataset.

### Executive Summary

The goal of this project is thus to

**develop interactive, reproducible and scalable analysis workflows** (tools)
that **provide uncertainty and quality estimators about the analysis** (quality)

This will enable the analyst to **understand** the behaviors hidden in complex datasets collected in large scale dynamic systems, and to proceed with **confidence**. We intend to leverage various techniques from information theory, data analytics, workflows methodologies and high performance computing to achieve this goal. This will take the form of deep investigations on the first two axis of this collaboration proposal, since they tackle problems at different levels : the low-level (analysis blocks) and the high-level (workflow methodologies).

An anticipated challenge is the convergence of these two axis towards a single framework. Tackling this challenge involves finding a good case study scenario that might encompass the whole analysis process. The three case studies we propose enable us to conduct a wide spectrum of analysis, serving perfectly our goal.

## 2.2 Work-program (for the first year)

The LICIA workshop is taking place in September 2018 at Grenoble, France. The coordinators will take this opportunity to engage in the first part of the work program. In the following, we detail the per-axis work program, the problems that we aim at solving, the very high-level methodology, the participants involved and the planned scientific missions. Table 3 shows how the permanent members are mapped to the three axis of this collaboration proposal. At the end of the first year, we plan a visit of Guillaume Huard to Porto Alegre to complete the integration of the Analysis Techniques with the Workflow methodologies and transfer the result on the case studies and transfer the whole process on the French side, especially on parallel application cases developed in Grenoble.

Table 3: Permament members mapped to the three axis.

| **Axis** | **Participants** |
|---|---|
| Analysis Techniques | Guillaume Huard, Jean-Marc Vincent, João Comba |
| Workflow methodologies | Guillaume Huard, Arnaud Legrand, Lucas Mello Schnorr |
| Evaluation | Arnaud Legrand, Lucas Mello Schnorr, João Comba |

### 2.2.1 Analysis Techniques

We plan a visit in Brasil involving Guillaume Huard to transfer the analysis techniques developed in Grenoble to Porto Alegre. More specifically this will concern the aggregation based on entropy techniques as well as new exploratory works on the characterization of information. These techniques are especially suited to spot anomalies or irregular patterns in a very large

dataset. We expect significant advances in the two first case studies (parallel applications and long term phenology) in which we want to differentiate expected behaviors from unexpected ones (anomalies, infrequent cases, pathological cases, ...) in large and mostly regular data.

#### 2.2.2 Workflow methodologies

Arnaud Legrand and Lucas Mello Schnorr already have a strong experience regarding the use of workflows for interactive and reproductible analysis. We plan a visit from Arnaud Legrand in Porto Alegre to further strengthen this common expertise. The goal will be to explore new techniques (parallel workflows, leveraging MapReduce paradigm) and validate then through the case studies. In particular, the first and the third cases (parallel applications and public datasets) will benefit from these advances because of their size and dimensionality. Parallel workflow will enable the analysis to scale while the interactive analysis will enable the exploration of high dimensionality datasets.

#### 2.2.3 Evaluation

Although case studies are already taken into account by the other axes, it seems important to mention that the third case study (public datasets) has already been involved in a common work regarding social impacts and dissemination between Jean-Marc Vincent and Lucas Mello Schnorr (resulting in a common master course between Grenoble and Porto Alegre). We plan a visit in Porto Alegre from Jean-Marc Vincent to pursue work in this direction. The objectives are two-fold: advance in the analysis of this case study and continue with the dissemination of the eventual results.

## 3 Budget

### 3.1 Budget (for the first year)

According to INRIA call we propose the same budget structure for each year. Table 4 provides a summary of the expected items: four visits per year with 9 days on average in Brazil, and a light support for the organization of an annual workshop in France. Whenever possible, we may also use the budget to support PhD students from Brazil to spent some weeks in Grenoble. The goal is to let the student get acquainted with the research environment and in preparation for longer term missions, funded by CAPES/Cofecub for instance. Such strategy has proven successful for the previous experiences, increasing student productivity from the first day.

