Distributed Automated Research Platform

ROBHOOT

Current Team:
Carlos J. Melián[1] Switzerland C1
Victor Eguíluz[2] Spain C2
Tentative Team:
Swiss Science Data Center[3] Switzerland C1
Eawag IT (Harald)[4] Switzerland C1
iDiv IT ()[5] Germany C3
Inca (Nterminal, Yupana)[6]
Benoit Baudry (DIVERSIFY-project)[7] France C4

1 Eawag Center for Ecology, Evolution and Biogeochemistry, Switzerland
2 Inst. de Física Interdisciplinar y Sistemas Complejos, IFISC (CSIC-UIB), Palma de Mallorca, Spain.
3 Tentative team

Contents

1 Funding scheme B

Call ID: FET-Open Challenging Current Thinking
H2020-FETOPEN-01-2018-2019-2020

Specific Challenge:

To lay the foundations for radically new future technologies of any kind from visionary interdisciplinary collaborations that dissolve the traditional boundaries between sciences and disciplines, including the social sciences and humanities. This topic also encourages the driving role of new actors in research and innovation, including excellent young researchers, ambitious high-tech SMEs and first-time participants to FET under Horizon 2020 from across Europe.

Scope:

Proposals are sought for cutting-edge high-risk / high-impact interdisciplinary research with all of the following essential characteristics ("FET gatekeepers"):

Radical vision: the project must address a clear and radical vision, enabled by a new technology concept that challenges current paradigms. In particular, research to advance on the roadmap of a well-established technological paradigm, even if high-risk, will not be funded.

Breakthrough technological target: the project must target a novel and ambitious science-to-technology breakthrough as a first proof of concept for its vision. In particular, blue-sky exploratory research without a clear technological objective will not be funded.

Ambitious interdisciplinary research for achieving the technological breakthrough and that opens up new areas of investigation. In particular, projects with only low-risk incremental research, even if interdisciplinary, will not be funded.

The inherently high risks of the research proposed shall be mitigated by a flexible methodology to deal with the considerable science-and-technology uncertainties and for choosing alternative directions and options.

The Commission considers that proposals requesting a contribution from the EU of up to EUR 3 million would allow this specific challenge to be addressed appropriately. Nonetheless, this does not preclude submission and selection of proposals requesting other amounts.

Expected Impact:

Scientific and technological contributions to the foundation of a new future technology Potential for future social or economic impact or market creation. Building leading research and innovation capacity across Europe by involvement of key actors that can make a difference in the future, for example excellent young researchers, ambitious high-tech SMEs or first-time participants to FET under Horizon 2020.

Deadline: ID: September 18, 2019 17:00:00 Brussels time FETOPEN-01-2018-2019-2020

2 Summary

Public funded data is the norm in science and engineering landscapes. Yet, distributed and open-source automated knowledge-inspired technology accounting for the interdisciplinary research cycle remains challenging in science and engineering. We propose a decentralized automated research platform to facilitate reproducibility in a knowledge-inspired society, technology and science. The platform will be implemented in two stages. First, a testnet biodiversity research prototype accounting for data integration and inference to visualization and reporting generation. Second, a mainnet platform integrating literature discovery, data integration, inference, validation, visualization and reporting generation. Decentralized and open automated research platforms can strongly contribute to a knowledge-inspired governance to help take informed decisions in solving complex social problems merging interdisciplinary science and technology.

Keywords: Knowledge-inspired society, Data integration, Multilayer networks, Deep explicability-based learning, Transparent governance, Bayesian inference. Probabilistic Bayesian networks.

3 State of the art

Radical vision:

- Decentralized, open and scalable automated research platform accounting for the research cycle, from literature and data integration to visualization and report generation.

Breakthrough technological target:

- Transparent automated data-rule-knowledge-inspired technology fully accounting for the interdisciplinary research cycle.

Ambitious interdisciplinary research:

- Mixing researchers, engineers and society to build an open and transparent data-rule- and knowledge-inspired science, technology and society.

