Paris-Saclay Center for Data Science

IDEX Paris-Saclay – Appel à projets recherche 2014

Project title[1]	Paris-Saclay Center for Data Science (Paris-Saclay CDS)
Co-PI	Balázs Kégl, DR/CNRS
Laboratory	Laboratoire de l'Accélérateur Linéaire (LAL)
Email	kegl@lal.in2p3.fr
Phone	01 64 46 85 95
Co-PI	Arnak Dalalyan, Pr/ENSAE
Laboratory/school	Laboratoire de Statistique (LS), ENSAE-CREST
Email	arnak.dalalyan@ensae.fr
Phone	01 41 17 65 33

Executive committee

Name	**Position**	**Laboratory/school**	**Email**
Florence d'Alché-Buc	Pr/UEvry	IBISC	florence.dalche@ibisc.univ-evry.fr
Michaël Aupetit	IC/CEA	CEA Tech/LIST	michael.aupetit@cea.fr
Frederic Chazal	DR/INRIA	INRIA Saclay	frederic.chazal@inria.fr
Laurent Barthes	MdC/UVSQ	LATMOS	laurent.barthes@latmos.ipsl.fr
Arnak Dalalyan	Pr/ENSAE	LS/ENSAE	arnak.dalalyan@ensae.fr
Gilles Faÿ	Pr/ECP	MAS	gilles.fay@ecp.fr
Elisabeth Gassiat	Pr/UPSud	LMO	elisabeth.gassiat@math.u-psud.fr
Cécile Germain	Pr/UPSud	LRI	cecile.germain@lri.fr
Alexandre Gramfort	MdC/Telecom	Telecom, CEA/Neurospin	alexandre.gramfort@telecom-paristech.fr
Balázs Kégl	DR/CNRS	LAL	kegl@lal.in2p3.fr
Erwan Le Pennec	Pr/Polytechnique	CMAP	erwan.le-pennec@polytechnique.edu
Stéphane Robin	DR/INRA	Agro	stephane.robin@agroparistech.fr
Michèle Sebag	DR/CNRS	LRI, INRIA	michele.sebag@lri.fr
Emmanuel Vazquez	MdC/Supelec	E3S/Supelec	emmanuel.vazquez@supelec.fr
François Yvon	Pr/UPSud	LIMSI	yvon@limsi.fr

[1] We changed our earlier name of Data Science Institute (DSI) to avoid a clash of acronyms with the Digital Society Institute and to better express the nature of the initiative.

Project summary

The subject of *data science* is the design of automated methods to **analyze massive and complex data** in order to **extract useful information** from them. Data science projects 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 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 propose to **provide initial resources** for helping collaborations to get off the ground, to **mitigate the non-negligible risk** taken by researchers venturing into interdisciplinary data science projects, and to **encourage** the use of unconventional forms of **information transmission and dissemination** essential in this communication-intensive research area (such as brainstorming sessions or data challenges). The CDS would fit perfectly in the **recent surge** of similar initiatives, both at the **international** and at the **institutional level**, and it would make the University one of the international fore-runners of data science. The CDS will naturally coexist and collaborate with existing structures, including six Labexes, doctoral schools, and M.Sc. programs. Although the primary focus of the initiative will be on **scientific data**, it will also be in a perfect position to play the role of a **contact point** to **industrial partners**.

Contents

1 Introduction

Scope. The subject of *data science* is the design of automated methods to **analyze massive and complex data** in order to **extract useful information** from them. 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 explosion of scientific knowledge in the last fifty years has initiated a process of specialization that manifested itself both in the institutional structure (hierarchical organization of mono-disciplinary laboratories) and in the career incentives that have compelled researchers to develop deep and narrow expertise from early on in their professional lives. The result of this natural tendency is that broad competences required by data science projects are rarely found in one single research group or laboratory. 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.

Big data and industry. Data science and big data are closely related but **not identical**. Whereas big data covers a broad spectrum of themes on capturing, transferring, storing, searching, securely sharing, archiving, and analyzing massive data, the **focus of data science is on the algorithmic and mathematical aspects of extracting knowledge from data**. In this sense, data science is a confined but indispensable component of big data. To keep the initiative focused, we deliberately chose to narrow the scope of the proposal to data science. In the same spirit, the primary focus of the initiative will be on **scientific data**. At the same time, we will be open to industrial partners whose participation will be essential to the long term success of the initiative, especially on the engineering aspects of big data and on dissemination and deployment of tools. The Center for Data Science (CDS) at the University of Paris-Saclay (UPSa), due to its size above the critical mass, will be in a perfect position to play the role of a **contact point** to these **industrial partners**.

Why now? Due to the unprecedented development of computational power, data storage, transmission, and cheap sensors in the last twenty years, collecting immense data quasi-automatically is as easy as ever. It has been transforming every-day life in front of our eyes, and it has been also changing the process of scientific discovery. Science has always been partially driven by data, but nobody could imagine the *scale* at which science is becoming data-driven. This phenomenon has been fundamentally transforming the scientific process: **data-driven science** has been rapidly becoming a fourth epistemological paradigm (besides the theory, experiments, and simulation). In some disciplines, such as particle physics, experiments assimilated computational paradigms a long time ago: both simulators and semi-automatic data analysis techniques are applied widely in today's large-scale experiments. Biology is another well-known example with the birth of bioinformatics, a distinct field that emerged from the interaction between biology and computer science. The sheer quantity of data poses a difficult challenge even in these fields, which have learned to cross the disciplinary aisles. On the other hand, massive data sets are now appearing in almost every field, ranging from life sciences to human sciences. These fields, novice to big data, create a huge demand for data science expertise that is hard to satisfy through the traditional mono-disciplinary channels. The **horizontal organization** envisioned by this proposal would be a perfect structure to accommodate this demand.

International integration. Data science is an emerging and essentially new research field. Nevertheless, in the last couple of years, similar initiatives to the CDS have been rapidly shaping the research landscape both at the **international** and at the **institutional level**. What would make the CDS unique among these initiatives is its size and depth of talent: the participating universities, laboratories, and engineering schools accommodate a large number of top researchers both on the data analysis and data producing sides. The CDS would be the **first such institute in France**, and creating the CDS now would **make UPSa one of the international fore-runners of data science**, helping the participants to gain visibility and to attract talented researchers. The goals of the initiative are also in perfect accordance with the scientific strategies of national research institutes.

Positioning. The CDS will naturally coexist and collaborate with existing structures, including **Labexes**, **doctoral schools**, and **M.Sc. programs**. Research teams and laboratories participating in the CDS are members of six Labexes (DigiCosme, Ecodec, Hadamard, LaSIPS, BIG, and P2IO) and the Département Hospitalo-Universitaire (DHU) Hepatinov, indicating that data science is partially represented in these structures. The mission of the CDS in this context will be to **support and motivate interdisciplinary data science research across the Labexes and the Schools** of UPSa.

Data science is an essential part of some of the existing and future M.Sc. programs and doctoral schools, and, to a lesser extent, it is also present in some of the science programs. The research supported by the CDS will be disseminated naturally by the teaching faculty which participates in both the CDS and these M.Sc. programs. In the long run, depending on the support and the mandate it will receive, the CDS can play a federating role in **catalyzing the data science curriculum** across these programs.

1.1 The structure of this document

The remaining two sections of this introduction are devoted to the **relatively new and somewhat unique challenges** we are facing in building a data science culture (Section 1.2) and to the recent **international response** to these challenges (Section 1.3). Section 1.4 contains the current list of UPSa laboratories participating in the CDS.

Section 2 describes the **management** of the CDS, the planned **actions** and the **budget** designed to finance these actions, and the **interface** of the CDS towards various related actors. Section 2.1 describes the actions designed to **foster synergy between data analysts and data producers**. The goal is not classical end-to-end financing of full projects, rather to **provide initial resources** for helping collaboration to get off the ground, to **mitigate the non-negligible risk** taken by researchers venturing into interdisciplinary data science projects, and to encourage the use of unconventional forms of information transmission and dissemination essential in this communication-intensive research area. Section 2.2 is devoted to the issue of **financial and institutional sustainability** of the CDS, including some **concrete avenues and resources** that we will explore and exploit during the first two years of the project. Section 2.3 describes the positioning of the CDS towards actors in big data and data science in **industry**. We provide a list of actions and our long-term vision related to **data science education** in Section 2.4. Section 2.3 describes our positioning towards **data centers**, **open data**, and **reproducible research** in data science. Section 2.6 contains the **management and governance** structure of the CDS, and Section 2.7 gives our planned yearly **budget**.

The scientific content of the project is divided into two parts. Section 3 groups data science application projects into ten themes, **biology and medicine** (Section 3.1); **astrophysics and cosmology** (Section 3.2); **neuroimaging** (Section 3.3); **particle physics** (Section 3.4); **chemistry** (Section 3.5); **music and text** (Section 3.6); **environment** (Section 3.7); **economy and social sciences** (Section 3.8); and **engineering** (Section 3.9). Section 4 gives a non-exhaustive panorama of data science themes an expertise of the participating teams. The summary is grouped into four subsections. We start with an introductory section on **fundamental data analysis** methodologies (Section 4.1), followed by three sections around three major data science challenges: **data complexity** (Section 4.2), **resource limitations** (Section 4.3), and **interactive visualization and experimental design** (Section 4.4).

1.2 Unique challenges in data science

In this section we summarize some of the unique challenges we face in building successful data science research collaborations in the near future. The **CDS will tackle** some of these challenges for which a local, **bottom-up approach** is adequate; other challenges will have to be addressed in a **top-down** fashion by **local and national institutions**. The section is based on our experience in observing and participating in large particle and astroparticle physics experiments, and it is also inspired by recent discussions on data science, in particular, by the presentations of Ed Lazowska, Saul Perlmutter, and Yann LeCun at the recent event on the Data to Knowledge to Action[2] initiative.

Data science is a **deeply interdisciplinary** research domain. On the top of the usual challenges of research projects involving experts of *two* distinct domains, potentially successful data science projects also have to include a *third* pole involving **software and system engineers** who can implement the methods developed by data scientists, maintain the tools, and sometimes run the software in production mode. Concretely, the following **distinct roles** must be filled if we want to implement state-of-the-art methods in order to solve new and data-intensive science problems.

1. The experimental **domain scientist** builds instruments and detectors, collects data, and analyzes the data in order to study new phenomena and to discover new laws of nature. He/she usually **works with engineers** both at the instrumentation and at the data acquisition interfaces. He/she may also be **acquainted by the methodological research and software engineering** aspects of the analysis chain, but his/her main **drive is scientific discovery** and his/her **carrier incentive is to publish scientific results**. He/she is less interested in advancing the state of the art in methodological research as long as the analysis gets done reasonably efficiently. He/she may be interested in developing and maintaining software tools which can be reused in other experiments, usually within the same domain.

2. The **data scientist** designs and analyzes algorithms. His/her main drive is to **propose new or improved methods** or **analyze them with new techniques**, where improvement is measured on **standard** and well known-problems and **benchmark data**. His/her carrier incentive is to **publish technical papers on methods**. He/she is less interested in solving actual problems as long as the **motivation to develop the method is plausible and/or well accepted** within the technical community, especially when a problem can be solved by existing techniques which are not publishable in technical venues. He/she is interested in building tools which are flexible enough to allow **wide methodological experimentation**, but which do not necessarily have the quality and efficiency to be used in large-scale production.

3. The **software engineer** implements existing techniques, designs and maintains software, and runs large-scale production on large data and large computational resources. His/her main drive and carrier incentive is to **build tools that are used by a large community**. Since (today) the **software engineer is often employed by domain scientists**, these tools often focus on the problems of a given scientific community. For the same reason, the software engineer may also be **acquainted by domain science**, but **rarely with the latest research in data science**. The demand and the methods to be implemented come from domain science, so most of the developed tools are not shared in a horizontal fashion among different domains.

4. This triangle is completed by the **system engineer** who builds and runs large computational infrastructure, occasionally working with software engineers who develop the middleware to provide flexible ways to access the infrastructure.

In industrial R&D, especially in multinational IT companies, covering all these aspects is rapidly becoming the standard way of tackling large-scale data science problems. In science, the closest example that comes to mind is experimental particle physics. Indeed, large **particle physics experiments** have both the manpower and the **experience and culture of managing large and complex projects**, so they can afford to have specialized experts on a continuous spectrum covering all four roles and even the interfaces between them. But even in these experiments most of the roles are filled by physicists who, for one reason or another, cross over into data science or software engineering.

[2] Blue texts are hyperlinks, clickable in the pdf file..

The fact that crossing over is possible in particle physics experiments also brings up another unique feature: specializing in physics is possible because all members of a collaboration co-author all science publications. Whereas it is probably not possible to implement this in other domains, in which research is carried out in small groups (often in fierce inter-group competition), it is possible and desirable to **design and implement new carrier paths** to incentivize researchers to delve into interfacing with domain scientists and software engineers, possibly spending precious time (not spent on publishing methodological papers) on understanding the domain science and developing tool-building skills.

Once such carrier paths are implemented, it is also possible to **hire people who are trained to fill these new roles**. However, another challenge we are facing in academic research is that engineers (and today even scientists) who can fill these roles are very much in demand in industrial R&D. Indeed the skillset required in data science research in industry and in academic research overlaps probably more than ever before. The problem is not only that salaries are higher in industry (it has always been the case), but also that the **problems raised by commercial applications are highly interesting and challenging**, and the **research environment and freedom** proposed by these companies is often **on par** with what we have in **academic research**.

To summarize, to develop a culture in which data science research can thrive, we need

- administrative structure and infrastructure to **temporarily co-locate scientists and engineers** of different backgrounds,
- incentives to allow researchers to follow **new carrier paths**,
- incentives to make people of different backgrounds **communicate and venture into "risky" interdisciplinary projects**,
- infrastructure, manpower, and management for the development and long-term maintenance of **software tools that can be shared** among several scientific domains,
- investment in **data science education** in order to satisfy the growing demand both in industry and in academic research, and
- new models for **openly disseminating knowledge** that comes in the form of **data and tools**.

1.3 The international scene

Providing an exhaustive list of national and international initiatives would be impossible. We enumerate here some of the most important actions to paint a landscape in which we wish to place the Paris-Saclay Center for Data Science.

- Following the announcement of the National Big Data R&D Initiative of the White House in 2012 in the US, both the national funding agencies (e.g., NSF, NIH, DOE, and DARPA) and individual universities engage in large-scale top-down actions in order to promote data science and research on big data.
- To summarize the latest trends and challenges on data science research, the US National Research Council's Committee on the Analysis of Massive Data releases an important document on Frontiers in Massive Data Analysis.
- The Research Data Alliance (RDA) is formed to accelerate data-driven innovation world-wide through research data sharing and exchange.
- NIST forms the Big Data Working Group to draw a Big Data Technology Roadmap.
- New York University opens its Center for Data Science.

- University of Washington founds its eScience Institute.
- Berkeley will launch its Institute for Data Science.
- The Moore and Sloan foundations announce a five-year 37.8M$ cross-institutional initiative to support the three previous institutes.
- Columbia opens its Institute for Data Sciences and Engineering.
- The University of Rochester announces a 100M$ commitment to create and house its Institute for Data Science.
- The University of Amsterdam announces the creation of its Data Science Research Center.
- Edinburgh University launches its Center for Doctoral training in Data Science.

1.4 Participating teams

The following table contains the current list of UPSa laboratories and the number of permanent researchers participating in the CDS.

Laboratory/school	**team(s)**	**contact(s)**	**number of permanents**	**themes**
Laboratoire de l'Accélérateur Linéaire (UMR8607, UPSUD / IN2P3-CNRS)	APPSTAT, ATLAS, AUGER, ILC, JEM-EUSO, LSST	Balázs Kégl, DR/CNRS	8	experimental particle and astroparticle physics, cosmology
Laboratoire de la Recherche en Informatique (UMR8623, UPSUD / INS2I-CNRS/INRIA)	A+O/TAO, LAHDAK/OAK , BIOINFO/AMIB	Cécile Germain, PR/UPSud, Ioana Manolescu, DR/INRIA, Christine Froidevaux, PR/UPSUD	15	machine learning, data mining, optimization, bioinformatics
Laboratoire Traitement et Communication de l'Information (UMR5141, TELECOM-PARISTECH / INS2I-CNRS)	STA, TII, AAO, IC2	Olivier Cappé, DR/CNRS, Alexandre Gramfort, MDC/TELECOM	19	machine learning, signal processing
Laboratoire IBISC (UEVRY / ENSIIE)	AROBAS	Florence d'Alché-Buc, PR/UEVRY	6	machine learning, optimization, bioinformatics
NEUROSPIN (CEA, INRIA)	INRIA PARIETAL, MEG LAB.	Bertrand Thirion, DR/INRIA, Gaël Varoquaux, CR/INRIA Alexandre Gramfort, MEG/CEA	3	computational neuroscience, brain imaging (fMRI / MEG), biomedical signal processing, machine learning
Département Hospitalo Universitaire (DHU) Hepatinov		Jean-Charles Duclos-Vallée, MD/HDR	3	pathogenesis and treatment of liver cancer
Centre de mathématiques et de leurs applications (ENS CACHAN)	Apprentissage statistique, Images et Signaux	Nicolas Vayatis, PR/ENS CACHAN	6	machine learning, signal processing
Laboratoire de Mathématiques (UPSUD, INRIA)	Probabilités, Statistique et Modélisation , INRIA SELECT	Pascal Massart, PR/UPSUD, Christophe Giraud, PR/UPSUD	21	statistics
Laboratoire de Statistique (ENSAE/CREST)	LS	Arnak Dalalyan, PR/ENSAE	6	statistics, machine learning
E3S (SUPÉLEC) and Laboratoire des Signaux et Systèmes (UMR 8506, SUPÉLEC/INS2I-CNRS/UPSUD)	SIGNAUX, SYSTÈMES, E3S	Matthieu Kowalski, MDC/SUPELEC, Pascal Bondon, DR/CNRS, Emmanuel Vazquez, MDC/SUPELEC,	7	statistics, signal processing, machine learning

Laboratory/school	**team(s)**	**contact(s)**	**number of permanents**	**themes**
Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (UPR 3251, INS2I-CNRS)	TLP, ILES, Venise	Nicolas Férey, MdC/UPSud, Guillaume Wisniewski, MdC/UPSud	10	speech processing, natural language processing, visualization
Centre de Mathématiques Appliquées (Polytechnique)	TSA, MEV, SIMPAS	Erwan Le Pennec, Pr/Polytechnique, Stéphane Gaiffas, MdC/Polytechnique	8	statistics, machine learning, signal processing
Laboratoire d'Informatique (Polytechnique)	DASCIM	Michalis Vazirgiannis, Pr/Polytechnique	8	machine learning, graph/text mining
Mathématiques et Informatique Appliquées (UMR518 Agro/INRA)	Stat&Génome, MORSE	Stéphane Robin, DR/INRA	6	statistics, machine learning
CosmoStat (CEA)	SparseAstro	Jean-Luc Starck, DR/CEA	6	computational cosmology
DTIM (ONERA Palaiseau)		Jean-Denis Muller, DDS/ONERA	10	machine learning, image processing
CEA Tech/LIST (CEA)		Michaël Aupetit, IC/CEA	10	machine learning, data mining
Cognitive Neuroimaging Unit (INSERM-CEA)	UNICOG	Christophe Pallier, DR/CNRS	12	cognitive neuroscience, neuroimaging
INRIA Saclay	Geometrica, Aviz	Frederic Chazal, DR/INRIA, Jean-Daniel Fekete, DR/INRIA	6	visualization, computational geometry and topology
Laboratoire Mathematics Applied to Systems (Centrale)	Statistique, FiQuant, Digiplante	Gilles Faÿ, Pr/Centrale	5	machine learning, statistics, biological data
Centre de Vision Numérique (Centrale)		Nikos Paragios, Pr/Centrale	5	image processing, machine learning
Institut d'Astrophysique Spatiale (UMR 8617, UPSud / INSU-CNRS)	IDOC	Marian Douspis, AA/CNAP	5	signal processing, data analysis, astrophysics
Laboratoire de Microéconométrie (ENSAE/CREST)	LMI	Xavier D'Haultfoeuille, Pr/ENSAE	6	économétrie, statistique
Mathématiques et Informatique Appliquées de Jouy (UR0341, INRA)	MegaDim, Animod, DynEnvie	Alain Trubuil, IR/INRA	6	statistics, bioinformatics, text mining, image processing, experimental
Mathématique, Informatique et Génome (UR1077, INRA)	Bibliome, BioSys, GenEvol	Claire Nédellec, CR/INRA	6	statistics, bioinformatics, text mining, image processing, experimental design

