Reproducibility and inference in automated research platforms

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

1 Department of Anthropology, University of Zurich, Switzerland
2 Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland
3 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 data coming from many sources is becoming standard in many scientific fields. Yet, inferring insightful patterns and processes integrating databases with analytical frameworks remain challenging in many disciplines. In this work we aim to introduce the main features of an semi-automated workflow to integrate data and pattern- and process-based inference accounting for many sources of uncertainty to take better informed decisions in research, management and investment landscapes.

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

2 Introduction

We are in a era of data accumulation and integration, yet obtaining insights from such an integration accounting for robustness and reproducibility in pattern and process inference is at a very incipient stage (Ioannidis, 2005). 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. Protocols should take into account the robustness of experiments, algorithms and inference to facilitate the reproducibility of solutions to questions given new data and experiments (Figure 1). Open semi-automated platforms might play a leading role in addressing these challenges. Open platforms might allow for maximizing reproducibility of pattern-processes detection along the many paths in the scientific enterprise (Figure 1). These platforms might also help to make the scientific method such a data, pattern, process and insight integration less hard to follow.

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

Reproducibiilty and robustness across the different stages of a research platform are two of the desire properties. Reproducibility might facilitate the accuracy of the results in future analysis. Many programming languages have tools to facilitate reproducibility (notebooks) and notebooks implementing many languages are already available (

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 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 (DI) ($\mathcal{DAADI}$)

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

3.2 Complexity Reduction (CR) ($\mathcal{GOCORE}$)

PCA family – High-dimensionality of Convex hull – Information metrics multilayer networks

Data dimension reduction is a second step to increase performance during the next stages of analysis. Complexity reduction in economics and in ecology has a long tradition mostly by looking at variance-covariance matrices. Portfolio theory in economics has a long tradition (Markowitz, 1991). The theory is rooted in the concept of efficient frontier. There are several packages in several languages to calculate efficient frontiers[2,3,4,5]. Most maths underlying portfoliio theory are based in matrix correlation patterns. In ecology, portfolio concept has also been used to predict the number of coexisting species in landscapes with highly fluctuating environments[6].

Many fields aim at predicting fluctuations of several time series at local and regional scales. The better the predictions are the better we know the ecosystem. Unfortunately, it is not easy to predict time series of a large number of interacting (ideally independent) variables. Given we can not predict most of the ideas' trends, we should build a minimum understanding on how to investigate ideas and build a diversified portfolio with a balance between risk and reward. Basic questions will always remain when discussing about predicting the future and diversifying portfolios. For example, in a complex ecosystem, which is the best strategy under complete ignorance? And under complete information? Should we invest in ideas following a random walk ? Should we produce a portfolio with

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

[2] http://www.quantcode.com/

[3] https://github.com/JuliaQuant/PortfolioModels.jl

[4] https://www.wikinvest.com/account/portfolio/register

[5] https://d1so5k0levrfcn.cloudfront.net/SigFig%20Investment%20Methodology.pdf

[6] Check references

neutral risk ? [7]. Given the basic maths underlying complexity reduction, which are the algorithms and models out there? Which one perform the best? Which is the mixed of models to minimize data complexity?

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

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

3.4 Validation (V) ($\mathcal{VATION}$)

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

3.5 Visualization (VI) ($\mathcal{VITION}$)

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).

4 Gaps

5 Discussion

[7] https://en.wikipedia.org/wiki/ARandomWalkDownWallStreet

6 Acknowledgments

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.

Markowitz, H., 1991. Portfolio selection. Blackwell Publishing, MA.

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.

7 Tables

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

8 Figures

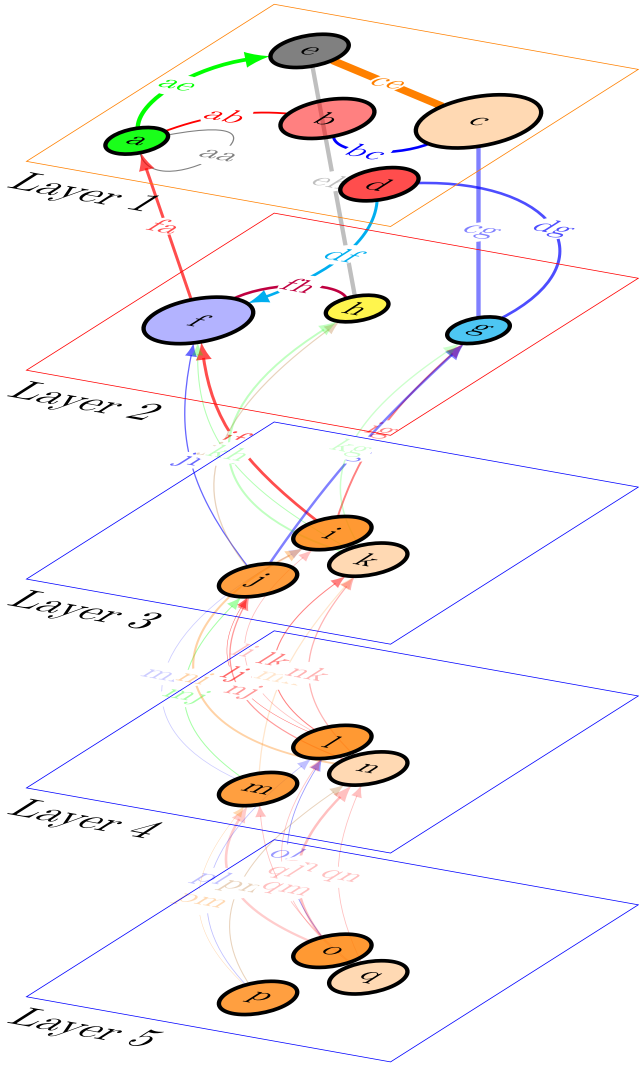

Figure 1: A five layer research platform: Data Integration (Layer 1), Complexity reduction (Layer 2), Pattern-process inference (Layer 3), Validation (Layer 4) and Visualization (Layer 5). Research platforms might play a leading role in accounting for bias and reproducibility in the pattern-process detection enterprise. a) A fully connected 5-layer research platform, and b) A specific path representing the best solution to solve a task.

