Improving Reproducibility Through Better Software Practices

Michael A. Heroux
Senior Scientist, Sandia National Laboratories
Director of Software Technology, US DOE Exascale Computing Project
Scientist in Residence, CSB|SJU

Better Scientific Software Tutorial
SC19, Denver, Colorado

License, Citation and Acknowledgements

License and Citation

- **The requested citation the overall tutorial is: David E. Bernholdt, Anshu Dubey, Michael A. Heroux, and Jared O'Neal, Better Scientific Software tutorial, in SC '19: International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, Colorado, 2019. DOI: 10.6084/m9.figshare.10114880**
- Individual modules may be cited as *Module Authors, Module Title*, in Better Scientific Software Tutorial…

Acknowledgements

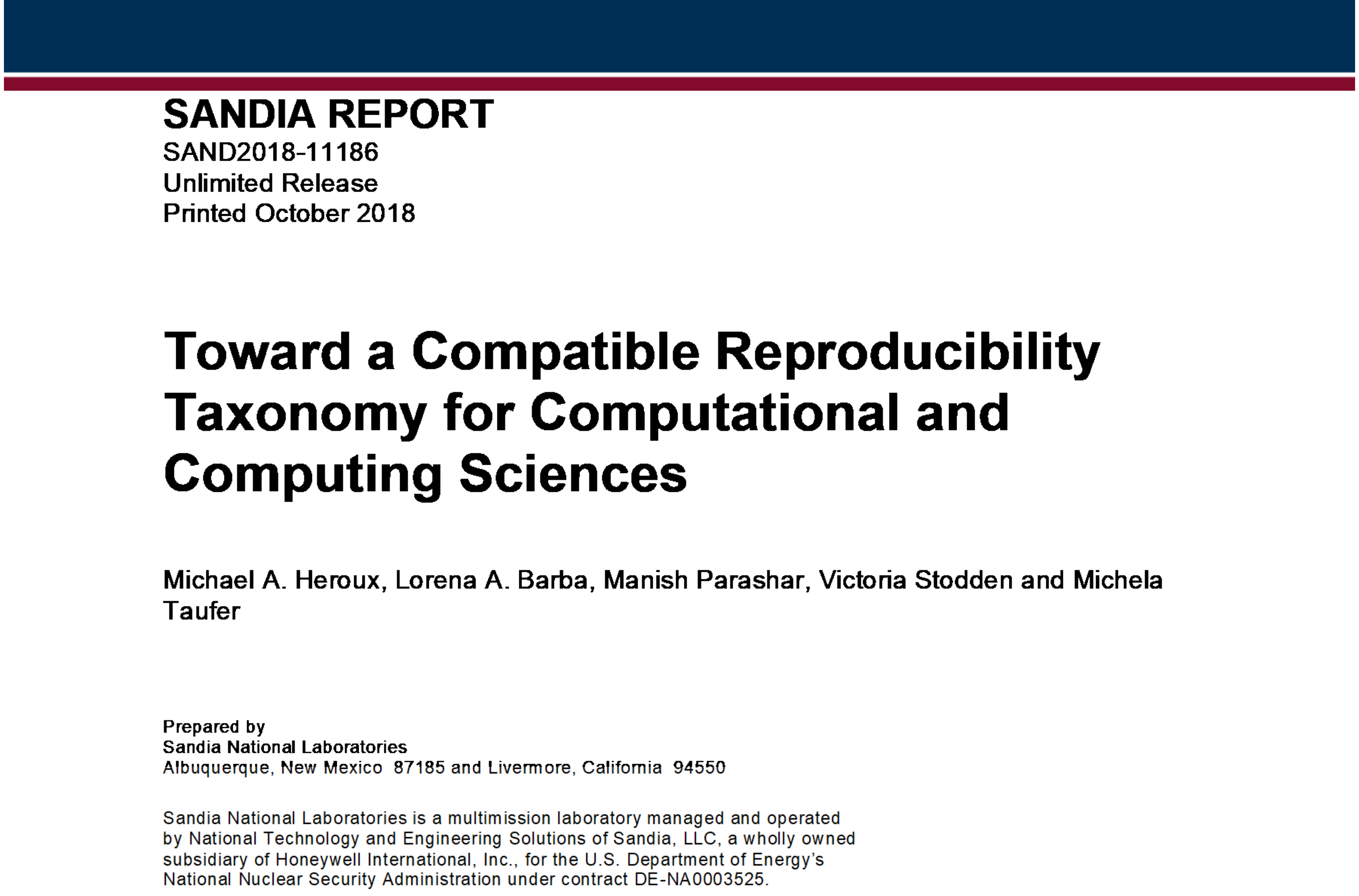
- Additional contributors to this this tutorial include: Alicia Klinvex, Katherine Riley, and James Willenbring
- This work was supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research (ASCR), and by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration.
- This work was performed in part at Sandia National Laboratories. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. SAND NO SAND2017-5474 PE

Outline

- Reproducibility taxonomies.
- Increasing focus on reproducibility.
- Role of better software practices.
- Publication requirements.
- Trustworthiness at Scale.
- Personal Productivity Commitment.