Laboratory/school	**team(s)**	**contact(s)**	**number of permanents**	**themes**
LATMOS (UMR8190, UVSQ/CNRS), LOCEAN (UMR7159, UVSQ/CNRS)	SPACE, ESTER, SHTI	Cécile Mallet, MdC/UVSQ, Laurent Barthès, MdC/UVSQ	7	data assimilation, image processing, data analysis of Earth, atmosphere and ocean observations
EA4041 (UPSud)	GCAPS	Sana Tfaili, MdC/UPSud	5	analytical chemistry
RITM (UPSud)	Network and Innovation, Globalization and Territories	Jose de Sousa, Pr/UPSud	10	economics and management
Laboratoire de Finance Assurance (ENSAE/CREST)	LFA	Jean-Michel Zakoian, Pr/ENSAE	5	finance, insurance, econometrics
IGM/I2BC (UPSud / CNRS)	Séquence, Structure et Fonction des ARN; eBio Bioinformatics platform	Daniel Gautheret, Pr/UPSud	5	biological data analysis, bioinformatics
CESP (UPSud / UVSQ / INSERM)	Methodology and clinical epidemiology in molecular oncology	Stefan Michiels, Pr/UPSud	6	medical and biological data analysis, bioinformatics
EA1611 (UPSud)	Droit et Sociétés Religieuses	François Jankowiak, Pr/UPSud	2	history of canon law and ecclesiastical institutions, church and state

2 Management, relations, actions, budget

2.1 Research actions

WORK GROUP: **Alexandre Gramfort**, Balázs Kégl, Arnak Dalalyan, Florence d'Alché-Buc, Erwan Le Pennec, Emmanuel Vazquez

The main mission of the CDS is to motivate, foster, and organize research around data science. The goal is not classical end-to-end financing of full projects, rather to **provide initial resources for helping collaborations to get off the ground**, to **mitigate the non-negligible risk taken by researchers venturing into interdisciplinary data science projects**, and to **encourage the use of unconventional forms of information transmission and dissemination** essential in this communication-intensive research area. Our actions are designed based on hands-on experience of some of the participants in data projects already running in scientific fields (Section 3). We also took inspiration from the actions proposed by similar recent initiatives in the international scene (Section 1.3). At each anniversary of the CDS, we will assess the obtained results using formal metrics (e.g., **number of papers co-authored** by researchers coming from different domains and **number of *submitted* projects** to financing agencies) and other evaluation techniques and re-adjust the allocation of the funds. The lessons we learn and the obtained research results will be summarized in a **yearly progress report**.

We propose to finance the following actions.

1. We will design and maintain an open **interactive web portal**. It will list the participants, broadcast calls and other news, and provide information on educational relations (Section 2.4) and data centers (Section 2.5). To surmount the initial hurdle of finding the right experts for a given problem, we will design and maintain a tool (possibly using existing social networking applications) where scientists can describe their expertise and their interests and initiate collaborations around particular data science projects. The web site should be up and running after six months.

2. In most of the successful data science collaborations, it is necessary that the participating parties learn each other's language and develop a solid understanding of the discipline where the data comes from. Experience shows that for a successful collaboration, the best way to proceed is to **embed researchers** in the data provider laboratory for an extended time period. This takes time and effort and it is a decision that implies considerable risk compared to a more "classical" carrier path, especially for young scientists. To mitigate this risk and to recompense researchers taking it, the CDS will support **full or partial sabbatical stays** in different partner laboratories. In addition, the CDS will also partially finance **projects that involve at least two teams** of different partner laboratories or schools, preferably one from the data provider side and one from the data analysis side. This axis could be partially covered through other IDEX and Labex calls (doctoral fellowships, postdoctoral calls, chairs). In our preliminary survey we have identified demand for about 15 Ph.D. and postdoctoral fellowships and one or two visiting fellowships, so we will be able to roll out this action in the spring of 2014.

3. The CDS will provide support for organizing scientific meetings, ranging from traditional **workshops** and summer schools to less formal multi-day **brainstorming sessions** (such as the Brainhack unconference series) in which researchers from different disciplines can gather to discuss the challenges and the possible solutions of a given data science problem. We have identified eight to ten themes around which we can organize meetings of various formats in the first year. One series on data science in particle physics, astrophysics and cosmology is already running.

4. The CDS will provide support for organizing **data challenges**, in which experts work together for providing simplified but meaningful public data sets (scientific or industrial) that can incite the data analysis community to adapt their tools to novel applications. Successful examples, including

the Netflix prize contest, challenges organized by the PASCAL Network of Excellence, the KDDCUP series, and competitions managed by Kaggle, demonstrate that a data challenge is an excellent way to orient data scientists to new fields, and to define, formalize and solve the arising problems. The Kaggle platform is best known for offering public contests but most of its revenue comes from private contests accessible to people who did well on previous public challenges. The CDS aims to promote such actions on a dedicated web page and support such initiatives financially. Should a the data challenge investigated during the workshop be proposed by an industrial partner, private funding is planned to be used as an additional resource. Besides Kaggle, datascience.net, a portal launched by Bluestone with some of the participants of the CDS, is also a potential partner in organizing challenges. We are in the process of launching a data challenge in the particle physics theme, and we foresee to organize three to four challenges per year.

5. **Software** plays a fundamental role for the success of data science projects. To be able to scale to larger datasets, to be ready for distributed environments such as clouds, or simply to be ran on laptops for rapid prototyping and interactive data exploration software has to be well written and maintained. Software can provide quasi-standardized solutions to typical data science problems and make the latest algorithmic/theoretical advances available for a wider audience. The emergence of peer-reviewed software venues, such as the Journal of Statistical Software, or the software track of JMLR, shows that the data science community now recognizes software as commendable form of scientific research output. The CDS will promote scientific software development by financing coding sprints and through initiatives similar to the Google Summer of Code. Such *UPSa summer of code* programs should have, as priority, the funding of **open source software projects** developed by the partners of the CDS but will be open for international and highly visible software projects. Some candidate projects are Scikit-Learn, multiboost, MNE and NiLearn. Applications by students of the different schools and universities of IDEX Paris-Saclay shall be favored but should not be a requirement. While Google offers $5,000 to each student for 4 months of work from June to September, the CDS plans to give each student 800€per months. Following the model of Google, students will be evaluated at mid-term to know if they are allowed to finish the program. We are ready to launch this action in the summer of 2014.

In terms of budget (Section 2.7), our most costly action is the second. To give incentives to the teams to participate in the other (less costly but more time consuming) actions, we will require from the participants of every financed project to actively engage in the collaborative actions 3-5.

2.2 Institutional relations and sustainability

WORK GROUP: **Balázs Kégl, Arnak Dalalyan**

The project received **overwhelming support and encouragement** both from the **participating UPSa laboratories/universities/schools** and **national institutions**. The overall consensus is that there will be support for sustaining the CDS in the long run if a bottom-up demand is demonstrated in the first two years, but concrete financial engagement is premature at this time. The following avenues and resources will be explored and exploited during the first two years of the project.

- We received strong commitments from the **host institutions of the two PIs** (LAL, UPSud, ENSAE, and GENES) both for **setting up, assisting, and complementing the management** structure of the CDS and for **sustaining the CDS** after the initial IDEX-financed period.

- The Mission Interdisciplinarité (MI) of the CNRS expressed a strong interest in the initiative. First, the goals of the CDS are fully aligned with the CNRS Mastodons Program; mutual interactions may lead to cross fertilization and possibly common supports for projects. Second, the objectives of the CDS correspond to the long-term national strategy of the MI around big data and data science; the

CDS can play a significant role in the new GDR project whose content is currently in discussion within the CNRS.

- We will actively explore possibilities to participate in **European and national calls**, e.g, ANR, Horizon 2020.
- We foresee to become a contact point to industrial projects around big data, and we will explore the possibility to **complement the financial support of the CDS from private sources** (see Section 2.3).
- We will negotiate **in-kind contributions from Labexes and institutions**. For example, we can insert the data science theme into existing calls (IDI theses, IDEX chairs, Labex postdocs). We will also propose to participating institutions the possibility of delegating to the CDS the selection of projects but only finance them when the candidate is joining their laboratories or teams.

Long-term sustainability Besides financial stability, finding the right long-term structure for the CDS is also a challenge we will tackle in the first two years. The main dilemma (see Section 1.2) is that the very nature of data science is in contrast with being closed within the strict boundaries of a classical laboratory with a mostly permanent research staff. The solution that seems to be emerging in international examples (Section 1.3) is to establish **research centers or "studios"** with a **small team of core members** (including **researchers** and **engineers**), a **larger circle of part-time members**, and an **infrastructure for temporarily co-locating interdisciplinary teams**. A structure defined and implemented by the MI of the CNRS that roughly corresponds to these requirements is the *Hôtel à projet (HAP)*. The HAP was mainly established as a structure that can manage the access to shared resources, but the model could also be easily adapted to tackle the challenges of research on data science (Section 1.2). We are also thinking about studying the possibility of collaborating with the SystemX IRT beyond industrial projects (Section 2.3): their model is ideal for managing interdisciplinary projects within a light and flexible structure, and it is not unimaginable to adapt their model for managing purely scientific (but interdisciplinary) data science projects. If we move towards a HAP model, permanent positions (researchers/teachers but also toolbuilder/software developer engineers) could come from the UPSa quota or from the interdisciplinary quota of national institutes.

2.3 Industrial relations

WORK GROUP: **Emmanuel Vazquez, Arnak Dalalyan**, Balázs Kégl, Michalis Vazirgiannis, Alexandre Gramfort, Erwan Le Pennec, Michaël Aupetit

Announced by the French government as one of the priorities in the matter of industrial development, Big Data is becoming an essential component of gaining a competitive advantage in business. Given the high-level expertise of the teams participating in the CDS on various fields of Data Science/Big Data (see Section 4), the CDS will be well-placed to become a privileged **contact point for both multinational corporations and small and medium enterprises**. There are already several examples of successful collaborations (industrial chairs, CIFRE theses, ...) on projects related to data science between an industrial partner and an academic laboratory involved in the CDS. We intend to build on these collaborations to develop industrial relations within the CDS.

Even if the primary focus of the CDS is on challenges coming from the scientific data, our conviction is that the CDS has an important role to play in establishing and maintaining relations with the industrial world, at least for two specific reasons. The first reason is that in a near future, it will be crucial to capture **funding sources from the industry**, in order to achieve financial independence of the CDS. The second reason is that in many scientific areas, such as in finance and insurance, the data is mainly collected and stored by private companies. Having **access to** these **data** is vital for researchers to ensure scientific progress while the findings of researchers may help the companies to increase their industrial competitiveness.

In its activities related to the collaboration with industrial partners, the CDS will work in tight relation with the existing structures such as the IRT SystemX and the French cluster for digital contents and services Cap Digital.

Following a meeting with the management of the IRT SystemX, at least two areas of collaboration between the CDS and the IRT SystemX were identified. First, if the contacts between the CDS and industrial partners lead to concrete **projects**, they can be **realized within the IRT SystemX**. Second, if the IRT SystemX is approached by some companies for realizing projects that require expertise of academic partners on data science, then the IRT SystemX will get in touch with the CDS for participating in the project.

We also intend to organize **knowledge dissemination events** to which representatives of private companies will be invited. Cap Digital kindly proposed to help us in contacting startups and enterprises interested in data science and in a potential collaboration with the CDS.

2.4 Educational relations and actions

WORK GROUP: **Elisabeth Gassiat**, Arnak Dalalyan, Balázs Kégl, Michèle Sebag, Michalis Vazirgiannis, Gilles Faÿ, Laurent Barthès, Nicolas Vayatis, Agnès Desolneux, Florence d'Alché-Buc

Data science is an essential part of some of the existing and future **mathematics**, **computer science**, **bioinformatics**, and **atmospheric science** M.Sc. programs and doctoral schools, and, to a lesser extent, it is also present in some of the other science programs. Although the primary goal of the CDS is research, it can play a natural federating role in **catalyzing the data science curriculum** across these programs. For the project duration, our goals are

1. to make the **M.Sc. programs with a data science context** visible across the different schools,
2. to list all **Ph.D. theses and M.Sc. internships** with data science content on the web page of the CDS,
3. to help **industrial partners** to find the appropriate programs in which they can propose internships or Ph.D. fellowships,
4. to **foster interactions** between professors and/or students of existing courses, and
5. to collect requests of specific **short courses focusing on data science** at the doctoral level, and to organize the courses.

In the long run, depending on the support and the mandate it will receive, the CDS can play a deeper role in catalyzing the data science curriculum across these programs, for example, by advising and encouraging students to take courses across these programs on a case-by-case basis.

For the project duration, we are planning to carry out the following concrete actions.

- During the first year, the CDS will provide support for **drawing a detailed map of all data-science-related programs and courses**. The list will be available on the web portal of the CDS in 2015, at the moment when the common M.Sc. programs of UPSa will start. The list will include M.Sc. programs ("parcours") which are mainly dedicated to data science, but also individual courses in disciplinary M.Sc. programs dedicated to data analysis or applied statistics. In M.Sc. Mathematics, large parts of the *Mathematics of randomness* program and of the *Mathematics, Vision, Learning* program concentrate on the mathematical grounds of statistical learning and computational statistics. Other programs in M.Sc. Mathematics, namely *Mathematics for life sciences*, *Financial mathematics*, and *Optimization* also contain individual courses focusing on data science. In M.Sc. Computer

Science, large parts of the four programs (*Machine Learning, Information, Content*; *DataScale*; *Data & Knowledge*; and *Decision Support and Business Intelligence*) in the *DataSense* theme focus on the computational aspects of data science. The multidisciplinary TRIED M.Sc. program and the new M.Sc. Bioinformatics are also largely devoted to data science. There are also programs focusing on data science in the engineering schools such as "la voie Data Scientist" at ENSAE and the Mastère Spécialisé Big Data at TÉLÉCOM.

- During the first year, we will also draw a **map of doctoral schools** and list them on the web portal of the CDS. The page will also contain **ongoing and proposed Ph.D. theses and M.Sc. internships** related to data science.
- The CDS will provide support for organizing an **annual meeting between professors from M.Sc. programs** with a data science content. The goal is to exchange information on what is taught in the different programs, to negotiate possible "exchanges" (students in one program taking credited courses in another program), and to discuss the common challenges in teaching this interdisciplinary subject.
- The CDS will provide support for organizing **short doctoral courses and summer schools**. The idea is to streamline the data science content to a targeted audience. An example of what we have in mind is the IN2P3 School of Statistics, a bi-annual spring school designed for Ph.D. students and young researchers in particle physics and astrophysics (2010, 2012, proceedings).
- The CDS will encourage **reading clubs and student seminars** given by and to Ph.D. students in data science.

2.5 Interface to data centers

WORK GROUP: **Cécile Germain**, Balázs Kégl, Michalis Vazirgiannis, Isabelle Blanc, Michel Jouvin

SUMMARY. For a long time, the national and local research institutions have deployed pooled (shared) IT infrastructures that enable sharing physical resources and expertise, creating key capabilities and experience in this process. A paradigm shift occurs with the current call of national research institutions for an integrated approach, with **interdisciplinary structures** embedded in the local environment; the structure is organized around a permanent high-level and open technological platform operated and animated by qualified staff. In our case, Data Science is based on Information Technology (IT). A **Data Science IT Platform** is the component of the IT system that should provide an effective and efficient platform for empowering the communities of researchers to create and share knowledge: develop, evaluate, exploit, and disseminate software, build and share new applications, share access to common data repositories, and create information sources and new services. The CDS will contribute to the definition, organization, and implementation of a Data Science IT Platform within UPSa at two levels: the CDS can be used by the governing bodies as a **transversal entry point** for organizational and infrastructure actions; and the CDS will develop prospective studies and proof of concept implementations for **advanced sharing and knowledge building**.

CONTEXT. The pooling of physical, immaterial, and human IT resources is driven by multiple factors. The creation of large, energy-efficient **data centers** is motivated by the changing structure of the operational costs that stems from technology trends. **Software sharing** is motivated by the need for avoiding redundant activities between science fields or between science fields and industry and preventing the technological balkanizations that precludes interoperability. **Data sharing**, beyond its general necessity to simply perform state-of-the-art research, is the first step to build the critical mass required for maintaining research data for reuse and preservation through its lifecycle of interest. Finally, the new frontier of IT for research is automated support for knowledge extraction and sharing, with a growing concern for **reproducibility.**

Assets. The CDS collaboration features exceptional assets to advance a Data Science IT Platform at UPSa.

Large scale infrastructures VirtualData is a data center housing infrastructure designed for high energy efficiency. The resources are exploited mainly under a grid or a cloud (StratusLab) model. VirtualData has been initiated by eight physics laboratories within the P2IO Labex; it pools 6000 cores and 3PB of disk, on two locations (University of Paris Sud and École Polytechnique). The University of Paris Sud is building a roadmap to exploit the housing infrastructure (with or without cloud technology) and more extensive resource pooling. IDRIS is one the High Performance Computing (HPC) national centers, and a major resource of competence for HPC activities, with a particular interest for us due to its historic interactions with other national and local facilities (Maison de la simulation, Meso-centre Ecole centrale) in the campus geographic area. BADAP is a project jointly led by the GENES and the IMT—funded by the program "investissements d'avenir"—of a Big Data platform oriented towards research and innovation. Besides several hundreds of terabytes of storage facilities and 4TB of memory, it also offers a small team of qualified specialists for accompanying scientific projects related to Big Data along with a technology (through the CASD) providing secure access to sensitive or confidential data.

Innovative exploitation A *Data Science IT Platform* calls for a systemic approach that exploits synergy among computer research communities who see it as an object of research, and other research communities who see it as a platform in service of research. A long-lasting scientific interdisciplinary collaboration already exists in this domain with the Grid Observatory project, as well as relevant basic research (see Section 3.9).

Reproducible research CMLA has promoted the foundation of a new kind of scientific journal, whose first working example is the prototypical journal Image Processing On Line (IPOL) founded in 2010. IPOL publishes peer-reviewed papers describing the algorithms in accurate literary form, coupled with code. Furthermore it allows scientists to check directly the published algorithms on line by providing a web execution interface on any uploaded image. An archive associated with each article permits researchers to share their experiments. Finally, data papers are encouraged, where researchers can publish papers linked to databases that they wish to share. The acquisition must be described accurately in the IPOL paper. With a growing experimental archive of more than 100000 original image data used online by researchers worldwide, IPOL demonstrates the efficiency and appeal of online execution to foster reproducible research and interdisciplinary communication. CMLA can therefore bring to the CDS its experience, technique and software in the online publication of shared data and software, with a set up that rewards researchers doing this effort.

