Benchmarking to Close the Credibility Gap: A Computational BioEM Benchmark Suite

J. W. MASSEY, C. LIU, and A. E. YILMAZ

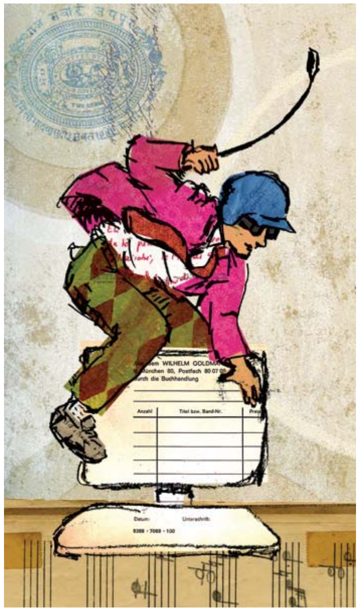
Institute for Computational Engineering & Sciences
Department of Electrical & Computer Engineering
The University of Texas at Austin

URSI Commission B
International Symposium on Electromagnetic Theory (EMTS 2016)
Espoo, Finland, 14-18 August 2016

Outline

- The Credibility Gap: A Present and Growing Challenge in Computational EM
  - Ubiquity of Error
  - Pillars of Science
  - One of the Hallmarks of Science: Independent Reproducibility
  - The Many Levels of Reproducibility
    - From Internal Repetition to Independent Corroboration
  - Really Reproducible Research: A Possible Approach to Closing the Credibility Gap
- Alternative to Closing the Credibility Gap: Benchmarking
  - Benchmarking to the Rescue?
  - 4 Key Ingredients
  - Better Benchmarking
- Example: Austin Computational BioEM Benchmark
- Conclusion

The Credibility Gap



"The traditional image of the scientist … is long obsolete. The more accurate image … depicts a computer jockey working at all hours to launch experiments on computer servers …"

"A rapid transition is now under way…that will finish with computation as absolutely central to scientific enterprise. However, … scientific computing has already brought us to a state of crisis …"

"The prevalence of very relaxed attitudes about communicating experimental details and validating results is causing a large and growing credibility gap. It's impossible to verify most of the results that computational scientists present at conferences and in papers."

D. L. Donoho *et al.*, "Reproducible research in computational harmonic analysis," *Comp. Sci. Eng.*, Jan.-Feb. 2009.

Original cartoons from:
http://www.123rf.com/photo_18862130_illustration--cartoon-character-scientist-in-laboratory-on-white-background.html
http://www.weliver.com/index.php/9-illustration/6-desk-jockey

The Credibility Gap

Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
UB Center for Computational Research University at Buffalo The State University of New York

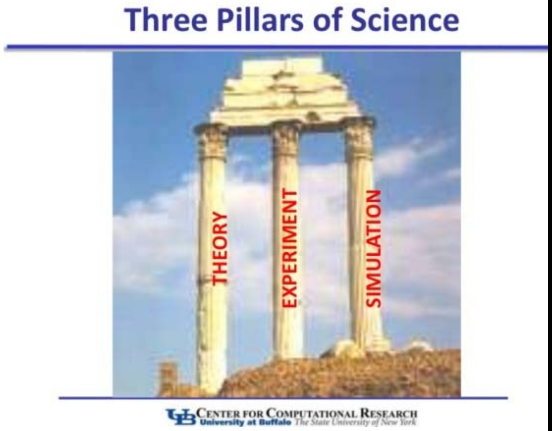
"Originally, there were two scientific methodological branches—deductive (e.g., mathematics) and empirical (e.g., statistical data analysis of controlled experiments). Many scientists accept computation (e.g., large-scale simulation) as the third branch … while computation is already indispensable, it does not yet deserve elevation to third-branch status because current computational science practice doesn't generate routinely verifiable knowledge."

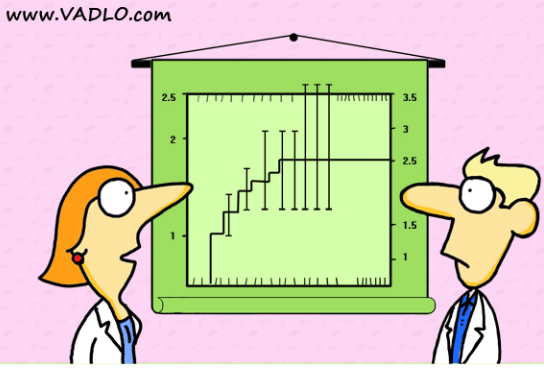
"The scientific method's central motivation is the *ubiquity of error*— … mistakes and self-delusion can creep in absolutely anywhere … the scientist's effort is primarily expanded in recognizing and rooting out error … Before scientific computation can be accorded the status it aspires to, it must be practiced in a way that accepts the ubiquity of error, and work then to identify and root out error."