Reproducible vs Replicable

Addressing Confusion in Taxonomies



SANDIA REPORT
SAND2018-11186
Unlimited Release
Printed October 2018

Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences

Michael A. Heroux, Lorena A. Barba, Manish Parashar, Victoria Stodden and Michela Taufer

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Table 1: Definitions of Reproducible and Replicable

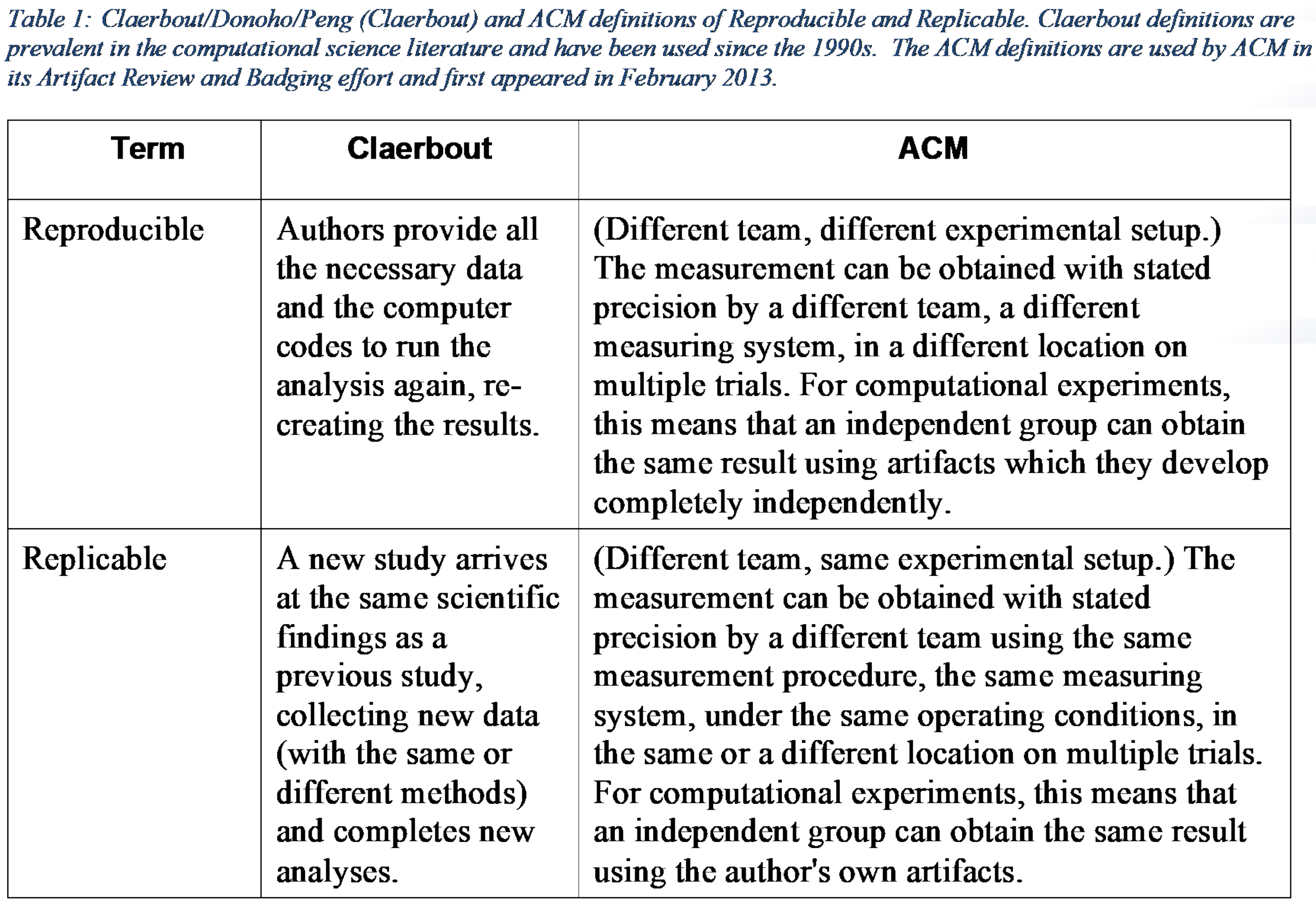


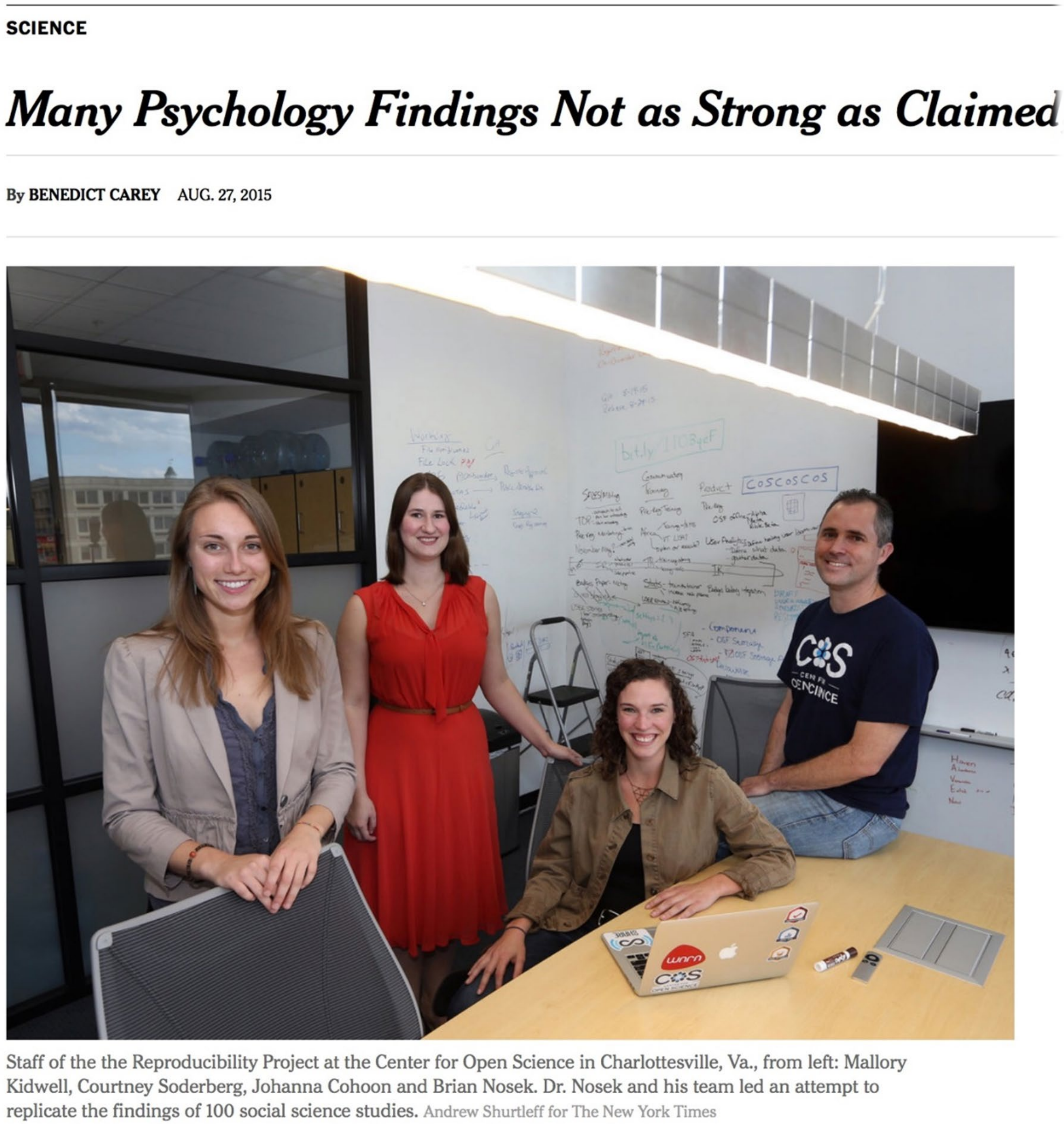
Table 1: Claerbout/Donoho/Peng (Claerbout) and ACM definitions of Reproducible and Replicable. Claerbout definitions are prevalent in the computational science literature and have been used since the 1990s. The ACM definitions are used by ACM in its Artifact Review and Badging effort and first appeared in February 2013.

Term	Claerbout	ACM
Reproducible	Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results.	(Different team, different experimental setup.) The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.
Replicable	A new study arrives at the same scientific findings as a previous study, collecting new data (with the same or different methods) and completes new analyses.	(Different team, same experimental setup.) The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.