A data-intensive scientific environment Most participants of CDS are involved in a rich ecosystem of data-oriented scientific environments, through thematic networks, Equipex facilities, and local data acquisition facilities. In many cases, they are involved in the policy design for scientific IT at their respective institutions. These positions will contribute to a short feedback loop between CDS and the relevant institutions, and contribute to a better integration of data-intensive IT at the UPSa level.

Objectives. The functionalities of a Data Science IT Platform can be organized along the 1.0/2.0/3.0 terminology: even if the underlying technologies and goals are not essentially web-oriented, the final goal is the same: make the experience of the user, in our case the research activity of data scientists, much easier and more productive.

1.0: Raising awareness. We will assess the **basic requirements** (computational power and storage) of the CDS partners in the short (project duration) and medium (5 years) term. We will also map the existing **data repositories** operated by CDS partners that are relevant for the interdisciplinary research of CDS, and their relations to wider scientific networks. The expected results in general are a better knowledge of resources and needs within the project, and promotion of pooling experiments, including cloud technologies and parallel systems. Specific outcomes can be expected both at the disciplinary level - new sharing of data, infrastructure or expertise - and at the interdisciplinary level, with a particular focus on making available truly large datasets for algorithmic research internally.

2.0: A collaborative interface. Scientists should be able to do more than just retrieve information, by interacting and collaborating as creators of user-generated content in a **virtual community**. At this level, the content should probably be limited to software and data; needless to say, to the extent that the researchers accept to share them. The essential envisioned capabilities are firstly to make these software and data actually usable by a human researcher, and second to facilitate the related interactions: information should be **richer, easier to find and more thoroughly categorized** than by the usual static ”portfolios”. The expected results are threefold: firstly, actual proof of concepts developments for cross-exploitation of codes and data, organized around a shared infrastructure; second, sharing of programming expertise; and finally a structured methodology derived from theses experiences.

3.0: The computer generates new information. The goal is to create a capability that provides seamless access to effective and personalized science aids. Within this wide and still speculative area, the CDS will explore two precise avenues. The first one is the Linked Data framework, by exploring how related data repositories operated by CDS partners could be seamlessly and meaningfully **queried** with the specific objectives of Data Science. The second is **reproducibility/repeatability** of experiments. This subject is receiving increasing attention; elaborating on the interaction of the existing exceptional experience at CMLA, the scalable cloud infrastructure at VirtualData, and the scientific expertise of CDS data providers and analysts would be a powerful instrument of cohesion for the project, and offer the perspective of a **“killer application”**.

ACTIONS.These goals will be implemented through the following activities.

Web portal Within the CDS general portal (action *interactive web portal*), a portal dedicated to the implementations of the above-mentioned goals will be created. The envisioned timeline is T0+6 months for the 1.0 goal, T0+12 and T0+24 for *experiments* on respectively the 2.0 and 3.0 goals.

Proof of concept developments Cross-exploitation of codes and data as well as the deeper integrations of 3.0 objectives are likely to raise two kind of adaptation issues: interoperability at the applicative level, and accessibility at the infrastructure level. The specific role of this action will not be to provide support for the strictly applicative developments that, at this stage, will better be realized directly within the research teams (possibly through CDS funding, either *coding sprints* or *data challenges* actions), but to advise on good practices and implement the tools required to make them sustainable. Accessibility encompasses the obstacles that can be specifically encountered with shared or cloud infrastructures, and should be supported by the *coding sprints* action.

Roadmap This activity targets the contribution of CDS to the wider UPSa initiative for organizing its research computing infrastructures. The WG will provide the scientific interface to the Data Science community by its feedback along the project, and by contributing to a roadmap for the Data Science IT within UPSa.

The technical developments will be realized by a research engineer, with strong skills in semantic data organization and collaborative technologies, for the duration of the project.

2.6 Management and governance

For the first two years we are implementing a light structure with an executive committee (EC), a chairman, and a deputy chairman. We have also formed work groups (WGs) around five issues related to the functioning of the CDS and its interfaces. The founding EC consists of 15 members (first page): the chairman, the deputy chairman, heads of the WGs, and (at least) one representative per participating institution.

- **The executive committee** will approve the budget, make decisions on financed projects, review project reports, and approve the yearly progress reports. Members will be delegated by the participating institutions, and the EC can also elect additional members. We will have monthly meetings (with possible electronic participation due to the geographic distances). Electronic votes will also be possible on urgent matter. The EC will elect the chairman for a duration of two years, ratify the deputy chairman, and approve the formation of WGs and WG heads.

- The **chairman** is responsible for proposing the budget, managing the administrative team, managing the project calls, and chairing the EC meetings. He/she will select the deputy chairman (approved by the EC). He/she assures the continuity of the executive structure between EC meetings. He/she may make budget decisions on urgent items for sums not exceeding 5K€. He/she can delegate responsibilities to the deputy chairman and to the work groups. He/she sends supporting documents and agenda a week before each EC meeting and summarizes the decisions after (with the help of administrative staff) a week after. The chairman represents the CDS in negotiations and outside discussions, and towards the IDEX.

- The number and composition of the **work groups** is flexible: any participating member can propose the creation and join any WG (which the EC then approves). For now, we have formed five WGs

  1. to define and evaluate **research actions** (Section 2.1),
  2. to design a plan for **long-term sustainability** of the CDS (Section 2.2),
  3. to define and implement a strategy vis-a-vis possible collaborations between the CDS and **industrial partners** UPSa (Section 2.3),
  4. to define and implement a strategy vis-a-vis **data science education** at UPSa (Section 2.4), and
  5. to define and implement a strategy vis-a-vis **computational infrastructures** at the UPSa (Section 2.5).

  The WGs will propose action items at the monthly EC meetings, and contribute to the yearly progress report of the CDS (approved by the EC).

In the first two years we will make an attempt to organize an **advisory committee** with outside members which can help to orient the CDS and to define its relations with other French and international initiatives. In the first two years we will also make an exception of the general rule: the chairman and the deputy chairman will change role at the end of the first year.

2.7 Budget

The following table describes a detailed estimate of the **yearly budget** of the CDS financing actions described in Sections 2.1, 2.4, 2.5, and 2.6.

management costs (project management, secretariat, engineering support, web portal)	100K€
approximately 14 sabbatical and visiting positions, postdoctoral and Ph.D. fellowships	800K€
2 summer schools and other short courses	100K€
UPSa summer of code for 20 students and coding sprints	100K€
workshops, brainstorming sessions	50K€
3 data challenges	50K€

3 Scientific themes I: data science in natural, human, and engineering sciences

This section groups data science application projects into ten themes, biology and medicine (Section 3.1); astrophysics and cosmology (Section 3.2); neuroimaging (Section 3.3); particle physics (Section 3.4); chemistry (Section 3.5); music and text (Section 3.6); environment (Section 3.7); economy and social sciences (Section 3.8); and engineering (Section 3.9). Some of the sections describe large scientific projects and collaborations. These projects are not, strictly speaking, part of the CDS. The role of these sections is to provide a panorama, to give the context in which data science can become a major ingredient, and to motivate data science research by describing some of the concrete challenges.

3.1 Biology and translational medicine

Author(s): **Florence d'Alché-Buc**, Stéphane Robin, Fariza Tahi, Daniel Gautheret, Arthur Tenenhaus, Alain Trubuil, Alain Denise, Christine Froidevaux

Team(s): AROBAS/IBISC, BioInfo/LRI, Stat&Génome/MIA, MORSE/MIA, Signaux/Supélec, MegaDim/MIAJ, Animod/MIAJ, DynEnvie/MIAJ, Bibliome/MIG, BioSys/MIG, GenEvol/MIG, CESP/UPSud, Hepatinov

For now two decades, biology has faced the so-called data revolution: spectacular advances of experimental technologies have provided access to the "omics" data (transcriptomics, proteomics, metabolomics), drastically changing research in molecular biology and genetics. Progress in the life sciences will depend on the ability to properly interpret the large-scale, high-dimensional and structured data sets that are generated by these technologies. Three big challenges central to bioinformatics and translational medicine are especially relevant for CDS teams:

- From genotype to phenotype,
- Biological Networks,
- Meta-genomics.

Moreover, an additional challenge, transversal to multiple biology and medicine fields, has also been identified as an important target: the multi-platform data analysis challenge.

3.1.1 From genotype to phenotype

Modern technologies in molecular biology give access to a variety of markers along the genome of plant or animal species. Such markers provide a quite comprehensive view of the genotype of a given individual. In the past few years, single nucleotide polymorphism markers (SNPs) have become the most popular ones. In human, several hundreds of thousands of such loci are known and can be routinely measured. The genotype of an individual also be defined in a broader way, including copy number variations (CNV) or epigenetic informations (methylation data). In many applications in medicine or agronomy, genotype information can also be combined with other exogenous information such as environmental or climate data.

Many studies in biology, medicine or agronomy now aim at understanding the relationships or associations that exist between such markers (constituting the genotype) and the phenotype of the individuals in a given species. When carried at the whole genome level, such studies are known as genome-wide association studies (GWAS).

Molecular signatures of diseases Biomarkers has become an important key-word in medicine in the recent years. Such markers are supposed to be associated with, say, some increased disease risk for the patients who carry them. Finding such markers is a typical variable selection problem that is to be considered in a 'large p, small n' framework as the number of patients is always much smaller than the number of available markers. The hope in such study is that a relevant set of markers can help to early diagnose diseases and better treat them. Although the studies are carried at the cohort level, genotype data being individual, a second hope is that they can help to adapt treatments to each single patient, resulting in personalized medicine. High-throughput genome analysis is the fastest growing area of cancer research. Genomics enables the discovery of new cancer driver genes against which new drugs can be targeted, and it is the basis of new therapeutic trials in which cancer treatments are adapted to a patient's specific genetic profile (personalized medicine). Partner groups are involved in the integrative analysis of cancer genomic data produced by clinical research, including exome and full genome sequences, transcriptome (RNA-seq) and genome-protein interactions (ChIP-seq). Integrating these different levels of information poses serious computational challenges in terms of data management and biostatistics. Several research projects aim at modelling this multi-level data to produce comprehensive molecular signatures of cancer subtypes and characterize the gene networks disrupted in cancer. Collaborations between clinicians, bioinformaticians, and other CDS teams should help to tackle the important technical bottleneck created by the sheer volume of the processed data.

Phenotype prediction for plants and animals Apart from medicine, the knowledge of the genotype-phenotype relationship has many other potential applications. For example, genomic selection aims at predicting the phenotype (or traits) based on the genotype. Such approaches are progressively replacing quantitative trait localization (QTL) approaches, as the determination of few isolated predictive markers may not be needed. In this field, many challenges arise such as the prediction of the phenotype of an individual based on the genotypes of its parents (or grand-parents), the transfer of the genotype-phenotype relation from one large, well-studied population to another similar but still different and smaller population, or the analysis of the way genotype and environment interact to affect the phenotype.

Complex phenotypes High-throughput phenotyping technologies also develop very rapidly, resulting in complex high-dimensional outputs, making the description and the analysis of an individual's phenotype challenging by itself. With the objective to better understand the gene by environment interaction in plant growth, two strategies have emerged for the acquisition of phenotypic data. The first one concerns the development of technological platforms of high-throughput phenotyping, for which a large number of genotypes are grown in controlled conditions, with automatic measuring devices, generally including image acquisition with digital cameras and image analysis, to determine the evolution of some key variables during plant growth (leaf surface, soil humidity, plant mass when plants are grown in individual plots, etc.). This first strategy allows a very regular monitoring of plant growth in a few environmental conditions (a few hydric stress levels for example). However, the limited number of experimental conditions that can be studied simultaneously is a restriction to infer the results of the genotypic analysis in a broad range of environments. For this reason, a second strategy is also explored, based on cheaper experimental protocols (far simpler observations on plant data, for example final yield, intermediate phenological stages, leaf area index via satellite images, etc.) repeated in a large variety of environmental conditions. Some open public data base (FAO, USDA, etc.) are available for this purpose. Both strategies generate huge data sets combining genotypic data, phenotypic plant data, soil data, and climatic data. Devising proper methods to analyze these data sets with parametric or non-parametric regression is a challenge for modelers, in order to improve the predictive capacity of the models describing the interaction of the gene with the environment. Information extraction (Section 3.6) from scientific papers is a complementary strategy to collect and formalize the published information about phenotypes related to given environmental properties and genetic specificities. The integration of knowledge from these various (experimental, curated, and predicted) sources, is a new and open challenge.

3.1.2 Biological networks inference and pathways analysis

The cellular response to internal or external input signals arises from the interaction of numerous components such as DNA, RNA, proteins, and small molecules. Biological networks are the simplest abstraction to describe those complex interactions. Identification and analysis of those networks, such as protein-protein interaction networks, gene regulatory networks, and signaling pathways, will not only lead to a better understanding of biological functions but also open the door to therapeutic targeting. Recently, a number of technological advances, such as DNA microarrays, RNA-Seq, liquid chromatography tandem mass spectrometry, and similarly liquid or gaseous chromatography mass spectrometry, have enabled biomedical researchers to collect large amounts of transcriptomic, proteomic, and metabolomic data. In addition, curated repositories and bibliographic databases containing both vast amounts of such data, as well as functional information, ontologies, gene and protein interactions, pathways, etc. are expanding at a fast pace (e.g. KEGG, IntegromeDB, BioGrid, GEO, NURSA, PubMed, WoS, etc.). The increasing availability of such high dimensional data and structured information have led to a number of novel learning problems, including that of *network inference*, *information extraction from text* (Section 3.6) and of various modeling issues. It is now well recognized that theoretical methods, such as statistical inference, discrete or continuous (dynamical) modeling, graph analysis, and automated reasoning are needed to make sense of this abundance of information.

Role of RNA Next generation sequencing (NGS) have paved the way for the mass production of genomic data in a limited time (the last generation sequencer Illumina HiSeq2000 can for instance generate 200Gb per week). As genomes are currently sequenced faster than they are processed and analyzed, new approaches to genomic data analysis are needed. One of the central issues in genome analysis concerns the analysis and discovery of RNA genes. RNA is a key regulator that controls central cellular functions in all organisms. A particular class of RNA, microRNA (miRNA) inhibits transcription or translation of target messenger RNAs and is involved in vital functions such as cell growth, differentiation, apoptosis, or metabolism and it has been linked to major human diseases such as cancer, Alzheimer, muscular pathologies and autism. A first important challenge in RNA research is the development of in silico methods for identification of miRNA and other RNA genes. Integrative approaches considering different types of data and methods (data mining, algorithm design and data integration) should be used for an efficient prediction (in term of results) as well as for a powerful prediction (in term of time calculation). Here parallel architectures, such as GPUs, should also be considered. A second challenge is the modelling of RNA structure, which is instrumental in the prediction of RNA-ligand interactions and the in-depth study of RNA functions. New computationally intensive RNA structure prediction methods relying on high-throughput sequence and 3D data should strongly benefit from collaborations between CDS teams.

Network inference Automated network inference from experimental data raises several issues among which, the limited number of observations compared to the large number of components, the presence of intrinsic and extrinsic noise in data and the absence of observations for numerous components. To face the high dimensionality issue, sparse modeling is mandatory while graphical models offer a probabilistic framework to encode noise as well as hidden variables. Other ways to compensate the lack of experimental points include prior knowledge incorporation in a wide sense and active learning strategies devoted to experimental design. Finally, the key question regarding network inference is the semantics associated with a direct edge in a regulation graph. Directed edges under certain conditions reflect causal relationships. Although estimating such relationships is known to be a very challenging task, it nevertheless represents a central challenge in network inference that may be addressed using perturbation data.

3.1.3 Meta-genomics

Communities of many species live as a whole and species from these communities can not be studied separately as they can not be isolated and grown away from their medium and from the other species. In the recent years, many scientific projects studied, for example, the microbial community hosted by the human gut, by ocean or lake water, or a portion of soil to name a few. Metagenomics aims at understanding the functioning of such communities through the analysis of the set of the genomes of all the present species considered as a whole, called metagenome.

Modern technologies of deep sequencing have allowed a huge step forward in metagenomics. A typical metagenomic experiment results in a huge set (typically 10^8 or 10^9) of small sequence samples in the genomes of all individuals from the community. Based on these sequences (called 'reads') we hope to be able to know which species are present in the medium (and in which proportion) or, from a functional view-point, which genes are there, whatever the species that carry them.

Multivariate analysis. When a reference genome is available for each species composing the community, we face a classical situation in ecology when interested in comparing the diversity or the composition of several communities observed in different medium or under different conditions. Classical multivariate techniques can then be considered, but they face novel issues, mostly due to the throughput of modern technologies. On the one hand, the number of considered samples will grow very rapidly in the near future. On the other hand, the number of possibly present species is already very large (10^3 or 10^4). For both reasons multivariate techniques do not scale and new relevant and efficient models and algorithms are needed.

Assembly and alignment Many communities are mostly composed of unknown species or from species for which no reference genome is available. Therefore, the presence of such species can not be easily assessed by the reads sampled in the metagenome. New species can be hopefully discovered by assembling reads using standard genome assembly techniques. Again, such approaches face new issues such as the relatively short length of the reads compared to usual genome sequencing projects or the facts that a large number of species are simultaneously present in the sample.

3.1.4 Transversal challenge: multi-platform data analysis

Modern technologies in molecular biology give access to a variety of markers along the genome and provide a full view of the state of a cell under a given experimental or developmental condition. These datasets are used to identify targets of signaling pathways and their components, can classify samples into phenotypic classes and can provide useful predictors of cellular fate. This is a discipline which has seen rapid improvements in the scale, speed, and resolution of data, and the evolution in technology platforms appears to continue unabated. The major hurdle for researchers wishing to exploit rapid technology developments in gene expression profiling is the inability to compare data derived on different experimental platforms, a barrier arising from difficulties in reconciling datasets which inevitably contain heterogeneities, both in recording conditions and in annotation resulting in 'batch-effects' which mask genuine biological signal.

All the challenges described previously (e.g., network inference, relationships between (possibly very complex) genotype and phenotype) have been developed for a single version of a single microarray platform, and required exhaustive data acquisition to develop robust analyses. Cross-platform analysis is a way to address such problems and demonstrates the community need to benchmark new samples against a collective community standard, but it already faces obsolescence as the platform manufacturer updates chip versions and changes technology formats.

3.2 Astrophysics, cosmology, and astrostatistics

AUTHOR(S): **Jean-Luc Starck, Marc Moniez**

TEAM(S): PLANCK-LSST/LAL, IDOC/IAS, STA/LTCI, MAS/CENTRALE, COSMOSTAT/CEA

The main tools for comparing theoretical results with observations in astronomy are statistical. However, the development of huge astronomical databases presents challenges of scale, and has initiated an active use of newly-developed statistical techniques in astronomy, notable examples being Bayesian analysis (Sections 4.1.1 and 4.3.6), sparsity and compressed sensing (Section 4.3.5). As examples, in the field of the microwave background, the resolution of PLANCK leads to a dataset whose size is so large that analysis even at the two-point level is non-trivial, and at higher order it is extremely challenging; PAN-STARRS1 will have a complete survey of 3π steradians of petabyte size; the Dark Energy Survey and the VST KIDS surveys will be well underway and offering similar difficulties of analysis, and the cosmological community will be preparing for LSST and for EUCLID, a survey of a large fraction of the sky at a resolution close to that of the Hubble Space Telescope. Wide-field spectroscopic cosmology surveys will be contemplating surveys of over 10 million objects with a spectral resolution of 5000, and the SKA precursors will be grappling with data challenges which are currently unsolved. These examples also highlight the big current role and even bigger future role of archival data in astrophysics research. Our teams plan to work on the following projects.

