The Value of Computational Transparency

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

Legal and Policy Issues Posed by Artificial Intelligence Advances
UC Berkeley
September 6, 2018
Key Issues

- source? academic vs industry vs government
- algorithmic transparency: who is the audience / what is the purpose? data/code?
  - Copyright on software / data. Transparency for re-use?
  - Levels of transparency: English, Math, Pseudocode; Code, (Data?).
- Computational uncertainty: we may not know everything when we know everything..
Remember Google Flu Trends?

In 2008 Google Flu Trends claimed it can tell you whether “the number of influenza cases is increasing in areas around the U.S., earlier than many existing methods”

In 2013 Google Flu Trends was predicting more than double the proportion of doctor visits for flu than the CDC.

Today:

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we’re excited to see what comes next. Academic research groups interested in working with us should fill out this form.

Sincerely,
The Google Flu and Dengue Trends Team.
What Happened?

- How did Google Flu Trends work? What was the data collection process? What was the algorithm?

- Why should we believe Google Flu Trends output? Many people did in 2008..

---

**The Parable of Google Flu: Traps in Big Data Analysis**

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

In February 2013, Google Flu Trends (GFT) made headlines. The algorithm in 2009, and this model has run ever since... with a...
The Reproducible Research Standard

The Reproducible Research Standard (RRS) (Stodden, 2009)

A suite of license recommendations for computational science:

• Release media components (text, figures) under CC BY,
• Release code components under MIT License or similar,
• Release data to public domain (CC0) or attach attribution license.

¬ Remove copyright’s barrier to reproducible research and,
¬ Realign the IP framework with longstanding scientific norms.
Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in the computational steps that produced results (1, 2). Over the past several years, a number of guidelines, from the Organization for the Advancement of Research Excellence and Productivity (OAREP) guidelines (1) and recommendations for field data (2), emerged from workshops and discussions among funding agencies, publishers, and journal editors, industry participants, and researchers representing a wide range of disciplines. These guidelines reflect the growing recognition that access to the computational steps taken to process data and generate findings is as important as access to data themselves.

Access to the computational steps taken to process data and generate findings is as important as access to data themselves.


All data, code, and workflows, including software written by the authors, should be cited in the references section (10). We suggest that software citation include software version information and its unique identifier in addition to the cited references.
Reproducibility Enhancement Principles

1: To facilitate reproducibility, share the data, software, workflows, and details of the computational environment in open repositories.

2: To enable discoverability, persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3: To enable credit for shared digital scholarly objects, citation should be standard practice.

4: To facilitate reuse, adequately document digital scholarly artifacts.

5: Journals should conduct a Reproducibility Check as part of the publication process and enact the TOP Standards at level 2 or 3.

6: Use Open Licensing when publishing digital scholarly objects.

7: Funding agencies should instigate new research programs and pilot studies.
Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) via the Constitution and subsequent Acts:

“To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.” (U.S. Const. art. I, §8, cl. 8)

**Argument:** both types of intellectual property are an imperfect fit with scholarly norms, and require action from the research community to enable re-use, verification, reproducibility, and support the acceleration of scientific discovery.
Copyright

• Original expression of ideas falls under copyright by default (papers, code, figures, tables..)

• Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original

• limited time: generally life of the author +70 years

• Exceptions and Limitations: e.g. Fair Use.
Patents

Patentable subject matter: “new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof” (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,

2. *Non-obvious*,

3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) “A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena” (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).
Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (i.e., Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.
Legal Issues in Data

- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).

- Copyright adheres to raw facts in Europe.

- the possibility of a residual copyright in data (attribution licensing or public domain certification).

- Legal mismatch: What constitutes a “raw” fact anyway?
Privacy and Data

• HIPAA, FERPA, IRB mandates create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/)

• Potential privacy implications for industry generated data.

• Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..
Ownership: What Defines Contribution?