————-How to solve it?———
1. The problem:
Why? Centralized, Non-transparent and non-reproducible vs decentralized, transparent and reproducible
Is it radikal?
Is it viable? Efficient?
Which are the Knowns?
Unknowns?
2. Our own plan
How much time?
Existing gaps? Why?
Data available? Methods? Main gaps
Our own method? Why?
3. Carry on the plan
Quantitative assessment
The maths
4. Solutions (Milestones)
5. Go backwards (Reproducibility)
Is it correct? Reproducible? Scalable? Decentralized? Safe?
Apply to other problems? Datasets?
————————————-

Automation is rapidly occurring in many fronts, from robotics and investments to gaming and ecommerce. Many technologies are in a era of massive data accumulation, integration and pattern detection. Yet, obtaining insights from such an integration accounting for open science, full reproducibility, inference and prediction power is at a very incipient stage (Ioannidis, 2005; Reichstein, M. et al., 2019). There are many challenges when aiming to integrate data, inference and prediction. For example, sampling design and experiments (Voelkl et al., 2018), randomizations to achieve solid statistics , and process- or pattern-based model selection and inference just to name a few. All these steps require many intermediate decisions that make the scientific process challenging to decentralize, repeat, replicate, and reproduce. Currently, there are many protocols and platforms automatizing partial steps of the (Table 1 summarize a non-exhaustive list of automated platforms).

Reproducibility, robustness and minimizing bias across the different stages of a research platform are two of the desire properties of automated research platforms. Reproducibility guarantees the future improvement of the results and many programming languages currently offer tools to facilitate reproducibility (i.e., Jupyter notebooks). Automated research platforms will facilitate tracking the explored paths (i.e., the within and between layer interactions, Figure 1) and can produce statistics on how close each path is to the empirical patterns.

The following are six features of open automated research platforms that might play a leading role in moving towards a information-inspired science, society and technology: 1) Building the science of science to provide quantitative statistics of interdisciplinary science (Figure 1 for the arquitecture of an automated research platform); 2) Identifying systematically and accounting for bias and uncertainty in inference; 3) Exploring prediction and explanatory gradients to gain sinergy between AI predictive approaches and explanatory power to complex problems (Figure 2); 4) Identifying gaps in patterns not explored consequence of lack of syntesis in interdisciplinary research, 5) Allowing for reusability, repeatibility, replicability and reproducibility along the many paths in the scientific enterprise, and 6) Building on rapidly evolving open-source computing programming languages to facilitate the decentralization, scalability and integration of the scientific process (Table 2 and Figure 3).

4 Milestones

4.1 Milestone 1: $\mathcal{ROBHOOT}$ testnet

Building a functional testnet case for an automated research platform. We will use a Biodiversity research case for this testnet as a showcase of how to gain insights about knowledge-inspired technology (See Robhoot testnet box).

ROBHOOT **testnet**
Develop automated packages for each layer (See Figure 3 and Table 2 for a prototype in the computing language Julia. The platform will have access to data from both centralized and decentralized platforms in Biodiversity research [a] .

4.1.1 Milestone 1.1: Data Integration (*DAADI*)

4.1.2 Milestone 1.2: Complexity Reduction (*GOCORE*)

4.1.3 Milestone 1.3: Pattern-Process Inference (*PROPENCE*)

4.1.4 Milestone 1.4: Validation (*VATION*)

4.1.5 Milestone 1.5: Visualization (*VITION*)

[a] https://github.com/melian009/Robhoot/blob/master/resources/databases.md

4.2 Milestone 2: *ROBHOOT* mainnet

Develop a platform fully integrated in decentralized network platforms [1]

ROBHOOT **mainnet**

4.2.1 Milestone 2.1: Data Integration

4.2.2 Milestone 2.2: Complexity Reduction

4.2.3 Milestone 2.3: Pattern-Process Inference

4.2.4 Milestone 2.4: Validation

4.2.5 Milestone 2.5: Visualization

[1] https://golem.network/

References

Ioannidis, J. P. A., 2005. Why most published research findings are false. PLoS medicine 2:e124. URL http://www.ncbi.nlm.nih.gov/pubmed/16060722.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, 2019. Deep learning and process understanding for data-driven earth system science. Nature 566:195–2024.