D. L. Donoho *et al.*, "Reproducible research in computational harmonic analysis," *Comp. Sci. Eng.*, Jan.-Feb. 2009.

Original images from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012
http://vadlo.com/cartoons.php?id=22

The Credibility Gap


Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
Center for Computational Research
University at Buffalo The State University of New York



"Did you really have to show the error bars?"

"Like deduction and empiricism, computation is also highly error-prone… In stark contrast to the sciences relying on deduction or empiricism, computational science is far less visibly concerned with the ubiquity of error. At conferences and in publications, it's now completely acceptable for a researcher to simply say, "here is what I did, and here are my results." Presenters devote almost no time to explaining why the audience should believe that they found and corrected errors in their computations. The presentation's core isn't about the struggle to root out error—as it would be in mature fields—but is instead a sales pitch: an enthusiastic presentation of ideas and a breezy demo of an implementation.

D. L. Donoho *et al.*, "Reproducible research in computational harmonic analysis," *Comp. Sci. Eng.*, Jan.-Feb. 2009.

Original images from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012
http://vadlo.com/cartoons.php?id=22

The Credibility Gap

Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
Center for Computational Research
University at Buffalo The State University of New York

> “Computational science has nothing like the elaborate mechanisms of formal proof in mathematics or meta-analysis in empirical science. Many users of scientific computing aren’t even trying to follow a systematic, rigorous discipline that would in principle allow others to verify the claims they make. How dare we imagine that computational science, as routinely practiced, is reliable!”
>
> D. L. Donoho *et al.*, “Reproducible research in computational harmonic analysis,” *Comp. Sci. Eng.*, Jan.-Feb. 2009.

Original images from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012

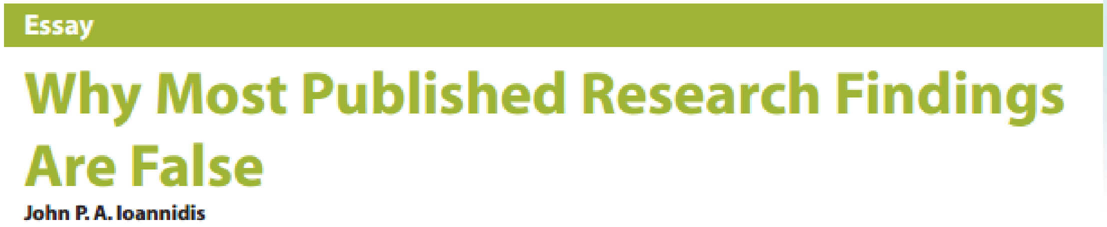
The Credibility Gap

Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
UB Center for Computational Research
University at Buffalo The State University of New York

> “Computational science has nothing like the elaborate mechanisms of formal proof in mathematics or meta-analysis in empirical science. Many users of scientific computing aren’t even trying to follow a systematic, rigorous discipline that would in principle allow others to verify the claims they make. How dare we imagine that computational science, as routinely practiced, is reliable!”

D. L. Donoho *et al.*, “Reproducible research in computational harmonic analysis,” *Comp. Sci. Eng.*, Jan.-Feb. 2009.

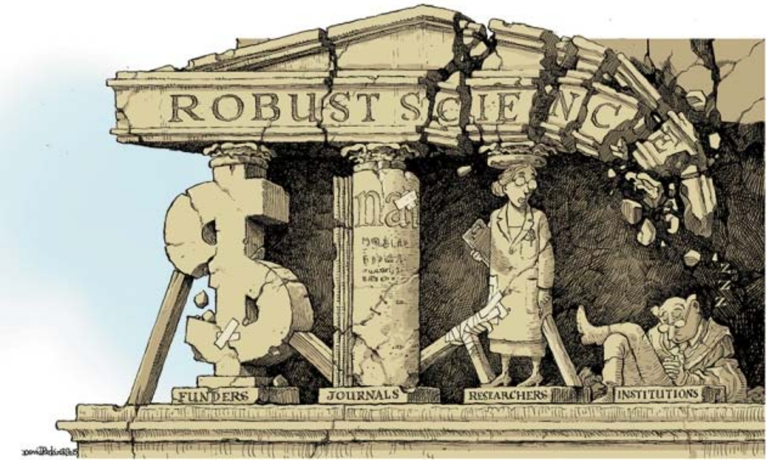
In practice, other pillars of science also suffer from reliability/verifiability/ reproducibility problems