Transparency & Reproducibility

Why we pursue better software practices

Transparency & Reproducibility



SCIENCE

Many Psychology Findings Not as Strong as Claimed

By BENEDICT CAREY AUG. 27, 2015

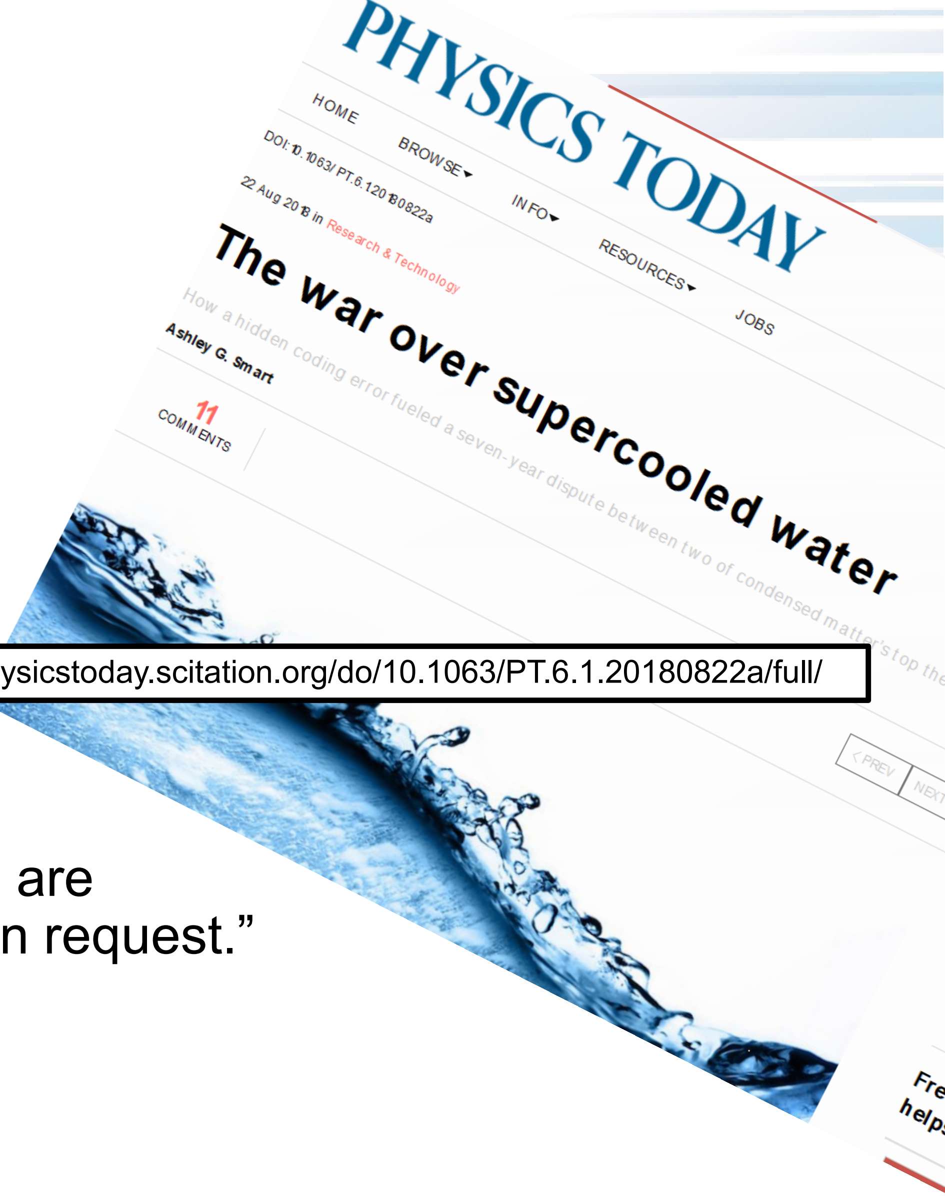
Staff of the the Reproducibility Project at the Center for Open Science in Charlottesville, Va., from left: Mallory Kidwell, Courtney Soderberg, Johanna Cohoon and Brian Nosek. Dr. Nosek and his team led an attempt to replicate the findings of 100 social science studies. Andrew Shurtleff for The New York Times

- NY Times highlights “problems”.
- Only one of many cited examples.
- Computational science ***had*** been spared this “spotlight”.

http://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html

Computational Science Example

- Behavior of pure water just above homogeneous nucleation temperature (~ - 40 C/F).
- Debenedetti/Princeton (2009):
  - 2 possible phases: High or low density.
- Chandler/Berkeley (2011):
  - Only 1 phase: High density.
- No sharing of details across teams until 2016:
  - Chandler in Nature: “LAMMPS codes used in refs 5 and 12 are standard and documented, with scripts freely available upon request.”
  - Debenedetti with colleague Palmer: ”Send us your code.”
  - Received code after requests and appeal to Nature.



