Improving Reproducibility Through Better Software Practices

SIAM CSE
Atlanta, GA
February 28, 2017

Tutorial slides available at: http://bit.ly/siam-cse17-mt3

Michael A. Heroux, Sandia National Laboratories

Acknowledgments

- Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2016-8466 C.

Outline

- Introduction
- Increasing focus on reproducibility.
- Role of better software practices.
- Practical ideas:
  - Testing.
  - Containers.
  - Other projects.

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

http://www.nytimes.com/2015/08/28/science/many-social-science-findings-not-as-strong-as-claimed-study-says.html?_r=0

Reproducibility

- NY Times highlights "problems".
- Only one of many cited examples.
- HPC has been spared this "spotlight" (so far).

- Lots of activity:
  - AAAS, ACM initiatives.
  - PPoPP, Supercomputing 2016.
- But what is reproducibility?

IDEAS productivity

Reproducibility Terminology

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. 2013. Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics. (2013). https://icerm.brown.edu/tw12-5-rcem/icerm_report.pdf

- **Reviewable Research.** The descriptions of the research methods can be independently assessed and the results judged credible. (This includes both traditional peer review and community review, and does not necessarily imply reproducibility.)
- **Replicable Research.** Tools are made available that would allow one to duplicate the results of the research, for example by running the authors' code to produce the plots shown in the publication. (Here tools might be limited in scope, e.g., only essential data or executables, and might only be made available to referees or only upon request.)
- **Confirmable Research.** The main conclusions of the research can be attained independently without the use of software provided by the author. (But using the complete description of algorithms and methodology provided in the publication and any supplementary materials.)
- **Auditable Research.** Sufficient records (including data and software) have been archived so that the research can be defended later if necessary or differences between independent confirmations resolved. The archive might be private, as with traditional laboratory notebooks.
- **Open or Reproducible Research.** Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.

ACM Transactions on
Mathematical Software

ACM TOMS

- TOMS RCR Initiative: Referee Data.
- Why TOMS? Tradition of real software that others use.
- Two categories: Algorithms, Research.
- TOMS Algorithms Category:
  - Software Submitted with manuscript.
  - Both are thoroughly reviewed.
- TOMS Research Category:
  - Stronger: Previous implicit “real software” requirement is explicit.
  - New: Special designation for replicated results.

IDEAS productivity

ACM TOMS Replicated 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:
  - Replicated Computational Results Designation.
  - The RCR referee acknowledged.
  - Review report appears with published manuscript.

RCR Process: Two Basic Approaches

1. Independent replication (3 options):
   A. Transfer of, or pointer to, author's software.
   B. Guest account, access to author's software.
   C. Observation of authors replicating results.

Or (Untested, rare)

2. Review of computational results artifacts:
   - Results may be from an unavailable system.
   - Leadership class computing system.
   - In this situation:
     - Careful documentation of the process.
     - Software should have its own substantial V&V process.

TOMS:

- First RCR paper in TOMS issue 41:3
  - Editorial introduction.
  - van Zee & van de Geijn, BLIS paper.
  - Referee report.
- Second: TOMS 42:1
  - Hogg & Scott.
- Third: TOMS 42:4.
- Several others in queue.

TOMACS

- Similar.

Big Picture of ACM RCR

Thank you for taking the time to consider our paper for your journal.

XXX has agreed to undergo the RCR process should the paper proceed far enough in the review process to qualify. ***To make this easier we have preserved the exact copy of the code used for the results (including additional code for generating detailed statistics that is not in the library version of the code).***

- Improve science.
  - Quality of prose: Good.
  - Quality of data: Poor.
- So bad now:
  - Trust comes from seeing a "cloud" of similar papers with similar results.
  - Which could still be wrong (built on a common bad piece).
  - Replicability: First step toward improvement.
- Engage a "dark portion" of the R&D community.
  - Reviewers not among typical reviewer pool.
  - Practitioners, users. Expert at use of Math SW.

