H.F.R.I.
Hellenic Foundation for Research & Innovation

Description of the funded research project

1st Call for H.F.R.I. Research Projects to Support Faculty Members & Researchers and Procure High-Value Research Equipment

Title of the research project: “Data Science for All”

Principal Investigator: Ioannis Tsamardinos

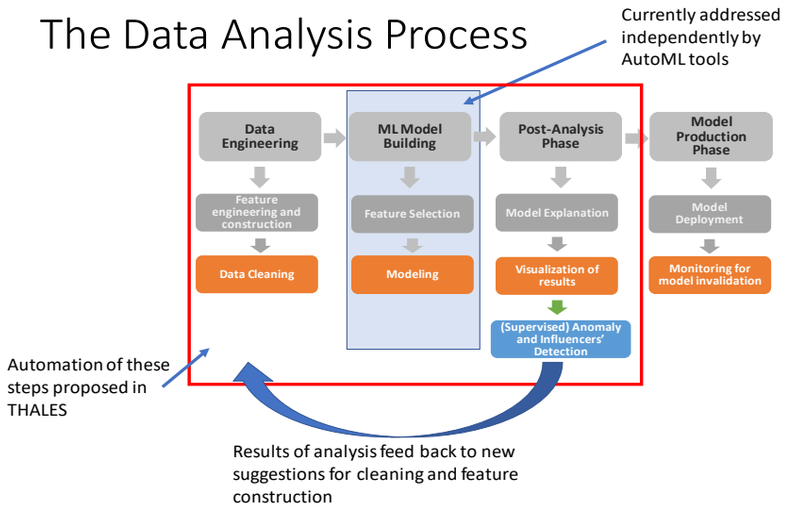
Reader-friendly title: “THALES”

Scientific Area: Mathematics and Information Sciences

Institution and Country: Greece

Host Institution: Institute of Computational and Applied Mathematics, Foundation for Research and Technology Hellas (IACM-FORTH)

Collaborating Institution: Computer Science Department, University of Crete



Budget: 169.515,50 Euro

Duration: 36 months

H.F.R.I. Hellenic Foundation for Research & Innovation

Research Project Synopsis

Data Science promises to disrupt modern science, business, and societies. Yet, it is still largely a manual, time-consuming, and error-prone process, requiring significant expertise. As a result, almost every data analytics software platform vendor, a wave of startup companies in the US and EU, as well as prominent academic research groups now focus on the automation of the underlying data analysis processes. Unfortunately, we are far from this goal. Particularly, Data Engineering (data cleaning, preprocessing, and feature construction) - considered to be the most time-consuming step of the analytic process - has not been automated. In addition, it is treated independently of the subsequent Data Analytics. In THALES we put forth a research program to bridge Data Engineering with Data Analytics into an integrated algorithmic framework, automate the full pipeline of analytics, and make Data Science accessible to All. Particularly, we focus on using predictive-model building to facilitate targeted cleaning of records, automated feature construction from Relational Data Bases, and design massively-parallel, scalable, multiple, feature selection algorithms. Most importantly, we emphasize integration of the above into a single algorithmic framework and Machine-Learning-as-a-Service implementation. Results are of scientific interest, but also highly commercially exploitable; the PI has already in place a University spin-off company to serve as the vehicle of commercial exploitation.

Project originality

There are several challenges and gaps in our theoretical, algorithmic, and engineering understanding addressed by THALES:

- Data Engineering and Model Building have been addressed in isolation by DB and ML communities using independent, sequential steps in data analysis pipelines. Thus, potentially huge volumes of data are cleaned unnecessarily even when they do not participate in the training of the final model; similarly, a plethora of features is usually constructed even if they have no predictive value and thus, they will not be selected by the feature selection step.
- Analytical processes need to be repeated through trial-and-error with every new cleaning operation or new idea for feature construction. Thus, huge volumes of data are potentially extracted again and again from a database to be stored as flat files and run the modeling algorithms anew.
- Analytical processes typically concerns all available data; thus, it is not scalable to Big Data. Ideally, one would like to process (automatically) only the data subsets that suffice for the ML algorithms to make robust decisions.

Long-Term Vision	Targeted scientific breakthrough
Bridge Data Engineering (Cleaning and Feature Construction) with Machine Learning Perspectives into an integrated algorithmic framework	Create theoretical and algorithmic underpinnings for interweaving solutions of Data Engineering and Model Building tasks to create Automated Data Cleaning and Feature Construction pipelines
Automate the entire pipeline of predictive analytics	Create algorithms for simultaneous Cleaning, Feature Construction, and Feature Selection that are fully automated, efficient, and scalable to Big Data
Make Data Science accessible to All types of users, increasing productivity by at least 10-fold	Encapsulate the algorithmic advances to industrial-strength, distributed, robust and versatile implementations that are offered as MLaaS

Expected results & Research Project Impact

Objective 1: Automate Data Cleaning in a closed loop. The main idea is to forgo untargeted cleaning of all data and values that do not affect the final machine learning model, before feeding the data to the learning algorithms in a subsequent, independent step. Instead, the goal is to design Model-aware cleaning algorithms that exploit the inter-play with the model-learning algorithms to identify both the training examples, as well as the features that are potential “dirty” and require some cleaning action to take place.

Objective 2: Automate Feature Construction for Relational Data Bases. The goal is to design algorithms that automatically constructs potentially useful and informative features, by performing SQL queries (joints) directly on the RDB schema, before feeding them to the feature selection and modelling algorithms. The feature construction will exploit statistical heuristics that remove the need from the human expert to guess which features to explicitly construct, through trial-and-error.

Objective 3: Scale-up Feature Construction and Selection for Big Data. The goal is to design algorithms that massively parallelize feature selection computations, while minimizing network communication and return results that well-approximate these computations performed in a centralized way. In addition, such algorithms should not have to go through all training examples because in very large datasets, this results in linear time algorithms as a lower bound, which is still unacceptable for Big Data. Instead, the algorithms should examine only the smallest portion of the data that suffices to decide which features to retain (select) and which features to filter out.

Objective 4: Integrate Data Cleaning, Feature Construction, and Feature Selection for Big Data, in a closed loop. The goal is to integrate the above steps into a single framework that fully automates a larger portion of the data analysis process than what is possible now, at least for data stored in an database.

The importance of this funding

Although, the scale and demand of data analysis does not cease to increase, data analysis remains mostly a labor intensive, time consuming, and error prone activity with technical experts in the loop. Despite significant R&D efforts to strengthen some self-service analytics by automating data preparation or model building, the development of high-quality Big Data analytics pipelines is still an inherent iterative process that requires highly skilled analysts combining statistics, optimization and machine learning knowledge with database and programming skills. The proposed work is ambitious in many aspects. First, it will bridge together two, currently separate and distinct, research perspectives and methods from database and machine learning communities. Second, it will explore their common goals regarding widespread and large-scale data analytics to build a research agenda of novel problems around on the fly model learning and data engineering. Third, it will capitalize the industrial momentum concerning automation of data science tasks expressed in terms of real use cases, datasets, etc. In this respect, it is expected to produce novel algorithms, theory, systems, and tools that will be published in first-tier venues but also organize specialized conferences, workshops, forums, and other research and scientific activities.

H.F.R.I.
Hellenic Foundation for Research & Innovation

COMMUNICATION

185 Syggrou Ave. & 2 Sardeon St. 2
171 21, N. Smyrni, Greece
+30 210 64 12 410, 420
communication@elidek.gr
www.elidek.gr