PHYSICS TODAY

HOME BROWSE▾ INFO▾ RESOURCES▾ JOBS

DOI: 10.1063/PT.6.1.20180822a

22 Aug 2018 in Research & Technology

The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top the

Ashley G. Smart

11 COMMENTS

< PREV NEXT

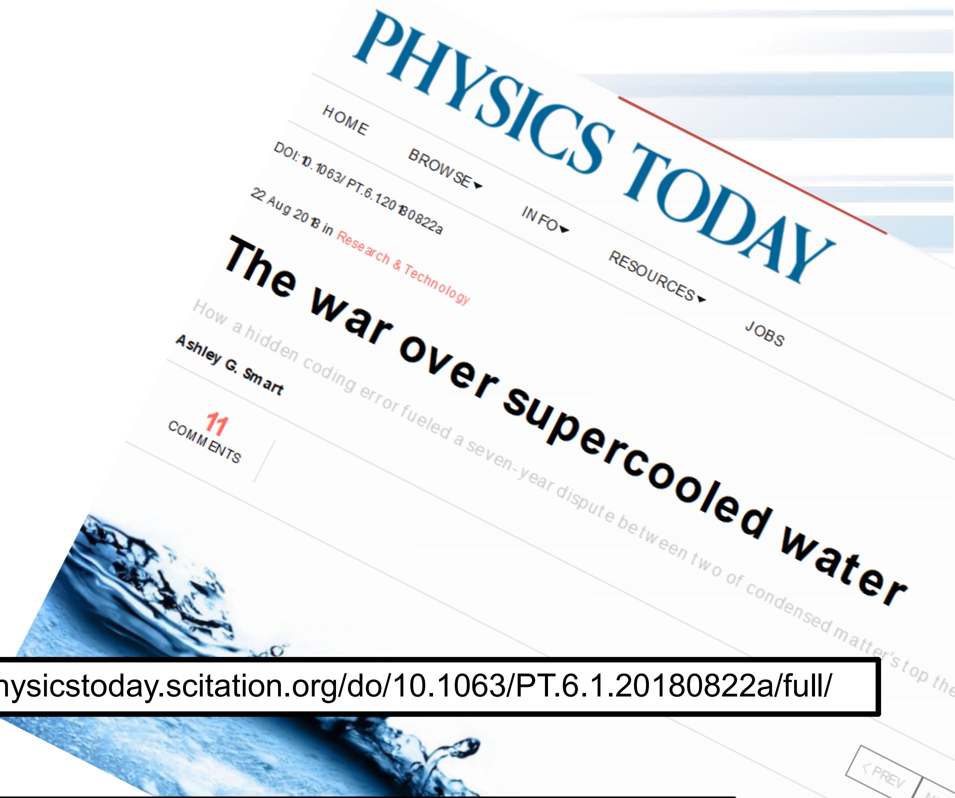
Fre

help

Source: https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full/

Computational Science Example

- Palmer located bug/feature in Berkeley code.
- Used to speed up LAMMPS execution.
- Replaced with more standard approach.
- Obtained result similar to Debenedetti 2009.
- Resolution took 7 years.



DOI: 10.1063/PT.6.1.20180822a

22 Aug 2018 in Research & Technology

The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top the

Ashley G. Smart

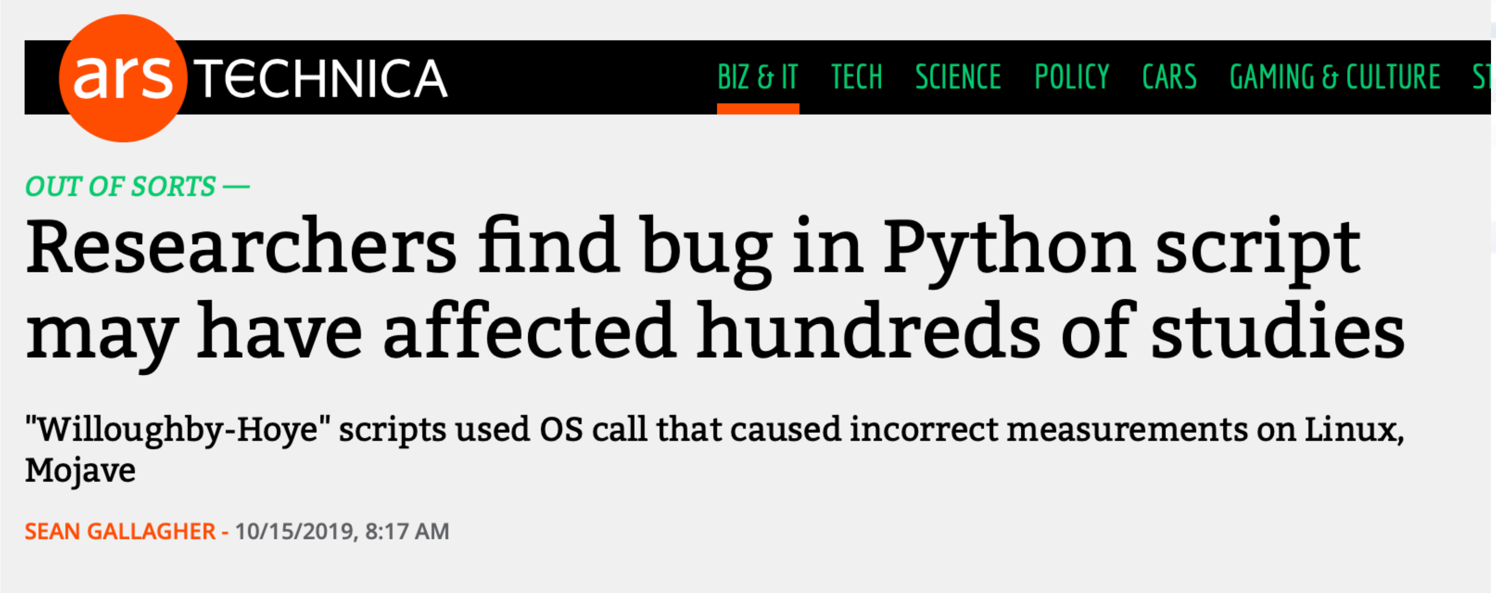
11 COMMENTS

Source: https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full/

> *For Palmer, the ordeal exemplifies the importance of transparency in scientific research, an issue that has recently drawn heightened attention in the science community. "One of the real travesties," he says, is that "there's no way you could have reproduced [the Berkeley team's] algorithm—the way they had implemented their code—from reading their paper." Presumably, he adds, "if this had been disclosed, this saga might not have gone on for seven years."*

Most Recent Example: My Response, “What about a unit test?”

