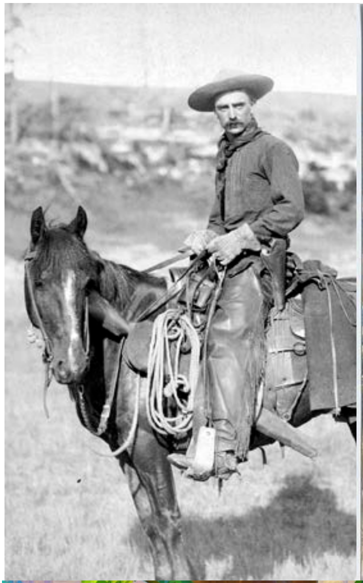
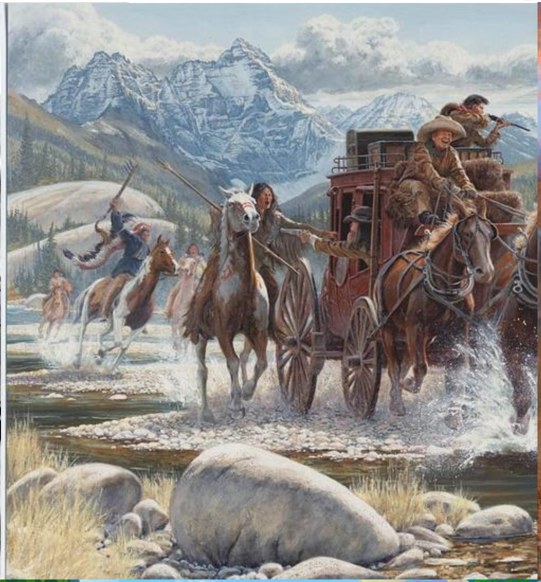
Benchmarking at the Frontiers of Computational EM

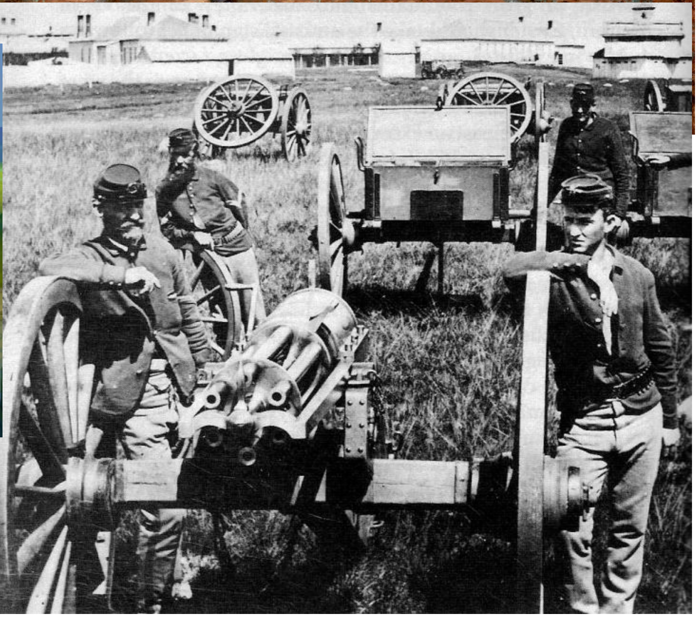
IEEE International AP-S Symposium
July 2017
San Diego
USNC-URSI Radio Science Meeting

AP-S/URSI Meeting
San Diego, CA, 9-4 July 2017

Benchmarking at the Frontiers of Computational EM







Original images from:

http://www.popularmechanics.com/military/a22451/history-gatling-gun/

http://jonmcnaughton.com/patriotic/wild-wild-west/

https://www.socialquantum.com/games/new_frontier

https://en.wikipedia.org/wiki/American_frontier#/media/File:Grabill_-_The_Cow_Boy-edit.jpg

https://images.fineartamerica.com/images-medium-large-5/trouble-for-the-overland-stage-kirk-stirnweis.jpg