- Issue for producers: credit and citation.
- What is the role of peer-review?
- Repositories adding meta-data and discoverability make a contribution.
- Data repositories may be inadequate: velocity of contributions.
- Future coders may contribute in part to new software, other software components may already be in the scholarly record. Attribution vs sharealike.
  - (at least) 2 aspects: legal ownership vs scholarly credit.
- Redefining plagiarism for software contributions.
Licensing in Research

Background: Open Source Software

Innovation: Open Licensing

- Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

Hundreds of open source software licenses:

- GNU Public License (GPL)
- (Modified) BSD License
- MIT License
- Apache 2.0 License
- ... see [http://www.opensource.org/licenses/alphabetical](http://www.opensource.org/licenses/alphabetical)
The Reproducible Research Standard

The *Reproducible Research Standard (RRS)* (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,  
- Release code components under **MIT License** or similar,  
- Release data to public domain (**CC0**) or attach attribution license.

➡ Remove copyright’s barrier to reproducible research and,  
➡ Realign the IP framework with longstanding scientific norms.
Testing the Claims: How Much of a Problem is Computational Reproducibility?
Study 1: Effectiveness of Artifact Access on Demand

February 11, 2011:

“**All data** necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. **All computer codes** involved in the creation or analysis of data **must also be available to any reader of Science**. After publication, **all reasonable requests for data and materials must be fulfilled**....”

- Obtained a random sample of 204 scientific articles with computational findings. Asked for the data and code!
Responses to Artifact Requests (n=204)

<table>
<thead>
<tr>
<th>Response</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>No response</td>
<td>26%</td>
</tr>
<tr>
<td>Contact to another person</td>
<td>11%</td>
</tr>
<tr>
<td>Asks for reasons</td>
<td>11%</td>
</tr>
<tr>
<td>Refusal to share</td>
<td>7%</td>
</tr>
<tr>
<td>Directed back to Supplemental Materials</td>
<td>3%</td>
</tr>
<tr>
<td>Unfulfilled promise to follow up</td>
<td>3%</td>
</tr>
<tr>
<td>Email bounced</td>
<td>2%</td>
</tr>
<tr>
<td>Impossible to share</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Shared data and code</strong></td>
<td><strong>36%</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

12% of the articles provided direct access to code/data
We were able to obtain data and code from the authors of 89 articles in our sample of 204,

- overall **artifact recovery rate** estimate: 44%, 95% confidence interval [0.36, 0.50]

Of the 56 articles we deemed potentially reproducible, we randomly choose 22 to attempt replication, and all but one provided enough information to do so.

- overall **computational reproducibility** estimate: 26%, 95% confidence interval [0.20, 0.32]
When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

The data files remains our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

We do not typically share our internal data or code with people outside our collaboration.

The code we wrote is the accumulated product of years of effort by [redacted] and myself. Also, the data we processed was collected painstakingly over a long period by collaborators, and so we will need to ask permission from them too.

Normally we do not provide this kind of information to people we do not know. It might be that you want to check the data analysis, and that might be of some use to us, but only if you publish your findings while properly referring to us.
Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code.

I’m sorry, but our computer code was not written with an eye toward distributing for other people to use. The codes are not documented and we don’t have the time or resources to document them. If you have a particular calculation you would like done and it is not a major extension of what we are presently set up to do, we might be able to run the codes for you.

R is a free software package available at www.r-project.org/ I used R for the [redacted] models. As you probably know, [redacted] and [redacted] are quite complicated. But I don’t have to tell you that given that you are a statistics student! I used Matlab for the geometry.
Our program [redacted] is available here [URL redacted] (documentation and tutorials were included)

If you go to [URL redacted], under the publications, I have a link to the GitHub repository. I don’t know if I have all of the raw simulated data, but I certainly have the processed data used to make the plots. What do you need? All of the simulated data could of course be regenerated from the code.

Please find attached a .zip file called [redacted].zip that has the custom MATLAB [redacted] analysis code. If you run Masterrunfigureone.m this will generate several panels from the paper.

In the next email I will enclose the custom image analysis software. This can also be accessed from [URL redacted] where there is a manual and tutorial.

