Fondation Mathématique Jacques Hadamard

Université Paris Saclay

Research initiative in industrial data science (IRSDI)

Call for projects – March 2018

Overview and scientific scope

The Research initiative in industrial data science (IRSDI - *Initiative de Recherche en Sciences des Données pour l'Industrie*) is a corporate patronage funded by EDF and THALES and operated by the Jacques Hadamard Mathematical Foundation (FMJH - *Fondation Mathématique Jacques Hadamard*).

It is part of the Gaspard Monge Program for Optimization, operational research and their interactions with data science (PGMO - *Programme Gaspard Monge pour l'Optimisation, la recherche opérationnelle et leurs interactions avec la science des données*), launched by EDF and the FMJH. The focus of this IRSDI initiative is on data science and the many ways it may help industry.

Mathematicians and computer scientists from both the academic and industrial worlds can benefit from it. Projects are open to all academic researchers with no restrictions due to administrative or geographic location. Nevertheless, small teams with few but committed researchers will be favored.

Projects to be funded should be relevant to the field of data science (including machine learning, statistics, and computer science in relation to data analytics) and should be focused on solving industrial problems in the fields of energy or complex systems. A detailed list of suggested topics and methods is included in appendix.

Objectives

The objective is to support research projects through collaborative actions between academic researchers and industrial researchers or practitioners, focused on solving industrial problems in the fields of energy or complex systems. These projects are encouraged to be a kick-off for a future partnership between academic and industrial researchers.

Each proposal is thus formed by a pair given by an academic team and a partner company. The academic team must clearly identify a scientific leader, whose lab will manage the funding for the rest of the team.

The partner company must identify a corresponding member and will have to write a support letter describing the industrial challenges to be addressed, the data sets to be studied and the expected benefits of the collaborative research to be undertaken. *The true research (and not only development) nature of the project should be underlined in this letter*.

The partner companies do not necessarily need to be EDF or THALES, though these two companies are extremely willing to build partnerships through this sub-program.

Call for projects: schedule

- March 2018: publication of this call for projects
- Before final submission: prospective candidates are encouraged to get in touch with the PGMO board (via Gilles Stoltz, gilles.stoltz@math.u-psud.fr) to get some pre-submission feedback on the proposal
- May 2018: deadline for submission of the projects (link indicated below)
- July 2018: notifications of acceptance or rejection to the project leaders (after recommendations issued by the scientific committee and final decisions made by the executive committee of the PGMO)

Submission of projects: via EasyChair, at the URL
https://easychair.org/conferences/?conf=pgmo2018

Template: a submission template is provided at
https://www.fondation-hadamard.fr/fr/pgmo-calls-projects/2018-call-project
(Note that a single PDF file describing the scientific content of the project as well as all required administrative information is expected; it can be written in French or in English.)

Call for projects: rules

What follows is only a summary of the general PGMO submission rules, fully detailed at https://www.fondation-hadamard.fr/fr/pgmo-calls-projects/2018-call-project

Necessity of an industrial partner; note on data and on the codes

An IRSDI project consists of a pair composed by an academic team and an industrial partner.

Projects must emphasize the link with real data. Projects based on public/open data or on the creation of public/open data resembling to industrial or confidential data will be particularly welcome. If the data sets to be studied need to be collected and created first, the project leaders must describe the methodology to be followed and provide a timeline.
Codes are encouraged to be made available publicly.

Funding expectations / Budget rules

Projects duration should typically last 1 year (typically, October 2018 – November 2019).
Funding per project will be typically from 10 to 15 kEuros, but might reach, in specific cases, 20 kEuros.
We expect to fund about 6 or 7 projects.

All typical research expenses such as travels, computers, internships, invitations of researchers, purchase of data, etc., can be covered. Upon funding, an agreement will be signed between the main lab for the project and FMJH, and the lab will handle the obtained money. This lab or research institution or teaching institution must be from the academic world.

Commitment by funded project teams

PGMO / IRSDI being a program of the FMJH, which is part of the Paris Saclay University, all teams of funded projects will be asked to participate to at least one research event in the Saclay area, e.g., at the end of the project: typically, the annual 2-day-long PGMO workshop in Fall 2019.