Benchmarking at the Frontiers of Computational EM

- Before the Break: 13:20-15:00
  - “Advancing Computational Electromagnetics Though Benchmarking”
  - “The Benefit of Simple Benchmarks to Highlight Problems in CEM Codes”
  - “Benchmarking Full Wave Analysis of Periodic Structures: Non Perpendicularity at Periodic Boundaries”
  - “Benchmarking Computational Electromagnetics with Exact Analytical Solutions of Canonical Electromagnetic Scattering Problems”
  - “On Higher Order Imperative in Computational Electromagnetics through Benchmarking of Boundary Element methods for Canonical Scattering Problems”
- Break: 15:00-15:20
- After the Break: 15:20-17:00
  - “Benchmarking the Solutions of Billion-Unknown Problems”
  - “Accurate and Efficient Solution of Bioelectromagnetic Models”
  - “On Computational Electromagnetic Code Testing and Benchmarking”
  - “Figure of Merit for Computational Electromagnetics Solvers”
  - “Austin Benchmark Suite for Computational Bioelectromagnetics: AIM Performance Data”

Advancing Computational Electromagnetics Research Through Benchmarking

A. E. YILMAZ

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

AP-S/URSI Meeting
San Diego, CA, 9-4 July 2017

Outline

- Motivation & Observations
  - What is Benchmarking?
    - Performance
    - Theory of benchmarking
    - Proto benchmarks vs. benchmarks
    - Types of benchmarks
  - Why?
  - Is CEM Ready as a Field?
- Conclusions

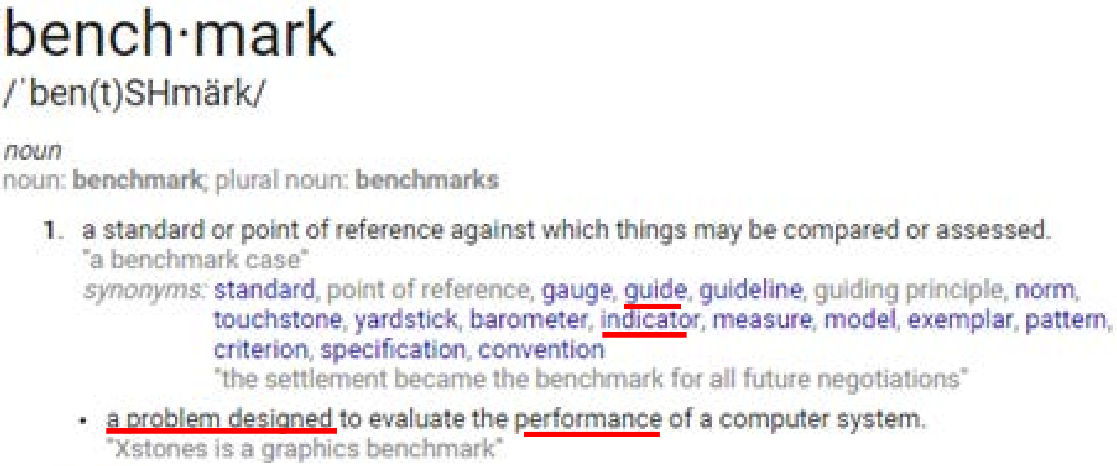
Outline

- Motivation & Observations
  - What is Benchmarking?
    - Performance
    - Theory of benchmarking
    - Proto benchmarks vs. benchmarks
    - Types of benchmarks
  - Why?
  - Is CEM Ready as a Field?
- Conclusions

What is Benchmarking?

A tentative definition…

Benchmarking: A (scientific) method to judge the "performance" of a (complex) system based on experiments & empirical evidence.