Please let me know if you have any troubles, or if there is anything else I can help with.
Study 2: Reproducibility in Computational Physics

- Are artifacts available (can we obtain them)? Do they replicate the published results?

<table>
<thead>
<tr>
<th>Artifact Access via Information in the Article (n=306)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No discussion in the article and no artifacts made available</td>
</tr>
<tr>
<td>Some discussion of artifacts none made available</td>
</tr>
<tr>
<td>Some artifacts made available</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluation Criteria</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A precise statement of assertions to be made in the paper</td>
<td>100%</td>
</tr>
<tr>
<td>Full statement (or valid summary) of experimental results</td>
<td>100%</td>
</tr>
<tr>
<td>Salient details of data reduction &amp; statistical analysis methods</td>
<td>73%</td>
</tr>
<tr>
<td>Necessary run parameters were given</td>
<td>86%</td>
</tr>
<tr>
<td>A statement of the computational approach and why it tests the proposed hypotheses</td>
<td>100%</td>
</tr>
<tr>
<td>Complete statements of, or references to, algorithms and salient software details</td>
<td>63%</td>
</tr>
<tr>
<td>Discussion of the adequacy of parameters such as precision level and grid resolution</td>
<td>76%</td>
</tr>
<tr>
<td>Proper citation of all code and data used, including that generated by the authors</td>
<td>4%</td>
</tr>
<tr>
<td>Availability of computer code, input and output data, with reasonable level of documentation</td>
<td>4%</td>
</tr>
<tr>
<td>Avenues of exploration examined throughout development, including negative findings</td>
<td>0%</td>
</tr>
<tr>
<td>Instructions for repeating computational experiments described in the article</td>
<td>79%</td>
</tr>
<tr>
<td>Precise functions were given, with settings</td>
<td>11%</td>
</tr>
<tr>
<td>Salient test environment details: hardware, system software, and number of processors used</td>
<td>24%</td>
</tr>
</tbody>
</table>
## Attempts to Replicate Results (n=55)

<table>
<thead>
<tr>
<th>Computational Reproducibility Evaluation (n=55)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Straightforward to reproduce with minimal effort</td>
<td>0%</td>
</tr>
<tr>
<td>Minor difficulty in reproducing</td>
<td>0%</td>
</tr>
<tr>
<td>Reproducible after some tweaking</td>
<td>9.1%</td>
</tr>
<tr>
<td>Could reproduce with fairly substantial skill and knowledge</td>
<td>16.4%</td>
</tr>
<tr>
<td>Reproducible with substantial intellectual effort</td>
<td>12.7%</td>
</tr>
<tr>
<td>Reproducible with substantial tedious effort</td>
<td>3.6%</td>
</tr>
<tr>
<td>Difficult to reproduce because of unavoidable inherent complexity</td>
<td>3.6%</td>
</tr>
<tr>
<td>Nearly impossible to reproduce</td>
<td>3.6%</td>
</tr>
<tr>
<td>Impossible to reproduce</td>
<td>50.9%</td>
</tr>
</tbody>
</table>
The LifeCycle of Data Science as a Framework
Lifecycle of Data

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire
- Create, capture, gather from:
  - Lab
  - Fieldwork
  - Surveys
  - Devices
  - Simulations
  - More

Clean
- Organize
- Filter
- Annotate
- Clean

Use/Reuse
- Analyze
- Mine
- Model
- Derive much more additional data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

Publish
- Share:
  - Data
  - Code
  - Workflows
  - Disseminate
  - Aggregate
  - Collect
  - Create portals, databases, and more
  - Couple with literature

Preserve/Destroy
- Store to:
  - Preserve
  - Replicate
  - Ignore
  - Subset, compress
  - Index
  - Curate
  - Destroy

Berman et al., “Realizing the Potential of Data Science,” CACM, April 2018
Lifecycle of *Data Science*

- Framework to incorporate data science contributions from different fields,
- Explicit emphasis on re-use and reproducibility,
- Explicit emphasis on computational tools (e.g. Kubernetes), hardware (e.g. Google Edge TPUs) and software (e.g. Jupyter Notebooks)
- Surfaces ethics (human subjects, privacy), social context (interpretations of “bias”), scholarly communication and reproducible research.
Lifecycle of Data Science: An Abstraction