- scripts' use of Python's glob module
- Generated different file lists in Linux and Mac Mojave
- Casts doubt on results in 150 papers.



OUT OF SORTS —

Researchers find bug in Python script may have affected hundreds of studies

"Willoughby-Hoye" scripts used OS call that caused incorrect measurements on Linux, Mojave

SEAN GALLAGHER - 10/15/2019, 8:17 AM

https://arstechnica.com/information-technology/2019/10/chemists-discover-cross-platform-python-scripts-not-so-cross-platform/

Publication Trends

Increased Emphasis on Transparency & Reproducibility

ACM TOMS Reproducible Computational Results (RCR)

- Submission: Optional RCR option.
- Standard reviewer assignment: Nothing changes.
- RCR reviewer assignment:
  - Concurrent with standard reviews.
  - As early as possible in review process.
  - Known to and works with authors during the RCR process.
- RCR process:
  - Multi-faceted approach, Bottom line: Trust the reviewer.
- Publication:
  - Reproducible Computational Results Designation.
  - The RCR referee acknowledged.
  - Review report appears with published manuscript.

SC19 Reproducibility Initiative

- Two appendices:
  - Artifact description (AD).
    - Blue print for setting up your computational experiment.
    - Makes it easier to rerun computations in future.
    - AD appendix will be mandatory for SC19 paper submissions.
  - Artifact Evaluation (AE).
    - Targets ”boutique” environments.
    - Improves trustworthiness when re-running hard, impossible.
- Details:
  - https://sc19.supercomputing.org/submit/reproducibility-initiative/

Improving Trustworthiness at Scale

What if we can’t re-run a computational experiment?

Reproducibility and Supercomputing

Scenario:
You compute a “hero” calculation using 5M core-hours on Mira and submit your results for publication. During the review process, a referee questions the validity of your results. What options are feasible:

- The reviewer re-runs your code on a laptop or cluster.
- The reviewer re-runs your code on Mira.
- You re-run your code on Mira.
- Your results are rejected.
- Your results are accepted, but with risk.

Sources for meta-computations

- Synthetic operators with known:
  - Spectrum (Huge diagonals).
  - Rank (by constructions).
- Invariant subspaces:
  - Example: Positional/rotational invariance (structures).
- Conservation principles:
  - Example: Flux through a finite volume.
- General:
  - Pre-conditions, post-conditions, invariants.

Can you think of something for your problems?

Example: HPCG Benchmark

- Symmetry:
  - For any linear operator A, $x^TAy = y^TA^Tx$.
  - If A symmetric $A = A^T$, so $x^TAy = y^TAx$.
  - And $\boldsymbol{x^TAy - y^TAx = 0}$.
- HPCG computes the above expression for:
  - User matrix and the preconditioner.
  - Numerical detail: Need to scale by vector & matrix norms.