Essay

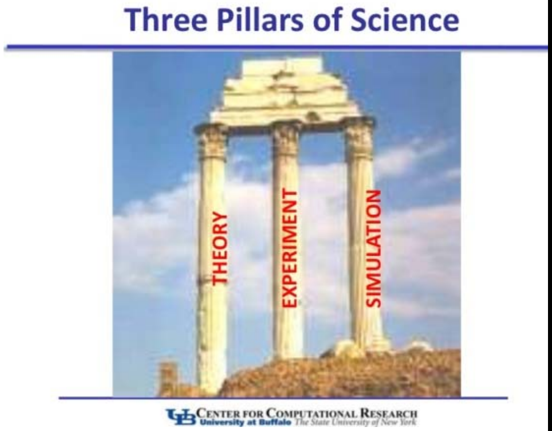
Why Most Published Research Findings Are False

John P. A. Ioannidis



Original images from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012
http://www.nature.com/news/robust-research-institutions-must-do-their-part-for-reproducibility-1.18259

One of the Hallmarks of (Empirical) Science: Independent Replication


Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
Center for Computational Research
University at Buffalo The State University of New York

In practice, other pillars of science also suffer from reliability/verifiability/ reproducibility problems

"Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into testable laws and theories. The success and credibility of science are anchored in the willingness of scientists to:

1. Expose their ideas and results to independent testing and replication by others. This requires the open exchange of data, procedures and materials.

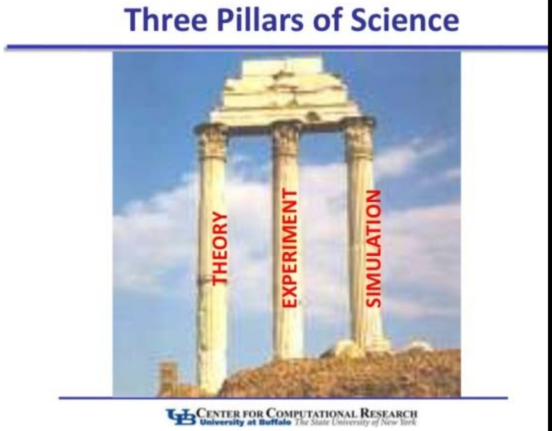
2. Abandon or modify previously accepted conclusions when confronted with more complete or reliable experimental or observational evidence.

Adherence to these principles provides a mechanism for self-correction that is the foundation of the credibility of science."

American Physical Society, "What is science?" adopted Nov. 1999.

Original image from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012

One of the (Fading) Hallmarks of (Theoretical) Science: Surveyability


Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
Center for Computational Research
University at Buffalo The State University of New York

In practice, other pillars of science also suffer from reliability/verifiability/ reproducibility problems

"The old four-color problem was a problem of mathematics for over a century. Mathematicians appear to have solved it to their satisfaction, but their solution raises a problem for philosophy… What is a proof? …

(b) Proofs are surveyable. … they can be checked by members of the mathematical community…Genius in mathematics lies in the discovery of new proofs, not in the verification of old ones…

(c) Proofs are formalizable…a proof is a finite sequence of formulas of a formal theory satisfying certain conditions….

…There is no surveyable proof of the lemma … there is a formal proof. Our knowledge of this is grounded, in part, in the results of a well-conceived computer experiment"