- **Application level**
  - Experimental design
  - Data generation and collection
  - Data exploration and hypothesis generation
  - Data cleaning and organization
  - Feature selection and data preparation
  - Model building and statistical inference
  - Simulation and cross-validation
  - Visualization
  - Notebooks
  - Workflow software
  - Artifact linking tools

- **Infrastructure level**
  - Notebooks and workflow software
  - Database structures
  - Workflow software and preregistration tools
  - Data management tools
  - Notebooks, workflow software, and inference languages
  - Notebooks, visualization software
  - Notebooks, visualization software

- **System level**
  - Hardware, cloud computing infrastructure, systems and system management, data structures, storage

- **The study of data science**
  - Ethics, documentation and metadata creation, best practices, policy
  - The science of data science

- **Tools and Technologies**
  - Notebooks
  - Workflow software
  - Artifact linking tools
  - Data management tools
  - Database structures
  - Workflow software and preregistration tools
  - Data cleaning and organization
  - Feature selection and data preparation
  - Model building and statistical inference
  - Simulation and cross-validation
  - Visualization
  - Notebooks
  - Hardware, cloud computing infrastructure, systems and system management, data structures, storage

- **Ethical and Policy Considerations**
  - Best practices
  - Policy
Example: AIM: An Abstraction for Improving Machine learning

- We developed *infrastructure* for comparative Machine Learning.
- Our goal: **List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their misclassification rates.**
AIM: Using Structured Containers

We compared models via classification rates:

<table>
<thead>
<tr>
<th>Paper/Classifier</th>
<th>Preprocessing / Feature Selection</th>
<th>Ave</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 WeightedVote</td>
<td>.91 .94 .97 .97 .89 .74 .90</td>
<td></td>
</tr>
<tr>
<td>3NN</td>
<td>.97 .94 .91 .94 .97 .97 .95</td>
<td></td>
</tr>
<tr>
<td>3 SVM Linear</td>
<td>.97 .97 .94 .97 .97 .77 .93</td>
<td></td>
</tr>
<tr>
<td>3 SVM Quad</td>
<td>.97 .88 .97 .97 .97 .91 .95</td>
<td></td>
</tr>
<tr>
<td>3 Adaboost</td>
<td>.91 .91 .97 .97 .91 .91 .93</td>
<td></td>
</tr>
<tr>
<td>6 Logit</td>
<td>.97 .97 .97 .97 .97 .88 .96</td>
<td></td>
</tr>
<tr>
<td>6 QDA</td>
<td>.94 .91 .94 .97 .97 .85 .93</td>
<td></td>
</tr>
<tr>
<td>9 NN</td>
<td>.97 .91 .85 .97 .94 .94 .93</td>
<td></td>
</tr>
<tr>
<td>9 Decision Tree</td>
<td>.91 .91 .97 .97 .91 .77 .90</td>
<td></td>
</tr>
<tr>
<td>9 Bagging</td>
<td>.94 .91 .97 .97 .97 .92 .91</td>
<td></td>
</tr>
<tr>
<td>9 Bagging (CPD)</td>
<td>.74 .85 .82 .91 .77 .68 .79</td>
<td></td>
</tr>
<tr>
<td>9 FLDA</td>
<td>.88 .88 .97 .97 .88 .88 .91</td>
<td></td>
</tr>
<tr>
<td>9 DLDA</td>
<td>.97 .94 .97 .97 .97 .88 .95</td>
<td></td>
</tr>
<tr>
<td>9 QDQA</td>
<td>.97 .94 .97 .97 .97 .88 .95</td>
<td></td>
</tr>
<tr>
<td>29 Bayes Network</td>
<td>.74 .88 .97 .97 .83 .62 .83</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>.92 .92 .95 .97 .92 .83</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Classification accuracy for the AML/ALL dataset, when preprocessing and feature selection steps are held constant for each classifier, allowing for the direct comparison of the classifiers presented in the five publications studied.
Challenges for the Research Community