- PLANCK: Cosmic Microwave Background (CMB) component separation, CMB non-Gaussianity studies and primordial power spectrum reconstruction.
- EUCLID: The EUCLID mission is now selected. We are strongly involved in the Weak Lensing activity (2D and 3D dark matter mass mapping, non-Gaussianities, etc). We are in charge of the management of the OU-LE3 (unit in charge of designing the algorithms to be used to derive the EUCLID products).
- CFHTLENS: The weak-lensing data from the Canada-France Hawaii Lensing Survey (CFHTLENS) was released in 2012. We are applying now to the CFHTLENS data several methods developed at CEA-Saclay (mass reconstruction, peak counts to constrain cosmological models, PSF and shape measurement).
- LOFAR: We are working on LOFAR image reconstruction from the measured visibilities using the compressed sensing theory concept.
- XXL: Our involvement in the XXL surveys is two-fold: First, our wavelet tools will be used for detection and classification of extended objects in very low signal-to-noise X-ray images. Second, we will use the CFHTLENS data to measure the weak-lensing signal of X-ray selected clusters, to obtain independent mass estimates.
- LSST will be a large-area, wide-field, ground-based telescope designed to provide deep images of roughly half the optical sky every few nights during 10 years of operations. We are involved both in the data management and the analysis.

The following two sections summarize EUCLID and LSST, two of the main future large-scale surveys involving huge data sets.

3.2.1 The EUCLID project

The EUCLID mission is the next major spatial survey of the European Spatial Agency. The launch of the telescope is planned for 2020. The scientific goal of this mission will be to shed light on the "dark" components of the Universe with a wide field imager in space. To study the dark Universe, EUCLID will make use of the weak gravitational lensing effect which provides a direct measure of the distribution of dark matter in the Universe. This is done by measuring the weak distortions induced by intervening large-scale structures on the images of distant galaxies. This can be used to measure cosmological parameters, and, in particular, the dark energy equation-of-state parameter which affects the growth of cosmic structures. The wide-field imager of EUCLID will circumvent atmospheric effects, which limit ground-based surveys, and provide both high statistics (i.e., more resolved galaxies) and low systematics (thanks to a small and stable PSF) for weak lensing.

The data reduction is extremely challenging. It requires to measure shape of galaxies with an extremely high accuracy. This implies to work on many data processing problems.

- Estimating the point spread function (PSF) of the instrument from undersampled and noisy observations of stars.
- Estimate the PSF variations along the field in order to allows to derive a PSF at any position of the field.
- Estimate the shape of galaxy from undersampled and noisy observations, and correct the measurement from the PSF effect.

Once all shape measurements are achieved, several other processing steps remain to be done.

- 2D Mass Mapping: this consists in reconstructing a projected mass map from the shear measurements. This is a ill-posed inverse problem that requires the development of sophisticated.
- 3D Mass Mapping: a 3D density mass map can be reconstructed. This is an inverse problem closely related to tomography. It has been shown that there are also some links between this problem and compressed sensing theory, developed in statistics.
- Power spectrum and 2PCF Estimators: we have to develop estimators of the power spectrum and two point correlation functions which are able to analyze catalogs with several billions of galaxies and on 15000 square degrees

The volume of data to be processed should be between 100 and 1000 petabytes.

3.2.2 The LSST project

The Large Synoptic Survey Telescope (LSST) will be a large-area, wide-field, ground-based telescope designed to provide deep images of roughly half the optical sky every few nights during 10 years of operations. The LSST survey will provide an unprecedented data set, both in quantity and quality, to study questions on Dark Matter and Dark Energy, and a critical resource for a variety of astrophysical investigations (e.g., studies of small bodies in the solar system, programs that map the outer regions of the Milky Way, and searches for faint optical transients on a wide range of time scales). With its 8.4 m primary aperture, the LSST will join the present generation of telescopes with "8-meter class" mirrors. The unique LSST 3-mirror optical design, combined with a large (65 cm diameter) focal plane, produces an extraordinary field of view (3.5° FOV $\sim$ 10 square degrees). The telescope will be equipped with a 3.2-billion-pixel camera and with a fast 2 s readout low-noise electronics that will record images of the sky spanning six photometric bands (0.3 to 1.0 micron). The images will be used to find time variations in the sky, and, after co-addition, to produce a huge catalog (of the order of 10^{10} objects) up to magnitudes 26.5-27 AB. Among the main scientific LSST products, 250 000 Type Ia supernovae ($z \leq 1$) will be detected each year, and prompt alerts will be issued to the international observing community for follow-up spectroscopic observations and observations in other wavelength bands. Surface brightness shapes of over 3 billion galaxies ($z \leq 3$) will also be measured in the course of the ten-year survey.

The LSST will acquire nearly 2000 images and produce $\sim$ 120 terabytes of raw and preprocessed image and catalog data per night. The data will be reduced in real time and the resulting images, database, search tools, and software will be made publicly available. Images will be acquired every 15 seconds, and image analysis for stringent quality control and detected transient alerts will be generated within 60 seconds. This dynamic range poses challenges to the design of the LSST data acquisition and management systems similar to those encountered in particle physics. The deep, very wide-field, multi-color imaging survey of the sky that LSST will produce will be a "goldmine" for astronomical investigations that will obviously benefit from the progress in data science.

Data management in LSST The Computing center of the IN2P3 will provide CPU and storage resources corresponding to 50% of the LSST needs for the Data Release Processing. Its role might be extended to cover Data Access Center and level 3 services for a broader European community. The following key numbers expected at the end of the survey give a flavor of the LSST computing challenges

- 70 petabyte long-term storage at the end of the survey.
- 25 petabyte of storage space on disk, used for image access processing cache.
- $\sim$ 900 teraflops of sustained computing power. This computing power requirement is dominated by the object-by-object measurement stage.

Data science in LSST The following themes have been identified has major challenges that can be addressed through the CDS.

- **Image analysis and visualization.** The data reduction will involve complex workflows with massive parallelism. Efficient image handling via emerging processing architectures and compression algorithms will thus be investigated. The data visualization at this scale presents also a challenge on its own, be it the visualization of images, data products, or the aggregation of both.
- **Data analysis and machine learning.** With billions of objects and trillions of detections, LSST will provide an unprecedented dataset for data mining in astronomy: associations between sources and objects (possibly moving and varying in luminosity), discovery and enumeration of characteristic features, concise data representation, astronomical object classification will challenge current computer assisted knowledge extraction techniques.

3.3 Neuroimaging

AUTHOR(S): **Alexandre Gramfort**, Christophe Pallier, Bertrand Thirion, Gaël Varoquaux
TEAM(S): INRIA PARIETAL, MEG/CEA-NEUROSPIN, LTCI, INSERM UNICOG, CMLA/CACHAN

Neuroimaging is a sub-field of medical imaging dedicated to the brain. Images of the brain can be anatomical, showing the structures and tissues, or functional, capturing the effects resulting from neural activations. The history of modern neuroimaging is marked by a few milestones: the first clinical applications of MRI magnetic resonance imaging in the last 70's, the invention of diffusion MRI (dMRI) by researchers among which Denis Le Bihan, head of the Neurospin facility at CEA, in the 80's, the first functional MRI (fMRI) recordings in 1992, the first full head magnetoencephalography (MEG) system also in 1992, the first scanner combining MRI and Positron Emission Tomography in the last three years, among others.

While the breakthroughs listed above are the consequences of the progress in physics and engineering, the last decade has seen a growing interest of other scientific communities for this field of research. The images and signals produced by brain imaging devices are all digital. An MRI scan is obtained after some clever physics and signal processing and consists of a cubic grid of volume elements, a.k.a. voxels. When doing fMRI such data are recorded every one or two seconds leading to large 4 dimensional files (3 spatial and 1 temporal axes) which leads to computational challenges when looking for statistical effects in the data. This simple illustrative example with fMRI demonstrates that the reality of today's research in the field of neuroimaging is highly interdisciplinary requiring skills in physics, mathematics, computer science, statistics and machine learning.

Already today the building blocks of what is called here Data Science (computer science, statistical machine learning, engineering) are pervasive in the context of neuroimaging. Still as demonstrated by recent international initiatives, namely the American *Human connectome project* and the European FET Flagship *Human Brain Project (HBP)*, the future of brain research, and in particular neuroimaging, will be more and more data centered. For example the Human connectome project, whose data are made freely available to the public, will contain the high resolution anatomical MRI, the dMRI and the fMRI scans of 1200 persons. It will also contain the MEG recordings of 100 persons. The size of the dataset produced for one subject reaches 18 GB in a compressed file format. The total for the 1200 subjects will lead to multiple terabytes of data, duplicated a few times due to the storage of post processed files. The HBP will similarly open terabytes of data ready to be explored by data scientists in close collaboration with brain researchers.

The data science challenges in the field of Neuroimaging fall into different categories:

- Neuroimaging leads to supervised learning problems. An example of such problems is the prediction from neuroimaging data (MRI, MEG, etc.) of the category of a patient, e.g. healthy or not. This task is is a binary classification task in the machine learning literature. When working with fMRI or M/EEG one may want to decipher the neural code by predicting from data some mental processes of the subject or what kind of stimulus he/she was being presented. Another supervised learning task relevant in the field of neuroimaging is ordinal regression and ranking (See section 4.2.4) problems where one may what to predict among an ordered set of values, e.g., healthy, Mild Cognitive Impairment (MCI), or Alzheimers disease (AD) in the clinical context of AD detection. The resolution of such problems involve the minimization of convex functions (See Section 4.3.1) with potentially sparse regularizations (Section 4.3.5).

- Neuroimaging raises also some unsupervised learning (Section 4.1.3) challenges such has the mining of task-free data (so called "resting state" fMRI or M/EEG). The ambition is here the extraction and characterization of so called "brain networks" that refer to a set of distributed brain regions that coactivate spontaneously. From such networks, brain graphs **??** can be extracted in a data-driven manner. The techniques employed rely on statistical models and the development of tractable inference techniques in order to propose scalable solutions. The use of online techniques (Section 4.4.3) to support out of core computation is relevant here.

- A third example, in this non-exhaustive list of data science challenges in the field of Neuroimaging, is the screening of large databases in order to detect acquisitions problems and artifacts. This ambition of automatic quality control and assessment is known in machine learning as "outlier detection" (Section 4.1.4).

In all the problems listed above statistical machine learning tools employed need to be adapted to the particularities of the data in order to succeed. For example variables are voxels with a 3D grid structure and local similarities with MRI, and observations are time series (Section 4.2.1) obtained from a linear mixing process with MEG or EEG etc. What is also common to these problems is the critical issue of model selection (Section 4.1.5), as well as tools to visualize massive data (Section **??**) as produced by brain imaging devices.

The Neurospin facility at CEA Saclay thanks to its imaging equipments (4 MRI scanners including high field, 1 MEG) is a key asset, and it already federates research efforts going from cognitive neurosciences (INSERM Unicog group lead by Stanislas Dehaene getting European funding with 2 ERC grants), to data analysis methods and statistics (CEA/INRIA Parietal Team lead by Bertrand Thirion, and Alexandre Gramfort at Telecom ParisTech / CNRS LTCI also affiliated with CEA). The Unicog group has made outstanding contributions to the the field neurosciences and the Parietal Team working with Telecom ParisTech as had over the last few years a real impact on practice of neuroimaging data processing. Since the creation of Neurospin the interdisciplinary environment of Neurospin has proven to be a key factor of success.

3.4 Particle and astroparticle physics

AUTHOR(S): **Balázs Kégl**, Roman Pöschl, David Rousseau, Sylvie Dagoret-Campagne
TEAM(S): APPSTAT/LAL, ATLAS/LAL, AUGER/LAL, ILC/LAL, ILC/LLR

The objective of particle physics is to study the basic constituents of matter, largely within the theory called the Standard Model and its possible extensions. The main experimental tools are particle accelerators and colliders in which beams of particles are accelerated to very high kinetic energy and collided into other particles. The particles resulting from the collision are then detected in particle detectors consisting mainly of track detectors (high-resolution devices in which the paths of individual particles can be separated) and calorimeters (measuring the energy of particles or groups of particles). From these raw measurements, different events (mainly particle decays and collisions) are reconstructed, the whole "picture" is compared to model predictions, and model parameters (for example, the existence and the mass of new particles) are inferred from comparing a large statistics of collision events to simulated events. Astroparticle physics studies particles of astronomical origin. It shares its goals both with astrophysics (Section 3.2; where these particles are coming from, what the acceleration mechanisms are, etc.) and particle physics. Cosmic ray particles reach energies of several orders of magnitude higher than in man-made accelerators. By observing the particle cascade generated by the collision of the cosmic ray particle and atmospheric particles, parameters of the first interaction can be inferred at energies not available in particle detectors.

Historically, the knowledge in particle physics has been built gradually, accessing higher and higher energies by building larger and larger accelerators and detectors. This means that, at any point in time, models describing the low-energy physics in the detector are largely known. In principle, full parametric generative models can be built based on these models, and so classical forward-fitting statistical methods (either maximum likelihood or Bayesian; Section 4.1.1) could be used. There are great advantages of using parametric generative models: the modeling "language" remains faithful to physical concepts, model and observational uncertainties can be handled formally, and the different levels in the hierarchical models can be connected in principled way. There are several practical issues that nevertheless make it difficult to use a full generative treatment in today's experiments. The full generative model is often hierarchical, with a handful of parameters to infer, up to a million observed signals, billions of observations, and several conceptual levels between observations and parameters of interest. Gaps between these levels are usually filled using simulations, and the classical treatment would require to build reduced phenomenological models based on these simulations. The consequence is that, in practice, nonparametric approaches (Section 4.1.2) are often used to tackle inverse problems directly, and classical statistical tests are only applied at the end of the reconstruction chain.

The rich and long tradition of operating within a data-driven paradigm makes particle physics one of the most interesting "consumer" of future developments of data science. Advanced numerical Bayesian techniques (large-scale MCMC techniques, likelihood-less Bayesian numerical methods, etc.; Section 4.3.6) could be used to inject more forward-building flavor into reconstruction chains. Stochastic optimization (Section 4.3.2) will be the main tool of maximum likelihood fitting of complex generative models. Nonparametric techniques (Section 4.1.2) will continue to be used both in the reconstruction phase and also in the on-line triggers of the detectors. Finally, simulation-based experimental design (Section 4.4.2) will likely become the principle technology for designing and optimizing accelerators and detectors.

In the rest of this section, we describe four concrete experiments involving data-science projects.

3.4.1 The Pierre Auger experiment

Since their initial discovery by Victor Hess, we have learned a lot about cosmic rays: we know that they are sub-atomic particles (electrons, protons, and nuclei of heavier elements up to iron and even uranium) with energies that vary on a large scale (from a few billion eV to more than 10^{20} eV – the energy of a tennis ball flying at 200 km per hour!). Cosmic rays are produced by known or unknown astrophysical mechanisms, so studying the composition, the energy, and the sources of these particles is important for understanding the universe tracing back to its origins. The most interesting and enigmatic cosmic particles are those with the highest energies. Whereas there are known mechanisms that produce particles up to 10^{15} eV, the acceleration mechanisms involved in producing the highest energy cosmic rays are still unknown. While they are interesting, high energy cosmic ray particles are also extremely rare: they arrive at a rate of a few per km^2 per century. The low rate makes direct detection (usually by high altitude air balloons) impossible. Fortunately, when one if these particles collides with the atmosphere, it generates a huge cascade (shower) of atmospheric particles,that covers several square kilometers on the Earth's surface.

The objective of the Pierre Auger experiment is to study the properties of ultra-high energy cosmic ray particles by observing the showers. To obtain reasonable statistics at the higher end of the spectrum, the detector has to be huge. Indeed, the Auger detector, built on the Argentinean Pampas, covers 3000 km^2. The detector contains two independent measuring devices, a surface detector (SD) consisting of 1600 water tanks placed on hexagonal grid at a 1.5 km resolution, and a fluorescence detector (FD) consisting of 24 fluorescence telescopes placed at the edges of the detector area, looking inside.

The goal of the statistical data analysis is to estimate the four generating parameters of the cosmic particle (two angular directions, energy, and type of nucleus) based on two independent measurements (surface detectors and fluorescence detectors). The analysis faces huge challenges: the air-cascade is an intrinsically probabilistic process and only partially understood, atmospheric effects are barely known, and real measurements deviate seriously from the simulations. The generative model is inherently hierarchical: the population of showers consists of events, and each events consists of a set of surface detectors; also, the model of individual events involves high-energy physics at the first interaction and low-energy physics in the shower development and in the detector. The most interesting data-science project is to build a full generative model and fit the population of showers using numerical Bayesian techniques. Despite the complexity of the model and the large number of events, building and fitting such a model is within reach.

3.4.2 The JEM EUSO experiment

The goal of the JEM-EUSO experiment is the same as that of the Pierre Auger experiment: to study the properties of ultra-high energy cosmic rays. Similarly to the fluorescence detector in Auger, we will observe the light emitted by the air-cascade in the Earth's atmosphere, but this time from the space. JEM-EUSO will be on orbit on the Japanese Experiment Module (JEM) of the International Space Station (ISS) at the altitude of approximately 400 km. The sensor is a super wide-field telescope that detects high energy particles with energy above 10^{19} eV, an order of magnitude higher than Auger. The observational aperture of the ground area is a circle with 250 km radius which means that the instantaneous aperture of JEM-EUSO is larger than the Pierre Auger Observatory by a factor of 50 to 250. The design of the on-board software faces extraordinary challenges. Its main purpose is triggering and selecting small number of candidate events (vs. background noise) which will be transmitted back to the ground. It has to be fast, computationally simple (strict limit on power consumption), and produce a low false positive rate (strict limit on transmission bandwidth) while missing the fewest possible high energy events. Our goal is to adapt budgeted learning techniques (Section 4.3.4), successfully deployed in real-time object detection applications, for the third-level selection of candidate events.

3.4.3 The ATLAS detector of the Large Hadron Collider (LHC)

The ATLAS and the Compact Muon Solenoid (CMS) experiments recently claimed the discovery of the Higgs boson, acknowledged by the 2013 Nobel prize in physics given to François Englert and Peter Higgs. The existence of the particle was predicted almost 50 years ago to have the role of giving mass to other elementary particles. It is also the final ingredient of the Standard Model of particle physics, ruling subatomic particles and forces. The experiment sits on the Large Hadron Collider (LHC) at CERN. It began operating in 2009, after about 20 years of design and construction, and it will continue operating for at least the next 10 years. The particle discovered is so far consistent with the Higgs boson, however, it has only been seen in three distinct decay channels. Finding it in other channels is a crucial step in proving that it is indeed the predicted Higgs boson.

The discovery of the new particle makes significant use of nonparametric classification techniques (Section 4.1.2) developed in the last two decades. Typically, standard classification algorithms are used for signal/background separation. The classifiers are trained on simulated signal and background events. The raw features are typically obtained in detectors, and standardized/aligned features are extracted "manually" based on background knowledge in particle physics and models of the detector.

The goal of classifier design is to find regions of the feature space where the signal is present or where it is amplified with respect to its average abundance. Once the subregion is found, we claim the discovery of a novel phenomenon (particle) when the number of events in the region is significantly higher than that predicted by the pure background hypothesis. The formal objective function is different from standard classification error, nevertheless, the standard practice is to learn a discriminant function using standard classification methods that minimize the (weighted) classification error, and then determine a classification threshold by maximizing the expected significance. Our goal within this project is to adapt classical learning methods to this new objective function. To involve the machine learning community in the project, we are preparing a data challenge.

