Science Gateways: A Path to Reproducibility

Nancy Wilkins-Diehr
San Diego Supercomputer Center
wilkinsn@sdsc.edu

In a 2005 report to the President, computational science was introduced as a third pillar of science, joining theory and experimentation, by the President's Information Technology Advisory Committee (PITAC) [Benioff, Lazowska, 2005]. The authors point to the information technology revolution (nearly 10 years ago now) as the impetus behind this development. They describe computer modeling of events such as stellar explosions and climate shifts that can only be observed via simulation.

The scientific method is defined as "principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses" [Merriam-Webster, 2014]. These steps must be repeatable and in fact a cornerstone of the scientific method has always been the ability to reproduce results [Stodden, 2010]. Interestingly, perhaps due to the novelty of the approach at that point, reproducibility is not mentioned as a concern in the 2005 report.

Today, reproducibility is a growing problem. A paper from the 2012 altmetrics conference [Iorns, 2012] reports several worrying findings:

1. *"Amgen found that 47 of 53 "landmark" oncology publications could not be reproduced [Begley, Ellis, 2012].*
2. *Bayer found that 43 of 67 oncology & cardiovascular projects were based on contradictory results from academic publications [Mullard, 2011].*
3. *Dr. John Ioannidis and his colleagues found that of 432 publications purporting sex differences in hypertension, multiple sclerosis, or lung cancer, only one data set was reproducible [Patsopoulos, Tatsioni, Ioannidis, 2007]."*

This has become a concern in computational science as well, as evidenced by several recent efforts to address the topic. The Yale 2009 Roundtable, referenced in the call for this event [Stodden, 2009], resulted in a code sharing resolution.. A mini-symposium at the SIAM conference in 2011 [SIAM, 2011], also focused on code sharing and open source and included lessons learned from specific projects. The Reproducibility in Computational and Experimental Mathematics workshop in 2012 [ICERM, 2012] was unique in that it gathered a broad cross section of computational scientists and mathematicians, including pure mathematicians to discuss how to improve the current state of practice. And a reproducibility workshop sponsored by the new eScience Institute [eScience, 2014] will feature panel discussions and breakout groups in addition to keynote talks. A Coursera course entitled Reproducible Research is being offered by Robert Peng of Johns Hopkins [Peng, 2014]. Though this course focuses only on the use of R, clearly it is an early example of what will certainly be many future courses on this topic. The number of journal topics addressing the topic is exploding as well.

The problem of reproducibility is understandable when computational scientists work in isolation, developing their own codes, running on their own machines. A great deal of care must be taken to record all the details necessary to reproduce results. While there are many pieces of

the reproducibility solution, software frameworks (web portals or science gateways) can clearly play a role.

The term science gateways [Wilkins-Diehr 2007] originated in the NSF's TeraGrid program in 2004. The web was growing increasingly capable, progressing from static HTML pages to CGI scripts to the fully functional, collaborative environments we take for granted today. The term science gateway refers to the web portals and client-server software solutions being designed for and by scientists. The web continues to grow in sophistication. Data in many fields are born digital. Against this backdrop, science gateways continue to emerge as the interface to remote and distributed resources including instruments, datasets and computers - in labs, universities and supercomputer centers. Gateways also support collaboration, allowing users to contribute and share data, computing results, curricula materials and more. Because gateways provide a framework with which to run computational jobs, they can be used to address reproducibility issues. They can record which versions of a code were used to run a job, what version of the operating system was installed at the time and they can allow others to re-run and confirm results.

The Galaxy gateway [Goecks et al. 2010; Blankenberg et al. 2010; Giardina et al. 2010] was created specifically to ensure reproducibility when the developers recognized the Wild West approach to computation emerging in bioinformatics. GenePattern [Mesirov, 2009] provides access to massive datasets and computational algorithms that are the foundation of genomic medicine. It delivers infrastructure that allows a reproducible approach to research of the past. The nanoHUB science gateways was cited in "Computational Reproducibility: State-of-the-art, Challenges, And Database Research Opportunities" [Freire et al., 2012] where it was described as an example of a repository that addresses some aspects of reproducibility.

But we find that gateway developers need advice too if gateways are to be designed with reproducibility in mind. And they need sustainable funding if they are to serve as lynch pins to reproducibility. The Science Gateway Institute [Wilkins-Diehr, et al., 2011] hopes to address reproducibility as well as other issues by serving as a clearing house for gateway development - providing instruction, development help in a centralized fashion to a raft of gateway developers. There are many components to sustainable, successful gateway design and operation and the ability to address the important issue of reproducibility is surely one of these.

References

[Benioff, Lazowska, 2005]
http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf

[Merriam-Webster, 2014] http://www.merriam-webster.com/dictionary/scientific%20method

[Stodden, 2010] http://hdl.handle.net/10022/AC:P:11417

[Iorns, 2012] http://altmetrics.org/altmetrics12/iorns/

[Begley, Ellis, 2012] Drug development: Raise standards for preclinical cancer research. Begley CG, Ellis LM. Nature. 2012; 483(7391):531-3.

[Mullard, 2011] Reliability of 'new drug target' claims called into question. Mullard A. Nat Rev Drug Discov. 2011; 10(9):643-4.

[Patsopoulos, Tatsioni, Ioannidis, 2007] Claims of sex differences: an empirical assessment in genetic associations. Patsopoulos NA, Tatsioni A, Ioannidis JP. JAMA. 2007; 298(8):880-93.

[Stodden, 2009]
http://www.stanford.edu/~vcs/Conferences/RoundtableNov212009/index.html

[SIAM, 2011] http://jarrodmillman.com/events/siam2011.html

[ICERM, 2012] http://icerm.brown.edu/tw12-5-rcem/

[eScience, 2014] http://escience.washington.edu/event/first-reproducibility-workshop

[Peng, 2014] https://www.coursera.org/course/repdata

[Wilkins-Diehr, 2007] N. Wilkins-Diehr, "Special issue: Science Gateways-- Common Community Interfaces to Grid Resources: Editorials", April 2007, Journal of Concurrency in Computation: Practice and Experience. Volume 19, Issue 6 (April 2007), pages 743-749.

[Goecks, Nekrutenko, Taylor, and The Galaxy Team 2010] Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010 Aug 25;11(8):R86.

[Blankenberg, Von Kuster, Coraor, Ananda, Lazarus, Mangan, Nekrutenko, Taylor, 2010] "Galaxy: a web-based genome analysis tool for experimentalists". *Current Protocols in Molecular Biology*. 2010 Jan; Chapter 19:Unit 19.10.1-21.

[Giardine, Riemer, Hardison, Burhans, Elnitski, Shah, Zhang, Blankenberg, Albert, Taylor, Miller, Kent, Nekrutenko, 2005l] "Galaxy: a platform for interactive large-scale genome analysis." *Genome Research*. 2005 Oct; 15(10):1451-5.

[Mesirov, 2009] Research Reproducibility through GenePattern.
http://themindwobbles.wordpress.com/2009/06/27/research-reproducibility-through-genepattern-ismb-dam-sig-2009/.

[Wilkins-Diehr, et al., 2011] http://www.sciencegateways.org