APPEL A MANIFESTATION D'INTERET INS2I 2015 Soutien Plateformes Science des Données (PlaSciDo)

Identification

PIs	Alexandre Gramfort and Balázs Kégl
Email	alexandre.gramfort@telecom-paristech.fr and kegl@lal.in2p3.fr
Project title	A data science platform in the Paris-Saclay Center for Data Science (Paris-Saclay CDS)
Acronym	DSP@PSCDS

Project summary

The Paris-Saclay Center for Data Science (PSCDS) is an interdisciplinary initiative of the recently inaugurated Université Paris-Saclay. It groups about **250 scientists in 35 laboratories**. Roughly half of us are **data scientist** (doing research in statistics, machine learning, signal processing, data visualization, databases), and half of us are **domain scientists** working with data (in physics, biology, environmental sciences, social sciences, and neuroscience).

The goal of this initiative is to establish an **institutionalized agora** in which these scientists can **find each other**, **exchange ideas**, **initiate and nurture interdisciplinary projects**, and **share their experience** on past data science projects. To foster synergy between data analysts and data producers we provide **initial resources** for helping collaborations to get off the ground and to mitigate the non-negligible risk taken by researchers venturing into interdisciplinary data science projects. Besides seed-financing concrete projects, we have also been **designing and learning to manage generic tools** to accompany data science projects with different needs. We organize innovation and strategy workshops, coding sprints, rapid analytics and model prototyping session, and data challenges.

The goal of this project is to define a **comprehensive platform to manage scientific data and scientific projects** in a data science ecosystem in general and in the context of the PSCDS.

Partners

Eric Chassefière	Pr/UPSud	eric.chassefiere@u-psud.fr	GEOPS/UMR8148
Alexandre Gramfort	EC/Telecom	alexandre.gramfort@telecom-paristech.fr	LTCI/UMR5141
Balázs Kégl	DR/CNRS	kegl@lal.in2p3.fr	LAL/UMR8607
Claire Nédellec	DR/INRA	claire.nedellec@jouy.inra.fr	MaIAGE/UR1404
Michèle Sebag	DR/CNRS	sebag@lri.fr	LRI/UMR8623
José de Sousa	Pr/UPSud	jose.de-sousa@u-psud.fr	RITM/UPSud
Sana Tfaili	MdC/UPSud	sana.tfaili@u-psud.fr	EA4041/UPSud
François Yvon	PR/UPSud	yvon@limsi.fr	LIMSI/UPR3251
Michalis Vazirgiannis	PR/Polytechnique	mvazirg@lix.polytechnique.fr	LIX/UMR7161
Emmanuel Vazquez	EC/CentraleSupelec	emmanuel.vazquez@supelec.f	SSE/E3S

1 Introduction

The subject of *data science* is the design of automated methods to **analyze massive and complex data** in order to **extract useful information**. Data science lies at the crossroads of computer science, applied mathematics, and statistics, and its raison d'être is the unprecedented growth of data that has been revolutionizing both science and industry for the last decade. Data science projects, by nature, require expertise from a vast spectrum of scientific fields ranging from **research on methods** (statistics, signal processing, machine learning, data mining, data visualization) through **software building** and maintenance to the **mastery of the scientific domain where the data originate from**.

The Paris-Saclay Center for Data Science (PSCDS) is an interdisciplinary initiative of the recently inaugurated Université Paris-Saclay. It groups about **250 scientists in 35 laboratories**. Roughly half of us are **data scientist** (doing research in statistics, machine learning, signal processing, data visualization, databases), and half of us are **domain scientists** working with data (in physics, biology, environmental sciences, social sciences, and neuroscience).

The goal of this initiative is to establish an **institutionalized agora** in which these scientists can **meet**, **exchange ideas**, **initiate and nurture interdisciplinary projects**, and **share their experience** on past data science projects. To foster synergy between data analysts and data producers we provide **initial resources** for helping collaborations to get off the ground and to mitigate the non-negligible risk taken by researchers venturing into interdisciplinary data science projects. Besides seed-financing concrete projects, we have also been **designing and learning to manage generic tools** to accompany data science projects with different needs. We organize innovation and strategy workshops, training and coding sprints, rapid analytics and model prototyping sessions, and data challenges.

Data science is a **deeply interdisciplinary** domain. Besides the usual challenges of projects involving experts of *two* distinct domains, successful data science projects also have to include a *third* (crucial) pole. It is the **software and system engineers** who can implement the methods developed by data scientists, maintain the tools, provide easy-to-use interfaces to scientist to access these tools, and train the scientists to use them. Figure 1 sketches a fully developed data science ecosystem, outlining the domains and the actors. There are numerous challenges at the interface between domain scientist and data scientist, but this document is naturally concerned by the engineering challenges at both the data scientist–tool building and domain scientist–tool building interfaces.