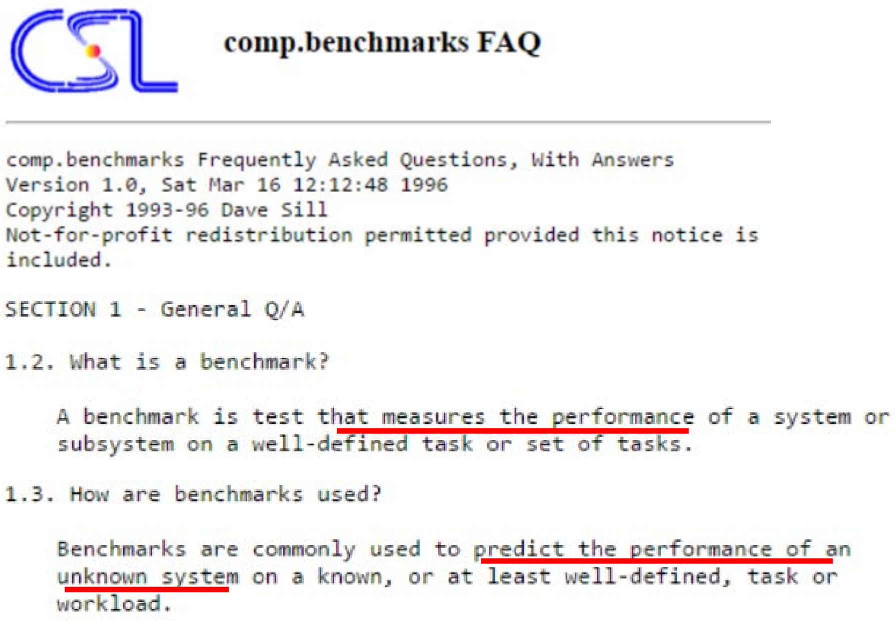
bench·mark

/ˈben(t)SHmärk/

noun

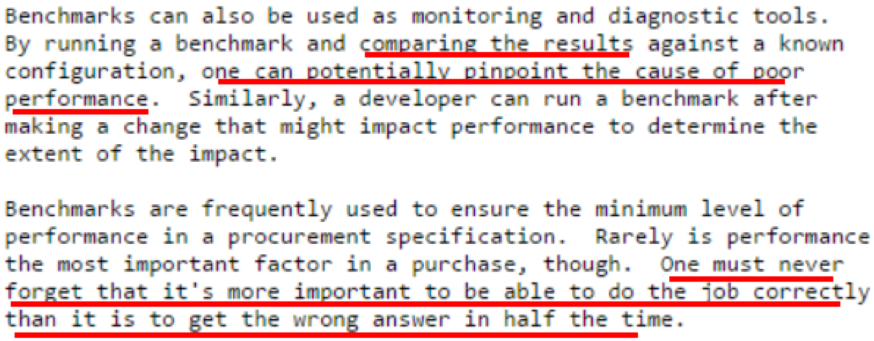
noun: **benchmark**; plural noun: **benchmarks**

1. a standard or point of reference against which things may be compared or assessed.
   "a benchmark case"
   synonyms: standard, point of reference, gauge, guide, guideline, guiding principle, norm, touchstone, yardstick, barometer, indicator, measure, model, exemplar, pattern, criterion, specification, convention
   "the settlement became the benchmark for all future negotiations"
   - a problem designed to evaluate the performance of a computer system.
     "Xstones is a graphics benchmark"



