DATA SCIENCE SYMPOSIUM

26-27 NOV. 2015
Campus de Beaulieu RENNES
Espace Conférences Inria Rennes-Bretagne Atlantique

1975 • 2015 IRISA

Science Data Ecosystem
Workshop

Workshop Organizers:

Christine Morin, *Senior Scientist and Head of the Myriads project team, IRISA/Inria*
Deb Agarwal, *Senior Scientist and Data Science and Technology Department Head, Lawrence Berkeley National Laboratory;* Senior Fellow, Berkeley Institute for Data Science; International Chair, Inria, Rennes

Workshop Description:
Data produced by scientific instruments (large facilities like telescopes or field data), large-scale experiments, and high-fidelity simulations are increasing in magnitude and complexity. Data analysis methods, tools and infrastructure are needed to provide the complete data management, collaboration, and curation environment suitable to manage these complex, dynamic, and large-scale data analysis ecosystems. During this workshop, methods and techniques developed to overcome the data management challenges faced by scientists from different disciplines will be presented. We will discuss the data management commonalities across various scientific domains and how computer scientists can better work together with other scientists to implement highly usable systems that will accelerate the pace of scientific discoveries.

Speaker: **Anne Siegel,** CNRS/IRISA
Title: Learning discrete dynamical systems for signaling biological process from experimental "data deluge".

Abstract:
Systems biology is a field of computer science and biology, which aims at describing the response of a cell to environmental perturbations. In the last decade, this field was confronted to technological evolutions, which simultaneously measure the time-series activity of hundreds of molecules. A priori, such data should be very interesting to learn and clearly identify dynamical systems, which explain the cell response. Nonetheless, it appears that although very large, experimental data deluge never seem to be sufficient to extract robust information. In this short talk, we will illustrate some of these drawbacks on the specific issue of signaling network identification.

Bio:
Anne Siegel is a research director at CNRS and head of the Dyliss team at Inria, Rennes and IRISA. Her research focuses on the development of a constraint-based approach to modeling the response of biological systems from large-scale heterogeneous experimental data.

Speaker: **Samuel Corgne**; LETG Rennes COSTEL, CNRS
Title: Land use and land cover monitoring with optical and radar satellite data: application to agriculture areas

Abstract:
In agricultural areas, land use and land cover monitoring at a fine scale represents an important stake for environmental management, as they have important impacts on water resources, soil erosion biodiversity, etc. Land use and land cover monitoring at a fine scale are generally realized with remote sensing data at high and very high spatial resolution. With recent spatial programs as Sentinel (Copernicus) which combined optical and SAR (Synthetic Aperture Radar) data at very high spatial and temporal resolutions, new scientific issues are discussed: the use of high spatial and temporal data for a better discrimination of crop (phenology, agricultural practices…), radar and optical data fusion for a better land use identification, etc. The presentation will focus here on land cover and crop characterization at a field scale with multi-temporal SAR data (TerraSar-X and Radarsat-2) and with optical and radar data fusion.

Bio:
Samuel Corgne is a Professor at the Department of Geography in Rennes 2 University (France) and a member of the LETG Rennes COSTEL lab (UMR 6554 CNRS). His research interests include land use and land cover monitoring with remote sensing data. His work focuses on optical and radar data fusion for land use modeling on agriculture areas characterized by high spatial changes and important environmental challenges (water pollution, soil erosion…).

Speaker: **Christian Barillot**, CNRS
Title: Science Data Ecosystem in Medical Imaging

Abstract:
I will talk about general issues of Medical Imaging as a Service in the context of the emerging infrastructures. I will first introduce the general context and then provide some examples of how these new services can be implemented, a short introduction of the Shanoir system we are providing in Rennes for these solutions and two examples of infrastructures that are proposing specific services for dedicated communities. I will finish with some perspectives and open questions.

Bio:
Christian Barillot is a senior scientist at CNRS. He is the scientific leader of the VisAGeS U746 research Unit, and since 2010 he has been the director of the Neurinfo imaging platform. His research interests concern the processing of multidimensional images applied to medicine including the aspects of medical image analysis, integration and fusion of medical images and their applications in brain pathologies.

