

# Reproducible Research: Tools and Strategies for Scientific Computing



**T**his special issue is comprised of articles contributed by participants in a workshop held in Vancouver in July 2011 called “Reproducible Research: Tools and Strategies for Scientific Computing.” The workshop was co-organized by Randall J. LeVeque (the Founders’ Term Professor of Applied Mathematics at the University of Washington), Ian M. Mitchell (an associate professor of computer science at the University of British Columbia), and myself. One of our aims was to improve the visibility of the nascent group of tool

builders working to facilitate really reproducible research in computational science.

This special issue focuses on tools and strategies for reproducible computational science and includes seven articles, each describing software development efforts. It begins with an introductory article by the three workshop co-organizers

1521-9615/12/\$31.00 © 2012 IEEE  
COPUBLISHED BY THE IEEE CS AND THE AIP

VICTORIA STODDEN  
*Columbia University*

motivating our goals for the workshop and its successes, and providing a context for the articles that follow.

Next, Juliana Freire and Claudio Silva describe VisTrails, software that supports the computational aspects of scientific discovery, publication, and review, and permits reuse of the software, data, and results. VisTrails manages provenance and workflow tracking from data acquisition to review and reuse of the published scholarship by other scientists.

Matan Gavish and David Donoho describe three dream applications that would be made feasible within a system of labeling and linking published computational results to the underlying code and data that produced them. The system they've developed, enabling verifiable computational results, affixes a hash to each published figure or table, permitting independent identification and search, thereby opening new avenues of research and enabling reproducible research through code and data sharing in the cloud.


Philip Guo introduces CDE, a Linux tool that discovers software dependencies used when producing computational results. These platform-specific dependencies are packaged such that they can be ported to an independent system and the results regenerated. This tool sidesteps the need to manually install libraries or configure software when replicating results, providing a lightweight alternative to virtual machines.

Bill Howe addresses the role of cloud computing in enabling reproducibility in computational research. Creating a virtual machine for complex computing environments hosted in the cloud, making it publicly available, and providing a link in the associated published paper, permits the

independent re-execution of experiments. He analyzes the costs and benefits of such an approach and provides numerous examples of challenges faced.

Patrick Vandewalle provides a perspective on citation, one of the traditional rewards for publication. He analyzes citation trajectories for published articles in several signal processing journals and finds that those for which the underlying code is available are more likely to be cited.

Finally, Andrew Davison builds on the notion of reproducibility from first principles, considers appropriate features to be implemented in software used toward scientific ends, and provides a toolset for the automated capture of computational details, called Sumatra, designed to incorporate reproducibility into the day-to-day computational life of a scientist. He also gives a detailed treatment of best practices to simplify reproducibility.

**T**hese articles represent groundbreaking efforts in software development and tool building that facilitate really reproducible computational research and help to lay a framework for future progress. I hope you enjoy this special issue. 

*Victoria Stodden is an assistant professor of statistics at Columbia University. Her current research focuses on how pervasive and large-scale computation is changing our practice of the scientific method—in particular, regarding the reproducibility of computational results and the role of legal framing for scientific advancement. Stodden has a PhD in statistics from Stanford University. Contact her at [vcs@stodden.net](mailto:vcs@stodden.net); <http://blog.stodden.net>.*