3.4.4 The pixel calorimeter of the future International Linear Collider (ILC)

Future lepton colliders with center-of-mass energies of around 1 TeV will play a key role in understanding the origin of electroweak symmetry breaking. This breaking mechanism is intimately coupled to the existence of the Higgs boson or of the mass hierarchy in the fermion sector of the Standard Model of particle physics. The new generation detectors for the lepton collider will include a high-resolution pixel calorimeter to precisely measure the trajectory and the energy of particles produced by the collisions. Technically, a pixel calorimeter will produce 4D data (three spatial dimensions and deposited energy) that will allow us to determine the topology of hadronic showers to unprecedented detail. In today's practice, high level features of the signals produced in the calorimeters are extracted "manually" from raw collision data, and machine learning techniques are only used on the resulting features to separate interesting signals from background events. The goal of our collaboration between the AppStat team and the lepton collider teams (the ILC groups at LAL and LLR) is to investigate the feasibility of deep representation learning techniques (Section 4.3.3) to alleviate and optimize the manual process.

3.5 Analytical chemistry

AUTHOR(S): **Sana Tfaili**, Danielle Libong, Ali Tfayli, Arlette Baillet-Guffroy, Pierre Chaminade

TEAM(S): GCAPS/EA4041

The Group of Analytical Chemistry of Paris-Sud (GCAPS) has a primary objective to promote basic and methodological researches in the field of lipids. Our studies focus on the development of analysis tools and on the data processing in the field of lipidomics in particular. Lipids present a very important molecular diversity; each lipid class is very heterogeneous. In fact, molecular structure and activity are related; it has been shown that the presence, the geometry, and the location of carbon-carbon double bonds in lipids can greatly influence their biological functions. Therefore, lipidome analysis and the necessity to access to the fine structure of lipids arise as a real analytical challenge, in terms of separation, detection and data processing. In this goal, the tools we develop include:

- Separation techniques, mainly chromatographic, with particular attention to the study of stationary phases and detection systems
- Coupled mass spectrometry techniques (LC ,GC, and GC×GC/MS)
- Vibrational spectroscopy (IR, NIR and Raman) and electronic (fluorescence) techniques
- Chemometric techniques for optimization and data processing

Our research covers four themes

1. Cell membrane lipidomics
2. Lipids in skin barrier
3. Lipids: from natural substances to heritage objects
4. Lipid analogues for diagnostic and therapeutic aims

A brief description of the first and second thematic is presented below:

3.5.1 Cell membrane lipidomics

This theme offers methodological developments in the field of separation techniques coupled with mass spectrometry (LC/MS) and of data processing. It began with the study of phospholipids in Leishmania membranes and the evaluation of the impact of treatment by hexadecylphosphocholine (miltefosine) on the membranes lipid composition. In addition to Leishmania donovani lipidome analysis, studies continue today with the analysis of the human erythrocyte (red blood cells) lipidome; and also with analysis of the membrane phospholipids and their impact on the efflux of cholesterol from macrophages in atherosclerosis.

As a first step, a LC/MS profiling is used for the analysis of lipids of interest. Basically, the profile of a sample obtained by LC/MS contains the relative distribution of the species and the molecular ion for each. The objective is to compare the distribution of membrane lipids between different populations (subjects/samples). Then a statistical comparison of the profiles obtained for several samples of each population can be highlighted by characteristic signals (signal over or under expressed in some groups).

In addition, access to databases and further analysis by mass spectrometry, permit to formally identify the compounds of interest. Concerning data analysis, we have to highlight that the evolution of chromatographic techniques requires the use of multivariate statistics and chemometrics to understand and manage the huge quantity of generated data.

The work strategy led to the development of different coupled mass spectrometry techniques for the separation of lipids, but also to the development of different chemometric analysis that can be transposed to several applications. These methodologies of GC or LC/MS profiling and chemometric analysis, allowed us to be partners in many projects; among others, the ANR Omegasomes, a current project directed by Maud Cansell from the University of Bordeaux.

3.5.2 Lipids in skin barrier

The theme of lipid tissue characterization was developed to improve the knowledge of skin barrier functions by understanding the lipid structures in the inter-corneocyte cement of the stratum corneum (SC) at the molecular and supramolecular levels. The main objectives are the development of spectroscopic descriptors of the supramolecular organization for SC lipids by vibrational spectroscopies, the development of spectroscopic descriptors related to physiological or pathological status of the stratum corneum (hydration, elasticity, oxidative stress, atopy), the development of methods for molecular analysis of complex lipid mixtures (lipids of the SC, hydro lipid film).

Molecular and supramolecular characterization of the skin, by defining cutaneous barrier descriptors using vibrational spectroscopies (Raman and infrared), has been validated in vitro and ex vivo on human skin biopsies. The current analytical challenge is to validate previously selected descriptors for direct measurements in vivo. Some problems are encountered in in vivo measurements starting from the quality of spectral measurement, interference removal, data normalization to the calculation of descriptors on a large number of spectra. Hence, the development of algorithms for the treatment of spectroscopic signal revealed to be necessary. Raman spectroscopic signals were related in an innovative way to the mechanical properties and to the hydration status of the skin. This work has been considered by other research teams and is a part of the project ANR-12-003-JSV5 CARE directed by Ali Tfayli.

Spectroscopic data are represented by point-to-point spectra, Z-profile spectra, or hyperspectral images. However, the big quantity of data has to be considered. For this, the development of multiparametric approaches, of algorithms and the use of multivariate statistical analysis were required to conclude on the data.

The nature of information provided by chromatographic and spectroscopic techniques are different, the treated subjects and themes too. The first (chromatographic) is destructive and provides structural information, the sample has to be in solution; and the second (spectroscopic) is non-destructive and gives information on the systems organization and their environment. Both analytical techniques are complementary and are used herein for lipidomics studies. Another important feature is that vibrational spectroscopic (Raman and IR) data are multidimensional in space (x,y,z) and each point of such 3D matrix contains an intensity versus wavelength spectrum. For the LC/MS the current trend is to associate orthogonal separation techniques before the sample enters the MS interface. The mass spectrometer itself is able to perform simultaneously MS, MS2 and MS3 fragmentation of ions. High dimensionalities are then also encountered with this technique. To date, statistical approaches were the key to manage data and to build results for each analytical tool and subject. However, the quantity of data will continue to grow faster, making some latent, potential information not really extracted from the registered data.

3.6 Text and Music

AUTHOR(S): **Guillaume Wisniewski**, Claire Nédellec, Michalis Vazirgiannis, Sophie Schbath, Hélène Papadopoulos, Matthieu Kowalski, François Jankowiak

TEAM(S): TLP/LIMSI, BIBLIOME/MIG, DASCIM/LIX, SIGNAUX/L2S, AAO/LTCI,DSR/UPSUD

Scientific knowledge can be extracted from different kinds of “big” data such as images (obtained, for instance, by functional magnetic resonance or by a telescope), real values captured by sensors or textual data such as medical records, scientific literature, web pages, patents or even tweets. Exploiting textual data raises specific challenges as it is expressed in natural language, the ambiguity of which is well-known, and it often do not present any kind of formal structure. However it contains valuable information that can, for instance, be used to detect influenza epidemics by mining search engine queries or identify the rise and fall of scientific fields thanks to the analysis of the digital libraries. Automatic information extraction from medical records is another striking example of the importance of textual data in scientific discoveries. For instance, a study has recently shown that comorbidity information[3] can be automatically determined from the analysis of electronic medical records. Researchers should therefore be equipped by tools that assist with the selection, the extraction and the formalization of the relevant information from texts, so that it can be then used in addition to other knowledge sources.

Huge volumes of textual data can be found in all scientific domains, often in documents of different qualities: curated documents (e.g. articles, patents, books), as well as grey literature (e.g. reports, tweets, web pages) or even book scans that require special treatment (like OCR recognition) before they can be automatically processed. For instance the archives of London's Old Bailey, made of 197,745 criminal trials held at London's central criminal court between 1674 and 1913, contains the largest body of texts (over 10 millions words) detailing the lives of non-elite people ever published and is a source of great interest for historians, linguistics and other researchers in Humanities. This new kind of resources and its exploitation with automatic methods, has lead to the development of a new branch of History called computational history. In the context of the University Paris Saclay, a similar initiative has been undertaken by the *Droit et Sociétés Religieuses* team of the Faculté Jean Monnet which, in the past five years, has developed the Gregorius on-line international bibliography for history of canon law and Roman Church institutions, covering a period from the early Christianity to the second Vatican Council (1962-1965). This database, created in 2007 and for which was elaborated a 3,000 thematic keywords thesaurus, includes over 2,200 detailed cards and has proven its value to the community.

As for experimental data the useful information must first be automatically gleaned from the textual source and represented in a form suitable for its analysis. Depending on the objectives (e.g. the size of the collection, the availability of external resources), this representation can be extracted, with **Information Extraction** or more general **Text Mining** techniques (Section 4.2.6), using only surface information or can rely, through the use of **Natural Language Processing** methods (Section 4.2.6), on a full syntactic and semantic analysis of the document that is required to compute high quality interpretation.

Both Text Mining and Natural Language Processing fields make intensive use of Machine Learning and Data Analysis methods. These methods are applied to the text representation for highlighting useful regularities among documents or words using unsupervised classification methods (Section 4.1.3) or for making predictions of new knowledge in unseen documents (e.g. predicting relations among entities) using supervised classification method (Section 4.1.1). As many linguistic information can be represented as sequences, trees or graphs, NLP is also a testbed for many structured prediction methods (Section 4.2.3). Due to the ambiguity of the natural language, complex pipelines with several interleaved Machine Learning and Natural Language Processing steps may be required for high quality data.

In the rest of this section, we describe two real-world examples of how text analysis methods can be help scientific discoveries.

3.6.1 Cochrane Systematic Reviews

Each day several hundreds scientific papers are published in human health related fields: PubMed, the main bibliographic database in the life science, medical and biomedical research, now contains more than 23 millions references and around one new paper is added every minute. Keeping track of the more recent results and and making sense of the large volumes of frequently conflicting data derived from primary studies is a daunting task for any researcher, not to mention practitioners or the general public.

The Cochrane Reviews provide systematic reviews of primary research in human health care and health policy, and are internationally recognized as the highest standard in evidence-based health care: given a clearly formulated question, such as *Can antibiotics help in alleviating the symptoms of a sore throat?*, the review collates and summarizes all published papers on that topic to establish whether or not there are conclusive evidences about a specific treatment. More generally, the Cochrane collaboration aims

[3] *Comorbidity* denotes the simultaneous presence of two chronic diseases or conditions in a patient

at helping people make well- informed decisions about health care by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions. Today more than 10,000 people from more than 80 countries contribute to the reviews and translate them into 5 languages.

Several problems studied by the NLP community, such as question answering, machine translation, automatic simplification or summarization, sentiment analysis or terminology extraction can be used to make their work easier. LIMSI has recently started a collaboration with the Cochrane Institute to provide them with such tools.

3.6.2 Gene regulation network

The study of gene regulation networks is a key step in the understanding of the cell mechanisms and then of the whole organism (Section 3.1). A large subset of the information is not described in structured databases but only in millions of scientific papers. Since the middle of the two thousand years, the automatic extraction of gene and protein interactions for the design of regulation networks has been the main challenge of Information Extraction in biology. The combination of NLP and ML methods achieve now high scores in the recognition of biological entities and their interaction relations as measured in the recent international competitions BioCreative and BioNLP, (the '13 series co-organized by LIMSI and MIG-INRA). Many on-line tools based on these technologies are now available and used by biologists together with other bioinformatics tools , (see for instance CoCitation on gene- protein interactions of the **Bacillus subtilis** model bacteria, integrated with the genetic information of the IGO platform).

3.6.3 Music Information Retrieval

Within the last few years, the huge explosion of online audio music collections has become a great source of attention in the music industry. The availability of millions of tracks on the Web has posed a major challenge in terms of searching, retrieving, and organizing music content. Techniques for interacting with those enormous digital music libraries at the song level are necessary. Content-based Music Information Retrieval (MIR) has thus become a very active field of research that opens a large number of perspectives for music industry and related multimedia commercial activities. The increasing number of projects funded by the European Community that involve MIR aspects, such as Semantic HIFI, QUAERO or 3DLife, shows that MIR is a key research area for the European scientific development. Content-based music information retrieval deals with the extraction and processing of meaningful information from musical audio. Many applications based on content-based indexing and retrieval have emerged, such as cover song detection or disc jockey (DJ) mixing. Most of these applications are based on the use of musical descriptors that are extracted from the audio signal, such as the key, the chord progression, the melody or the instrumentation.

The proliferation of the emerging MIR and Music Digital Library techniques and technologies demands the creation of the necessary resources for their development (in order to derive knowledge directly from the data or to train systems) and benchmarking/evaluation. Indeed, in speech processing, the numerous projects of database collection (e.g. the EU funded projects Speech Dat, Speech Dat2, Speech Dat car, Speecon), and the diffusion of huge multilingual databases by institutions responsible for providing annotated corpora (such as ELRA or LDC), have favored a strong increase in the performances of the developed technologies. The MIR field would strongly benefit from such efforts too, as it is illustrated by the numerous discussions on the "Music Information Retrieval Evaluation eXchange" (MIREX) wiki. Recent work have shown the necessity of establishing and supplying common evaluation databases and metrics for content-based estimation, using methodologies that will ensure sustainability, usability, and sharing of the corpora. It is thus of primary importance to supply annotated audio corpora, and define evaluation metrics and criteria to bring new evaluation methods to the MIR community.

The collection and creation of audio music annotated data require a deep synergy between the process of producing the data and their practical use. On the one hand, the data of interest are very complex and the annotation process requires a precise methodology (collecting a set of audio items, creating and attaching related annotations, documenting and storing the results to ensure sustainability and sharing) according to the users needs. On the other hand, the creation of such resources is a key step in the understanding of the various music attributes, as theoretical but also perceptual attributes. Direct exchanges between the producers of audio data and meta-data and the users of such data would help making progress in various aspects of music processing such as music information retrieval, computational musicology or cultural music issues.

3.7 Environment, atmosphere, oceanology

AUTHOR(S): **Laurent Barthès**, Edwige Vannier, Cécile Mallet, Sylvie Thiria, Yvon Lemaitre
TEAM(S): LATMOS/SPACE, LOCEAN/MMSA

The objective of the theme "Modélisation Mathématiques et Statistique de l'Environnement" (MMSE) is the analysis and modeling of environmental data from the observation of natural environments. The natural environment observed through in-situ or remote sensing sensors presents extreme spatial and temporal variability in a wide range of scale. The data available to model this complex environment are heterogeneous and characterized by the presence of instrumental noise and limited resolution and repeatability. Current research is facing problems such as extracting information from observations or reproducing an observable from a mathematical or statistical description on a wide range of spatial and temporal scales. These tasks are an important part of the work of researchers for development of new methods. The major difficulty lies in the fact that the proposed methods must provide accurate solutions to practical problems, i.e. robustness with respect to the various sources of error. Moreover, in general, the variability of the geophysical processes is governed by nonlinear dynamic equations covering a wide range of scales ranging in some cases from the global scale to the millimetric scale. Thus, in the field of numerical modeling, the range of scale of the processes is often too large to be explicitly represented due computational and memory costs. Research faces Many challenges including the estimation of transfer functions, inversion of satellite data, classification, pattern recognition, prediction, data assimilation, statistical downscaling, extreme modeling, data fusion, multi-scale estimation using variational techniques or Markov chains

3.7.1 Oceanology

An example theme in oceanography is the assimilation of sea surface data (temperature, altimetry) in oceanic Global Circulation Models in order to improve their accuracy. Another theme is the estimation of the content of phytoplankton in seawater or the reconstruction of the 3D ocean constitution by the mean of observation satellites. Indeed, the signal received by the satellites is often degraded due to the presence of clouds and aerosols in the atmosphere. Robust methods must be implemented to counteract degradation due to atmosphere.

3.7.2 Atmosphere

Concerning the atmosphere, we focus on precipitation which are extremely heterogeneous and variable processes whatever the considered spatial or temporal scale. A realistic description using stochastic models based on scale invariance assumptions are used. The objective is to better represent the interrelationships in the spatial and temporal domain and to improve methodologies for rain maps retrieval, rain events simulation; "downscaling" or modeling IDSF (Intensity, Duration, Size, Frequency) curves relationships. From a general point of view, the identification of the effect of dynamic processes on rainfall behavior requires the development of multiscale tools based on multidimensional analysis. An important field of research concerns the determination of the weather types or the circulation types that allows taking into account the phenomena with the large scale and the meso scale.

A new device dedicated to the observation of the rain at small scale is currently being tested. The measurement of the attenuation of an electromagnetic wave transmitted from geostationary satellites is used to estimate rain rate along the path link. Assimilation technics using rain cells advection scheme are then used to estimate rain maps.

3.7.3 Soil

Because the soil surface occurs at the boundary between the atmosphere and the pedosphere, it plays an important role for geomorphologic processes. Soil irregularities at small scale, such as aggregates, clods and interrill depressions, influence water outflow and infiltration rate. They have also an influence on remote sensing studies, by producing scattering and shadowing effects. In order to link the remote sensing observations to scattering physical models as well as for modelling purpose, key features of the soil microtopography should be characterized. However, this characterization is not fully understood and some dispersion of roughness parameters can be observed in the same field according to the methodology used. The proposed approaches are detecting (by segmentation methods) and characterizing (statistically) some of the soil surface irregularities that are clods and big aggregates.

3.7.4 Hydrology

In the field of hydrology, we have improved the operational models through the determination of a machine learning procedure for determining, for each watershed its internal parameters. This allows improving flood forecasting by controlling the initial conditions of water levels in the basement.

3.7.5 Fluids Mechanics, Heat and Mass Transfer

AUTHOR(S): **Patrick le Quéré**, Christian Tenaud, Caroline Nore, Yann Fraigneau, Nicolas Férey
TEAM(S): LIMSI

Fluids Mechanics, Heat Transfer and Mass Transfer are key scientific disciplines at the heart of many crucial societal challenges in the domain of energy, transportation, and environment. Indeed, achieving more efficient, more reliable, more environment friendly means of converting or using energy, of transporting people and goods, requires a better identification of the corresponding technological bottlenecks and in turn a deeper knowledge of the involved physical mechanisms in all their intrinsic complexity and mutual interactions. It also requires a continuous progress in numerical modelling and simulation capabilities that are instrumental to mastering and optimizing the technological processes and that stand at the heart of a progressive substitution of empirical know-how by a deterministic approach in the conception and design processes. Fluid mechanics has profoundly evolved over the last decades through the increasing availability of techniques or tools, either numerical or experimental, allowing for a deeper understanding of its unsteady characteristics, and by the development of tools aiming at mastering this unsteadiness, either through manipulation or control, in order to achieve predefined objectives. Data set in fluid mechanics are designed to supply relevant information to characterize a flow. They relate to a very large number of quantities whose the relevance mainly depends on the nature of the fluid (inviscid or not, compressible or not....), the nature of the flow (forced convective flows, thermo-convective flows...) and the state of the flow (laminar, unsteady, turbulent). These quantities can be of scalar type (mass, pressure, temperature, viscosity...), vector type (velocity, vorticity...), or more rarely tensor type (Reynolds stress tensor). A data set can have different dimensional structures depend on the aim of study and approaches used (specific experimental techniques, numerical simulations). Commonly, the data layout can be mono dimensional (i.e. time series provided by probes), 2D (cut plans) or 3D (data acquisition on a part or on the whole domain of study). 3D data set are generally associated to a spatial representation of the flow. For unsteady flows, the temporal aspect can be related to a group of data set where each of them is associated to a snapshot of the flow. With the increasing progress of experimental techniques and computers used in high performance computing, the volume of data is more and more large, especially about 3D unsteady flows. This leads to two specific issues : the compression of data to limit the storage effort and the capability to exploit data set, especially using Visualisation (section 4.4.1) and analysis.