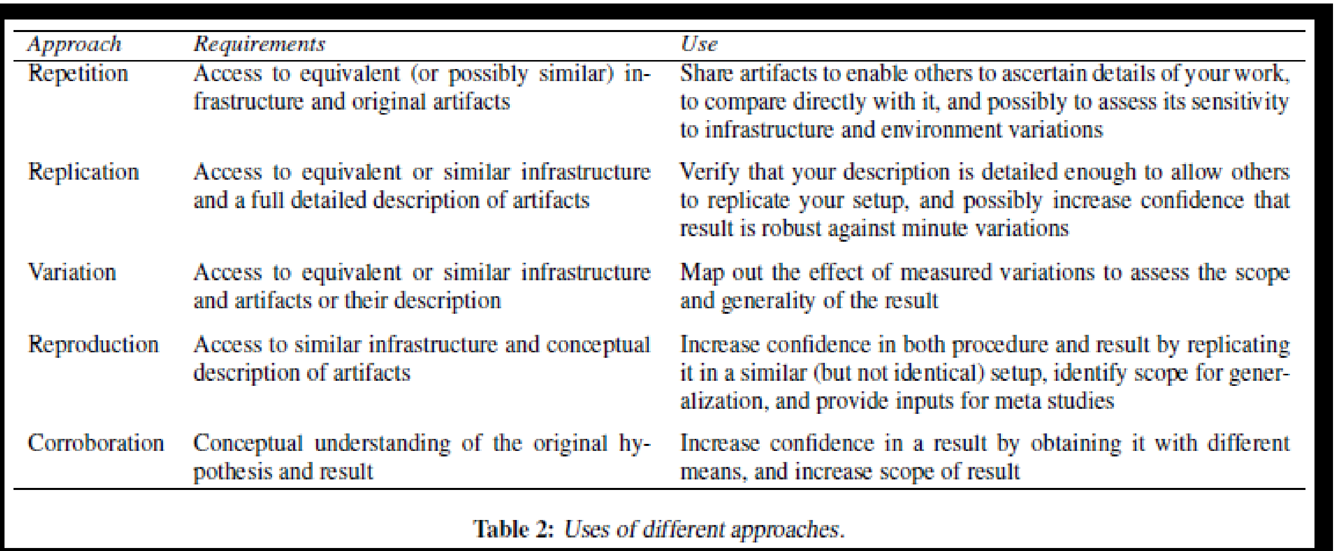
T. Tymoczko, "The four-color problem and its philosophical significance," *The Journal of Philosophy,* Feb. 1979.

Original image from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012

The Many Levels of Reproducibility: From Exact Repetition to Corroboration

Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
UB Center for Computational Research
University at Buffalo The State University of New York

In practice, other pillars of science also suffer from reliability/verifiability/ reproducibility problems



Approach	*Requirements*	*Use*
Repetition	Access to equivalent (or possibly similar) infrastructure and original artifacts	Share artifacts to enable others to ascertain details of your work, to compare directly with it, and possibly to assess its sensitivity to infrastructure and environment variations
Replication	Access to equivalent or similar infrastructure and a full detailed description of artifacts	Verify that your description is detailed enough to allow others to replicate your setup, and possibly increase confidence that result is robust against minute variations
Variation	Access to equivalent or similar infrastructure and artifacts or their description	Map out the effect of measured variations to assess the scope and generality of the result
Reproduction	Access to similar infrastructure and conceptual description of artifacts	Increase confidence in both procedure and result by replicating it in a similar (but not identical) setup, identify scope for generalization, and provide inputs for meta studies
Corroboration	Conceptual understanding of the original hypothesis and result	Increase confidence in a result by obtaining it with different means, and increase scope of result

