Automated research platforms

Ali R. Vahdati[1, *]
Charles N. de Santana[2]
Alejandro Rozenfeld[3]
Carlos J. Melián[4, *]

1 Department of Anthropology, University of Zurich, Switzerland
2 Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland
3 Conicet-cificen-intelymec, University of the Center of Buenos Aires, Argentina
4 Department of Fish Ecology and Evolution, Center for Ecology, Evolution and Biogeochemistry, EAWAG, Kastanienbaum, Switzerland

* These authors contributed equally to this work.

Contents

1 Abstract

High-resolution and heterogeneous data coming from many sources is standard in science, engineering and investment landscapes. Yet, automated inference providing insightful patterns and processes integrating databases with analytical frameworks remains challenging. In this work we propose a prototype of an automated research platform to discuss the challenges for automated workflows to integrate data and pattern-process-based inference accounting for many sources of uncertainty. Automated research platforms can strongly contribute to the science of science to take better informed decisions in research, management and investment landscapes.

Keywords: data integration, multilayer networks, deep process-based learning, approximate Bayesian computation, inference.

2 Introduction

Automation is rapidly occurring in many fronts, from robotics and investments to gaming and ecommerce. What about science? Science is in a era of massive data accumulation, integration and pattern detection. Yet, obtaining insights from such an integration accounting for 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 require many intermediate decisions that make the scientific process challenging to repeat, replicate, and reproduce. Currently, there are many protocols and platforms automatizing partial steps of the scientific cycle (Table 1). Here, we summarize automated platforms to analize the existing gaps with the aim to automate the whole scientific cycle (Figure 1). Open automated research platforms might play a leading role in addressing at least the five following challenges: 1) Helping in the science of science by providing quantitative statistics (Fortunato et al., 2018), for example, the many paths with solutions to specific questions; 2) Identifying systematically bias and uncertainty in inference; 3) Exploring prediction and explanatory gradients to gain sinergy between predictive and explanatory power to complex problems; 4) Identifying gaps in patterns not explored consequence of lack of syntesis within and between disciplines, and 5) Allowing for reusability, repeatability, replicability and reproducibility along the many paths in the scientific enterprise (Figure 1).

1. Testing science: Helping select the best paths in responding to a question? ARP can provide a distribution of solutions by classifying the topologies of the multilayer networks.
2. Identifying bias and uncertainty in inference.
3. Exploring predictions-explanatory gradients to gain sinergy between predictive and explanatory power.
4. Identifying gaps in patterns not explored consequence of lack of integration within and between disciplines, and
5. Facilitating the 4R in open science: reusability, repeatability, replicability, and reproducibility.

The design or research platforms is still in its infancy. Many factors are involved in research platforms: the programming language, the number of packages and their interactions, their efficiency and functionality, etc.

Many questions in science strongly depend on our own bias, lab inertia in the methods and data explored. Therefore, exploring new paths would require new efforts to lead to new methods or new collaborations...

Reproducibility and robustness across the different stages of a research platform are two of the desire properties. Reproducibility guarantees the future improvement of the results in future analysis. Many programming languages have tools to facilitate reproducibility (notebooks) and notebooks implementing many languages are already available (jupyter...). Automated research platforms track the explored paths (i.e., the within and between layer interactions) and outlines how close each path is to the empirical patterns accounting for

Sampling desing and experiments...

Randomizations to achieve solid statistics...

One of the most discussed challenges nowadays is how to balance pattern and process inference. Many problems might not require a mechanistic understanding to make predictions. Recent examples are AI algorithms playing chess and go. They do not require a theory of mind to win. On the other side, there can be problems that might require a solid mechanistic understanding to make accurate predictions. Examples of these problems can be global warming or astrophysics. Therefore platforms that learn to combine AI and process-based methods

3 The structure of automated research platforms

In this section we outline the steps to develop a research platform. We introduce each of the layers outlined in Figure 1, data collection and integration (DC), complexity reduction, pattern-process inference, validation and visualization. The second part introduces a simple example using the $\mathcal{ROBHOOT}$ package.