3.8 Economics, finance and insurance, social sciences and networks

AUTHOR(S): **Arnak Dalalyan**, Xavier d'Haultfoeuille, Jean-Michel Zakoian, Jose de Sousa, Ioana Manolescu, Stephane Gaiffas, Bogdan Cautis

TEAM(S): LMI/ENSAE, LFA/ENSAE, RITM/UPSUD, MEV/CMAP, LAHDAK-OAK/LRI, CMLA/CACHAN, MAS/CENTRALE

Statistical and machine leaning tools are used extensively in almost all the fields of economic sciences ranging from micro-econometrics and financial economics to labor economics and the economics of education. Empirical investigation is systematically conducted for performing a broad variety of tasks such as building and testing economic models, helping decision-making, analyzing the effects of monetary and fiscal policies, characterizing the behavior of economic agents, predicting economic indicators, assessing the risks, etc.

Unlike in physical sciences, controlled experiments are uncommon in economics since the latter often studies the behavior of agents over periods that are too long to allow to keep some parameters fixed. Time series (Section 4.2.1), cross-sectional data and, more generally, multidimensional panel data are the most frequently used types of datasets in economics. The term panel data—a.k.a. longitudinal data in biostatistics—refers to data containing observations of multiple characteristics over multiple time periods of one or several entities (people, firms, countries). Analyzing this type of data requires statistical techniques taking into consideration the time-dependent nature of the observations and their inherent heterogeneity. In particular, recent advances in nonparametric (Sections 4.1.2 and 4.1.3) and high-dimensional statistics (Section 4.3.5) are constantly exploited in econometrics to produce models with increased flexibility and better predictive power.

It is also important to mention that economics is not only a consumer of data science but also an important data producer. Data are usually collected by means of various types of surveys. One specificity of these data is their confidentiality. There are fortunately now very specific equipments, in particular within the Centre d'accès sécurisé distant aux données (Section 2.5), that allow the researchers from different institutions to access these data for scientific purposes within the data security laws and guidelines.

We describe below several very different concrete projects of economics and social sciences in which statistics and machine learning play an important role.

3.8.1 Young Lives project

The Young Lives (YL) project is a long-term study of childhood poverty being carried out in Ethiopia, India (in the state of Andhra Pradesh), Peru, and Vietnam. The broad objective of the YL project is to improve understanding of the causes and consequences of childhood poverty and to examine how policies affect children's well-being. Extensive child, household and community level questionnaires are administered to capture information on various aspects of the child's life including household demographics, care-giver background, child health (both physical and mental), economic shocks, household consumption, as well as social, economic and environmental context of each community. The YL survey involves tracking 12,000 children (two cohorts) growing up in the four developing countries over 15 years. For example, one can currently use information from several rounds of data collection for Andhra Pradesh, India. In Round 1, 2000 children aged around one (the "younger" cohort) and 1000 children aged around eight (the "older" cohort) were surveyed in 2002. Following up, Round 2 involved tracking the same children and surveying them in 2006 at age five and twelve, respectively. Data was collected through household questionnaires, child questionnaires, and a community questionnaire.

The two cohorts allow the researchers to investigate two distinct periods of childhood. In particular, the data on older children make it possible to specifically analyse the dynamics of cognitive and non-cognitive skill formation and influences exerted by the child's immediate environment. Given that cognitive and non-cognitive skills and parental input are unobserved, one has to treat them as latent variables. Statistical methods used in such kind of investigation include nonparametric (conditional) density estimation, latent variable models, parametric and nonparametric auto-regression, etc.

3.8.2 Compustat database

Published by Standard and Poor's, Compustat is a database of accounting information about a large number of companies throughout the world. It covers the Income Statement, Cash Flow Statement, and Balance Sheet, and contains company data going back 40 to 50 years on over 65000 securities. It also provides historical information on companies that no longer exist because of merger or bankruptcy, known as inactive or "research" companies.

Bankruptcy prediction is a typical problem tackled in the financial economics literature using data science methodology applied to this dataset. The goal is to construct a model that takes as input the historical values of various factors (total asset, inventories, total dividends, earnings before interest and taxes, net income, etc) of a given company over a time period $[t - \Delta_1, t]$ and outputs 1 if the company is likely to undergo bankruptcy in the time period $[t, t + \Delta_2]$ and 0 otherwise. This is a typical problem of binary classification, see Section 4.1.2, and is usually addressed in the financial economics literature using off-the-shelf classifiers (logistic regression, decision trees, neural networks, support vector machines) with a few tens of factors. Applying more advanced algorithms to this problem may allow us to use a significantly larger number of factors and, hopefully, to improve the prediction accuracy. In addition, model and variable selection algorithms may produce predictive models with increased interpretability and to put forward the factors that are most strongly related to the bankruptcy of a company.

3.8.3 Finance and insurance

The last twenty years have witnessed a considerable increase of the number of available data for financial and insurance applications. In particular, the development of electronic markets has favored the collection, storage and modeling of observations that are collected at a much finer time scale than the day. Such high-frequency data possess characteristics that pose interesting challenges to the econometric modeling: the huge number of observations, the fact that such data are often recorded with error, the fact that they are often irregularly spaced over time, and the presence of intra-day periodicities.

One challenge is to develop an econometric approach able to combine information stemming from low and high frequencies. Several classes of volatility models (GARCH, stochastic volatility, etc.) designed for low frequency data, daily say, have been proposed and extensively studied in recent years. On the other hand, the classical analysis of high frequency financial data is achieved by means of continuous time diffusion models, in which observed trading prices correspond to a discretization of the latent process with a small time unit. While this kind of approach may be convenient from a mathematical point of view, it is far from the reality of financial data. For instance, the closure of the markets between consecutive days is not taken into account. It is therefore necessary to introduce new modeling approaches including both the intraday dynamics of asset prices and the dynamics of closure prices from one day to the next one.

Another challenge is to study the so-called granularity effects. The large size of portfolios, which can include several thousands of contracts, make difficult the risk analysis of credits or life insurance contracts. The granularity principle has been introduced in the Basel II regulation for credit risk to solve this difficulty when computing the reserves. Asymptotic expansions with respect to the number of assets should allow to study the behavior of conditional risk measures in this framework.

3.8.4 Smartphones, Social Network Sites (SNS) and collection of personal data

Consumers have so far little information about the amount and the nature of their personal data which is transmitted by smartphones to various data aggregators. Do consumers change their behavior as they learn about the collection of their personal data? Moreover SNS may shape the decisions of their participants and in particular their choices about privacy. The rising of the SNS give access to many newly available data that may help addressing these questions: do consumers ”imitate” the network or do they take decisions on their own?

3.8.5 High-Speed-Rail networks and spatial disparities across cities

The mobility and dynamics of local population are some traditional topics of interest in social sciences. In particular, demographers have emphasized the migrations from the countryside to large cities in France after the Second World War. Consecutive censuses have been used to get some information on general migration patterns which showed the growth of large cities and the desertification of countryside. Since then, the interest has shifted to the analysis of specific groups (such as socio-professional categories) and specific mobility reasons (such as housing, job and family-related motives). New types of mobility and residential arrangements are currently at the center of the debate. These include long-distance commuting and living in two distant dwellings located on the places of work and residence. These practices often involve transport between two large cities, in which the high-speed train has a major role to play. The increase of a city activity related to a better connection to the transportation network might be enough to trigger growth as new job opportunities may attract workers. However, it may also cause an increase in land prices or give incentives to firms and population to delocate. In this research project, we plan to assess the benefits and drawbacks of the development of the High-Speed Rail network for French cities, thanks to the collection of a large-scale dataset on TGV travel time-tables and train frequencies since 1981.

3.8.6 Discovering and exploiting user profiles

As users interact with content (or data) management systems, their actions and profiles can be exploited to develop useful applications. For information access - search or recommendation - preference profiles help better personalize content provided to users as a result of a search query or as a recommendation. For intelligent (also called “expert”) crowdsourcing, where highly specialized expertise is being called upon, profiles help better assign task to users. Data-centric applications stand to benefit from a learning process that “closes the loop”, continuously accounting for user feedback, actions, evaluations and interactions, in order to better analyse and extract data, index it and address users’ information needs. Therefore, preferences and expertise need to be discovered over time via interactions with users, using a principled learning approach. For instance, the ALICIA ANR Contint project (2014-2017) aims at investigating adaptive learning algorithms for online, highly dynamic, user-centric environments, such as Multi-Armed Bandits algorithms (Section 4.4.3).

3.8.7 Social, structured, and semantic search

Interactions between users, between users and the content they produce (author), and the relationships between related pieces of electronic content provide valuable hints for the exploitation of these content fragments. In particular, we consider users creating structured content in a social context, a general setting which captures and generalizes popular applications, such as Web blogs and comments, micro-blogging (e.g., Twitter), social network applications, etc. The content produced collectively by the users can be then analyzed to derive relationships between them, and to help answer each user's questions in a way that most accurately reflects their interest, semantic profile, and social connections in the network.

3.9 Engineering sciences, "man-made data"

AUTHOR(S): **Cécile Germain, Kaouthar Benameur**

TEAM(S): TAO/LRI, LTCI, CEA Tech/LIST, DTIM/ONERA, LAHDAK/LRI, BIOINFO/LRI

Various aspects of data science are extensively used in a number of problems within the engineering sciences. We describe below two projects: analyzing data on the behavior of the EGI (European Grid Infrastructure) grid and global processing of information recorded by large sensor networks.

3.9.1 Globalized Computing Systems

Globalized computing (data centers, grids and clouds) provides new examples of complex systems with emergent collective behavior. Understanding, optimizing and designing these systems require models of their dynamics that cannot be built a priori, but must be inferred from behavioral data. Accordingly, research in distributed systems is now enthusiastically catching up with data science through applications of methods from modern optimization, game theory, machine learning and statistical time series. However, the gap between research - e.g. Autonomic Computing - and engineering practice remains significant. The general challenge is to *demonstrate* the quality of the scientific ideas in a high-dimensional decision landscape: policy design (e.g. scheduling, green optimization, file location) explores a very complicated space, with multiple feedback loops and significant externalities. Moreover, even the more constrained settings are multi-objective optimization problemsdifficulties, two operational issues contribute to challenge the researcher. First, real world experimentation is hardly possible. Second, significant experiments with simulators require large data sets and manpower. In practice, comparative evaluations are rare, and experiments on high level concepts such as autonomic programming models are extremely difficult to conduct. An alternative to experimenting on real, large, and complex data is to look for well-founded and parsimonious representations, with the unavoidable approximations implied. Of course, this is routine for many other sectors of engineering science, but they have since long built the necessary physical and intellectual tools and culture to do so: simulation, and more generally *in silico* experiments as the third pillar of science.

A first issue is data availability. The Grid Observatory initiative, which we lead, automatically collects, organizes and publishes on its portal the monitoring traces of the flagship European Grid Infrastructure EGI and of the University cloud. The Grid Observatory is a unique facility in that it provides data about e-science practice at real scale. This project has been or is currently supported by EGI, Digiteo, INRIA (Action Développement Technologique), CNRS (PEPS program), and University of Paris Sud (MRM program).

Three other fundamental issues—concerning interpretation—can be identified: intelligibility, non-stationarity and validation.

Globalized systems, like social networks, can be studied as complex graphs by the empirical description of their spatial and temporal properties. *Intelligibility* is the need to go further, by exhibiting the causal structure of the observations, in a situation in which exact prior knowledge is unreachable (as Lamport formulated decades ago for fault management). Two approaches are especially promising for tackling this issue. The first one considers the traces for what they actually are, namely texts, opening their analysis to text mining techniques, for instance latent Dirichlet allocation models. The benefits of the approach are twofold: some level of explanation revealed by the data and a relatively scalable strategy for capturing the dynamics, while retaining the full dimensionality of the problem. The second approach exploits the massive redundancy of monitoring to elucidate the underlying structure trough aggregative methods, either direct clustering or Collaborative Prediction. Both approaches have demonstrated their effectiveness even with oversimplifying assumptions.

Non-stationarity seriously complicates both model selection and policy design. In a nutshell, there is no solid reason to assume stationarity for globalized systems: the underlying causes, for example users' activity, middleware systems or usage, evolve over time beyond trends or seasonality and exhibit ruptures. A general framework for efficient off-line model selection must include non-stationarity, with the considerable supplementary complexity associated , and on-line policies must discover and adjust efficiently to breakpoints. For each of the above-mentioned approaches (traces as texts and aggregation), specific methods do exist. The challenge is to integrate them with non-stationarity in a conceptual framework that would be both principled and usable, with scalability as a major requirement. We expect significant interactions on this issue with the *Finance and Insurance* and *User profiles* actions in Section 3.8, as well as with the methodological themes in Sections 4.2.1 and 4.1.4.

Finally, *validating* the models is both essential to their transfer to exploitation, and difficult because no reference interpretation is available. The CDS collaboration offers the exceptional opportunity to confront the theoretical arguments with benchmarking on representative data and operational evaluation from practitioners (system managers).

3.9.2 Sensor networks

Sensor networks have been deployed for a number of monitoring and control applications, such as target tracking, environmental monitoring, manufacturing logistics, geographic routing, and precision agriculture. For many target tracking applications such as surveillance, anomaly detection, species distribution, the main purpose of the sensor network is to locate and track changes in remote environments. For example, for surveillance applications, the sensors must be able to locate where the intruders or the vehicles are moving in the network. As another example, in secure protocol and network routing it is critical to track anomalies such as denial of service attacks in the network.

In many of these applications, the phenomena of interest are global in the sense that they are not discernible at the level of individual sensors or nodes, and require corroborative input from many sensors, that is, events only become observable if sensed data of many nodes support them. Indeed, this issue of making global inferences from local data is characteristic of many distributed systems and is more known under the terminology of data fusion.

Depending on the applications, mainly surveillance and situation assessment, sensor networks are set with different kind of sensors, homogeneous, heterogeneous, passive or active. Recently, tiny, inexpensive sensor devices that can measure and observe limited and various environmental parameters, often in hazardous or humanly inaccessible places, thereby allowing real-time and fine-grained monitoring of physical spaces around us, start to compose large-scale networks. Many potential applications of this technology relate to surveillance or environmental monitoring spanning wide-spread geographical areas. Since a centralized data processing approach, where a central processor continuously collects signals from all the nodes and learns the state of the network, does not scale well with the size and the complexity of large scale sensor networks, more efficient solutions are of major interest. In a nutshell, one promising approach consists in building upon new data acquisition formalism, in which compression plays a fundamental role. From a signal processing standpoint, one can think about a procedure in which signal compression is carried out at different nodes, thereby reducing the amount of required observations and giving rise to tracklets or artificial measurements.

A Compressive Sensing Kalman filter (CSKF), as an approximation to Bayesian Compressive Sensing scheme, has been already introduced in the literature with the advantage of providing sequential statistics. Based on these developments, we consider that there are two major issues which should be addressed so to consolidate techniques for data association and tracking with compressed data. The first one concerns the definition of the artificial measurement approximation: how to define tracklets, what to consider in their definition communication constraints, geometry, environment and context, mission priorities, etc. The second one concerns the reduced order recovery of states: how to deal with data association in a sequential approach with sparse data. How to generate (or stitch) tracks to ensure continuity and complete recovery.

4 Scientific themes II: data science in computer science and mathematics

This section gives a non-exhaustive panorama of data science themes an expertises of the participating teams. The summary is grouped into four subsections. We start with an introductory section on fundamental data analysis methodologies (Section 4.1), followed by three sections around three major data science challenges: data complexity (Section 4.2), resource limitations (Section 4.3), and interactive visualization and experimental design (Section 4.4).

4.1 Fundamental data analysis methodologies

In this section we describe some of the basic data analysis approaches and tools. These methodologies form the core of most of data science and serve as an introduction to more advanced and specialized topics developed in subsequent sections.

4.1.1 Classical statistics: parametric probabilistic models, maximum likelihood and Bayesian inference

AUTHOR(S): **Arnak Dalalyan**

TEAM(S): LMI/ENSAE-CREST, STI/LTCI, APPSTAT/LAL, LMO

Classical parametric statistics is concerned with the theory of inferring the unknown features/parameters of probabilistic models describing stochastic phenomena. The main distinctive characteristic of the parametric statistic is that the observations $X_1, \dots, X_n$ are assumed to be drawn from a probability distribution that is parameterized by a parameter $\theta \in \mathbb{R}^d$, with a dimension d much smaller than the sample size n. It is well understood now that under some regularity assumptions the maximum likelihood estimator and the Bayesian posterior mean are asymptotically efficient in the sense that their risk measured by the quadratic loss is asymptotically equivalent to $(nI(\theta))^{-1}$, where $I(\theta)$ stands for the Fisher information. It has also been recognized that in non-regular models Bayesian posterior mean is a better estimator of the unknown parameter than the maximum likelihood. Analogous results have been established for the problem of hypotheses testing as well. Robustness of statistical procedures to the outliers is another important issue that received a lot of attention in statistical literature on parametric models. Although the parametricness assumption $d \ll n$ is somewhat restrictive, the ideas and tools developed in parametric statistics are often used as building blocks for designing statistical procedures in nonparametric statistics and machine learning.

4.1.2 Supervised learning: nonparametric multivariate classification and regression

AUTHOR(S): **Balázs Kégl**

TEAM(S): TAO/LRI, AppStat/LAL, LMO, STA/LTCI, AROBAS/IBISC, AIS/CACHAN, LS/ENSAE, L2S/SUPÉLEC, CMAP/POLYTECHNIQUE, LIX/POLYTECHNIQUE, CEA Tech/LIST, STATISTIQUE/CENTRALE, CMLA/CACHAN

Multivariate regression and classification is one of the best-studied problem in machine learning, with a plethora of well-tested and well-performing algorithms. The goal is to infer a function $g : \mathcal{X} \rightarrow \mathcal{Y}$ from a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ comprised of pairs of *observations* $\mathbf{x}_i$ and *labels* y_i. The setup is often used when labels have to be inferred from observations without a parametric model

(Section 4.1.1) $p(y|\mathbf{x})$ or when such a model exists but Bayesian or maximum likelihood inference is computationally infeasible. There are several families methods are available to solve a classical classification (the label comes from a finite set and the goal is to minimize the classification error) or regression (the label is a real number and the goal is to minimize the squared error) problem: support vector machines, neural networks, AdaBoost and random forests, or Gaussian processes, just to mention a few. The best method can depend on several factors, for example, the size n of the data set, the dimensionality of the input space and the sparsity of the solution (Section 4.3.5), whether the label space has a complex structure (Section 4.2.3), whether the input features have deep hierarchical interactions (Section 4.3.3), whether the final model has computational or other resource limitations (Section 4.3.4), whether the goal is to rank a set of instances rather than simply classifying them (Section 4.2.4). Supervised learning and optimization is connected through several links: convex optimization (Section 4.3.1) is one of the main engines behind some of the learning techniques, stochastic optimization (Section 4.3.2) and experimental design (Section 4.4.2) are important tool in practical model selection (Section 4.1.5) and hyperparameter optimization, and online learning (Section 4.4.3).

4.1.3 Unsupervised learning

AUTHOR(S): **Christophe Giraud**

TEAM(S): TAO/LRI, LMO, LS/ENSAE, CEA Tech/LIST, AROBAS/IBISC

The goal of unsupervised learning is to infer hidden classes or structures from unlabeled data. Unsupervised learning is much more tricky than supervised learning since the nature of the classes and even their cardinality are unknown. This situation arises in many different fields where data clusters different but unknown groups: co-expressed genes in transcriptomics, communities in social networks, segmentation in imaging or geology, etc.

