Learning from Scientific Data in Energy

IFPEN / Rueil-Malmaison (France) - 16-17 January 2018

DataSciEn'2018, co-hosted by IFPEN and Inria, held at IFPEN Rueil-Malmaison from 16-17 January 2018 brought together 110 participants – 50% from academia and 50% from industry - such as multi-disciplinary researchers, software editors and industry practitioners from the energy sector working on problems with large and complex datasets.

The use of scientific and technical data to successfully solve industrial problems is a specific challenge that requires close collaboration between the different professionals involved in data analysis. However, these professionals often come from different communities so that events bringing them together are few and far between. Thus the objective was to include, in addition to traditional academic presentations, technical presentations from software developers as well as industry practitioners to reflect the need for applied research dedicated to solve industry problems.

Session 1 : Frameworks and Platforms

The recent spike of interest in data science is also largely due to the development of frameworks freely available, allowing not only researchers but also engineers to easily use the large amount of data available in order to solve complex problems. These frameworks evolve quickly, with many appearing and disappearing over the course of a single year. This session objective was to discuss recent developments in this area to efficiently gather and analyze data.

Olivier Grisel (Inria) gave a keynote on the recent developments regarding the machine learning library scikit-learn, of which he is one of the main developers. The set of tools developed in scikit-learn are used worldwide in a diverse range of companies, from AirBnB to Spotify. The session also included presentations from NVIDIA, EDF, Cityzen Data and IFPEN on their own frameworks dedicated to other specific tasks, such as uncertainty quantification or geo-time series management.

Session 2 : Advances in Algorithms

Recent successes in thus far very complex problems - such as image recognition, computer vision and natural language processing - come from breakthrough in machine learning algorithms that happened in the last decade. The objective of this session was to present these recent advances and their application to solving scientific problems.

The session's keynote speaker **Patrick Gallinari** (UPMC) summarized the evolution of machine learning since its inception in the 1950's to the advent of deep learning today. Following

presentations from researchers from UTC, UPMC, INRIA, IFPEN and ENSMSE further extended these concepts to paint a picture of tomorrow's state of the art research.

Session 3 : Industrial Session

The most publicized applications of data science today still concern the general public. However, data science is also quickly transforming the industry sector with applications such as predictive modeling and maintenance, automatic decision making and data driven design. This session focuses on practical successful applications of data science to solve complex industry problems.
Our last keynote speaker **Michel Lutz** (Total) presented a detailed overview of how data science capabilities are efficiently and seamlessly implemented in a large corporation. His talk strongly resonated with the talk from Balázs Kégl (CNRS), who also advocated that the priority should be the focus on solving meaningful existing problems with data instead of trying to force business people to use data science. Successful examples of these ideas were presented in this session, from companies such as Total, Engie, EDF and Schlumberger.

Labs

In addition to the traditional sessions consisting of 30-minute presentations, DataSciEn'2018 allowed the participants to enroll in hands-on sessions in which they were able to manipulate some of the frameworks introduced in session 1 on real data sets. The two labs proposed were led by Olivier Grisel on scikit-learn and Frederic Pariente on Nvidia's DIGITS framework, with a very good review from the participants.

Challenge

DataSciEn'2018 was also the opportunity for IFPEN to launch its own Data challenge: it consists in predicting the oil residual saturation in a porous media solely based on 3D tomography images, with end applications in EOR processes. The challenge is hosted on the ENS challenge data platform (https://challengedata.ens.fr/en/home) and will run over the course of 2018, with a prize attributed to the best score at the end of this year's session !