Support by PGMO / IRSDI will have to be acknowledged in publications relative to funded projects.

A follow-up committee composed of representatives of the funding companies EDF and Thales may visit the project teams during the 2018-19 year.

Contacts

From the PGMO executive board:
- Gilles Stoltz (CNRS / Université Paris Sud, gilles.stoltz@math.u-psud.fr)

PGMO / IRSDI industrial sponsors
(May help to build projects based on the list of suggestions provided in appendix)
- Georges Hébrail (EDF, georges.hebrail@edf.fr)
- Camille Baysse (Thales, camille.baysse@fr.thalesgroup.com)

Scientific committee: Its composition can be found at the bottom of the page
https://www.fondation-hadamard.fr/fr/pgmo

List of suggested topics

Several topics are welcome, including but not limited to the following ones (listed with no order of priority). They were suggested by the industrial sponsors EDF and THALES, in the fields of electricity production, optronic systems, and C2 (Command and Control) Centers.

Axis 1: Time series (analysis and forecasting of time series)

Subproblem 1.1

Electricity load/production forecasting at small scale and short term; in particular, probabilistic forecasts; forecasts using large scale exogeneous data; and spatial forecasts

Efficient integration of a significant amount of renewable energy sources (solar and wind) in electricity markets is a present-day concern and key to success is the accuracy of forecasts. Such consumption and production forecasting has to be done locally since it depends heavily on local weather conditions and one could expect interesting gains in particular at short term horizon (e.g., a few minutes to days ahead) by modelling spatio-temporal dynamics. We propose to consider small geographic areas and exploit more available data at this scale (e.g., finer weather data and forecast, data related to human activity locally, images from ground fisheye images, satellite images, open data, data from social networks like Twitter or Facebook). The graph structure connecting the different local electric sub-networks can also be exploited. To improve the forecasts at different space and time scales, we propose to take into account the synergy between data sources and develop data fusion algorithms, leveraging when possible already learned representation of each data type.

For optimal bidding on markets, power grid operations such as dispatching, unit commitment or storage, probabilistic forecasts (density forecasts, uncertainties quantification, confidence intervals…) are essential. We suggest to develop probability forecasts at different times and geographical scales, and potentially connect it with stochastic or robust optimization.

New metering infrastructures as smart meters provide new and potentially massive information about individual (household, small and medium enterprise) consumption. Utilities now face the opportunity to improve modelling of individual consumptions and exploit this new source of information for forecasting,

profiling, designing new marketing offers or demand response programs. We propose to work on statistical modelling or machine learning approaches to achieve that. Both centralized and decentralized architectures for data storage and processing can be considered.

Subproblem 1.2.

Disaggregation of electricity demand at individual level (e.g., household/building) or small aggregates

With the deployment of smart meters in France beginning end of 2015 (ENEDIS Linky project), it will be easy and unexpensive to measure the consumption of each household or building every 10 minutes. Moreover, more detailed measures can be done inside buildings at a higher rate (e.g., every second) but still measuring the consumption of the whole building. There is a strong need to decompose this global consumption into the different usages of electricity (e.g., heating, water heating, washing machine, refrigerator, etc.), for instance to perform better demand-response (see below) or to give advice to customers about the energy efficiency of their equipment. There are mainly two ways to do so: either installing submetering devices for each appliance - which is quite expensive - , or applying machine learning/data analytics to the global consumption in order to perform this decomposition automatically. This is what we call disaggregation, which can be defined as several problems from the prediction of the presence of one equipment in the premises to the extraction of the consumption of each equipment continuously (for instance every hour). Deep learning approaches appear to be quite efficient to solve these problems.

Subproblem 1.3.