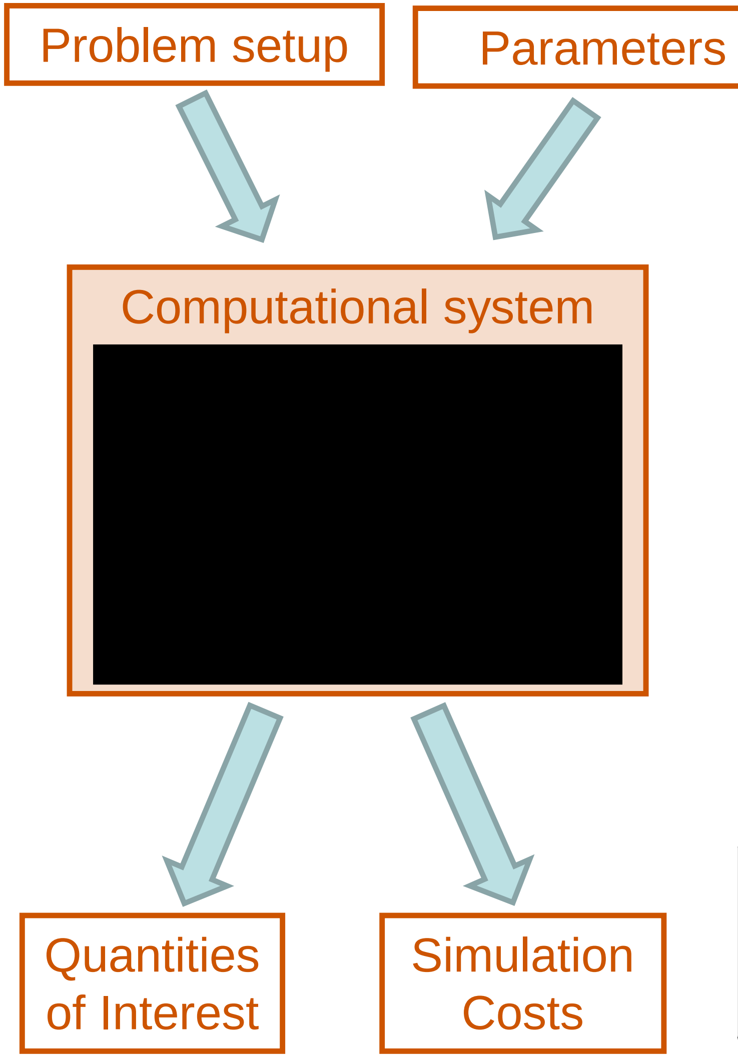
comp.benchmarks FAQ

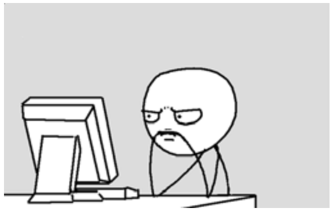
```
comp.benchmarks Frequently Asked Questions, With Answers
Version 1.0, Sat Mar 16 12:12:48 1996
Copyright 1993-96 Dave Sill
Not-for-profit redistribution permitted provided this notice is
included.

SECTION 1 - General Q/A

1.2. What is a benchmark?

    A benchmark is test that measures the performance of a system or
    subsystem on a well-defined task or set of tasks.

1.3. How are benchmarks used?

    Benchmarks are commonly used to predict the performance of an
    unknown system on a known, or at least well-defined, task or
    workload.
```



```
Benchmarks can also be used as monitoring and diagnostic tools.
By running a benchmark and comparing the results against a known
configuration, one can potentially pinpoint the cause of poor
performance.  Similarly, a developer can run a benchmark after
making a change that might impact performance to determine the
extent of the impact.

Benchmarks are frequently used to ensure the minimum level of
performance in a procurement specification.  Rarely is performance
the most important factor in a purchase, though.  One must never
forget that it's more important to be able to do the job correctly
than it is to get the wrong answer in half the time.
```