Coming to Your World Soon: Reproducibility Requirements

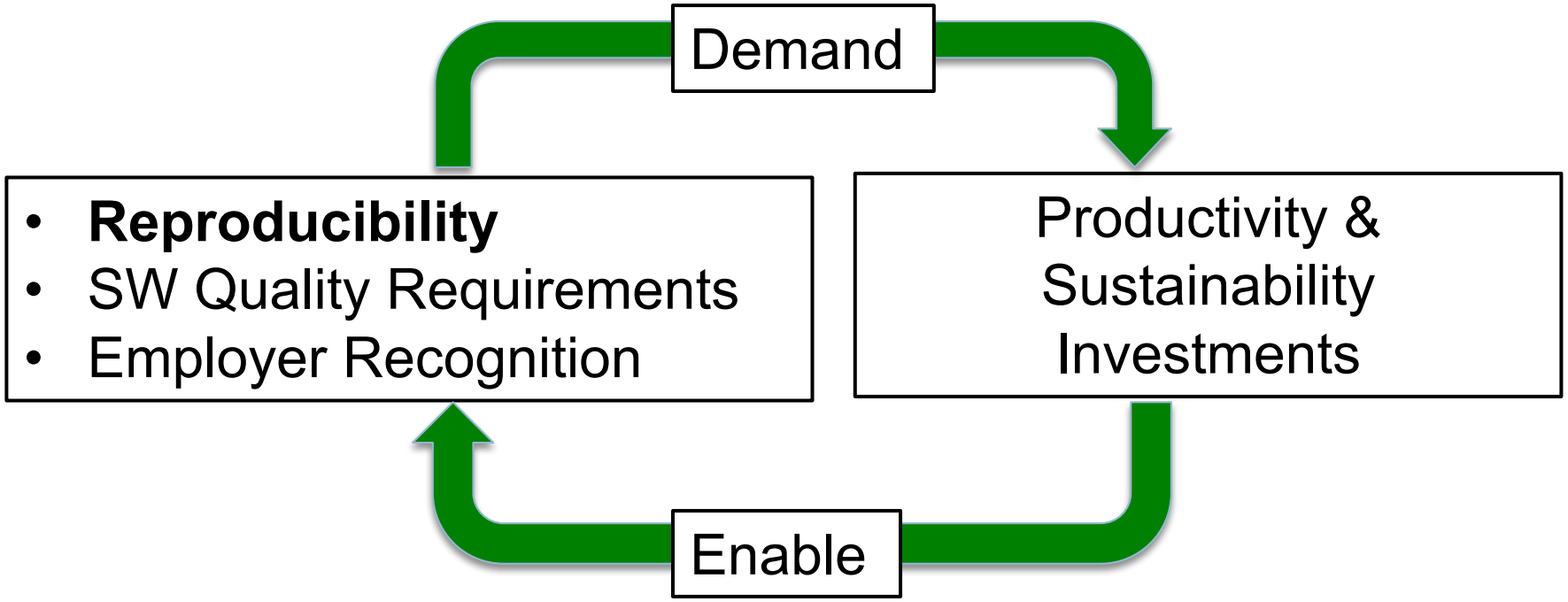
- These conferences expect artifact evaluation appendices (most optionally):
  - CGO, PPoPP, PACT, RTSS and SC.
  - http://fursin.net/reproducibility.html
- ACM Replicated Computational Results (RCR).
  - ACM TOMS, TOMACS.
  - http://toms.acm.org/replicated-computational-results.cfm
- ACM Badging.
  - https://www.acm.org/publications/policies/artifact-review-badging

How can you prepare?

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 software sustainability.

Incentives To Change



Common statement: "I would love to do a better job, 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.