Privacy-preserving analysis of individual energy consumption (clustering, forecasting, scoring on anonymized/encrypted data)

As mentioned above, the deployment of Linky smart meters and new devices able to measure the consumption of electricity at a high rate (up to 1s) will enable data analytics processing with many valuable applications. However, such individual data is sensitive and may reveal private information about the life of occupants of the premises. The standard way to process data analytics on such data is to centralize data in a data warehouse/cloud and to apply algorithms on them. We are looking here for approaches which enable to apply some standard methods (e.g., clustering, forecasting and prediction) in a way that preserves privacy. This covers running data analytics on encrypted data, ensuring results preserve privacy, and anonymizing individual data before applying data analytics. Proposed approaches can be either centralized or distributed (for instance in a P2P environment): distributed data and computation require to design specific learning

algorithms (especially concerning the optimization part). Anonymization techniques which enable the publication of such individual and/or aggregated data are also of great interest.

Subproblem 1.4.

Generation of synthetic data featuring electrical networks in operation

When the new infrastructures of the electrical network are not developed yet, there is a need for generating synthetic data to design and study new algorithms. The need lies both in the generation of networks descriptions (topology, equipment, etc.) and in the generation of operations on the networks (consumption data at the individual and/or aggregated levels, network events, weather events, etc.).

In order to achieve a good electrical network generation we can consider data from existing network topology data, where probability values are generated from topological assets occurrences. This method is currently the easiest choice if the process can parse a large amount of existing network data. Another possibility is to generate artificial networks for very specific contexts and apply machine learning methods to distinguish real networks from artificial ones (adversarial learning).

Subproblem 1.5.

Operational data analysis in C2 (Command and Control) Centers

As main functionality for a C2 center, a radar system is a real-time device which provides information about the current air situation. C2 centers gather a lot of data coming from surface radars: first detections given plots from primary radars or target tracks in same cases. On the other hand, contextual data are available, e.g., meteorological data, cartographic data (elevation, nature of ground) or flight plans for both civilian and military aircrafts. Data analysis is necessary to propose new functionalities for the supervision of the system and also to propose new decision aids for operators. We can assess the global detection capability of the system, taking into account the evolution of detection probability area for each radar, for different periods of time. Patterns of normal behaviour could be established based on all tracks received by the system for a long period of time and anomalies could be found (plots that should be present in an area or on the contrary plots that should not be there because of the lack of detection in some areas). So, by using machine learning algorithm over a large period of time, it should be possible to access to a deep understanding of the radar behaviors and performance.

For that, the set of training data can consist of several millions of sample data to be used by KDEs (Kernel Density Estimators). Those KDEs have to be tuned in order to perform a trade-off between accuracy and CPU performance. A new approach could be the improvement of those trade-offs using flexible architectures allowing to deploy computation to one or more CPUs or GPUs in several devices.

Subproblem 1.6.

Real-time data analysis and online learning for optimal decision in energy markets

In an ever more competitive environment, energy market actors have to cope with new information in real time. Being able to gather new sources of information, parse them efficiently and integrate them into decision-making processes is a key activity. This includes, e.g., data stream processing techniques, online learning algorithms, reinforcement learning, etc.

Axis 2: Predictive maintenance

Subproblem 2.1

Predictive maintenance in power plants and electrical networks

Maintaining the high reliability of power plants (especially nuclear plants) and electrical networks requires efficient and timely access to information on asset performance and condition, as well as processing of this information to make cost-effective decisions regarding maintenance priorities. The expectation of continued constraints on maintenance resources and staffing within the power industry calls for the establishment of a strategy for the effective condition-based maintenance of components. In general, this entails:

- the timely detection of abnormal operation conditions (condition monitoring)
- the identification of the causes of abnormality (fault diagnosis)
- the prediction of the remaining useful life in the given abnormal conditions (fault prognosis)

The main challenges to address in this context are the following:

- Automatic or computer-assisted modelling of the various data sources (description of equipment and linked equipment, context of operation, maintenance operations, monitoring information, alarms, logs, sensor values, computation results from physical models and simulation codes, expert knowledge, etc.), in order to enable easier integration of such data. Beyond data fusion methods, semantic web approaches like RDF/OWL are good candidates to do so (i.e. to build a semantic data lake).
- Automatic population of semantic data lakes from existing sources of data, this includes for instance extracting links between entities and/or objects from unstructured information (e.g., textual data).
- Easy navigation in semantic data lakes, including exploratory data analysis/unsupervised learning from complex and heterogeneous data (graph data, textual data, multi-dimensional time series from sensors, images, video, physical models, expert knowledge, etc.).
- Monitoring, diagnostics and prognostics approaches using predictive models and based on the exploitation of complex and heterogeneous data (graph data, textual data, multi-dimensional time series from sensors, images, video, physical models, expert knowledge, etc.). A key point is the presence of several problems in data: missing/not-applicable/censored data, inconsistencies, uncertainty, presence of noise, too small dataset, etc.
- For some systems, no or only few sensor data is available or it is difficult to clearly identify the link between monitoring data and the physical behaviour of the equipment. Such problems involve for example - but are not limited to - signals or images coming from non-destructive testing of complex