Voelkl, B., L. Vogt, E. S. Sena, and H. Würbel, 2018. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biology 16.

5 Tables

Table 1	
Automated platforms	**Webpage**
Nakamoto Terminal	https://www.nterminal.com
BigQuery	https://cloud.google.com/bigquery/
Automated statistician	https://www.automaticstatistician.com/index/
Modulos	http://www.modulos.ai/
Google AI	https://ai.google/
Iris	https://iris.ai

Table 2: Prototyping a script workflow in $\mathcal{ROBHOOT}$

SUMMARY==========
This is a prototype for a script workflow to automate interactions among data search, parsing, integration, database, cleaning, data complexity reduction, pattern and process inference, validation and visualization. The script is based in two types of packages: backbone and specialized packages. Backbone packages (B) connect intra- and inter-layer algorithms to automatically run the workflow. Specialized (S) packages feedback with backbone packages to run specific tasks: parsing, likelihoods, inference, plotting, visualizing, etc.

==========
Layers==========
DATA INTEGRATION: D
COMPLEXITY REDUCTION: C
PATTERN-PROCESS INFERENCE: P
VALIDATION: VA
VISUALIZATION: VI
==========
EXAMPLE with julia ==========
Julia packages:
https://github.com/melian009/Robhoot/blob/master/packages.md

WORKFLOW NETWORK----------

data.search D S − − − − − −− > Retriever.jl **M1**
parsing.data D S − − − − − −− > Query.jl **M1**
data.to.table D S − − − − − −− > MySQL.jl SQLite.jl Clickhouse.jl **M1**
data.julia D S − − − − − −− > DataFrames.jl **M1**
table.comp.reduction C B − − − − − −− > TensorFlow.jl lm4.jl Clustering.jl OnlineAI.jl LightGBM.jl
pattern.detection P S − − − − − −− > TensorFlow.jl DataVoyage.jl DataFitting.jl Mocha.jl DeepQLearning.jl Flux.jl AnomalyDetection.jl
proccess.simulation P S − − − − − −− > Simjulia.jl Agents.jl JuliaDynamics.jl Zygote.jl
pat.proc.infer P S − − − − − −− > mads.jl temporal.jl GlobalSearchRegression.jl BlackBoxOptim.jl JuMP.jl GeneticAlgorithms.jl NaiveBayes.jl Mamba.jl ABC.jl ApproxBayes.jl DynamicHMC.jl
validation.pat.proc VA S − − − − − −− > mads.jl LearningStrategies.jl Mamba.jl ABC.jl Measurements.jl
visualiztion.pattern.process − − − − − −− > Makie.jl VegaLite.jl