SC17 Reproducibility Initiative

- Two appendices:
  - Artifact description (AD).
    - Blue print for setting up your computational experiment.
    - Makes it easier to rerun computations in future.
  - Computational Results Analysis (CRA).
    - Targets ”boutique” environments.
    - Improves trustworthiness when re-running hard, impossible.
- Details: http://sc17.supercomputing.org/submitters/technical-papers/reproducibility-initiatives-for-technical-papers/

Example: HPCG Benchmark

- Exploit two properties:
  - Spectral properties of CG:
    - Eigenvalue clustering.
    - CG convergence related to number of *distinct* eigenvalues.
  - Operator symmetry:
    - Compact Finite Difference operator is symmetric.
    - Multigrid is symmetric.

Example: HPCG Benchmark

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

https://github.com/hpcg-benchmark/hpcg/blob/master/src/TestSymmetry.cpp


```
// Test symmetry of matrix
// First load vectors with random values
FillRandomVector(x_ncol);  FillRandomVector(y_ncol);
double xNorm2, yNorm2;  double ANorm = 2 * 26.0;  double xtAy = 0.0;

// Next, compute x'*A*y
ComputeDotProduct(nrow, y_ncol, y_ncol, yNorm2, t4, A.isDotProductOptimized);
int ierr = ComputeSPMV(A, y_ncol, z_ncol); // z_nrow = A*y_overlap
ierr = ComputeDotProduct(nrow, x_ncol, z_ncol, xtAy, t4, A.isDotProductOptimized); // x'*A*y

// Next, compute y'*A*x
// …
testsymmetry_data.depsym_spmv = std::fabs((long double) (xtAy - ytAx))/((xNorm2*ANorm*yNorm2 +
yNorm2*ANorm*xNorm2) * (DBL_EPSILON));
if (testsymmetry_data.depsym_spmv > 1.0) ++testsymmetry_data.count_fail;  // If the difference is > 1, count it
wrong

// Test symmetry of multi-grid
// Compute x'*Minv*y
// …
```

Example: HPCG Benchmark

- Eigenvalue clustering:
  - HPCG matrix is 27-point finite difference stencil.
    - -1 off diagonals, diagonally dominant, zero Dirichlet BCs.
    - Max diagonal value – 27.
  - Idea: Temporarily replace diagonal values.
    - For i=1:9 A(i,i) = i*1.0E6
    - For i>9 A(i,i) = 1.0E6
- Questions:
  - How many distinct diagonal values?
  - How many CG iterations?
  - How many preconditioned CG iterations?

https://github.com/hpcg-benchmark/hpcg/blob/master/src/TestCG.cpp


```
// Modify the matrix diagonal to greatly exaggerate diagonal values.
  // CG should converge in about 10 iterations for this problem, regardless of
problem size
  for (local_int_t i=0; i< A.localNumberOfRows; ++i) {
    global_int_t globalRowID = A.localToGlobalMap[i];
    if (globalRowID<9) {
      double scale = (globalRowID+2)*1.0e6;
      ScaleVectorValue(exaggeratedDiagA, i, scale);
      ScaleVectorValue(b, i, scale);
    } else {
      ScaleVectorValue(exaggeratedDiagA, i, 1.0e6);
      ScaleVectorValue(b, i, 1.0e6);
    }
  }
  ReplaceMatrixDiagonal(A, exaggeratedDiagA);
// …
```

Sources for CRA metrics

- 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?

Container based workflows

- https://github.com/systemslab/popper/wiki/Intro-to-Popper

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

Outline

Part I: 9:10-10:50 am

- [10 min] **Background, introductions, objectives, setup**
- [15 min] **Why effective software practices are essential for CSE projects**
- [25 min] **Software licensing**
- [50 min] **Effective models, tools, processes, and practices for small teams, including agile workflow management**
  - Interactive exercises

Part II: 1:30-3:10 pm

- [25 min] **Reproducibility**
- [75 min] **Scientific software testing**
  - Automated testing and continuous integration
  - Interactive exercises for code coverage
    - Access to Linux environment with Git and GNU compiler suite