There are mainly two different approaches for unsupervised learning. The first approach is geometric in the sense that it does not relies on any probabilistic model. The groups are then determined from the data according to some geometric criteria. This approach includes some popular clustering algorithms like k-means, spectral clustering or hierarchical clustering. The second approach is model based, in the sense that it relies on a probabilistic modeling of the data. The data is modeled by a mixture of simple statistical models, each of them representing a group. The clustering problem is then recast in an inference setting. Among the popular models we mention the mixture of Gaussian distributions, the Hidden-Markov Models and also the Bayesian Hierarchical Models.

The main statistical issues in unsupervised learning are choosing the modeling and the clustering procedure and selecting the number of clusters and evaluating the clustering results. These issues can be handled in the model-based setting by model selection and estimator selection theory. Implementing large scale clustering requires to develop some efficient optimization algorithms. This issue is crucial in high-dimensional settings where the classical algorithms suffer from a high computational cost.

4.1.4 Outlier and novelty detection

AUTHOR(S): **Alexandre Gramfort**

TEAM(S): LTCI, LS/ENSAE, CEA Tech/LIST

Many applications require being able to decide whether a new observation belongs to the same distribution as existing observations (it is an inlier), or should be considered as different (it is an outlier). Often, this ability is used to clean real data sets. Two important distinctions must be made:

1. novelty detection: The training data are not polluted by outliers, and we are interested in detecting anomalies in new observations.

2. outlier detection: The training data contain outliers, and we need to fit the central mode of the training data, ignoring the deviant observations.

Novelty detection Consider a data set of n observations from the same distribution described by p variables. Consider now that we add one more observation to that data set. Is the new observation so different from the others that we can doubt it is regular? (i.e. does it come from the same distribution?) Or on the contrary, is it so similar to the others that we cannot distinguish it from the original observations? This is the question addressed by the novelty detection tools and methods. In general, it is about to learn a rough, close frontier delimiting the contour of the initial observations distribution, plotted in embedding p-dimensional space. Then, if further observations lay within the frontier-delimited subspace, they are considered as coming from the same population as the initial observations. Otherwise, if they lay outside the frontier, we can say that they are abnormal with a given confidence in our assessment. The One-Class SVM is a popular method used for that purpose. It requires the choice of a kernel and a scalar parameter to define a frontier. The RBF/Gaussian kernel is usually chosen.

Outlier detection Outlier detection is similar to novelty detection in the sense that the goal is to separate a core of regular observations from some polluting ones, called "outlier". Yet, in the case of outlier detection, we don't have a clean data set representing the population of regular observations that can be used to train any tool. One common way of performing outlier detection is to assume that the regular data come from a known distribution (e.g. data are Gaussian distributed). From this assumption, we generally try to define the "shape" of the data, and can define outlying observations as observations which stand far enough from the fit shape. A strategy is to fit a robust covariance estimate to the data, and thus fit an ellipse to the central data points, ignoring points outside the central mode. For instance, assuming that the inlier data are Gaussian distributed, it will estimate the inlier location and covariance in a robust way (i.e. without being influenced by outliers). The Mahalanobis distance obtained from this estimate is used to derive a measure of outlyingness.

4.1.5 Model and estimator selection

AUTHOR(S): **Christophe Giraud**
TEAM(S): LMO

Model selection and estimator selection are central issues for analyzing complex data. When the data are the outcome of a complex system, there is no simple model describing the data. The data analyst then considers different candidate models (more or less complex) and aims to select the best model from the data. Roughly speaking, the best model is the one achieving the best trade-off between bias and variance. Actually, if a model is too rough, it will not be able to capture the main features of the data. Conversely, estimation in a very complex model will be very instable and will provide poor predictions. This last issue is especially important for high-dimensional data, where estimation is hopeless without identifying the main underlying structures of the data. A good model must then simultaneously be quite simple and reflects at best the unknown structures of the data. The goal of model selection is to select from the data (almost) the best model among the candidate ones.

Estimator selection is also an unavoidable step when analyzing data. The data analyst has at disposal a plethora of estimators. No estimator is universally better than the others, so a selection step is necessary. Furthermore, most of the statistical procedures have one or several tuning parameters and their preferences depend heavily on the choice of these parameters. Estimator selection aims to select from the data the most efficient estimator with the best choice of its tuning parameter(s). Since different estimators may corresponds to different structures of the data, estimator selection is tightly link to model selection.

There are mainly two families of model selection and estimator selection procedures. The first family gathers the procedures based on resampling like jacknife, bootstrap or cross validation. They try to infer directly from the data the variability of the estimators and their estimation accuracy. The second family gathers procedures based on complexity penalization, like Mallows' C_p, BIC, MDL, and some more recent criteria suited to high-dimensional data. The most sophisticated criteria rely on a non-asymptotic analysis of the risk of the estimators based on concentration inequalities. This mathematical analysis gives access to a tight estimation of the relative risk of the estimators.

4.1.6 Model aggregation and ensemble methods

AUTHOR(S): **Arnak Dalalyan**

TEAM(S): LS/ENSAE, AppStat/LAL, CMLA/CACHAN, AROBAS/IBISC

The classical approach in statistics consists in modeling the phenomenon of interest based on our a-priori knowledge with the help of probabilistic tools and then in estimating the parameters of the model, or in making a decision, by model-based statistical methods. However, for describing and analyzing complex objects and phenomena encountered in different applications, it appears to be more efficient to deal with a large pool of probabilistic models, each one of which is well suited to describe the phenomenon of interest under some specific conditions. For each probabilistic model, there may be one or several statistical procedures that are guaranteed to yield (nearly) optimal results if the model is correct or plausible. This leads to a large family of statistical procedures which can be used for solving the problem at hand. A challenging question that was the core of many recent studies is how to combine all these procedures in order to get one procedure that performs almost as well as (and even better than) the best one in the family, without knowing which model is the best. Any possible solution to this problem is called aggregation strategy or ensemble method.

There is growing empirical evidence of superiority of aggregated strategies, with respect to "pure" ones. Since their introduction in the 1990s, famous aggregation procedures such as *Boosting*, *Bagging* or *Random Forest* have been successfully used in practice for a large variety of applications. Moreover, most recent Machine Learning competitions such as the Pascal VOC or Netflix challenge have been won by procedures combining different types of classifiers/predictors/estimators. On the other hand, strong theoretical results have been recently obtained for aggregation strategies that stem from the PAC-Bayesian approach. Although it is recognized that these methods share a large number of features with those investigated in convex and stochastic optimization, there are still many challenging open questions related to assessing the computational complexity and establishing theoretical guarantees on the resulting procedures. Another important problem that received very little attention is that of aggregating testing procedures in order to build powerful statistical tests that may hopefully extend to the setting of multiple hypotheses testing.

4.1.7 Information theory and algorithmic probability

AUTHOR(S): **Yann Ollivier**

TEAM(S): E3S/SUPELEC, TAO/LRI

Algorithmic probability provides a general framework for data understanding and inductive reasoning, based on the equivalence between probabilistic modeling, prediction, and compression. Its practical implementation was long considered next to impossible, but the field has seen a revival, in particular through minimum description length techniques. These ideas have already proven to be useful for statistical learning (in particular model selection) and data compression.

This viewpoint is deeply related to information theory, and, in particular, to information geometry, which deals with the underlying geometric structure present in families of probability distributions via

Fisher information. The resulting algorithms are, in theory, optimal for statistical learning. They are, however, often too heavy for practical use. Progress in information geometry currently allows for more lightweight but still theoretically principled approaches. Neural networks are an important application: their information geometry is now better understood, which has led to improved algorithms in particular for recurrent neural networks.

4.2 Data complexity

Classical multi-variate analysis usually deals with tables constituted of many records (instances) of a limited number of fields. Data collected today are now more and more complex and they require new methodologies to extract meaningful information. For instance, data could be of **high dimensionality** with only a **few instances** as in bioinformatics or medical research. Some fields could be made of long **time series**, large **images** or more generally data living on a manifold, for instance in fMRI brain imaging. Other fields involve data with an explicit **structure** such as graphs, trees, sequences, or composite objects. Although important progress has been made in dealing with these data as inputs in a prediction system, some problems like **network inference** or **structure prediction** call for extending regression methods to structured outputs. Data could also be **textual**, requiring **natural language** understanding as in a web search request. Finally, automatic handling of **missing** or **corrupted entries** is also a major issue.

4.2.1 Time series and panel data

AUTHOR(S): **Pascal Bondon, Arnak Dalalyan**

TEAM(S): TAO/LRI, L2S/SUPÉLEC, LFA/ENSAE-CREST, LMI/ENSAE-CREST, CMLA/CACHAN, AROBAS/IBISC

A time series is defined as a set of quantitative observations arranged in chronological order. We generally assume that time is a discrete variable. The analysis of time series helps to detect regularities in the observations of a variable and derive "laws" from them, and/or exploit all information included in this variable to better predict future developments. A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely related than observations further apart. During the last 30 years, time series analysis has become one of the most important and widely used branches of mathematical statistics. Its fields of application range from neurophysiology to astrophysics and it covers such well-known areas as economic forecasting, study of biological data, control systems, and signal processing.

In economic sciences and bio-informatics, it is often of interest to analyze multiple multidimensional time series corresponding to the observation of several characteristics over a given time period for several individuals. This type of datasets are termed panel data or longitudinal data and it is of central interest to develop statistical techniques for dealing with such data in a global manner instead of analyzing the time series corresponding to each individual separately. An additional source of difficulty is that in many cases panel data are unbalanced, which means that time periods or observation instances for different individuals are different.

4.2.2 Directional data

AUTHOR(S): **Gilles Fäy**

TEAM(S): MAS/CENTRALE, LMO, COSMOSTAT/CEA, PLANCK-LSST/LAL, IDOC/IAS

Directional data refers to points living on the sphere $\mathbb{S}$ of some Euclidean space $\mathbb{R}^d$, typically the circle ($d = 2$) or the sphere ($d = 3$), or more generally on some smooth manifold. Directional data is ubiquitous in many applications, such as astrophysics or observational cosmology, computer vision etc. Observations of events such as particle shower induced by incident ultra-high energetic cosmic rays on the Earth, or neutrinos (relatively bigger data sets) are archetype of such data. A close category of data are direction-indexed data (signal on the circle or on the sphere, modelled for instance as a $\mathbb{S}$-indexed stochastic process). This is exemplified by the anisotropies of the cosmic microwave background. Note that in those two cases, namely $\mathbb{S}$-valued or $\mathbb{S}$-indexed data sets, a long pipeline of data pre-processing may be necessary to yield a proper set of directional data; see for instance the *map-making* step made necessary because of the scanning strategy followed by, say, the Planck mission of the European Space Agency. Many modern approach of statistics (including *e.g.* sparsity enforcing methods, non-parametric estimation or detection techniques, multiscale approaches) have been applied to this kind of data and are still investigated and developed.

4.2.3 Structured data analysis and structured prediction

AUTHOR(S): **Florence d'Alché-Buc, Arthur Tenenhaus**
TEAM(S): AROBAS/IBISC, E3S/Supélec

Many application fields especially in sciences (biology, chemistry, imaging) involve structured data either with an explicit structure such as sequences, trees, graphs or any composite objects or with implicit structure such as functions, curves or signals. When structured data are used as input in an unsupervised or supervised setting, mainly two approaches are developed: kernel-based approaches than rely on an appropriate design of the kernel and on its learning and probabilistic graphical models that capture the underlying structure of data by modeling the conditional distribution dependence structure of the random variables at hand. An even more challenging problem concerns structured data as output of a prediction system: link prediction, multi-task regression, structured classification, functional regression are all examples of structured prediction. The main issue is to learn functional dependences between structured inputs and structured outputs and still require new development both from a theoretical angle and from a practical one. However, promising research directions span already a large spectrum including large margin approaches, maximum entropy approaches and kernel methods such as joint kernel map and operator-valued kernel-based models. Regularization and constraints appropriate to structured data also play here a central role.

4.2.4 Ranking

AUTHOR(S): **Stéphane Clémençon**
TEAM(S): CMLA/CACHAN, STA/LTCI, LRI/TAO, APPSTAT/LAL

Recommendation systems and search engines are becoming ubiquitous. Operating continuously on still more content, use of such tools generates and/or exploits more and more massive data. The design of machine learning algorithms, tailored for these data, is crucial to optimize the performance of such systems (e.g., rank documents by degree of relevance for a specific request in information retrieval, propose a sorted list of items/products to a prospect she/he is most liable to buy in e-commerce). The scientific challenge essentially stems from the nature of the data feeding or being produced by such algorithms: input or/and output information generally consists of (partial) rankings/orderings, expressing *preferences*. Because the number of possible rankings explodes with the number of instances to be ranked, traditional methods in machine learning and statistics become quickly intractable and the approaches proposed these last few years to deal with preference data, though corresponding to significant methodological advances, can hardly be implemented in large-scale settings.

The design of ranking algorithms to optimize the performance of recommendation systems or (meta-) search engines involves the processing of *preference/ranking data*. This is a real mathematical/computational challenge, for two reasons. First, such data express a global, and thus non local, property: it is not about labeling the objects (*e.g.* relevant *vs.* irrelevant), objects are ranked (*e.g.* by degree of relevance) and modifying the rank of an object may possibly affect that of many other objects. In addition, the number of objects to be ranked is very high in most modern applications, whereas preferences are generally observed for a small number of objects only. When developing computational tools to analyze rank data, it should be thus kept in mind that any procedure of linear complexity with respect to the number of rankings is quite intractable. This calls for new concepts and algorithmic methods. A current stream of research on these topics investigates the use of tools from computational harmonic analysis to obtain efficient/sparse representations of rankings so as to achieve aggregation or prediction of rankings.

4.2.5 Topological and geometric inference

AUTHOR(S): **Frédéric Chazal**
TEAM(S): GEOMETRICA/INRIA, CEA Tech/LIST

Often data can be seen as point clouds embedded in Euclidean spaces, or as more general metric spaces, e.g., when data are just given as a matrix of pairwise distances/similarities between points which is often the case for sensor networks or social networks data. These point clouds carry some intrinsic topological and geometric structure. For example, when they are sampled in some space, they are usually not uniformly distributed in the embedding space but lie close to some lower-dimensional geometric structure (manifold or more general stratified space) which reflects important properties of the "systems" from which the data have been generated. Identifying, extracting, and exploiting these underlying geometric structures has become a problem of fundamental importance for data analysis and statistical learning. There exist various statistical and machine learning methods that intend to uncover the geometric structure of data but most of them assume the underlying structure to have a very simple geometry — diffeomorphic to a disc or isometric to an open set of a Euclidean space. Furthermore the only topological information they usually seek for is connectivity. With the emergence of distance-based approaches and persistent topology, geometric inference and computational topology have recently known an important development towards data analysis, giving birth to the field of topological and geometric data analysis. Despite its youth, and thanks to strong mathematical foundations, this new field has already found successful applications in various domains including neuroscience, bioinformatics, shape classification, clustering and sensor networks, to name a few. Moreover, it starts to provide efficient topological and geometric data analysis tools that can be successfully used by scientists and researchers from other communities.

4.2.6 Natural Language Processing and Text Mining

AUTHOR(S): **Guillaume Wisniewski**, Michalis Vazirgiannis, Claire Nédellec
TEAM(S): LIMSI, DaSciM/POLYTECHNIQUE, BIBLIOME/MIG

Natural Language Processing (NLP) is the technology for dealing with our most ubiquitous product: human language, as it appears (just to mention the written productions) in newspaper stories, scientific articles, product descriptions, emails, web pages, tweets, and social media, in thousands of languages and varieties. **Text mining** aims at extracting information out of its written form without necessarily trying to interpret it as a language. For instance, in the field of Information Retrieval (IR) and for its main application Web search, the research community have been proposing for decades methods to successfully retrieve relevant textual documents as response to human queries without the need of a full understanding of the meaning of text through lossy representations such as the bag-of-word. The main goal of NLP is to derive meaning from human or natural language input by building abstract structured representations of the sentence such as syntactic parse trees or dependencies graph, which both rely on the prediction of the 'meaning' of each word and of the relations between the different words in a sentence. However, depending on the task at hand, simpler analyses, such as recognizing named entities (e.g. names of places, people or organizations) or classifying documents into broad categories (text classification), may be sufficient.

Another goal of both NLP and Text Mining researchers is to solve real-word tasks aiming at making the interaction with textual content easier. Examples of such tasks include Web search, automatic summarization, machine translation, speech recognition, sentiment analysis (also known as opinion mining) or text simplification. Since the nineties, both communities have developed a strong evaluation culture: for each task, evaluation campaigns are organized regularly to compare new and existing approaches and measure the progress made including the TExt Retrieval Conference (TREC) and its European counterpart the Conference and Labs of the Evaluation Forum (CLEF) or the WMT evaluation campaign that evaluates each year Machine Translation systems.

NLP uses stochastic, probabilistic and statistical methods to model and predict linguistic information and the algorithms designed there are therefore closely related to many Machine Learning (ML) methods both supervised (Section 4.1.1) and unsupervised (Section 4.1.3). Similarly, Text Mining, just like the broader Data Mining field, has been using a lot of ML to extract meaningful patterns, be it on the raw input (textual features) or the output (e.g. learning to rank in IR). As many linguistic information can be represented by sequences, trees or graphs, NLP is also one of the main motivation for the development of structured prediction methods (Section 4.2.3). However, when dealing with human language, the peculiarity of textual data must be taken into account: for instance, the observed distributions are generally unbalanced and typically contain many rare events. At the lexical level, because of their Zipfian distribution, many words only occurs a few times in any finite sample, no matter how large they are: this makes learning harder. Another problem stems from the fact that existing corpora often only contain partial and weak supervision information. Finally, as the vocabulary, style and structure of a document vary greatly across writers, domains or kinds of documents, the most common situation is to deal with non-homogeneous sources of data where train and test distribution differ in several ways, a situation known as Covariate Shift in the statistical literature.

4.2.7 Graph Mining

AUTHOR(S): **Michalis Vazirgiannis**

TEAM(S): DASCIM/POLYTECHNIQE, CMLA/CACHAN, CEA Tech/LIST, AROBAS/IBISC

Graph Mining is an area within the domain of data mining focusing on data that can be represented by graphs i.e. nodes connected by edges. A wealth of data types fall in this category and prominent examples include protein interaction networks, metabolic networks, the World Wide Web, co-author networks and social networks. Depending on the data and the task graph mining can take different forms and semantics.

A general form of graph mining is characterizing statistical properties of the graph such as node degrees and their distribution, diameter/average path length and the transitivity within a graph. These general properties are used to detect anomalies in graphs, distinguish "real" graphs from artificial or create synthetic but realistic models of the original graphs.

A more specific topic in graph mining is that of frequent sub-graph mining aiming to extract all the frequent subgraphs (subsets of nodes and their connections), in a given data set, whose occurrence counts are above a specified threshold. These patterns can provide useful insights -into the relations formed by individuals within the original network- that could be used directly (e.g. intrusion detection, program control flow analysis) or could be utilized by other graph mining tasks (e.g. clustering, graph comparison).

Another task of graph mining is that of clustering; a distance is required for this procedure. Depending on what kind of clusters one is looking for, the distance can be defined in various ways. One approach is to cluster nodes by their common "neighbors", one example in this case is that of users liking "product" of some kind. This created a bipartite graph on "customer" nodes being connected to "product" nodes. Another approach towards clustering is by identifying the denser parts of graphs - also referred to as community detection in cases where the underlying network bares such semantics (e.g. an online social network).

As a graph might not be static, link prediction is also an active area of study in graph mining. The general idea is to study the current connections within a graph (and their history of appearance if available) in order to predict connection that will emerge or will be discovered in the future. Application here lie in predicting future collaboration is social networks, web page improvement (links that should exist) and biological networks (e.g. discovering new protein interactions).