Table 4: Budget for one year.

| **Description** | **Budget** |
|---|---|
| 4 air tickets for Brazil (1200€ each) | 4800 € |
| 36 accommodation expenses for one person in Brazil (112 € per day) | 4032 € |
| Workshop organization | 1200 € |
| TOTAL FOR EVERY YEAR | 10032 € |

### 3.2 Strategy to get additional funding

The members of the associated team proposal have a solid experience in securing funding. They already acted as coordinators of CAPES/Brafitec and CAPES/Cofecub projects, the latter being one of the most succesful since it provides long-term PhD stays in France funded

by the Brazilian government (1.5 year with a cotutelle agreement). As of today, some of the members of the associated team proposal are involved in a CAPES/Cofecub project called "Group Formation, Analysis, and Visualization in Big Data" and focused on the analysis of social network data.

We currently lack solutions to provide additional funding on the research theme proposed by this associated team. The LICIA Laboratory, that used to fund short-term visits in both directions, comes to an end in 2018 and the EU/Brasil H2020 project called HPC4E2 has been put in the waiting list with minor chances to get some funding from the European Comission. Of course, the members keep attentive to Brazil-Europe calls that can be used to take extra funding to support member's research.

# 4 Added value

João Comba (UFRGS) is an expert in visual analytics, which is precisely an expertise that Polaris members lack. The French side, in particular Jean-Marc Vincent and Arnaud Legrand, has a strong theoretical background in statistical learning and information theory while the Brazilian side (both Lucas Mello Schnorr and João Comba) has a more practical background on parallel workflows and data manipulation. All members have a common interest for reproducible research and have several practical contributions in this domain (e.g. through the systematic adoption of laboratory notebooks and of R pipelines). Both sides have experience in developing trace visualization tools and analysing data from different application domains. For instance, Guillaume Huard has been the main architect of FrameSoc [22] while Lucas Mello Schnorr developed Triva [13] and other related tools.

Our teams are thus quite complementary and, at the same time, build on a common expertise. This combination enables very fruitful collaborations.

# 5 References

## 5.1 Joint publications of the partners


[1] Guilherme Rezende Alles, João L.D. Comba, Jean-Marc Vincent, Shin Nagai, and Lucas Mello Schnorr. Measuring phenology uncertainty with large scale image processing. *Ecological Informatics*, 59:101109, 2020.

[2] David Beniamine, Matthias Diener, Guillaume Huard, and Philippe O. A. Navaux. Tabarnac: Visualizing and resolving memory access issues on NUMA architectures. In *Proceedings of the 2Nd Workshop on Visual Performance Analysis*, VPA '15, pages 1–9, New York, NY, USA, 2015. ACM.

[3] D. Dosimont, R. Lamarche-Perrin, L. M. Schnorr, G. Huard, and J. M. Vincent. A spatiotemporal data aggregation technique for performance analysis of large-scale execution traces. In *2014 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 149–157, September 2014.

[4] Alef Farah, Lucas Mello Schnorr, and Jean-Marc Vincent. Frequency-based overhead compensation in HPC application traces. In *XIV Workshop de Processamento Paralelo e Distribuido*, 2016.

[5] Vinicius Garcia Pinto, Lucas Mello Schnorr, Luka Stanisic, Arnaud Legrand, Samuel Thibault, and Vincent Danjean. A Visual Performance Analysis Framework for Task-

based Parallel Applications running on Hybrid Clusters. *Concurrency and Computation: Practice and Experience*, 30(18):1–31, April 2018.

[6] Rafael Keller Tesser, Lucas Mello Schnorr, Arnaud Legrand, Fabrice Dupros, and Philippe O A Navaux. Using Simulation to Evaluate and Tune the Performance of Dynamic Load Balancing of an Over-decomposed Geophysics Application. In *Euro-Par 2017: 23rd International European Conference on Parallel and Distributed Computing*, page 15, Santiago de Compostela, Spain, August 2017.

[7] R. Lamarche-Perrin, L. M. Schnorr, J. M. Vincent, and Y. Demazeau. Evaluating trace aggregation for performance visualization of large distributed systems. In *Performance Analysis of Systems and Software (ISPASS), 2014 IEEE International Symposium on*, pages 139–140, March 2014.

[8] Robin Lamarche-Perrin, Lucas Mello Schnorr, Jean-Marc Vincent, and Yves Demazeau. Agrégation de traces d'exécution pour la visualisation de grands systèmes distribués. *Technique et Science Informatiques*, 33(5-6):465–498, 2014.

[9] Lucas Leandro Nesi, Lucas Mello Schnorr, and Arnaud Legrand. Communication-aware load balancing of the lu factorization over heterogeneous clusters. In *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*, Los Alamitos, CA, USA, dec 2020. IEEE Computer Society.