6 Figures

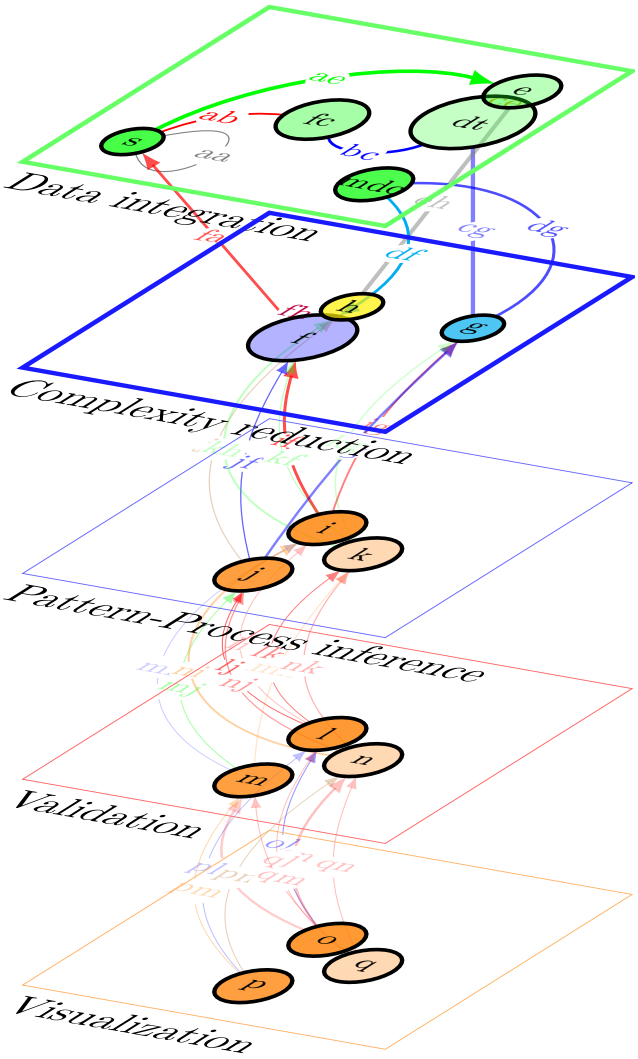

Figure 1: A cartoon of a five layer automated research platform: Data Integration, Complexity reduction, Pattern-process inference, Validation, and Visualization. Nodes and links represent algorithms and interactions between two algorithms, respectively. For example, the figure shows five algorithms in the layer Data integration (**a**, **b**, **c**, **d**, and **e**). Algorithm **a** interacts with algorithm **b** and **e** in the same layer (intra-layer connections) and with algorithm **f** from the second layer (inter-layer connection), Complexity reduction. The cartoon represents many intra- and inter-layer connections to solve a problem. The paths can be quantified by many metrics each producing a distribution of automated solutions. This distribution can be analyzed with the ones used for a specific domain in science, the science of science of a domain, to quantify properties as robustness, reproducibility and bias of a domain.