Another issue of great interest is the diffusion and propagation within a graph. This could refer to information diffusion, influence propagation or even virus propagation. The main idea here is to study how some sort of information (or some form of energy e.t.c.) spreads through a network. This is useful in identifying points of interest for protection or targeting. For example an advertising company could be interested in creating a "viral" campaign in a social network and for this reason it may require to find the best "targets" that will make the spreading most efficient.

4.2.8 Computer vision

AUTHOR(S): **Nikos Paragios**, Arnak Dalalyan

TEAM(S): CVN/CENTRALE, TII/LTCI,LS/ENSAE

The ultimate goal of computer vision as a scientific discipline is to make the computer see. Video frames and digital images acquired by various types of devices are the main visual objects the understanding of which is the core objective of computer vision. More specifically, the field of visual understanding pertains to the development of automatic techniques to obtain a semantic interpretation of visual data, such as object detection in natural images or foreground-background segmentation of medical images.

The problem of visual understanding is highly challenging due to the ambiguity inherent in the data. For example, consider a natural image containing object categories with high intra-class variations (change in appearance of humans due to clothing) and inter-class similarities (similar shapes of various vehicles such as cars, buses and trucks). Another example is a medical acquisition, such as an MRI scan, where the organs of interest are similar in appearance to the background. In order to deal with the ambiguity, it is common practice to formulate problems using probabilistic models, that is, models that represent the probability of each putative interpretation of the given data. Efficient fitting and tuning such models to the specific datasets studied in computer vision is a challenging task that requires leading-edge tools from optimization, Bayesian computations, etc.

4.3 Resource limitations

As **data size** grows, sometimes faster than the **computational resources** (CPU time, memory, communication) available to process them, algorithmic/computational issues become as important as inference questions traditionally dealt with in the statistical paradigm. **Real-time** constraints could also become important both in the model-building (learning) step (e.g., information should be extracted with the lowest possible delay when dealing with online anomaly detection) or in the inference step (e.g., in web-page ranking a learned decision rule should be computed as quickly as possible on a limited CPU budget). To solve these issues, standard methods have to be adapted to the **new computational platforms** (GPU farms, multicore machines, clusters, and grids), and **new algorithms** (*sparse* learning and inference, *deep learning*, *sub-sampling* techniques, *stochastic optimization* and *high-dimensional numerical integration* methods) have to be developed.

4.3.1 Convex optimization

AUTHOR(S): **Matthieu Kowalski**
TEAM(S): E3S/SUPÉLEC

Within the last few years, the quest for efficient algorithms in signal/image processing and machine learning has know great developments, motivated by

- increasing the amount of data to be processed and the size problems;
- development of sparse recovery or compressed sensing for which objectives to be optimized are constrained, non smooth.
- the development of new architectures (GPU) that constrained the numerical algorithms;
- the ”rediscovery” of a large literature of 60-70s in convex optimization, relatively unknown to a large part of the community of signal/image processing and machine learning in the 90s.

“Proximal point”-like methods of are now widely used to address non smooth problems such as LASSO or Basis pursuit using sparsity and L1 norm, or in imaging with the total variation. Developed in the 70s, they know a new growth due to their ease of implementation and ability to handle large problems without the need to have a differentiable objective function. Although the second order methods such as interior points methods are possible to obtain more accurate solutions, the first order methods allow one to obtain approximate solutions quickly, usually sufficient for the visual system in imaging.

However, the computational time can still be too long in practice to obtain an acceptable solution, which is the main limitation for the use of these recent approaches by the “applicative” communities. Nevertheless, several options can be used to accelerate the practical convergence of these algorithms. These include, among others:

- Splitting of monotone operators with variable metric: splitting methods of maximal monotone operators solve a large variety of convex problems. They provide simple algorithms that can be implemented naturally on parallel architectures such as GPUs. This can lead to significant performance gains.
- Active set: for very high dimensional problems, where the data can not be fully loaded in memory, stochastic methods seem the most relevant approach. On intermediate size problems, where the data can be loaded into memory but involve heavy computations at each iteration, a successful strategy are “active sets” methods. These methods exploit the expected sparsity of the solution to work on a limited size problem, thereby saving time and memory consumption with ”out of core” computation.
- Incremental methods: this kind of methods is crucial for large-scale problems, involving an objective written as the sum of a large number of functional. These methods operate at each iteration on a sub-set of coordinates, or on each individual function, instead of the entire objective which turns to be extremely expensive. In the literature, one can find both deterministic and stochastic strategies, with classes of methods applicable to the not differentiable but convex case.

4.3.2 Stochastic optimization

AUTHOR(S): **Niko Hansen**
TEAM(S): TAO/LRI

Black-box optimization is the last resort for ill-posed optimization problems, which lack the desirable properties of convexity or differentiability or even computability. Such problems are frequently met in design problems where the underlying objective function involves heavy computations (e.g., based on Finite Element Methods in Numerical Engineering) or can only be captured from direct interaction with the physical system.

Among the major directions of research for black box optimization algorithms are the search for invariance properties with respect to monotonous transformations of the objective function and with respect to affine transformations of the representation space. Such invariance properties are desirable as they are at the core of exceptionally robust and generic algorithms, since the good behavior on some problem instances translates to a whole set of problems. The relationships between such invariance properties and the natural gradient search in the space of distributions (see Section 4.1.7) on the solution space have been established recently, providing *a posteriori* the theoretical basis to explain the proven track record of algorithms such as Covariance-Matrix Adaptation–Evolution Strategies (CMA-ES).

Among optimization problems is the algorithm selection and parameter tuning problem, searching for the optimal portfolio algorithm/hyper-parameter values depending on the problem instance at hand. Not only is the parameter tuning problem a key bottleneck for the deployment and transfer of algorithmic platforms from research labs to industry. It also paves the way toward defining a typology of problem instances, and better understanding the difficulty factors.

4.3.3 Deep learning

AUTHOR(S): **Michéle Sebag**, Balázs Kégl

TEAM(S): TAO/LRI, APPSTAT/LAL, TLP/LIMSI

Finding an appropriate representation of the data is a key step for machine learning and for (scientific) modelling at large. While this step has long been dealt with manually (the so-called *feature engineering*), deep learning has emerged since 2006 as an automatic, often unsupervised, and tractable approach for learning new representations from large datasets.

Deep learning is credited with the winning approaches to recent challenges ranging from computer vision (ImageNet) to chemistry and molecular activity (Merck), improving by a significant margin on the competing approaches. Such successes indicate that deep learning uncovers general learning principles, although the theory is lagging behind the practice.

In particular, deep learning has been applied on unprecedently large datasets; at this scale, it appears that unsupervised learning might play a bigger role than was considered so far. The relationship between unsupervised and supervised learning currently is among the hottest topics of machine learning.

4.3.4 Budgeted learning

AUTHOR(S): **Balázs Kégl**

TEAM(S): APPSTAT/LAL

Most of the time the main computational bottleneck in designing large classifiers or regressors (Section 4.1.2) is the training phase. In some applications, however, there are constraints on the learned function itself. Most of the time the constraint is on the computational time as in real-time object detection or web-page ranking, but some applications may limit the memory to store the model (e.g., in mobile applications) or even the energy consumption of the hardware executing the function (Section 3.4.2). The classical solution for budgeted classifiers is the cascade design, popularized by the famous Viola-Jones cascade object detection framework. Since then the area has been active with the appearance of new and improved designs and several recent workshops (1, 2) at major machine learning conferences.

4.3.5 High-dimensional statistics and sparsity

AUTHOR(S): **Erwan Le Pennec**

TEAM(S): LMO, POLYTECHNIQUE, LS/ENSAE, CMLA/CACHAN, AROBAS/IBISC

Complex high-dimension data needs complex high-dimensional models, but high-dimensional model parameter estimation requires in general a number n of observation which is at least of order of the number p of parameters in order to be accurate. This is a strong limitation in practice where the number of observations is generally limited. In the **sparse** approach, one still uses high-dimensional models but assuming that it exists good approximating models in which most of the parameters are equal to 0. In that case, if we denote by k the number of non-zero parameters, roughly speaking, the number of required observation to identify that set of non-zero parameters and estimate their values if k was known is now of order $k \log p$ instead of p. Scientist in this area are trying to build models for which an efficient estimation procedure allows to compute such a non zero-coefficient set and to estimate their values. They often relate the performance of their method to the estimation using the best subset of coefficients knowing the functions and prove that they are of the same order. The most famous instance of this framework is the Lasso estimator but many other exist and are used. The challenge is this setting is to simultaneously propose a high-dimensional model adapted to the data, construct a computationally efficient method (often based on convex optimization technique) to select a solution with few non-zero parameters and to ensure the good performance of the algorithm either theoretically or numerically.

4.3.6 Numerical integration in Bayesian inference

AUTHOR(S): **Nicolas Chopin**
TEAM(S): LS/ENSAE, STI/LTCI, APPSTAT/LAL

As datasets become bigger and more complex, so are statistical models. One way to compensate for this increased complexity is to add some form of "structure" (such as sparsity, or a hierarchical structure, or some form of expert knowledge). Bayesian inference currently receives a lot of attention in Machine Learning and Statistics, because it makes it possible to introduce such a structure in a principled way through a prior distribution on the (finite or infinite-dimensional) set of unknowns. A good example is the field of "computer experiments", where one tries to infer the response of a complex system (say the whole universe in Physics) simulated on a computer as a smooth function of many inputs using Gaussian processes. In the same vein, there has been a lot of attention paid recently to pseudo-Bayesian inference, in particular PAC-Bayesian inference, which generalizes Bayesian inference to situations where it is difficult to define or compute a likelihood function.

Although Bayesian and pseudo-Bayesian inference are very appealing approaches, their numerical implementation ("Bayesian computation") is challenging: obtaining numerically the posterior expectation of certain functions, which amounts to compute an integral of a possibly large dimension, is a hard problem. In the nineties, the introduction of MCMC (Markov chain Monte Carlo) represented a breakthrough for Bayesian computation, but the performance of MCMC tends to be unsatisfactory in large or complex scenarios that become prevalent today.

Fortunately, Bayesian computation is currently going through a "second wave", thanks to several very recent breakthroughs. One approach, favored in machine learning, is to do away with Monte Carlo entirely, and instead resort to fast approximation schemes, such as variational algorithms, or the expectation-propagation algorithm. This type of approach opens many interesting areas of research: (a) how to adapt or generalize these fast approximations to different problems; (b) how to establish the properties of the so-obtained approximations (in particular the statistical properties of such an approximation), and so on. A second approach is to consider a new set of Monte Carlo algorithms, termed SMC (sequential Monte Carlo), or particle filters. Although the historical motivation of SMC was for problems with a sequential structure, there is growing evidence that SMC may be used as an alternative to MCMC even in non-sequential problems. There are, however, many open questions regarding the properties and practical implementation of SMC in large dimension. A third approach is to adapt certain ideas from physics to develop much better MCMC schemes, such as Hamilton Monte Carlo, which shows impressive performance in certain problems. Much needs to be done, however, to gain an understanding on how to better adapt these methods to statistical problems. Finally, there is also a lot of research on so-called "likelihood-free" scenarios, where the model is so complex that it is only possible to simulate from it; examples of such scenarios abound in population genetics for instance, but also in neuroscience, ecology, or particle physics. One may then use so called ABC (approximate Bayesian computation) algorithms, which sample from the model in some way until the data simulated from the model is close enough the actual data. A more general question is how these apparently very different approaches may be combined in order to produce even more powerful algorithms.

4.3.7 Massively parallel data processing

AUTHOR(S): **Ioana Manolescu**

TEAM(S): OAK/Inria, LAHDAK/LRI, DTIM/ONERA, DASCIM/POLYTECHNIQUE, E3S/SUPÉLEC

Massive data scale requires a complete re-design of data processing software, with an emphasis on massive parallelism, reducing inter-site data transfers, parallel graph processing, and smart cloud-based stores. We are particularly interested in smart stores and processing primitives for complex- and mixed-structure data, in particular documents, Semantic Web graphs, and other classes of graph data of interest in particular applications. For instance, we investigate the usage of structured documents with semantic annotations for representing structured text enriched with semantic annotations; we hope to further the study of (parallel) algorithms for such analysis in collaboration with our colleagues involved in **??** and 4.2.6. We develop such massively parallel processing techniques within the Europa/Stratosphere project (2011-2014) funded by EIT ICT Labs.

As another example of novel questions to answer through massively parallel algorithms, we collaborate within the Datalyse *Investissement d'Avenir* project (2013-2016) with data mining experts interesting in expressive and efficient tools for exploiting in a massively parallel fashion, large volumes of patterns identified through mining.

4.4 Interactive visualization and experimental design

Data collection is often an expensive process when, e.g., data comes from an experimental device or when it is generated using complex numerical simulations. Data science has a strong interaction with statistics, where it is well known that the efficiency of a decision procedure depends on how experiments are chosen. To deal with this problem, it is important to construct efficient and tractable **experimental designs** that **adapt** these choices to what has been seen before. Developing efficient **visualization** techniques (especially **interactive** visualization tools) is also an important goal for analyzing experimental data and simulations, and to explore and collaborate on large data sets.

4.4.1 Visualization and Interaction

AUTHOR(S): **Jean-Daniel Fekete, Nicolas Férey, Frédéric Vernier**

TEAM(S): VENISE/LIMSI, AVIZ/INRIA, CEA Tech/LIST

To produce knowledge from large amounts of data, it is nowadays critical that human expertise and computing power complement each other through well-adapted graphical representations, interaction techniques, and analysis tools. Indeed our visual senses can **extract information from graphical representations** efficiently and effectively, provided these visual representations are well-tuned to the **human perceptual system**. Yet, even the most optimized visual representation cannot convey more than a few million data points, due to limitations of the human perceptual system, as well as limitations in **human cognition**. Therefore, to address the problem of understanding large amounts of data, static **visualizations** should be coupled with **interactive data analysis** techniques. Again, this visual representations and interaction techniques should be coupled in a manner compatible with human capabilities. For example, an analytical inquiry performed by an analyst requires fast exploration to rapidly generate, validate or invalidate hypotheses. If analytic computations take more than a few seconds, the analyst can lose track of his/her generated hypotheses due to limitations of short-term memory. While the exploration process can still be performed, it becomes much less efficient. Therefore, it is essential to combine visualization with efficient analytical methods through Visual Analytics approaches, that embeds **visualization** and **interaction** methods, **information extraction** techniques, **knowledge representation** and **data mining** approaches to provide users with tools and techniques to synthesize information, detect the expected structures and discover the unexpected knowledge from massive data. Besides, we have reached a state where a large number of visualization techniques exist to explore temporal snapshots of data or produce animation of temporal phenomena, however the state of the art is not sufficient to explore patterns among large time series or at different scales. This statement now calls for new techniques to address the challenge of revealing all the temporal richness of data.

Going beyond the scope of past research which has largely focused on visualization for desktop-based work settings, there are also alternatives to the classical desktop context in order to display, explore, manipulate and share large amounts of data to provide with a more adapted new contexts of work. Indeed, in many science fields, classical applications in desktop context, commonly used in our labs, have reached their limits and cannot adequately deal with the size and complexity of today's problems. This reduces our ability to explore and understand complex data. On the one hand, novel interaction techniques for data including tangible user interface, touch user interface or multimodal interaction, encourage alternative forms of data exploration. The goal of these new techniques is not only to provide effective data representations, but also effective interactions that let analysts explore datasets according to multiple perspectives, and control and steer analytical algorithms. On the other hand, novel display devices for visualization including wall displays, virtual and augmented reality devices, large touch-based displays, or small mobile displays, provide new data analysis environments to explore very large datasets such as graphs with millions of vertices or complex 3D datasets such as medical scans of the human brain, fluid mechanics or molecular dynamics simulation results .

4.4.2 Experimental design

AUTHOR(S): **Emmanuel Vazquez**
TEAM(S): E3S/SUPÉLEC, CMLA/CACHAN, AROBAS/IBISC

Today, the design of new products in the industry is largely based on numerical simulations. Numerical simulations make it possible to test configurations, to optimize the design of a system and to assess its robustness with respect to failures. However, complex simulations require long computations, which sets a limitation on what can be learned in reasonable time.

The domain of design and analysis of computer experiments aims at defining what should be chosen for the inputs of a numerical model in order to achieve a prescribed objective. In particular, one may want to: (i) predict the behavior of a numerical model from the results of a small number of runs (ii) optimize the response of a numerical model; that is, to determine the values of the inputs corresponding, for example, to a highest performance or a smallest cost (iii) estimate the variability of a response as a function of that of the inputs (also known as sensitivity analysis) (iv) estimate a probability of failure in presence of uncertainties

Whereas space-filling designs are commonly used for the first objective, different types of designs may be more relevant in other situations. Sequential strategies (or active learning) that construct a model of the numerical simulator step by step, are especially attractive. Today, research questions focus on the definition of design criteria related to a given objective, the construction of efficient algorithms for the determination of optimal experiments, the investigation of asymptotic properties of designs, the construction of designs for dealing with simulators with several levels of predictive accuracy.

4.4.3 Online learning and sequential decision making

AUTHOR(S): **A. Dalalyan, O. Cappé**

TEAM(S): TAO/LRI, STA/LTCI, LS/ENSAE, CMLA/CACHAN

In many applications, data are acquired sequentially and it is necessary to learn a concept or to make decisions on the fly. This set-up may concern both supervised and unsupervised learning. In the first case, for instance, at each step t, we only observe the component x_t of the feature-label pair (x_t, y_t), randomly drawn from the space $\mathcal{X} \times \mathcal{Y}$. The goal is to predict the unobserved label y_t. Once a prediction $\hat{y}_t$ is made, the true label y_t is revealed and the process goes on with the next step $t + 1$. This is the typical set-up of the online learning as opposed to the conventional batch learning.

Thus, an online learning algorithm requires to update the decision rule at each step based on the new observations and a central issue is to design algorithms with updates having low computational complexity. Most famous algorithms of machine learning, such as support vector machines, expectation-maximization, boosting, etc have currently their online counterparts. From a theoretical point of view, one of main differences between online learning and batch learning is that in the former, the goal is not to design decision rules that generalize well to all the features of the feature space $\mathcal{X}$, but only to those features that are observed during the data collection process. Therefore, instead of the generalization error or the expected risk of prediction, the performance of an algorithm in online learning is measured by the cumulative regret with respect to what would have been, in retrospect, the best static prediction.

Sequential decision making refer to problem where the agent (or user) has several options that can possibly be activated at each time step. Here the situation is even more challenging as the actions of the agent influence the data collection itself and could result in severe inference bias in the worst case. In this situation, optimal strategies aim to reach a proper balance between exploration (trying all possible options to evaluate their average outcome) and exploitation (gradually focusing only on the best actions). The UPSa teams involved in the CDS project have a recognized expertise on this topic either in the context of basic independent sequential decision making (so-called multi-armed bandit models) but also in tree-structured sequential decision making.

4.4.4 Active learning

AUTHOR(S): **Nicolas Vayatis**

TEAM(S): CMLA/CACHAN, TAO/LRI

Active learning procedures are sequential learning algorithms which dynamically recommend evaluation points in order to improve the estimate of the decision rule. In classification, this amounts to select evaluation points that get closer and closer to the frontier between observations from different labels. Through efficient active learning, faster convergence to optimal decision rules is expected. Available methods are either model-based and seek for high variance or low information regions, or nonparametric and hit in the maximal discrepancy region for a committee of available decision rules.

From the viewpoint of applications, applicability of active learning requires the capacity to fully control the sampling mechanism for data collection. It can be viewed as a modern perspective on the well-known problem of experimental design and brings enough machinery to address high dimensional problems. Recent applications of active learning algorithms have led to breakthroughs in the control of numerical and physical experiments, as well as in complex systems design. Practical fallouts are cost reduction and innovation.