Empirical Approach



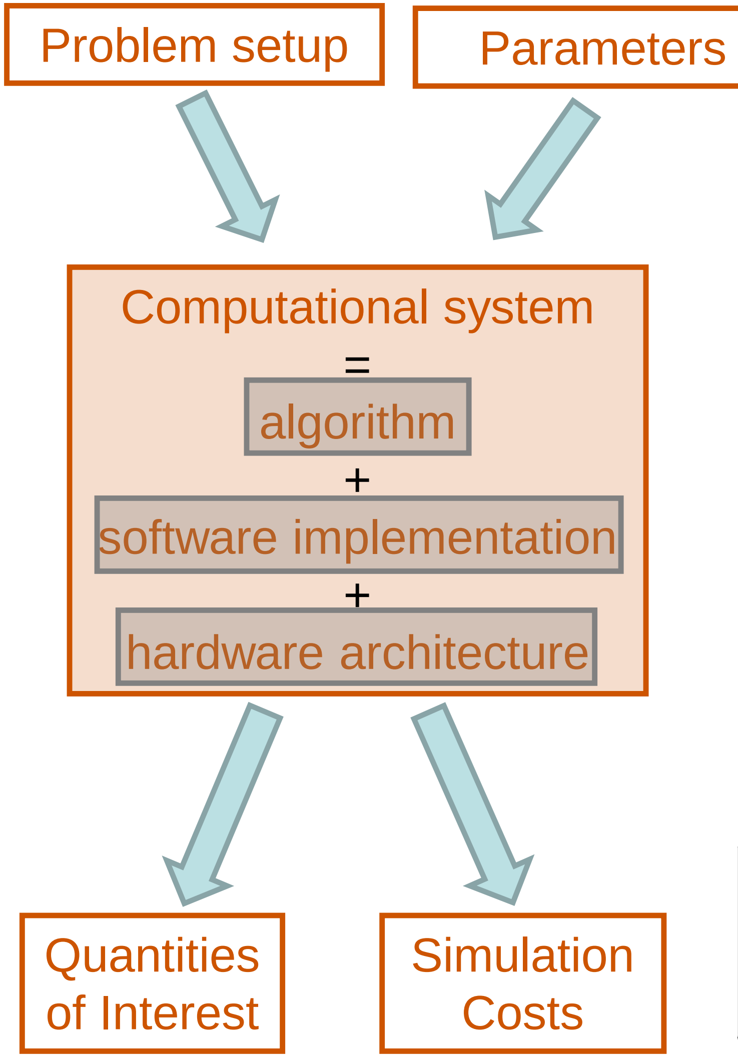


A tentative definition…

Benchmarking: A (scientific) method to judge the "performance" of a (complex) system based on experiments & empirical evidence.

Original images from:
https://www.pinterest.com/yafguimaraes/stickman/
knowyourmeme.com/photos/415209-computer-reaction-faces
http://www.richcontractor.com/my_weblog/2008/11/how-to-become-a-contractor-13-steps-to-becoming-your-own-boss.html

Empirical Approach



Problem setup

Parameters

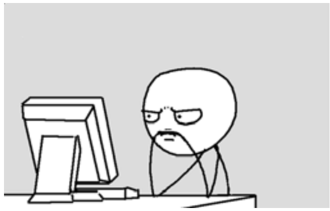
Computational system
=
algorithm
+
software implementation
+
hardware architecture