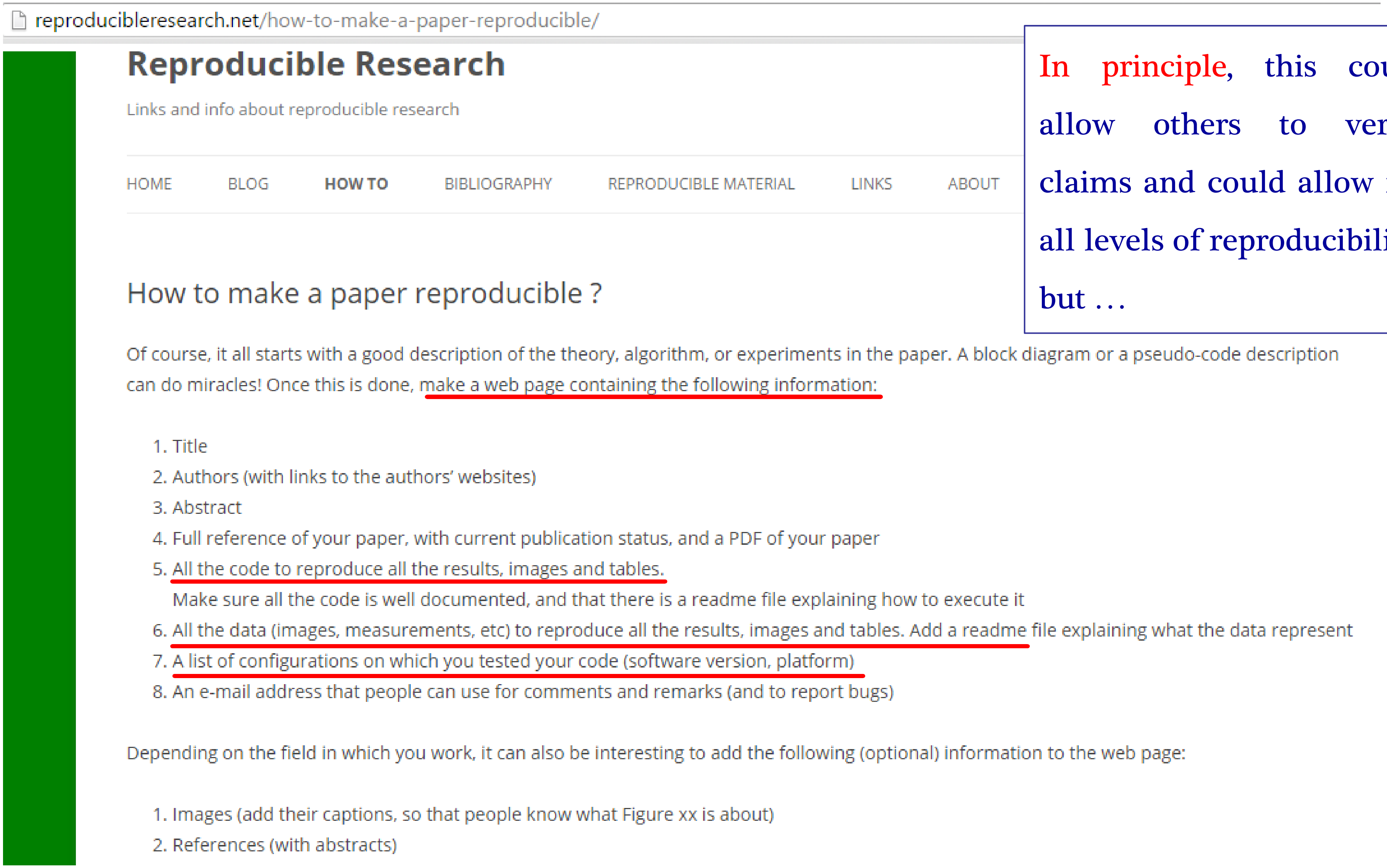
Table 2: *Uses of different approaches.*

"Being able to repeat experiments is considered a hallmark of the scientific method…but this can take many forms… Using "reproducibility" as a catch-all term loses fidelity. There are several levels of redoing previous experimental work, with differences in generalizability and scope (see Table 2)."

D. G. Feitelson, "From repeatability to reproducibility and corroboration," *ACM SIGOPS Oper. Sys. Rev.*, , Jan. 2015.

Original image from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012

"Really Reproducible Research": A Possible Approach to Closing the Credibility Gap?



reproducibleresearch.net/how-to-make-a-paper-reproducible/

Reproducible Research

Links and info about reproducible research

How to make a paper reproducible ?

Of course, it all starts with a good description of the theory, algorithm, or experiments in the paper. A block diagram or a pseudo-code description can do miracles! Once this is done, make a web page containing the following information:

1. Title
2. Authors (with links to the authors' websites)
3. Abstract
4. Full reference of your paper, with current publication status, and a PDF of your paper
5. All the code to reproduce all the results, images and tables.
   Make sure all the code is well documented, and that there is a readme file explaining how to execute it
6. All the data (images, measurements, etc) to reproduce all the results, images and tables. Add a readme file explaining what the data represent
7. A list of configurations on which you tested your code (software version, platform)
8. An e-mail address that people can use for comments and remarks (and to report bugs)

Depending on the field in which you work, it can also be interesting to add the following (optional) information to the web page:

1. Images (add their captions, so that people know what Figure xx is about)
2. References (with abstracts)

In principle, this could allow others to verify claims and could allow for all levels of reproducibility, but …

But …

Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
UB Center for Computational Research University at Buffalo The State University of New York