- Funders are now funding cyberinfrastructure more expansively in addition to traditional foundational research;
- More and more fields (e.g. cybersecurity (LASER2014), networks (SIGCOMM2017)) are becoming empirical, not just transformed by opportunities due to data;
- Leveraging cyberinfrastructure and methods across fields (e.g. Computational Photo-Scatterography); how to reward, promote, fund;
- New research areas: Datasets as discovery drivers (ImageNet; Wiki* text datasets); Scientific software resilience and data preserve/destroy decisions;
- Technology transfer beyond the university.

Managing massive computational projects requires better, more transparent tools; and such tools will enable much more ambitious computational experiments.
The Future of Data Science

The future: a major effort to develop infrastructure that supports the entire Lifecycle of Data Science, from ethics through applications, to systems research to hardware such as special-purpose processor design.

Infrastructure promotes good scientific practice downstream like transparency and reproducibility.

People will use such infrastructure not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work, and used because it enables efficiency and productivity, and discovery.
In 'The Human Use of Human Beings' (1950), Norbert Wiener postulates a hypothetical that a computer could run experiments to understand the impact of various stimuli on people, thereby learning to control them.

At the end of the book he then says:

“The thing about this book is that this hypothetical might seem scary, but in order for it to happen, there’d have to be some sort of global computing capacity with wireless links to every single person on earth who keeps some kind of device on their person all the time and obviously this is impossible.”
Executive Order -- Creating a National Strategic Computing Initiative

EXECUTIVE ORDER

CREATING A NATIONAL STRATEGIC COMPUTING INITIATIVE

By the authority vested in me as President by the Constitution and the laws of the United States of America, and to maximize benefits of high-performance computing (HPC) research, development, and deployment, it is hereby ordered as follows:
NSCI Sec. 2. Objectives.

1. Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.

2. Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing.

3. Establishing, over the next 15 years, a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post-Moore's Law era").

4. Increasing the capacity and capability of an enduring national HPC ecosystem by employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, accessibility, and workforce development.

5. Developing an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the United States Government and industrial and academic sectors.
From a technical requirements perspective, infrastructure for data-intensive science needs to consider data acquisition, storage and archiving, search and retrieval, analytics, and collaboration (including publish/subscribe services). Recent NSF requirements to submit data management plans as part of proposals signal recognition that access to data is increasingly important for interdisciplinary science and for research reproducibility. Although the focus is sometimes on the hardware infrastructure (amount of storage, bandwidth, etc.), the human and software infrastructure is also important. Understanding the software frameworks that are enabled within the various cloud services and then mapping scientific workflows onto them requires a high level of both technical and scientific insight. Moreover, these new services enable a deeper level of collaboration and software reuse that are critical for data-intensive science.

changing scientific workflows extend to the human side of scientific computing as well. Especially in regards to data-intensive science, reproducibility will be challenging. These requirements will often be as important as the traditional technical requirements of CPU performance, latency, storage, and bandwidth.

deciding how much data to save is a trade-off between the cost of saving and the cost of reproducing, and this is potentially more significant than the trade-off between disks and processors.
Community Infrastructure Innovations

Research Environments

- Verifiable Computational Research
  - knitR
  - Collage Authoring Environment
    - Sumatra
    - Galaxy
- SHARE
  - Sweave
  - SOLE
  - GenePattern
  - torch.ch
- Code Ocean
  - Cyverse
  - Open Science Framework
  - IPOL
  - Whole Tale
- Jupyter
  - NanoHUB
  - Vistrails
  - Popper
  - flywheel.io

Workflow Systems

- Taverna
- Wings
- Pegasus
- CDE
- binder.org
- Kurator
- Kepler
- Everware
- Reprozip

Dissemination Platforms

- ResearchCompendia.org
- DataCenterHub
- RunMyCode.org
- ChameleonCloud
- Occam
- RCloud
- TheDataHub.org
- Madagascar
- Wavelab
- Sparselab
Figure 2: A historical perspective of values of a few particle properties tabulated in this Review as a function of date of publication of the Review. A full error bar indicates the quoted error; a thick-lined portion indicates the same but without the "scale factor."
A (Very) Brief History..
Yale 2009


