Research initiative in industrial data science (IRSDI)

Call for projects – March 2020

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 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, within the fields of energy.

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.

Projects to be funded should be relevant to the field of data science (including machine learning, statistics, optimization and computer science in relation to data analytics) and should be focused on solving industrial problems in the field of energy systems.

Objectives

The objective of the 2020 IRSDI call is to support one research project by granting a one year postdoc position (possibly extendable next year), in order to develop collaborative actions between academic researchers and industrial researchers or practitioners, focused on addressing the challenge of ***automated machine learning*** in industrial applications within the field of energy.

The subject of ***automated machine learning*** is obviously quite large, both in terms of data types (time series, text, images, videos, graphs, …), of scientific approaches (choice of machine learning models, optimization of learning parameters, Neural Network Architecture Search, transfer learning, meta-learning,…) and of industrial use cases. A list of topics of interest is given at the end of this document.

Proposers are strongly encouraged to contact the experts at EDF R&D on each subject in order to have a thorough knowledge of the issues and research work already completed or planned for on each topic. To this end, please contact the PGMO board (mailto: pgmo@fondation-hadamard.fr).

The academic team must clearly identify a scientific leader, whose lab will manage the funding for the rest of the team.

Call for projects: rules

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

Note on data and on the codes

Projects must emphasize the link with real data in relation with EDF R&D. Projects based on public/open data or on the creation of public/open data resembling to industrial or confidential data will also be welcome. If the datasets 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

Within this call for project, we expect to fund One year of Postdoc (salary+ environment). Details on the general PGMO call.

Commitment by funded project teams

All funded projects will be asked to participate, at the end of the project, to the annual review of IRSDI projects within the PGMO workshop in Fall 2021 (typically lasts one morning and part of the afternoon).

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 company EDF will get in touch with and may visit the project teams during the 2020-21 year.

Contacts

PGMO / IRSDI industrial sponsor

(Will help to build projects by pointing to interested members of EDF)

- Georges Hébrail (EDF, georges.hebrail@edf.fr)

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

Scientific topics

As already mentioned, the subject of ***automated machine learning*** is obviously quite large, both in terms of data types (time series, text, images, videos, graphs, …), of scientific approaches (choice of machine learning models, optimization of learning parameters, Neural Network Architecture Search, transfer learning, meta-learning, …) and of industrial use cases.

The proposed project should address generic approaches that are applicable to several use cases of interest in the field of energy.

Scientific approaches

Automated machine learning can be considered for every family of models either supervised or unsupervised, from classical models (ex. decision trees, logistic regression, random forests, SVM, forecasting models, clustering, …) to more recent models (ex. deep learning, representation learning, …). We suggest to focus on recent deep learning approaches as proposed by Google[1] or Baidu [2]. In this context, there are several generic approaches to consider:

- Automated data processing/cleansing, model selection or hyperparameter selection
- Network Architectures Search (NAS) with different approaches (random/grid/evolutionary/greedy search, bayesian optimization, reinforcement learning) or architecture embedding methods (differentiable NAS)
- Transfer learning approaches with the construction and the use of pre-trained networks, representation learning and domain adaptation

Both supervised and unsupervised approaches are within the scope (ex. for instance anomaly detection, automatic tagging/clustering within a large set of time series). Resource-aware (green AI) and privacy-aware AutoML methods will also be welcome.

Data types

All data types are present in the energy domain, we list them by decreasing priority:

- Time series representing electric power consumption (either individually or aggregated) or time series produced by sensors (ex. installed in power plants or electric networks), potentially at a high temporal resolution

[1] See https://cloud.google.com/automl?hl=fr and https://arxiv.org/abs/1806.09055
[2] See https://baiduautodl.com/

- Textual data resulting from customer relationship management (CRM) or present in maintenance reports and equipment description, news (financial news, energy related news…)
- Images or videos describing either EDF installations or customer/cities equipment
- Graph data describing either the electric network or social networks between customers or employees of the company

Use cases

We list below some typical use cases but we recall that we expect generic approaches applicable to several of these use cases:

- Electricity load and production forecasting at small scale and short term
- Disaggregation of electricity consumption by usage at individual level or small aggregates (e.g., household/building, NILM – Non-Intrusive Load Monitoring)
- Power plants 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, videos, …).
- Image recognition and indexing for power plants inspection, building energy efficiency analysis from building pictures, short-term solar energy forecasting with fisheye cameras and/or satellite images
- Text understanding and generation in Customer Relationship Management (CRM)
- Demand-response management (optimal sequential selection of customers who are asked to reduce their consumption in order to facilitate load balancing).
- Construction of meta-models or reduced models using numerical simulations based on physical models