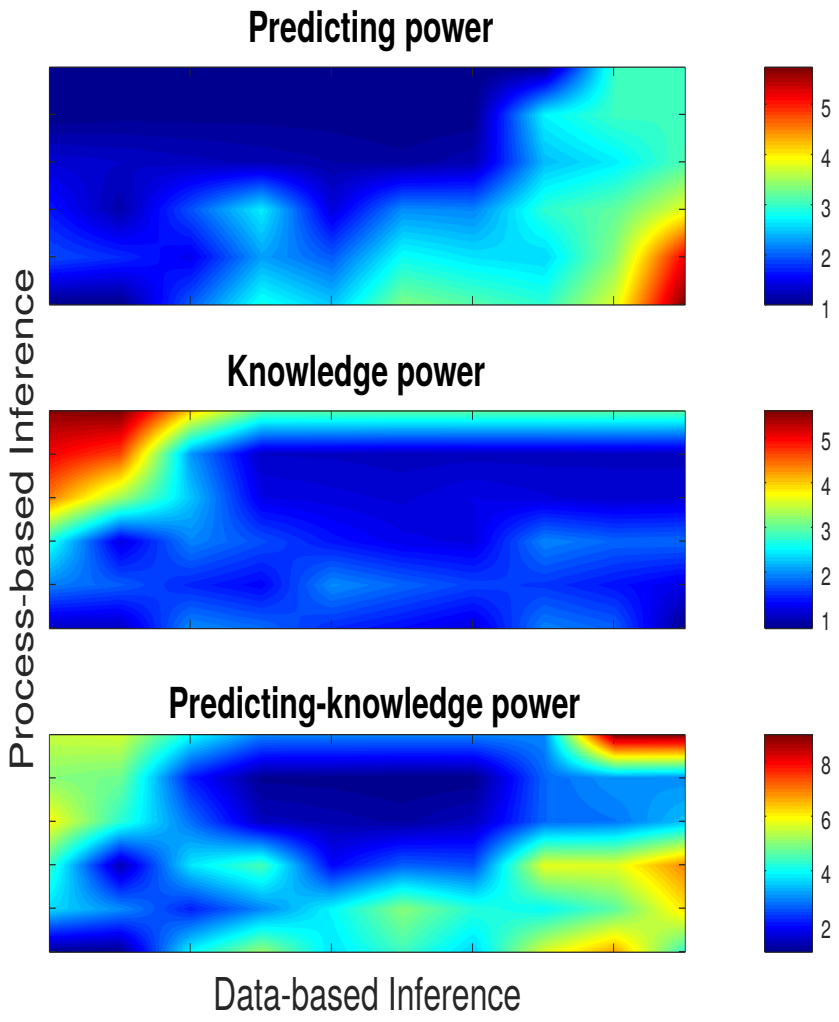

Figure 2: Predicting and knowledge power can be joined to drive open knowledge-inspired societies, technology and science. This figure shows a cartoon of the predicting power (top), the knowledge power (middle), and the predicting-knowledge power (bottom). x- and y-axis represent data-based inference (i.e., gradient of AI methods from low (left) to high (right) predictive power) and process-based inference (i.e., gradient of process-based methods from low (bottom left) to high (top left) knowledge power). The gradient of predicting power (top) shows a hot spot red area in the bottom right highlighting the region where AI methods best predict any given empirical data. The gradient of knowledge power (middle) shows a hot spot red area in the top left highlighting the region where the best mechanistic understanding of a problem occur. The predicting-knowledge power (bottom) integrates the sum of the two previous maps highlighting a red hot spot where interdisciplilnary research joining predicting and knowledge power of the empirical data occur.

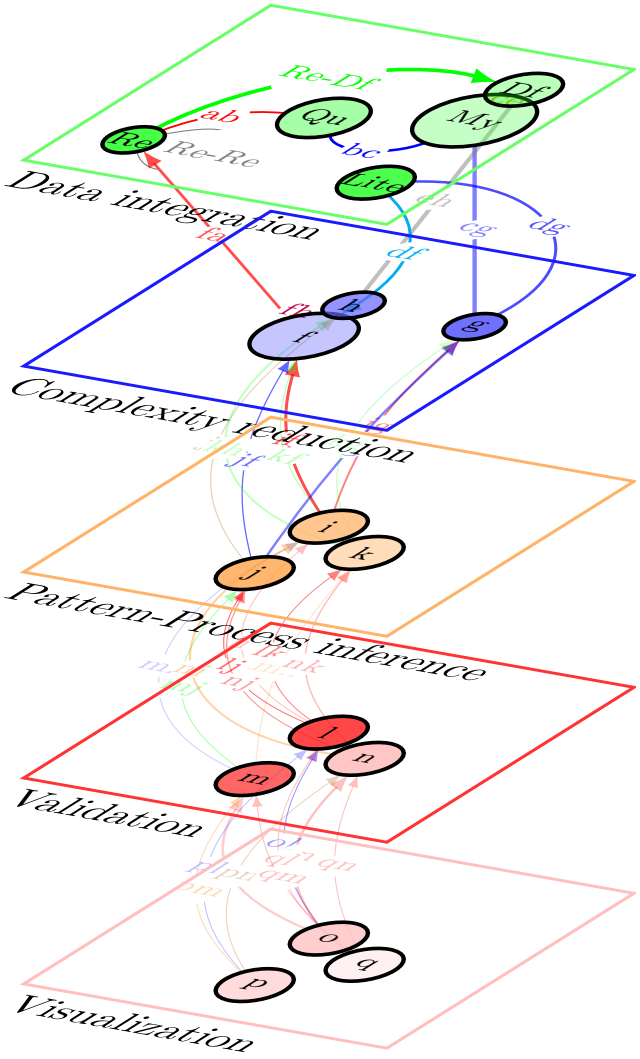

Figure 3: A julia prototype of an automated research platform. Nodes and links in each layer represent julia packages and interactions between two packages, respectively. The figure shows julia packages within each layer. For example, the layer Data integration contains the packages "Retriever.jl" (**Re**), "Query.jl" (**Qu**), "MySQL.jl" (**My**), "SQlite.jl" (**lite**), and "DataFrames.jl" (**df**).

Index