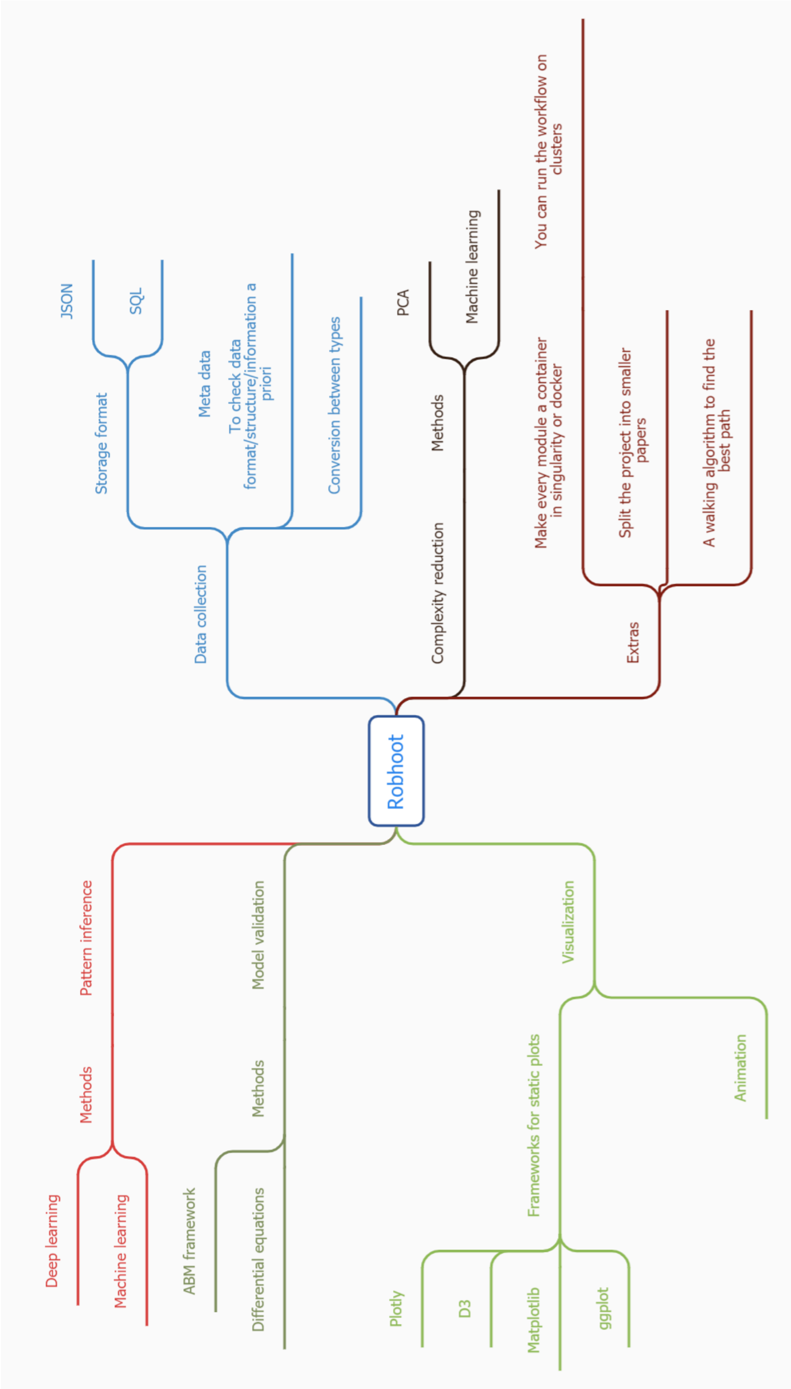

Figure 2: A mind map outlining the different methods to be used within each layer.

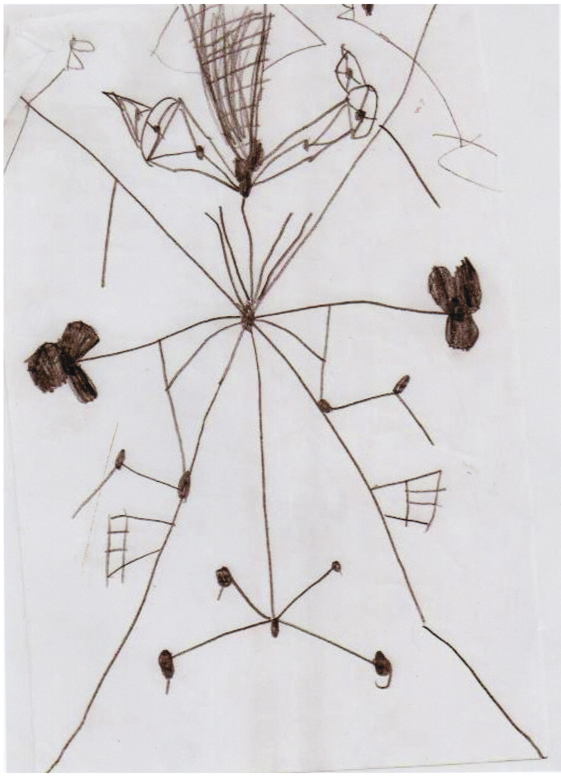

Figure 3: *ROBHOOT*
– An open multilayer platform for data integration, inference and prediction

Index