Coming to Your World Soon: Reproducibility Requirements

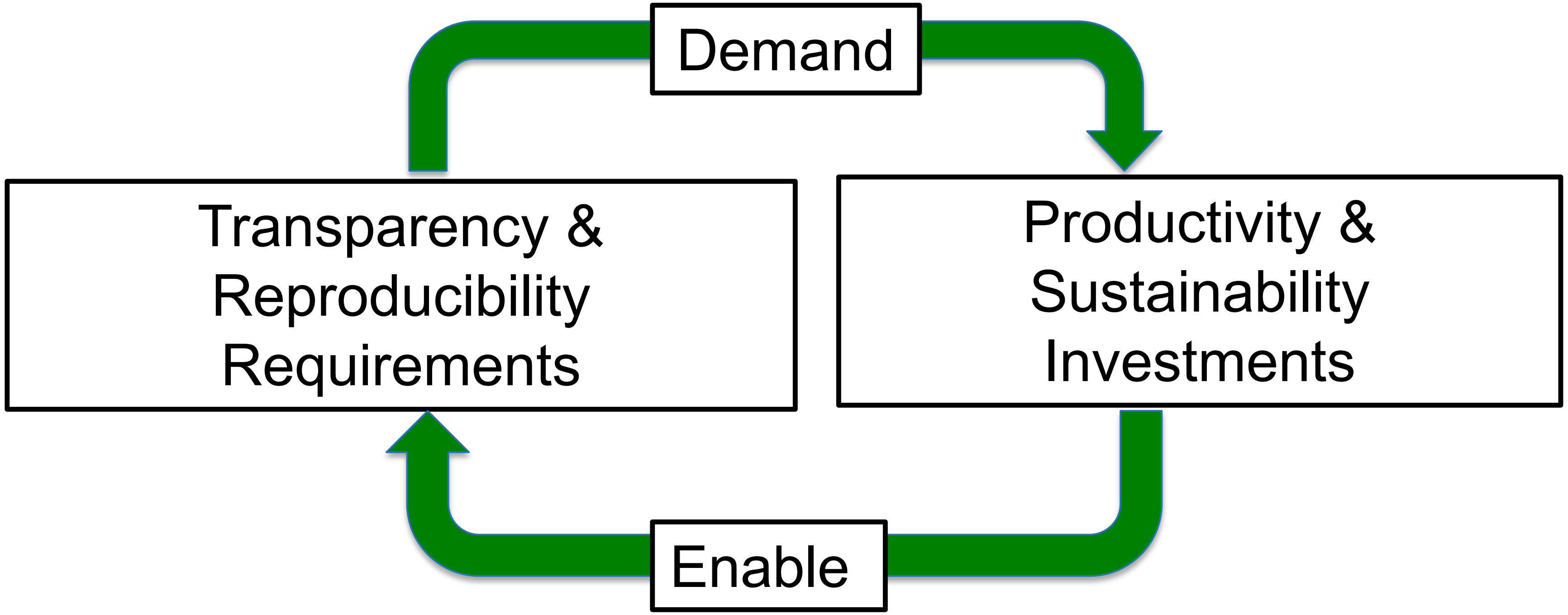
- These conferences have artifact evaluation appendices:
  - CGO, PPoPP, PACT, RTSS and SC.
  - http://fursin.net/reproducibility.html
- ACM ~~Replicated~~ Reproducible Computational Results (RCR).
  - ACM TOMS, TOMACS.
  - http://toms.acm.org/replicated-computational-results.cfm
- ACM Badging.
  - https://www.acm.org/publications/policies/artifact-review-badging
- NISO Committee on Reproducibility and Badging.
  - https://www.niso.org/niso-io/2019/01/new-niso-project-badging-scheme-reproducibility-computational-and-computing
  - Publishers: ACM, IEEE, figshare, STM, Reed Elsevier, Springer Nature

Questions, comments?

Better Productivity and Sustainability

Essential for affordable reproducibility

Incentives Demand Investments, Enabled by Investments


Demand
Transparency & Reproducibility Requirements
Productivity & Sustainability Investments
Enable

Common statement: “I would love to do a better job on my software, but I need to:

- Get this paper submitted.
- Complete this project task.
- Do something my employer values more.

Goal: Change incentives to include value of better software, better science.

Tradeoffs: Better, faster, cheaper

- “Better, faster, cheaper: Pick two of the three.”
  - Scenario: (Today)
    You are behind in developing a sophisticated new model in your software that you want to use for results in an upcoming paper.
  - Which of these could be reasonable choices?
    - Develop a simpler model for the paper.
    - Set other work aside and spend more time on development.
    - Ask for an extension on the paper deadline.
    - Develop sophisticated model, but don’t test its correctness.
    - Develop sophisticated model, but don’t document it or check it in.

Improved developer productivity

“Better, faster, cheaper: Pick all three.” – Near term.

Scenario: (6 months later)
After investing in **developer productivity improvements**, you are on time in developing a sophisticated new model in your software that you want to use for results in an upcoming paper.

Invest in developer tools, processes, practices.

Improved software sustainability

“Better, faster, cheaper: Pick all three.” – Long term.

Scenario: (3 years later)
After investing in **software sustainability improvements**, you are on time in developing **several** sophisticated new models in your software that you want to use for results in upcoming papers.

Invest in testing, documentation, integration for long-term software usability.

Which of These Enhance Reproducibility?

- Code written by first-year, untrained grad student.
- Tuning for high performance.
- Dynamic parallelism of modern processors.
- Better software testing.
- Source code and versioning management.
- Investing in developer productivity.
- Investing in software sustainability.

Personal Expectations

Calling out the best in team members

Final Thoughts: Commitment to Quality

Canadian engineers' oath (taken from Rudyard Kipling):

My Time I will not refuse;
my Thought I will not grudge;
my Care I will not deny
toward the honour, use,
stability and perfection of
any works to which I may be
called to set my hand.