For any given question, there are different methods within each layer that can complete the task. Ideally, one should be able to choose the best method from each layer and connect them to reach insightful patterns and predictions from the data. How many paths are there? Which of these minimize bias? Which topology within and between layers give the best response to our question?

3.1 Data Integration

Data access platforms within and qacross disciplines are highly scattered across the web [1]. Researchers have to deal with a highly complex set of intermediate stages and regulations before having access to the raw data. Having "easy" access to the information in a "perfectly informed market" should be simple and efficient, but unfortunately, it is not. Data integration in research platforms is rapidly evolving and there are many platforms that can have access and deliver real time data plots (Table 1).

Data Integration and standarization – Size effects – N labs vs N samplings per lab: Accuracy and uncertainty: How do initial distributions change accuracy and uncertainty? Trade-offs experimental vs big data

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

3.2 Complexity Reduction

3.3 Pattern-Process Inference

Outline classical variance-covariance matrices, AI algorithms and process-based methods.

3.4 Validation

Describe briefly Bayesian Inference, Approximate Bayesian computation, AIC and BIC model comparison methods. Gibbs sampling – Bayes factors

3.5 Visualization

3.6 An example with $\mathcal{ROBHOOT}$

In this section, we illustrate a semi-automated tool combining access to data from both centralized and decentralized platforms and integrating the datasets to infer insights and predictions obtained from analyzing patterns in the datasets (Figure 1). We aim to develop $\mathcal{ROBHOOT}$ in two stages. The first stage will be to develop the free-access platform to have access to integrated databases. The second stage will be to run it automatically to produce insights and pattern inference given specific questions (Figure 1).

3.6.1 Data Integration ($\mathcal{DAADI}$)

3.6.2 Complexity Reduction ($\mathcal{GOCORE}$)

3.6.3 Pattern-Process Inference ($\mathcal{PROPENCE}$)

3.6.4 Validation ($\mathcal{VATION}$)

3.6.5 Visualization ($\mathcal{VITION}$)

4 Discussion

5 Acknowledgments

References

Fortunato, S., C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, and A.-L. Barabási, 2018. Science of science. Science 359. URL `http://science.sciencemag.org/content/359/6379/eaao0185`.

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.

6 Tables

Table 1	
Data platforms	**Webpage**
Nakamoto Terminal	https://www.nterminal.com
BigQuery	https://cloud.google.com/bigquery/

Script Workflow Automated Research Platform
Sensu Renku (SDSC) knowledge graph

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. There are at least five properties automated ARP can provide to science:

==
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
parsing.data D S ——> Query.jl
data.to.table D S ——> MySQL.jl SQLite.jl Clickhouse?
data.julia D S ——> DataFrames.jl
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
FIN ===