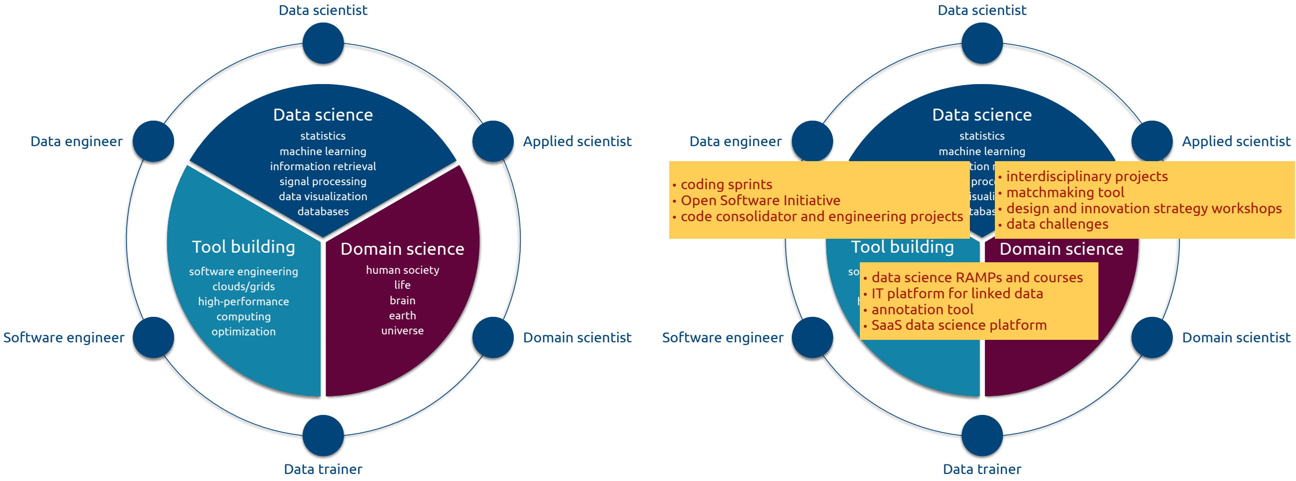

Figure 1: A fully developed **data science ecosystem**, and a set of tools that makes the ecosystem work.

In Section 2 we will briefly describe various tools that we have been building at the PSCDS to accompany data science projects with different needs. In Section 3 we describe our priorities and outline the ideal engineering profile that we are looking for to satisfy these needs.

2 Data science platform

One of the main bottlenecks of developing successful data science/domain science projects today is the **scarcity of data scientists motivated to participate in such interdisciplinary endeavors**. The two main focuses of the PSCDS are i) to **design management and engineering tools to make efficient use of the time of data scientists**, and ii) to **train scientists, novice to data science, to use data analysis tools**. In a longer term, our objective is to evolve the set of tools into a **complete platform** in which we can manage scientific data and data science projects.

In this section we outline a **subset of the tools** we have been developing, concentrating on the **tools that will require significant engineering effort**. We describe each tool, we summarize our activities, and we **define the role of a data engineer** in the development and maintenance of each tool.

2.1 Distributed collaborative development of (typically) open source software toolboxes

Most **research software is forgotten once its author moves** on to another position or another research topics, which means that most of the data science research output **does not end up in the hands of domain scientists** as usable tools. Some researchers invest time and energy to polish their code into a usable (but usually mono-purpose) software, maybe developed and maintained in a single research lab. Experience shows that the **most successful toolboxes follow the distributed collaborative open source paradigm** (e.g., scikit-learn, torch, or shogun). These toolboxes are **developed and maintained by a large community** with a light hierarchy, using appropriate development tools such as git and github. The PSCDS supports this effort with several actions. We have a **strong presence in scikit-learn** through our 4-5 core developers. We launched an **Open Software Initiative** to motivate PhD students to get involved with open source development. We also provide **code consolidator fellowships** for PhD students and postdocs to give them time to focus on integrating their code into general-purpose toolboxes. The data engineer will naturally join this ecosystem. If not yet involved, he/she will get acquainted with the open source development scheme, and will play an integrating role within the PSCDS community, **assisting data scientists to integrate their research code into general-purpose toolboxes**.

2.2 Data challenges

A data challenge is a recently developed unconventional dissemination and communication tool. Typically, a scientific or industrial data producer arrives with a **well-defined problem** and a corresponding **annotated data set**, defines a **quantitative goal**, and **makes the problem** and part of the data set (the **training set**) **public** on a dedicated site. Data science experts then take the public training data and **submit solutions for a test set** with hidden annotations. Submissions are **evaluated numerically** using the quantitative measure. Contestants are **listed on a leaderboard** with a score computed on part of the test set (the **public test** set). After a predefined time, typically a couple of months, the **final results** are revealed (using the complementary **private part of the test set**) and the **winners are awarded**.