Quantities of Interest

Simulation Costs

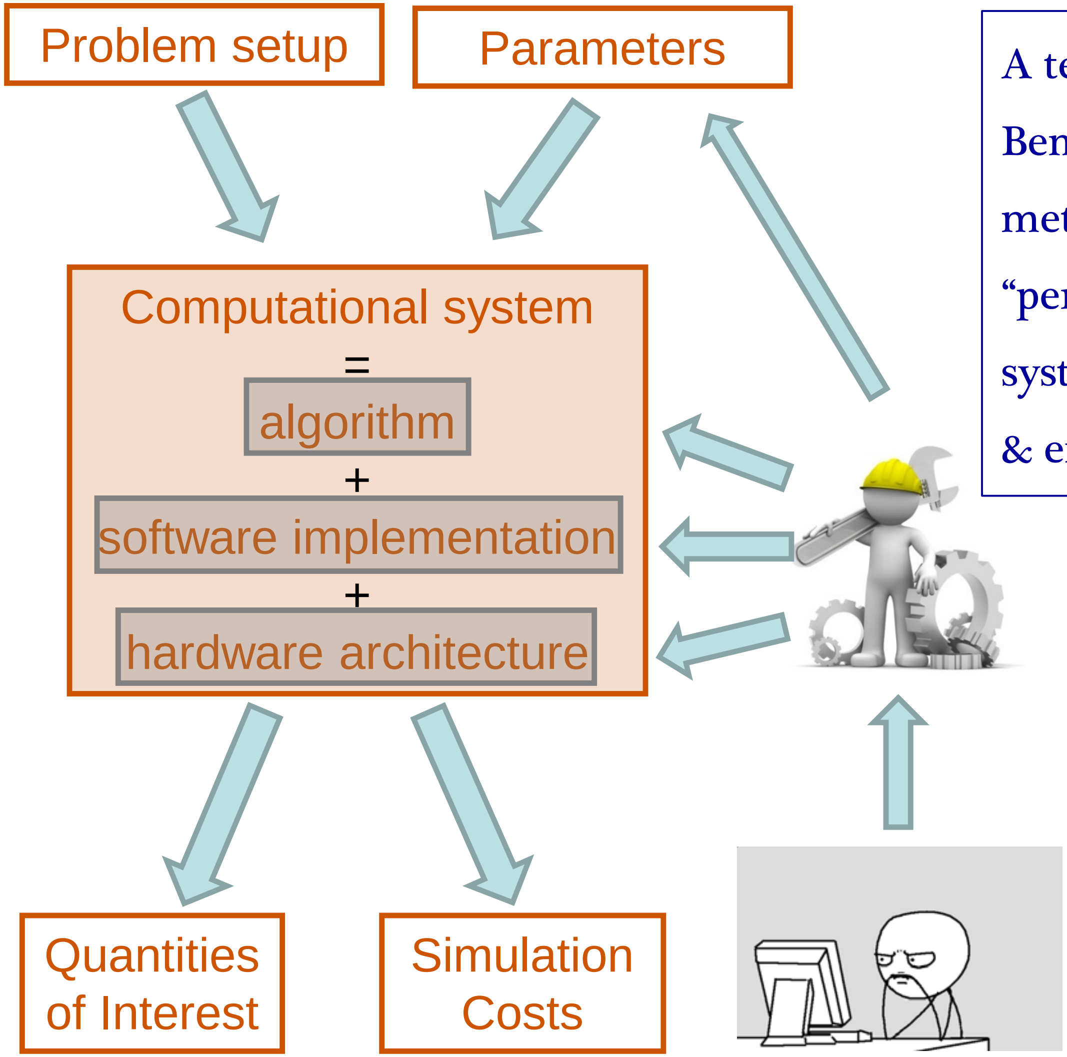
A tentative definition…

Benchmarking: A (scientific) method to judge the “performance” of a (complex) system based on experiments & empirical evidence.



Original images from:
https://www.pinterest.com/yafguimaraes/stickman/
knowyourmeme.com/photos/415209-computer-reaction-faces
http://www.richcontractor.com/my_weblog/2008/11/how-to-become-a-contractor-13-steps-to-becoming-your-own-boss.html

Empirical Approach



A tentative definition…

Benchmarking: A (scientific) method to judge the "performance" of a (complex) system based on experiments & empirical evidence.

Original images from:
https://www.pinterest.com/yafguimaraes/stickman/
knowyourmeme.com/photos/415209-computer-reaction-faces
http://www.richcontractor.com/my_weblog/2008/11/how-to-become-a-contractor-13-steps-to-becoming-your-own-boss.html

A Theory of (Community) Benchmarking

IEEE
Software
Developing Scientific Software
JULY | AUGUST 2008
8 | What Semantic Wikis Offer
89 | The Way We Program
92 | Developers' Motivation: A Study
IEEE
IEEE computer society
www.computer.org/software

"Our theory is concerned primarily with benchmarks that are created and used by a technical research community…The community of interest may include participants from academia, industry, and government, but they are all primarily interested in scientific research..."

We define a benchmark as a test or set of tests used to compare the performance of alternative tools or techniques."

S. E. Sim, S. Easterbrook, R. C. Holt, "Using benchmarking to advance research: A challenge to software engineering," *Proc. Int. Conf. Software Eng.,* May 2003.

A Theory of (Community) Benchmarking

IEEE
Software
Developing Scientific Software
JULY/AUGUST 2008
8 What Semantic Wikis Offer
89 The Way We Program
92 Developers' Motivation: A Study
IEEE
IEEE computer society
www.computer.org/software

“A benchmark has three components:

Motivating comparison…The purpose of a benchmark is to compare, so the comparison that is at the heart of a benchmark must be clearly defined. The motivation aspect refers to the need for the research area, and in turn the benchmark itself and the work on it.

Task sample…tests…should be representative sample of the tasks that the tool or technique is expected to solve in actual practice…a selection of tasks acts as surrogates.

Performance measures…measurements can be made by a computer or by a human, and can be quantitative or qualitative. Performance is not an innate characteristic of the technology, but is the relationship between the technology and how it is used. As such, performance is a measure of fitness for purpose.”

S. E. Sim, S. Easterbrook, R. C. Holt, “Using benchmarking to advance research: A challenge to software engineering,” *Proc. Int. Conf. Software Eng.,* May 2003.