7 Figures

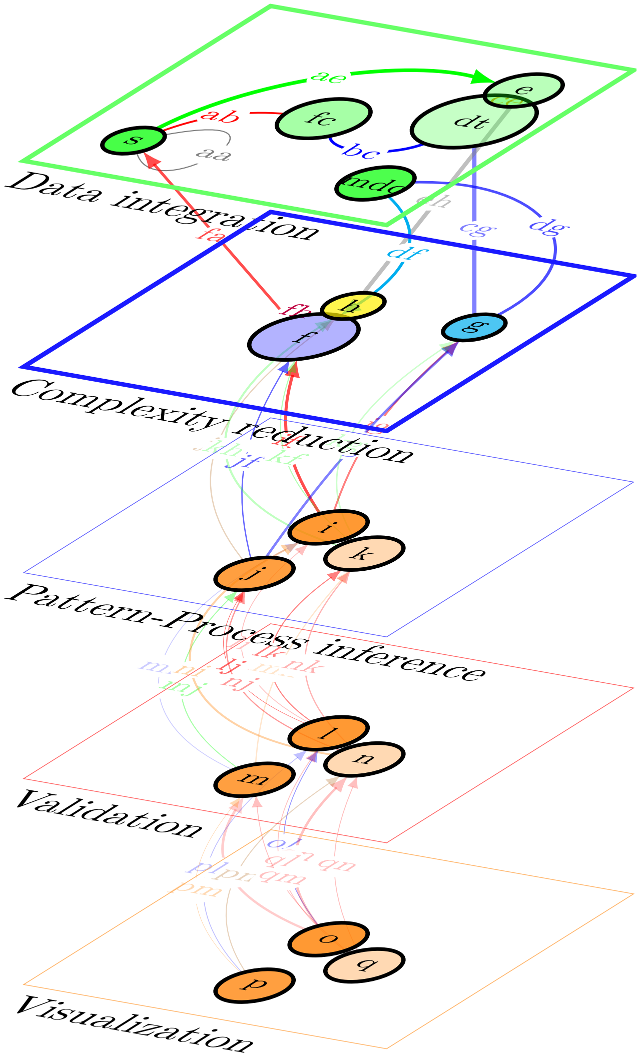

Figure 1: A cartoon of a five layer automated research platform: Data Integration, Complexity reduction, Pattern-process inference, Validation, and Visualization. Nodes and links represent algorithms and interactions between two algorithms, respectively. For example, the figure shows five algorithms 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.

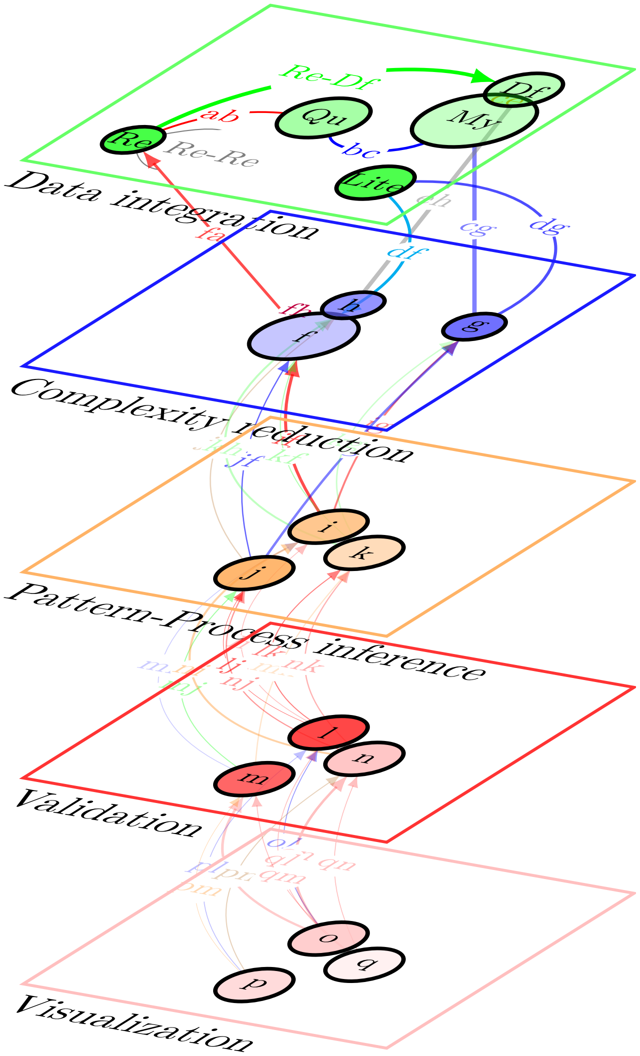

Figure 2: A julia prototype of an automated research platform. Nodes and links in each layer represent julia packages and interactions between two packages, respectively. The figure shows julia packages within each layer. For example, the layer Data integration contains the packages "Retriever.jl" (**Re**), "Query.jl" (**Qu**), "MySQL.jl" (**My**), "SQlite.jl" (**lite**), and "DataFrames.jl" (**df**).

Index