We collectively produced the Data and Code Sharing Declaration including a description of the problem, proposed solutions, and dream goals we’d like to see.
Reproducibility in Computational and Experimental Mathematics (December 10-14, 2012)

Description
In addition to advancing research and discovery in pure and applied mathematics, computation is pervasive across the sciences and now computational research results are more crucial than ever for public policy, risk management, and national security. Reproducibility of carefully documented experiments is a cornerstone of the scientific method, and yet is often lacking in computational mathematics, science, and engineering. Setting and achieving appropriate standards for reproducibility in computation poses a number of interesting technological and social challenges. The purpose of this workshop is to discuss aspects of reproducibility most relevant to the mathematical sciences among researchers from pure and applied mathematics from academics and other settings, together with interested parties from funding agencies, national laboratories, professional societies, and publishers. This will be a working workshop, with relatively few talks and dedicated time for breakout group discussions on the current state of the art and the tools, policies, and infrastructure that are needed to improve the situation. The groups will be charged with developing guides to current best practices and/or white papers on desirable advances.

Organizing Committee
» David H. Bailey
  (Lawrence Berkeley National Laboratory)
» Jon Borwein
  (Centre for Computer Assisted Research Mathematics and its Applications)
» Randall J. LeVeque
  (University of Washington)
» Bill Rider
  (Sandia National Laboratory)
» William Stein
  (University of Washington)
» Victoria Stodden
  (Columbia University)
Setting the Default to Reproducible

Reproducibility in Computational and Experimental Mathematics

Developed collaboratively by the ICERM workshop participants

Compiled and edited by the Organizers

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein

Abstract

Science is built upon foundations of theory and experiment validated and improved through open, transparent communication. With the increasingly central role of computation in scientific discovery this means communicating all details of the computations needed for others to replicate the experiment, i.e., making available to others the associated data and code. The “reproducible research” movement recognizes that traditional scientific research and publication practices now fall short of this ideal, and encourages all those involved in the production of computational science—scientists who use computational methods and the institutions that employ them, journals and dissemination mechanisms, and funding agencies—to facilitate and practice really reproducible research.

Set the Default to “Open”

Reproducible Science in the Computer Age. Conventional wisdom sees computing as the “third leg” of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all of science. Massive computation is often required to reduce and analyze data; simulations are employed in fields as diverse as climate modeling and astrophysics. Unfortunately, scientific computing culture has not kept pace. Experimental researchers are taught early to keep notebooks or computer logs of every detail: design, procedures, equipment, raw results, processing techniques, statistical methods of analysis, etc. In contrast, few computational experiments are performed with such care. Typically, there is no record of workflow, computer hardware and software configuration, or parameter settings. Often source code is lost. While crippling reproducibility of results, these practices ultimately impede the researcher’s own productivity.

The State of Experimental and Computational Mathematics. Experimental mathematics—application of high-performance computing technology to research questions in pure and applied mathematics, including

"It says it's sick of doing things like inventories and payrolls, and it wants to make some breakthroughs in astrophysics."

physicists, legal scholars, journal editors, and funding agency officials representing academia, government labs, industry research, and all points in between. While

Society for Industrial and Applied Mathematics

SIAM NEWS >

“Setting the Default to Reproducible” in Computational Science Research

June 3, 2013

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.

Victoria Stodden, Jonathan Borwein, and David H. Bailey
Issues from ICERM

• The need to carefully document the full context of computational experiments including system environment, input data, code used, computed results, etc.

• The need to save the code and data in a permanent repository, with version control and appropriate meta-data.

• The need for reviewers, research institutions, and funding agencies to recognize the importance of computing and computing professionals, and to allocate funding for after-the-grant support and repositories.

• The increasing importance of numerical reproducibility, and the need for tools to ensure and enhance numerical reliability.