A Theory of (Community) Benchmarking

IEEE
Software
Developing Scientific Software
JULY / AUGUST 2008
8 | What Semantic Wikis Offer
89 | The Way We Program
92 | Developers' Motivation: A Study
IEEE
IEEE computer society
www.computer.org/software

“A benchmark has three components:

Motivating comparison…

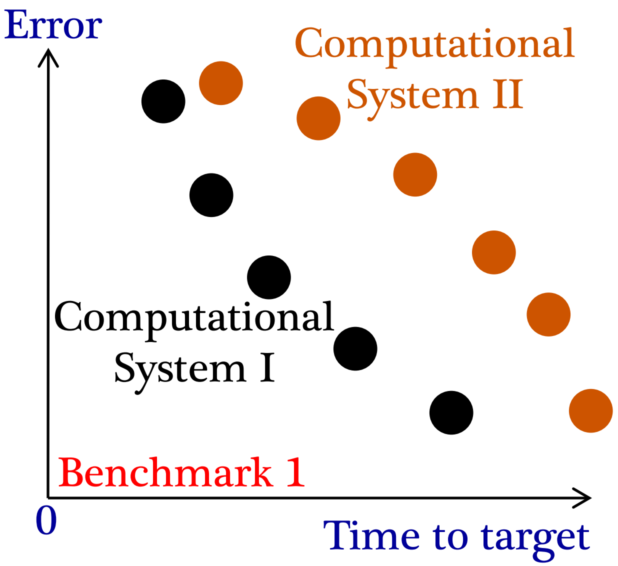
Task sample…

Performance measures… performance is a measure of fitness for purpose.

A proto-benchmark is a set of tests that is missing one of these components. The most common proto-benchmarks lack a performance measure and are sometimes called case studies or examplars. These are typically used to demonstrate the features and capabilities of a new tool or technique, and occasionally used to compare different technologies in an exploratory manner.”

S. E. Sim, S. Easterbrook, R. C. Holt, “Using benchmarking to advance research: A challenge to software engineering,” *Proc. Int. Conf. Software Eng.,* May 2003.

Performance Definition for Advancing CEM R&D



Computational system

=

algorithm

+

software implementation

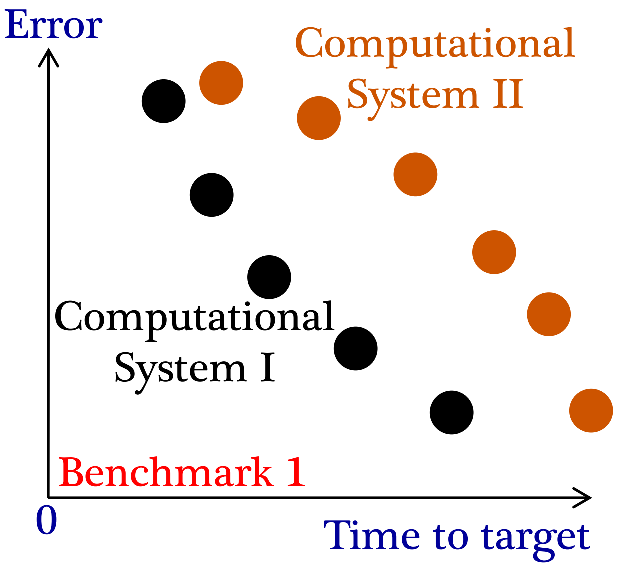
+

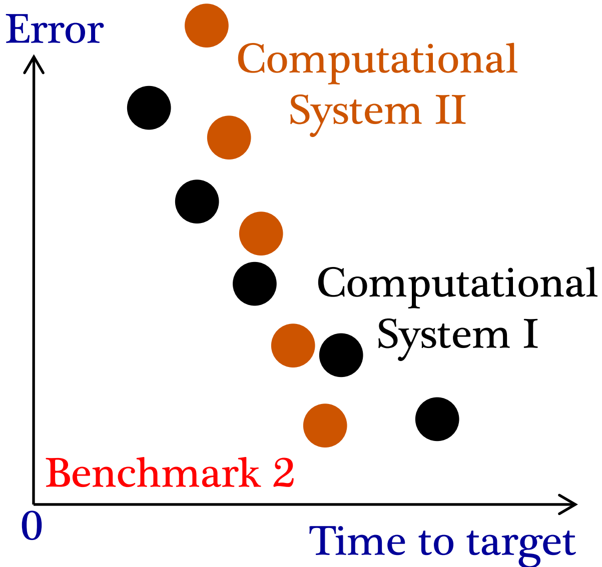
hardware architecture

Benchmarking: A (scientific) method to judge the “performance” of a (complex) system based on experiments & empirical evidence.