material, or transient signal measures at specific time in the process. These material exhibit either complex structure such as concrete, complex geometry or complex degradation type.

An additional topic which is specific to electrical networks is related to crisis management, still using complex and heterogeneous data:

- Forecast and identification of crises from past data
- Models to predict impact of weather events

Subproblem 2.2

Predictive maintenance of complex systems fleets

Similar issues as described above apply to complex systems, which are nowadays also equipped with several sensors (e.g., several dozens of sensors in complex optronic observation systems or in scientific lasers). These sensors are mainly used to control equipment but can also be used to reduce maintenance costs by applying data analytics to historical sensor data. As indicated above, applications cover detection, diagnostic and prognosis for maintenance tasks using multi-dimensional data analytics approaches, for example the identification and selection of data or combinations of data which have consequences on the behaviour or on some "health" states of the system. We encourage approaches which take into account additional information to sensor data, in particular, events and textual data (e.g., maintenance logs), or the links which exist between the different pieces of equipment. Applications also cover, for those data which have been identified as decisive for the system health, the development of dynamical models for the links between the data and the system health, and the use of these models for assessing the "best" maintenance operations at the "best" date.

In some applications, the maintenance problem can be defined in terms of an availability ratio for the customer, provided that the customer has several systems (typically some dozens or some hundreds). The above applications (detection, diagnosis and prognosis) need then to be generalized to that case.

Subproblem 2.3

Certification, verification and validation of machine learning methods

The best predictive methods are usually the worst understood (i.e. deep learning) but for critical tasks such as maintenance or control in the industry sector we need certified approaches. There is a strong need for techniques enabling to verify/validate/certify the behaviour of algorithms which involve learning, optimisation and statistical approaches.

Axis 3: Optimization and control of complex systems

Subproblem 3.1

Design of experiments/novelty search

In health monitoring systems, detection of actual causal data, as explained above, can be difficult due to the lack of relevant data. Analysis/modelling involves identifying areas of the search space in which the behavior of a system is poorly known, and in which the surface response is misspecified. This analysis has to be done in restricted conditions, due to cost and usage constraints.

Subproblem 3.2

Demand-response management

As mentioned above, the development of renewable energies (solar/wind) introduces instability in the electric power system. Since electricity storage remains expensive, a way to balance demand and production is to ask some customers to reduce their consumption in real time, for instance when there is a decrease of solar generation. Demand-response management is the process of selecting the customers who are asked to reduce their consumption, sending them the right signal (control, price…), of monitoring their response, and of updating continuously this process. Problems raised for this topic may find solutions in various domains, such as complex event processing and stream mining, game theory with exploration/exploitation dilemma and optimization techniques.

Subproblem 3.3

Optimization in the context of stochastic forecasts/uncertainties; application to unit commitment, dispatch, capacity expansion planning

The balance between electricity demand and generation is mainly operated in 2 steps: (1) forecasting of demand and more recently forecasting of generation with the development of renewable energies; (2) optimization of the ON/OFF planning of the different power plants (nuclear, thermal, hydraulic plants). Recent progress enables to consider stochastic forecasting (see above) and stochastic optimization algorithms. Since forecasting and optimization are used jointly to solve this problem, a better integration between the two components should improve the performance of the whole process. Another direction is to

use machine learning approaches in addition to optimization algorithms to cope with very local, specific or unexpected constraints (using historical data featuring both the optimization output and the manually adjusted schedules).