https://www.egbc.ca/Member-Programs/Students/Iron-Ring

A Few Concrete Recommendations

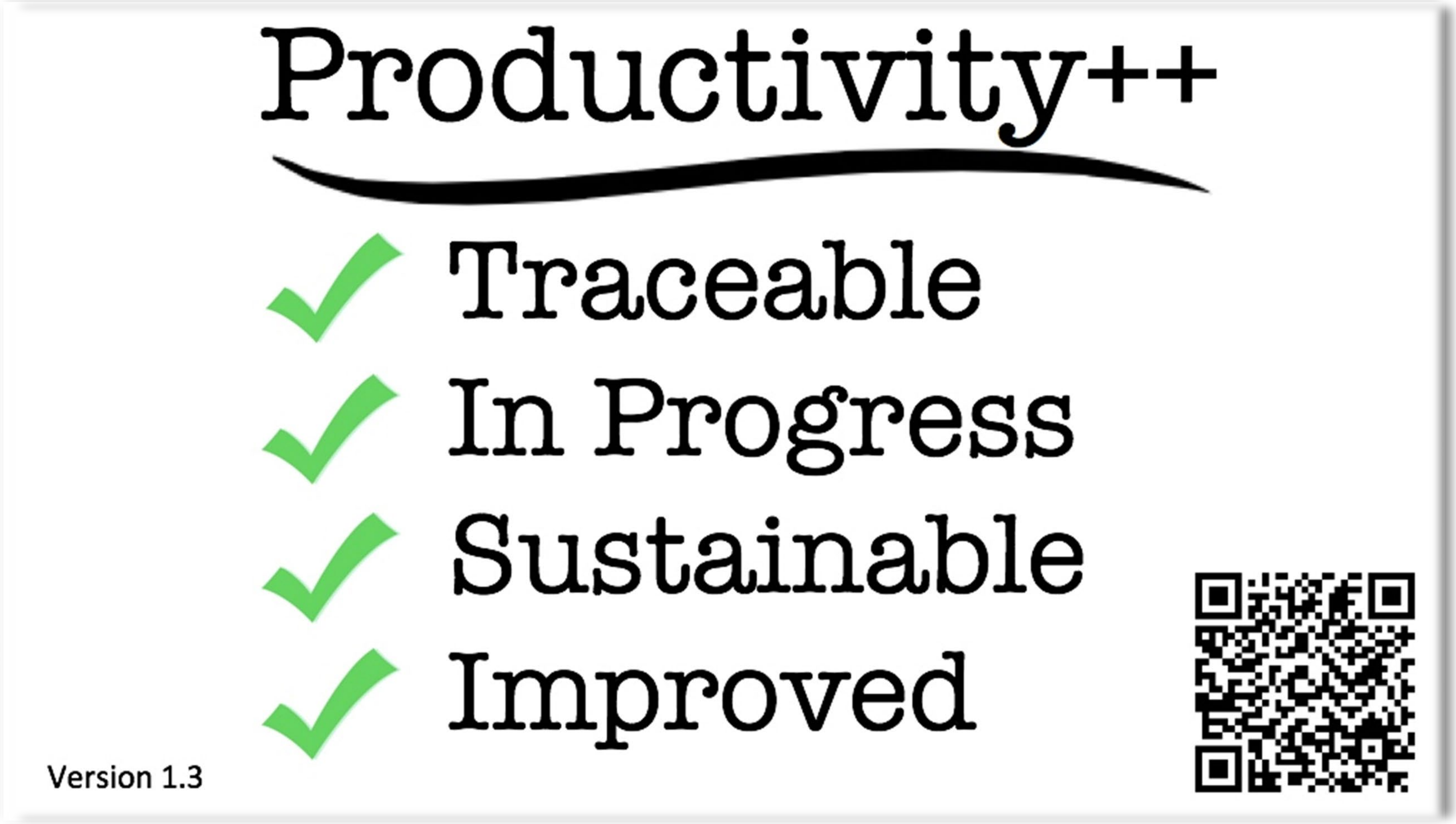
Show me the person making the most commits on an undisciplined software project and I will show you the person who is injecting the most technical debt.

- GitHub stats: Easy to find who made the most commits.
  - Some people: Pride in their high ranking.

- Instead, be the person who ranks high in these ways:
  - Writes up requirements, analysis and design, even if simple.
  - Writes good GitHub issues, tracks their progress to completion.
  - Comments on, tests and accepts pull requests.
  - Provide good wiki, gh-pages content, responses to user issues.

(Personal) Productivity++ Initiative

Ask: *Is My Work ________ ?*



Productivity++

- ✓ Traceable
- ✓ In Progress
- ✓ Sustainable
- ✓ Improved

Version 1.3

https://github.com/trilinos/Trilinos/wiki/Productivity---Initiative

Summary

- Reproducibility demands are coming.
  - Conferences first, journals slower.
- HPC software is particularly challenging:
  - Hardware variation.
  - Code optimization.
  - Dynamic parallelism.
- Better software practices:
  - Improve chances for reproducibility.
  - Lower its cost.
- Many tools emerging to enable reproducibility.

Other resources

Editorial: ACM TOMS Replicated Computational Results Initiative. Michael A. Heroux. 2015. *ACM Trans. Math. Softw.* 41, 3, Article 13 (June 2015), 5 pages. DOI: http://dx.doi.org/10.1145/2743015

Enhancing Reproducibility for Computational Methods. Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer Science (09 Dec 2016), pp. 1240-1241