Speaker: **Yann Busnel**, ENSAI
Title: Project INSHARE (INtegrating and Sharing Health dAta for REsearch): Putting Health Data into Big Data

Abstract:
Health Big Data (HBD) is a promising step towards decreasing research costs, increasing patient-centered research, and speeding the rate of new medical discoveries. Sharing and exploiting efficiently HBD requires tackling the following challenges: HBD is sensitive: reusing HBD must comply with data protection governance taking into account legal, ethical and deontological aspects which enables a trust, transparent and win-win relationship between researchers, citizen and data providers. HBD has a limited level of interoperability:

data are compartmentalized and are so syntactically and semantically heterogeneous. HBD is of variable quality depending of the source. Clinical Data warehouse (CDW) technologies have emerged as one of the solutions to address HBD exploitation.

The INSHARE project aims to demonstrate the feasibility and the added value of an IT platform based on CDW, dedicated to collaborative HBD sharing for medical research. The specific objectives of the INSHARE project are: (i) To design the governance of data sharing. (ii) To implement and to develop a prototype of trusted third-party HBD sharing platform. (iii) To evaluate the prototype on 3 different real world use cases: Registry enrichment: Comorbidities and drug exposure in End Stage Renal Disease patients; Characterizing the healthcare trajectories of children (and their mother) included in a Birth Defect Registry; Cross-domain study: Cancer, Diabetes and ESRD in a cohort study.

Bio:
Yann Busnel is head of the Computer Sciences department at Ensai, the national school for Statistics and Information Analysis and co-head of the MsC in Big Data. He is a member of CREST (Research Center in Economics et STatistics), Laboratory of Statistics and Models. He is also an associated member of Inria Research Center Rennes - Bretagne Atlantique, in the Dionysos team and associated member of LINA (Computer Science Laboratory of Nantes Atlantic). His research topics are (but not limited to): large-scale distributed data streams and distributed system models.

Speaker: **Gabriel Antoniu**
Title: Damaris: Jitter-Free I/O Management and In Situ Visualization of HPC Simulations using Dedicated Cores

Abstract:
Large-scale simulations running on leadership class supercomputers generate massive amounts of data for subsequent analysis and visualization. As the performance of storage systems shows its limits, an alternative is to embed visualization and analysis algorithms within the simulation code (in situ visualization). In this context, we present the benefits of using Damaris, a middleware for I/O forwarding and post-processing using dedicated cores to offload in situ visualization while sharing resources with the running simulation. Damaris fully hides the I/O variability as well as all I/O-related costs, which makes simulation performance predictable; it increases the sustained write throughput by a factor of up to 15 compared with standard I/O approaches; it allows almost perfect scalability of the simulation to over 9,000 cores; through its extension, Damaris/Viz, it enables a seamless connection to the VisIt visualization software to perform in situ analysis and visualization in a way that does not impact the performance of the simulation, nor its variability. Damaris/Viz was evaluated with the CM1 atmospheric simulation on Grid'5000 and on NCSA's Blue Waters with up to 6400 cores, and with the Nek5000 CFD simulation.

Bio:
Gabriel Antoniu is a Senior Research Scientist at Inria/IRISA, Rennes. He leads the KerData research team, focusing on storage and I/O management for Big Data processing on scalable infrastructures (clouds, HPC systems). He received his Ph.D. degree in Computer Science in 2001 from ENS Lyon. He leads several international projects in partnership with Microsoft Research, IBM, Argonne National Lab, University of Illinois at Urbana Champaign.

Speaker: **Olivier Dameron,** IRISA/Université Rennes1
Title: Semantic Web contribution to integrating data from different nature and different scales

in life science: application to TGF-beta in cancer

Abstract:

The recent joint evolution of data acquisition capabilities in the biomedical field, and of methods and infrastructures supporting data analysis (grids, the Internet...) resulted in an explosion of data production in complimentary domains (*omics, phenotypes and traits, pathologies, micro and macro environment...). This life science "data deluge" is part of the recent "Big data" phenomenon, with the specificities that data are usually structured and highly inter-dependent. The bottleneck that once was data scarcity now lies in the lack of adequate data processing and data analysis methods.

The Semantic Web is an extension of the current Web that recognizes the need to represent data on the Web in machine-readable formats and to combine them with ontologies. It aims to support fine-grained data representation for automatic retrieval, integration and interpretation. This presentation will demonstrate how Semantic Web technologies support the analysis of TGF-beta signaling pathways.

Bio:

Olivier Dameron is an associate professor at Université Rennes1. He is a member of the Dyliss team at IRISA where he works on ontology-based methods for analyzing life science data.