No more *Believe Me*: Make Your Informatics Experiments Reproducible [⋆]

Helmar Burkhart, Danilo Guerrera, and Antonio Maffia

University of Basel
Department of Mathematics and Computer Science
CH-4051 Basel, Switzerland
`http://www.unibas.ch`

Abstract. Being able to check and reproduce research results is a major scientific principle. Science disciplines in general, including computational disciplines, still find it hard to guarantee this. We explore the difficulties and general targets for reproducible research, discuss current Informatics approaches, and sketch a vision for 2025 by which many of the existing problems are assumed to be solved.

Keywords: Science principles, experiments, verification, reproducibility, trust

1 Experimental Science is in a Reproducibility Crisis

> ”Trouble at the Lab: Scientists like to think of science as self-correcting. To an alarming degree, it is not.”

This is the headline of an article dated from 2013 about the critical state of experimental science (see `www.economist.com`). Indicators for such a crisis have been detected in several science fields:

- **Psychology**: 9 separate experiments could not verify a famous study in the field.
- **Cancer research**: Out of 53 studies only 6 could be reproduced.
- **Pharma research**: Only 25% of 67 seminal studies could be verified.

Such lack of trust in scientific results has also been reported elsewhere[1].

> ”It is impossible to believe most of the computational results presented at conferences and in published papers today. Even mature branches of science, despite all their efforts, suffer severely from the problem of errors in final published conclusions.”

But being able to check and reproduce research results is a major scientific principle. To speak with Karl Popper [2] :

> ”Non-reproducible single occurrences are of no significance to science.”

2 Does Informatics Perform Better?

On the one hand Informatics has a better position because a communication infrastructure i.e. the Internet exists, which in principle makes every compute environment accessible on a global basis. On the other hand, technological progress is fast and systems are changing rapidly and the difficulty in reproducing computational research is in large part caused by the difficulty in capturing every last detail of the software and computing environment, which is what is needed to achieve reliable replication [3]. As can be found, articles often do not have a sufficiently detailed description of their experiments, and do not make available the software used to obtain the results claimed. A study in terms of reproducibility on a wide variety of papers has been carried out: each of such papers has been analyzed and classified according to the identified lacks [4]. Anyway we can already identify promising efforts in several Informatics fields.

2.1 Collaboratory on Experimental Evaluation of Software and Systems in Computer Science

In 2010 a community effort was established to develop a "Canon" which is a collection of readings on experimental evaluation and "good science". A web site (`http://evaluate.inf.usi.ch/`) serves as a resource and a hub for everybody interested in understanding and improving the state of practice in experimental evaluation. The Canon provides a bibliography on experimental evaluation and a list of venues focusing on experimental evaluation.

Recently, major conferences in the field of Programming Methodologies and Languages such as OOPSLA, POPL, and PLDI have started evaluating artifacts underlying the papers (http://evaluate.inf.usi.ch/artifacts). Artifacts (software, tools and data sets used in an experiment) accompany the paper submission process. A separate Artifact Evaluation Committee (AEC) checks submissions with respect to reuse, consistency, completeness, and documentation quality and honors successful authors by providing a certificate.

2.2 Algorithmic Engineering

The ACM Journal of Experimental Algorithmics `http://www.jea.acm.org/` stimulates research in algorithms based upon implementation and experimentation, distributes programs and testbeds throughout the research community and provides a repository of useful programs and packages to both researchers and practitioners. Authors are asked to make every effort to simplify the verification process by including instructions for installing the programs, clearly describing platform dependencies, creating sample input and output files, fully documenting the source code, and organizing and labelling files neatly in subdirectories. Referees are asked to evaluate the software and (at least partially) verify the experimental results.

2.3 Artificial Intelligence

Within the Artificial Intelligence community the following mission has been stated:

> ”If we can compute your experiment now, anyone can recompute it 20 years from now.”

A manifesto containing six theses has been published (`http://recomputation.org/manifesto`). Emphasis is put on virtual machine usage (”The only way to ensure recomputability”) and runtime performance is considered a secondary issue.

2.4 Computational Sciences

Several tools have been developed to address reproducibility issues and provide services such as up-to-date documentation by means of executable documents, and provenance support for programs and data. Such tools follow three different approaches (workbench, version control and virtualization) but do not address performance benchmarking (HPC orientation). In [5] we demonstrate case studies of repeatable benchmark experiments using our tool PROVA!.

2.5 Parallel and Distributed Computing

REPPAR (`www.reppar.org`) is a workshop series concerned with experimental practices in parallel computing research. The interest is in research works that address the statistically rigorous analysis of experimental data and visualization techniques of these data. Researchers are encouraged to state best practices to conduct experiments and papers that report experiences obtained when trying to reproduce or repeat experiments of others.

2.6 Assessment

Individual Informatics disciplines already see the relevance of the problem and have so far developed particular ad-hoc solutions. SNSF[1], in its Multi-Year Program 2017-2020, aims to introduce new measures for improving research data management and help to ensure good scientific practice, including the reproducibility of research results. It is now time for Informatics as ONE discipline to address this issue and help to overcome today’s weakness within one decade.

[1] Swiss National Science Foundation

3 Outlook 2025

In 2025 at the latest, the following statements should reflect established practice.

Research

- Funding agencies only support experimental science projects if applicants integrate reproducibility mechanisms.
- Leading scientific publishers and conference organizers only accept publications which make research results verifiable.
- Peer review of scientific work has been standardized, defining a set of requirements to be checked, in order to ensure reproducibility.
- Failures of the experiments are stored and shared together with the final results presented in the papers.

Curriculum

- In week 1 (titled Studium Generale Week) of their bachelor studies all science and engineering students learn about the basic scientific principles including the necessity of making results reproducible.
- In semester 1 of their bachelor studies all informatics students attend a mandatory course in applied statistics and data visualizations.
- Throughout the curriculum all lab results are fully documented and made repeatable for teaching staff.

Technology

- Tools which support reproducible Informatics research are widespread.
- Web sites and archival systems for trusted experiments exist.
- Provenance of software and data is securely managed and maintained.

Ethics and Society

- All professional Informatics organizations address the aim of reproducible research in their "Code of Ethics".
- Trust certificates are given for reproducible research, which has an impact to ranking purposes.
- Informatics and related computational disciplines receive positive response for their active role in making research results reproducible.

References

1. Stodden Victoria. Trust your science? Open your data and code. *Amstat News*, 2011.
2. K. Popper. *The Logic of Scientific Discovery.* Routledge, 1959.
3. A.P. Davison. Automated capture of experiment context for easier reproducibility in computational research. *Computing in Science Engineering*, 14(4):48–56, July 2012.
4. Christian Collberg, Todd Proebsting, and Alex M. Warren. Repeatability and benefaction in computer systems research. Technical report, University of Arizona, 2015.
5. Antonio Maffia, Helmar Burkhart, and Danilo Guerrera. Reproducibility in practice: Lessons learned from research and teaching experiments. In *Euro-Par 2015: Parallel Processing Workshops.* Springer International Publishing, 2015.