• The need to encourage publication of negative results as other researchers can often learn from them.

• The re-emergence of the need to ensure responsible reporting of performance.
reproducibility@XSEDE: An XSEDE14 Workshop

Overview

The reproducibility@XSEDE workshop is a full-day event scheduled for **Monday, July 14, 2014 in Atlanta, GA**. The workshop will take place in conjunction with XSEDE14 (conferences.xsede.org), the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE), and will feature an interactive, open-ended, discussion-oriented agenda focused on reproducibility in large-scale computational science. Consistent with the overall XSEDE14 conference theme, we seek to engage participants from a broad range of backgrounds, including practitioners whose computational interests extend beyond traditional modeling and simulation as well as decision-makers and other professionals whose work informs and determines the direction of computation-enabled research. We hope to help
Standing Together
for
Reproducibility in Large-Scale Computing

Report on reproducibility@XSEDE
An XSEDE14 Workshop
July 14, 2014
Atlanta, GA

Developed collaboratively by the reproducibility@XSEDE workshop participants

Principal Editors:
Doug James, Nancy Wilkins-Diehr, Victoria Stodden, Dirk Colbry, and Carlos Rosales

Finalized 17 Dec 2014

Abstract. This is the final report on reproducibility@xede, a one-day workshop held in conjunction with XSEDE14, the annual conference of the Extreme Science and Engineering Discovery Environment (XSEDE). The workshop's discussion-oriented agenda focused on reproducibility in large-scale computational research. Two important themes capture the spirit of the workshop submissions and discussions: (1) organizational stakeholders, especially supercomputer centers, are in a unique position to promote, enable, and support reproducible research; and (2) individual researchers should conduct each experiment as though someone will replicate that experiment. Participants documented numerous issues, questions, technologies, practices, and potentially promising initiatives emerging from the discussion, but also highlighted four areas of particular interest to XSEDE: (1) documentation and training that promotes reproducible research; (2) system-level tools that provide build- and run-time information at the level of the individual job; (3) the need to model best practices in research collaborations involving XSEDE staff; and (4) continued work on gateways and related technologies. In addition, an intriguing question emerged from the day's interactions: would there be value in establishing an annual award for excellence in reproducible research?
Supercomputing

SC16 Explores Reproducibility for Advanced Computing Through Student Cluster Competition by Michela Taufer

March 16, 2016 — Leave a Comment

Data sets and software are important by-products of research in fields that depend upon data-intensive and high performance computing. But these elements are typically absent when research results are recorded in a journal article or conference proceedings. There is a growing sense in the computational community that this gap needs to be filled if we are to create a stable base of research upon which reliable advances may be built. In short, we need to ensure that computational results are as reproducible as those from experiments.

Computational Reproducibility at Exascale: CRE2017

Synopsis

Where: Part of SC17, Denver, CO
When: Sunday afternoon, Nov 12, 2017
Submit: https://easychair.org/conferences/?conf=cre2017
Deadline: Friday, September 15, 2017
Notifications: Monday, October 2, 2017
Full Papers: Monday, October 9, 2017
Organized by: Walid Keyrouz (NIST), Miriam Leeser (NEU), and Michael Mascagni (FSU & NIST)
Registration: handled by SC17 (http://sc17.supercomputing.org/)

Motivation and Previous Offerings

This workshop combines the Numerical Reproducibility at Exascale Workshops (conducted in 2015 and 2016 at SC) and the panel on Reproducibility held at SC16 (originally a BOF at SC15) to address several different issues in reproducibility that arise when computing at exascale. The workshop will include issues of numerical reproducibility as well as approaches and best practices to sharing and running code and the reproducible dissemination of computational results. The workshop is meant to address the scope of the problems of computational reproducibility in HPC in general, and those anticipated as we scale up to Exascale machines in the next decade. The participants of this workshop will include government, academic, and industry stakeholders; the goals of this workshop are to understand the current state of the problems that arise, what work is being done to deal with this issues, and what the community thinks the possible approaches to these problem are.

Efforts by SIGHPC, SIGMOD, SIGCOMM