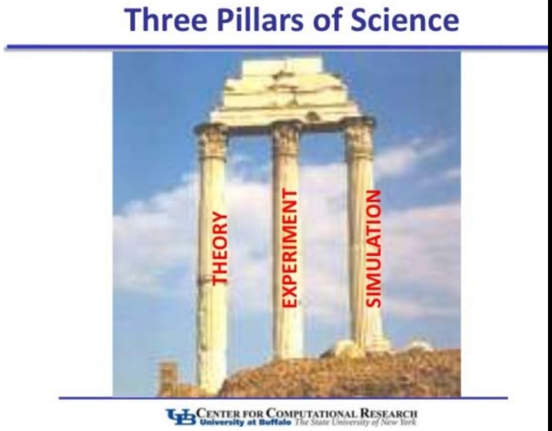
In practice, other pillars of science also suffer from reliability/verifiability/ reproducibility problems

"… separate … reproducibility, a generally desirable property, from replicability, its poor cousin … there are important differences between the two … crux of the matter is … reproducibility requires changes; replicability avoids them. A critical point of reproducing an experimental result is that unimportant things are intentionally not replicated … Although reproducibility is desirable … the impoverished version, replicability, is one not worth having. It would cause a great deal of wasted effort by members of our community… sharing of all the artifacts from people's experiments is not a trivial activity… at best, it would serve as little more than a policing tool, preventing outright fraud … there may be other virtues for having repositories of software … scientific reproducibility is not one of them."

C. Drummond, "Replicability is not reproducibility: nor is it good science," *4th Workshop Evaluation Methods Machine Learn.*, June 2009.

Original image from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012

But …


Three Pillars of Science
THEORY
EXPERIMENT
SIMULATION
Center for Computational Research
University at Buffalo The State University of New York

In practice, other pillars of science also suffer from reliability/verifiability/ reproducibility problems

"4. Determinism in numerical computing will be gone.

In fifty years, though the answers you get will be accurate without fail to the prescribed precision, you will not expect to duplicate them exactly if you solve the problem a second time…In the last fifty years, the great message communicated to scientists and engineers was that it is unreasonable to ask for exactness in numerical computation. In the next fifty, they will learn not to ask for repeatability, either.

7. Multipole methods and their descendants will be ubiquitous.

… Times have changed, and we are all asymptotickers…The success of multipole methods will exemplify a general trend. As time goes by, large-scale numerical computations rely more on approximate algorithms…more robust than exact ones and …also often faster."

L. N. Trefethen, "Predictions for scientific computing fifty years from now," *Mathematics Today*, Jan. 2000.

Original image from:
http://www.slideshare.net/ultrafilter/trends-challenges-in-supercomputing-for-eitaeitc-2012

Alternative to Closing the Credibility Gap: Benchmarking

- Benchmarking to the rescue?
- 4 Key Ingredients
- Better Benchmarking

Benchmarking to the Rescue?

- Verification, validation and performance benchmarks can
  - + help systematically combat the ubiquity of error
  - + inform public and researchers in the field about state of the art
  - + lower barriers to entry of new researchers/methods/tools
  - + reduce importance of subjective factors when judging simulation tools
  - + increase credibility of claims made by computational scientists and engineers
- Benchmark suites must
  - contain problems, quantities of interest, reference solutions, performance metrics
  - be many: each emphasizing/exercising features of computational methods most relevant to applications in sub-field of interest
  - strike balance between specialization (to be useful to applications in sub-field) and generalizability (to be predictive/representative for the different types of problems in sub-field)

4 Key Ingredients

- A precisely defined list of problems representative of a larger set of problems

  + problems should span different difficult levels, e.g., from basic and moderate to hard and challenge problems

  + list should evolve

- Clearly defined quantities of interest and reliable reference solutions for them

  + CEM benchmarkers are lucky: Analytical results for canonical shapes

  + for more complex problems, other computational or experimental results must be used as (unreliable) references

- Performance (error and computational cost) measures

  + must also quantify computational power available to the simulation and normalize costs across platforms