[10] Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Eric Simon, Fabian Colque Zegarra, João LD Comba, and Viviane Moreira. Userdev: A mixed-initiative system for user group analytics. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–8, 2019.

[11] Cicero AL Pahins, Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, Valérie Siroux, Jean-Louis Pepin, Jean-Christian Borel, and Joao LD Comba. Coviz: a system for visual formation and exploration of patient cohorts. *Proceedings of the VLDB Endowment*, 12(12):1822–1825, 2019.

[12] Vinicius Garcia Pinto, Lucas Mello Schnorr, Arnaud Legrand, Samuel Thibault, Luka Stanisic, and Vincent Danjean. Detecção de Anomalias de Desempenho em Aplicações de Alto Desempenho baseadas em Tarefas em Clusters Híbridos. In *WPerformance 2018 - 17º Workshop em Desempenho de Sistemas Computacionais e de Comunicação*, pages 1–14, Natal, Brazil, July 2018.

[13] Lucas Mello Schnorr, Guillaume Huard, and Philippe Olivier Alexandre Navaux. Triva: Interactive 3d visualization for performance analysis of parallel applications. *Future Generation Comp. Syst.*, 26(3):348–358, 2010.

[14] Luka Stanisic, Lucas C Mello Schnorr, Augustin Degomme, Franz C Heinrich, Arnaud Legrand, and Brice Videau. Characterizing the Performance of Modern Architectures Through Opaque Benchmarks: Pitfalls Learned the Hard Way. In *IPDPS 2017 - 31st IEEE International Parallel & Distributed Processing Symposium (RepPar workshop)*, Orlando, United States, June 2017.

[15] Fabian Colque Zegarra, Juan C Carbajal Ipenza, Behrooz Omidvar-Tehrani, Viviane P Moreira, Sihem Amer-Yahia, and João LD Comba. Visual exploration of rating datasets and user groups. *Future Generation Computer Systems*, 105:547–561, 2020.

## 5.2 Main publications of the participants relevant to the project


[16] Guilherme Alles and Lucas Schnorr. Parallel workflow support for starvz using drake. In *WSPPD 2018 – XVI Workshop de Processamento Paralelo e Distribuído*, pages 1–4, Porto Alegre, Brazil, September 2018.

[17] Claude Grasland, Robin Lamarche-Perrin, Marion Le Texier, Hugues Pecout, Sophie De Ruffray, Angelika Studeny, and Jean-Marc Vincent. Territoire, territorialité et territorialisation des événements médiatiques. In *CIST2016 - En quête de territoire(s) ?*, pages 207–213, Grenoble, France, March 2016. Collège international des sciences du territoire (CIST), Collège international des sciences du territoire (CIST).

[18] Robin Lamarche-Perrin, Yves Demazeau, and Jean-Marc Vincent. Building Optimal Macroscopic Representations of Complex Multi-agent Systems. *LNCS Transactions on Computational Collective Intelligence*, September 2014.

[19] Roger A. Leite, Lucas Mello Schnorr, Jurandy Almeida, Bruna Alberton, Leonor Patricia C. Morellato, Ricardo da Silva Torres, and João Luiz Dihl Comba. Phenovis - A tool for visual phenological analysis of digital camera images using chronological percentage maps. *Inf. Sci.*, 372:181–195, 2016.

[20] Vinicius Machado, Roger Leite, Felipe Moura, Sergio Cunha, Filip Sadlo, and João L.D. Comba. Visual soccer match analysis using spatiotemporal positions of players. *Computers & Graphics*, 68:84 – 95, 2017.

[21] Guilherme N. Oliveira, Jose L. Sotomayor, Rafael P. Torchelsen, Cláudio T. Silva, and João L.D. Comba. Visual analysis of bike-sharing systems. *Computers & Graphics*, 60:119 – 129, 2016.

[22] Generoso Pagano, Damien Dosimont, Guillaume Huard, Vania Marangozova-Martin, and Jean-Marc Vincent. Trace Management and Analysis for Embedded Systems. Research Report RR-8304, INRIA, May 2013.

[23] Luka Stanisic, Arnaud Legrand, and Vincent Danjean. An Effective Git And Org-Mode Based Workflow For Reproducible Research. *Operating Systems Review*, 49:61 – 70, 2015.


## 5.3 Other references


[24] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

[25] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2014.

[26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2009.