Research Reproducibility

Manish Parashar, Scientific Computing and Imaging Institute, University of Utah
Michael A. Heroux, Sandia National Laboratories and St. John's University
Victoria Stodden, University of Southern California

Digital Object Identifier 10.1109/MC.2022.3176988
Date of current version: 2 August 2022

Reproducibility has a foundational role in ensuring robust and trustworthy research, but achieving reproducibility can be challenging. This theme issue explores these challenges along with research and implementations across communities addressing them, with the goal of understanding the impact of existing solutions and synthesizing lessons learned and emerging best practices.

Reproducibility is foundational to scientific and technical research—the ability to repeat research is a key approach for confirming the validity of a new scientific discovery and an essential element toward ensuring the trustworthiness of the research.[1] As computing and data play an increasingly important role across all of science and engineering, ensuring the reproducibility of computational and data-enabled research is critical; it is a prerequisite to replicability and other activities that improve trustworthiness. However, as the complexity of these applications and the underlying computing and data platforms increases, achieving the many dimensions of reproducibility can be challenging.

The goal of this theme issue focused on research reproducibility is to explore these challenges as well as research and implementations across communities addressing these challenges, understand the impact of existing solutions on applications and their reproducibility, and synthesize lessons learned and emerging best practices. This issue includes four articles, which are introduced next. These articles bring together different perspectives and insights to the focus theme.

In "CUF-Links: Continuous and Ubiquitous FAIRness Linkages for Reproducible Research," Foster and Kesselman focus on mechanisms that can effectively capture basic information about every data product consumed or produced in a project with the goal of supporting reproducibility. They propose extending the scope of findable, accessible, interoperable, and reusable (FAIR) data to enable the creation of a continuous chain of continuous and ubiquitous FAIRness linkages (CUF-Links) from inputs to outputs provides a strong foundation for documenting the provenance linkages that are essential to reproducible research. The article provides specific examples of mechanisms that can achieve these goals.

In "Advancing Reproducibility at the NSF," Halbert provides a perspective on the increasing efforts regarding the advancement and funding of cyberinfrastructure for scientific research, in particular describing the emphasis on open data, provenance, the FAIR reporting standards, and how these support reproducibility efforts in scientific research. A review is given of the types of projects the National Science Foundation has recently funded on reproducibility-related research.

In "Defining the Role of Open Source Software in Research Reproducibility," Barba traces the relationship between the open source software development model and its success in the software development community as well as open source scientific code development and the relationship to scientific advancement. In particular, the unique roles of both community and reproducibility in scientific research are examined as scientific research itself increasingly leverages software and code in its advancement, and the article draws novel parallels and contrasts with open source software development.

Finally, in "Benchmarking Data Science: 12 Ways to Lie With Statistics and Performance on Parallel Computers," Hoefler provides a humorous and highly practical discussion of the ways in which data science results are reported that lead to incorrect or incomplete conclusions. Each of Hoefler's enumerated 12 Ways is succinctly stated and then followed by a

AS COMPUTING AND DATA PLAY AN INCREASINGLY IMPORTANT ROLE ACROSS ALL OF SCIENCE AND ENGINEERING, ENSURING THE REPRODUCIBILITY OF COMPUTATIONAL AND DATA-ENABLED RESEARCH IS CRITICAL.

ABOUT THE AUTHORS

MANISH PARASHAR is the director and chair in computational science and engineering at the Scientific Computing and Imaging Institute, Salt Lake City, Utah, 84112, USA; and a professor at the School of Computing at the University of Utah, Salt Lake City, Utah, 84112, USA. His research interests are in the broad areas of parallel and distributed computing and computational and data-enabled science and engineering. Parashar received a Ph.D. in computer engineering from Syracuse University. He is the founding chair of the IEEE Technical Consortium on High Performance Computing and editor in chief of *IEEE Transactions on Parallel and Distributed Systems*. He is a Fellow of the American Association for the Advancement of Science, Association for Computing Machinery, and IEEE. Contact him at manish.parashar@utah.edu.

MICHAEL A. HEROUX is a senior scientist at Sandia National Laboratories, Albuquerque, New Mexico, 87123, USA; director of software technology for the U.S. Department of Energy Exascale Computing Project, Oak Ridge, Tennessee, 37830, USA; and a scientist in residence at St. John's University, Collegeville, Minnesota, 56321, USA. His research interests include all aspects of large-scale, parallel scientific computing. Heroux received a Ph.D. in applied mathematics from the Colorado State University. He is a Fellow of the Society for Industrial and Applied Mathematics, a Distinguished Scientist of the Association for Computing Machinery, and a Senior Member of IEEE. Contact him at https://maherou.github.io or maherou@sandia.gov.

VICTORIA STODDEN is an associate professor in the Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, California, 90089, USA. She is an internationally recognized leader in improving the reliability of scientific results in the face of increasingly sophisticated computational approaches to research: understanding when and how inferences from data are valid and reproducible, what it means to have replicated a result, the effect of big data and computation on scientific inference, the design and implementation of scientific validation systems, standards of openness and transparency for data and code sharing, and resolving legal and policy barriers to disseminating reproducible research. Stodden received a Ph.D. in statistics from Stanford University and a law degree from Stanford Law School. Contact her at victoria@stodden.net.

discussion of the fallacy in detail and ways to rectify it. Some are even done with lighthearted humor. For example: Way #5, titled "Report highest (exa)op/s rates," states:

> "Floptimization is about reporting the highest flop/s rates! Thus use the smallest datatypes—after all your laptop can do 1018 bit flip (exa-)ops per second! Maybe you can even get away with reporting flop/s that you never do?"

The use of humor does not diminish the clarity but instead inspires the reader to look for and correct such fallacies when writing and reviewing content.

As noted previously, reproducibility has always been fundamental to scientific and technical research. However, it has taken on an increased urgency impacting society's trust in research results. For example, the research community is seeing a growing number of "fake papers," often enabled by artificial intelligence-based tools, that manage to get past review processes. At the same time, review and publication processes are evolving to include a trend toward greater unreviewed preprint release. This is in addition to the papers that are discovered to be disingenuous in their results and findings and have to be retracted. As a result, the perspectives and insights presented in the articles in this issue are very timely.

We hope that the articles in this special theme issue will help advance the broader conversation on these and other issues, and in particular, the role of open and democratized cyberinfrastructure services and expertise in addressing these issues.

REFERENCE

1. National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science*. Washington, DC, USA: National Academy Press, 2019.