Well-organized challenges are immensely useful for **generating visibility** in the data science community about novel application domains. Our HiggsML challenge attracted almost **2000 participants** and generated **30000 visits** on our main site. They are also useful for **benchmarking** in a fair way state-of-the-art techniques on well-defined problems. In an industrial setting, they are also helping companies to **find talented data scientists** to hire. On the other hand, they are **not adapted to solving complex and open-ended data science problems** in realistic environments. Neither the challenge host nor the organizers have **automatic access to the software** that generated the solutions or to the **participants** who are typically spread around the world. Indeed there are very few success stories on porting the winning solution into a final product or data analysis pipeline. They also emphasize competition between individuals or small teams, so they are diametrically opposed to the open source scheme where collaboration is the main driver.

The engineering needs of organizing challenges depend on the arrangement. If the technical issues of running the challenge is outsourced to a company (such as Kaggle or datascience.net), the engineering consists in **preparing the data** or developing an **annotation tool** and overseeing the annotation process. Running the challenge requires expertise in dynamic web site development and databases.

2.3 Rapid analytics and model prototyping (RAMP)

RAMPs are **single-day coding events**. They are combining distributed open-source development and data challenges. Similarly to data challenges, the preparation starts by **identifying a problem** and working with domain scientists and data scientists to **formalize the problem** and to **prepare a corresponding data set**. We invite about **thirty to a hundred participants** and organize the logistics (typically one big room with ample space to collaborate). We send them **preparation material** (the problem description, the data set, the tools required to solve the problems) before the RAMP. In the morning of the RAMP, the **data provider explains the problem and the data set**, then the participants tackle the problem during the day, **guided by coaches** (the number of coaches is about 10-20% of the number of participants).

The **technical backend** of a RAMP is crucial for its success. Similarly to a challenge, participants are submitting competing solutions, but instead of a simple prediction, they **submit code**. This means that the organizers have the opportunity to evaluate and combine the solutions, opening the door to various schemes that, besides evaluating the participants, can also motivate them to **collaborate and to build on each other's solutions**. We have started to build this backend, but developing and maintaining it will require a considerable engineering effort.

The second role of the data engineer in the RAMPs is to **deploy tools on various hardware resources** (clouds, GPU clusters, grids, personal computers). First, since participants submit code, we will have to **evaluate the code** that will require serious production engineering. Second, participants will be provided **transparent interfaces** to these tools that hide from them the technical details of the implementation and deployment.

We have been running RAMPs since the beginning of 2015, building both the backend and the tool set. It has been a great success. The first "warm-up" RAMP used the **HiggsML challenge** (Section 2.2) and the Kaggle engine. Our backend was deployed in the second RAMP where we tackled a **medical prediction** task. We have already identified three domains to fill the schedule until the summer: an **astrophysics** RAMP to categorize variable stars (with astrophysicists from LAL), a **chemistry** RAMP to verify drug quality (with chemists from the EA4041 lab), and a **conservation biology** RAMP on pollinating insect image data (with biologists from the Muséum National d'Histoire Naturelle).

While the development of the RAMP backend and the tool deployment is motivated by the success of our first RAMPs, they also constitute a **long term investment to assist domain and data scientists in their tool development and tool use**. A RAMP can be considered as the opening day of a collaborative data challenge. The site can stay open so participants can keep improving their models. All submitted codes are available to all the participants but also to the data provider, who can **reuse and recombine them to deploy a solution to the original problem**. A continued deployment and maintenance of data science tools is even more important: we will provide a transparent interface to **accompany domain scientists in their data analysis efforts**, and to accompany **data scientists in their methodological research**.

2.4 Training sprint

Initially we thought that bootcamps would work for both problem-solving and training. Because of the heterogeneity of our participants (which is a great asset) we decided to split our bootcamps into RAMPs (Section 2.3) and training sprints. Training sprints will typically consist of **short hands-on courses** to introduce to scientists, novice to data science, the tools that we deploy in RAMPs, and that they can use in their data analysis efforts outside RAMPs. The data engineer will have a crucial role in **designing the courses and the practical material** for the courses.

3 The data engineering profile for the PSCDS

We realize that it will be difficult to find a data engineer who can fill all the roles described in the previous section. We estimate that, in the long term, we **will need two to three engineers** for the development, maintenance, and organizational effort around PSCDS projects. Some of the effort can naturally be shared between data engineers and data scientists. For the data engineer profile in this call, we define the following **roles, in order of priority**.

1. **Installing and maintaining existing data science toolboxes** on various computational tools (clusters, HPC, GPUs, etc.), and **developing working relationships** with the various actors who manage these computational tools across the Saclay campus. Learning and applying **cloud and virtualization tools** (e.g., juju and SlipStream). Providing **transparent interfaces** to scientists to access these tools.

2. **Bringing data science research software into professionalized toolboxes**. Getting involved with **open source development**. Assisting data scientists (students, postdoctoral fellows, permanent researchers) to **develop their software engineering skills** and to get them involved with open source development.

3. **Training scientists to use the tools**. Accompany **domain scientists** in their **data analysis** efforts. Accompany **data scientists** in their **methodological research**. **Designing training sprints** and practical material for the courses.

4. **Animating** the developer community around data. Forming a core team. Finding and grouping a larger circle of (existing) data engineers at Université Paris-Saclay.