- Online databases

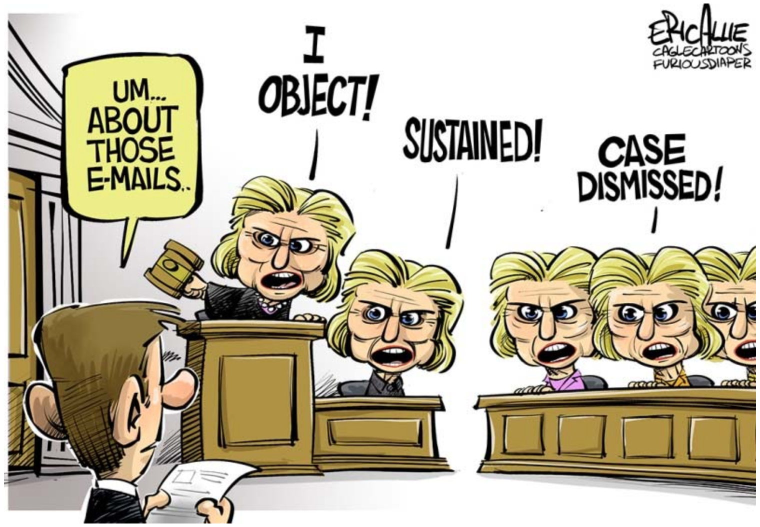
  + openness of benchmark results and exposure are important to build confidence

Better (External) Benchmarking

- Pitfall: Methods are often evaluated primarily by the same researchers who developed them

Better (External) Benchmarking

- Pitfall: Methods are often evaluated primarily by the same researchers who developed them


UM... ABOUT THOSE E-MAILS..
I OBJECT!
SUSTAINED!
CASE DISMISSED!
ERICALLIE CAGLECARTOONS FURIOUSDIAPER

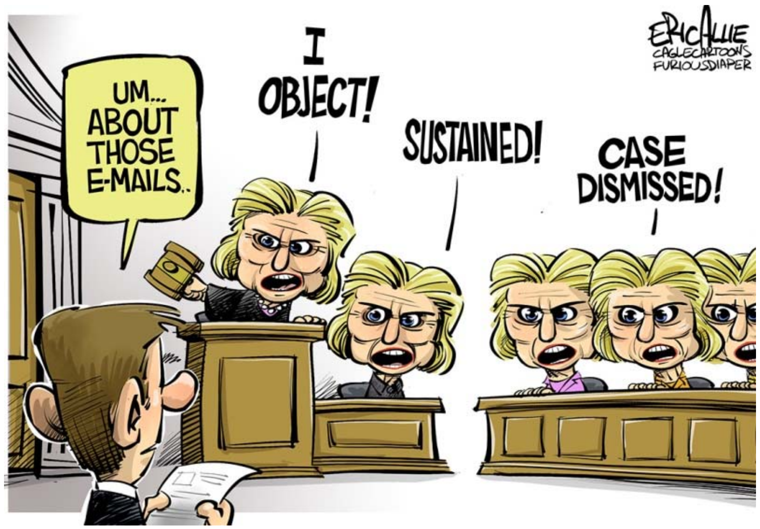
Original cartoons from:
http://www.ocregister.com/articles/ocregister-39908-left-margin.html
http://furiousdiaper.com/wp-content/uploads/2015/03/150320emailTFD.jpg

Better (External) Benchmarking

- Pitfall: Methods are often evaluated primarily by the same researchers who developed them

+ Competition-based or challenge-based benchmarking can help (but have myriad limitations and costs)

+ Must blind method developers to part (not all) of the benchmarking process



Original cartoons from:
http://www.ocregister.com/articles/ocregister-39908-left-margin.html
http://furiousdiaper.com/wp-content/uploads/2015/03/150320emailTFD.jpg

Example: Austin Computational BioEM Benchmark

http://web.corral.tacc.utexas.edu/BioEM-Benchmarks/

- Problem Set
- Quantities of Interest and Reference Solutions
- Error and Cost Definitions
- Online Database
- Example Comparison in Benchmark

Conclusion

- Computational science and engineering faces a “large and growing” credibility gap
  - similar to other branches: independent repetition, understanding, corroboration difficult
  - how important are (external) repetition, replication, variation, reproduction, and corroboration of ideas and results?
  - should we/can we perform *really reproducible research* in computational EM?
    - Our answer: Aim for *internal* repeatability/replicability (e.g., using really reproducible research principles) and *external* reproducibility/corroboration (e.g., through benchmarking)
      - + publicly available data can already identify the norm and the outliers
      - + extraordinary claims/results/performance requires extraordinary evidence, e.g., ask claimer to participate in benchmark

- Publicly available verification, validation, and performance benchmarks can
  - + help increase reproducibility without placing undue burdens of (perfect) replication
  - + reduce importance of subjective factors when judging methods
  - + benchmarks should be (partially) blinded to method developers
  - + example: http://bit.ly/BioEM-Benchmarks