Subproblem 3.4

Learning of models for Command/Control or tracking

Learning of models for Command/Control or tracking can be viewed as a dynamic optimization problem which is often corrupted with noise. In this context, the problem can be approached by stochastic black-box optimizers like the state-of-the art CMA-ES algorithm of Hansen et al. (2001, 2009). Invariance of those type of approaches inherited from information geometry makes them particularly robust to tackle dynamic noisy optimization problems. Yet, open questions in the case of a static environment with noise are related to what are good statistical measures to detect and quantify the noise and more critically how to handle the noise efficiently once the quantification has been done. Other questions are related to how the specific dynamic environment of control problems affect the treatment of the noise.

Subproblem 3.5

Markov Decision Process to learn optimal control of unknown dynamics, or strategies for repeated complex or combinatorial optimization

Recent advances in reinforcement learning, statistical high dimensional approximation and in algorithmic exploration have led to amazing successes such as AlphaGo or Libratus. In the energy sector such methods could be used to find optimal control strategies for unknown dynamics, for instance power plants involving uncertainty such as weather or changing performance of machines, or smart home control. It could also be used to leverage learning for repeated complex or combinatorial optimization, for instance optimal core design of nuclear power plants, energy planning smart grid management involving multi energy systems and electric vehicles.

Axis 4: Data visualization, images

Subproblem 4.1

Visualization of energy consumption time series and multi-sensors in power plants, households and buildings and electric networks

When large amounts of data are available, it is now recognized that the combination of data analytics algorithms and visualization techniques is valuable. We are looking here for visual analytics approaches when the end user is an expert of the domain (e.g., an engineer, a researcher) and for data mainly available in the form of multidimensional time series: electric power consumption, sensors in power plants, devices in households.

Subproblem 4.2

Image recognition and indexing, in particular for power plants inspection (inside/outside), for building energy efficiency analysis, for photovoltaic power generation forecasting, for detecting tree pruning needs around power networks

The last few years have shown much progress in image recognition and indexing, for instance for face recognition in pictures (e.g., Facebook) or address recognition (e.g., from StreetView). This progress is mainly due to the increase of available data, the improvement of machine learning algorithms (e.g., deep learning) and processing capabilities. Several similar problems have been identified in the electricity fields, for instance: recognition of machine or parts identifiers from plaque pictures, solar irradiance forecasting using fisheye and satellite images, identification of thermal properties of houses from their pictures, estimation of tree pruning needs around power lines from satellite or drone pictures, or inspection of power plants with robots and drones. These images have specific properties with respect to standard image recognition databases since they are taken in various potentially degraded environments (underwater, low light or bad weather conditions and with various, often bad quality cameras).

Subproblem 4.3

Image indexation and classification and image data basis augmentation for anomaly detection

Anomaly detection in more or less structured images can be used for target detection in images of a complex background (cloud, ground, urban). Some previous works have shown that, starting with real images, it is possible to classify the images for improving the detection task. It is also desirable to artificially augment the image data basis, because real images in large panels of conditions could be expensive to acquire.

List of suggested methods

We also suggest relevant methods; please keep in mind that this list is not exhaustive, and any unexpected but meaningful methodology is more than welcome:

- Ensemble methods, in particular online and adaptive ones;
- Integration of stochastic optimization and probabilistic forecasts;
- Spatial and temporal forecasts;
- Probabilistic forecasts;
- Deep learning, convolution networks;
- Cryptographic machine learning (supervised, unsupervised, semi-supervised);
- Bandit methods, exploration/exploitation dilemma;
- Hadoop and/or scaling environments for data science;
- Frameworks for data flow graphs processing;
- Fast machine learning with indexing;
- Typical and atypical patterns; breakpoint detection; weak signals;
- Multivariate time series, including Hidden Markov Models (HMM): analysis of causality, graph methods, functional links between components
- « Hybrid » models for time series, e.g. Piecewise Deterministic Markov Models
- Design of experiments;
- Markov decision process;
- Visual analytics;
- Image statistical models, synthesis, classification;
- Integration of human experts knowledge and analysis;
- Novelty search, Gaussian processes, design of experiments with constraints;
- Model-based clustering;
- Model selection in high dimension; functional data analysis.