Performance definition should include error, cost, and trade-off between error and cost.

Performance Definition for Advancing CEM R&D





Computational system
=
algorithm
+
software implementation
+
hardware architecture

Benchmarking: A (scientific) method to judge the "performance" of a (complex) system based on experiments & empirical evidence.

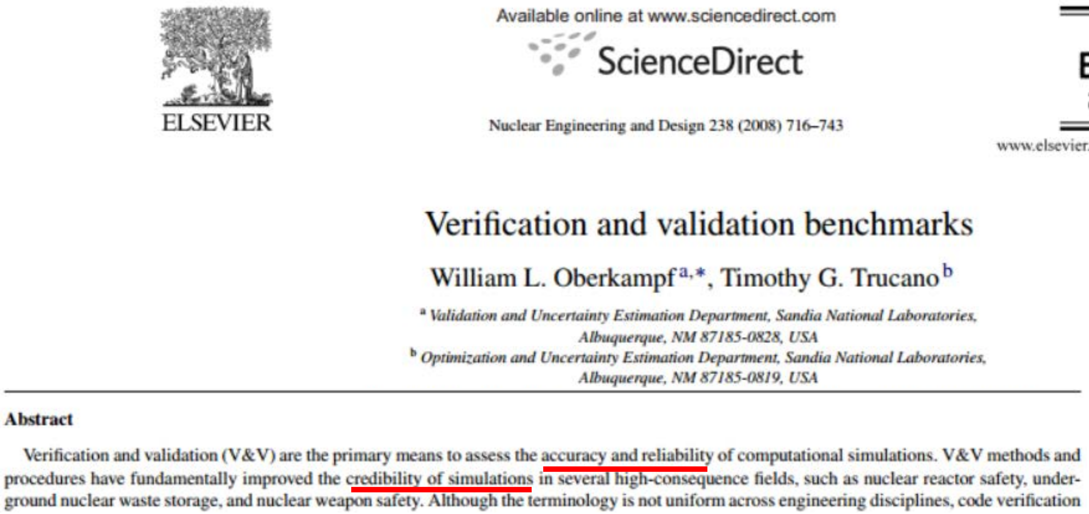
Performance definition should include error, cost, and trade-off between error and cost.

No universal best system. Corollaries:

Different computational systems ⇔ different trade-offs between error and cost.

Relative performance of systems will change from benchmark to benchmark.

Proto-Benchmarks in Computational Engineering & Science



ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Nuclear Engineering and Design 238 (2008) 716–743

Nuclear Engineering and Design
www.elsevier.com/locate/nucengdes

Verification and validation benchmarks

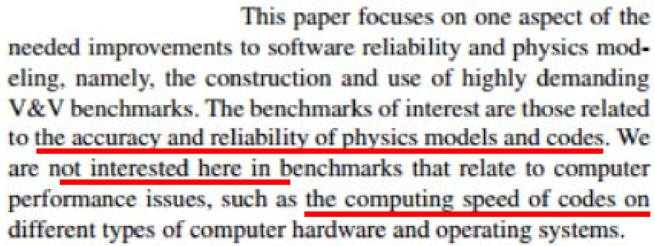
William L. Oberkampf[a,*], Timothy G. Trucano[b]

[a] Validation and Uncertainty Estimation Department, Sandia National Laboratories,
Albuquerque, NM 87185-0828, USA
[b] Optimization and Uncertainty Estimation Department, Sandia National Laboratories,
Albuquerque, NM 87185-0819, USA

Abstract

Verification and validation (V&V) are the primary means to assess the accuracy and reliability of computational simulations. V&V methods and procedures have fundamentally improved the credibility of simulations in several high-consequence fields, such as nuclear reactor safety, underground nuclear waste storage, and nuclear weapon safety. Although the terminology is not uniform across engineering disciplines, code verification

Benchmarking: A (scientific) method to judge the "performance" of a (complex) system based on experiments & empirical evidence.

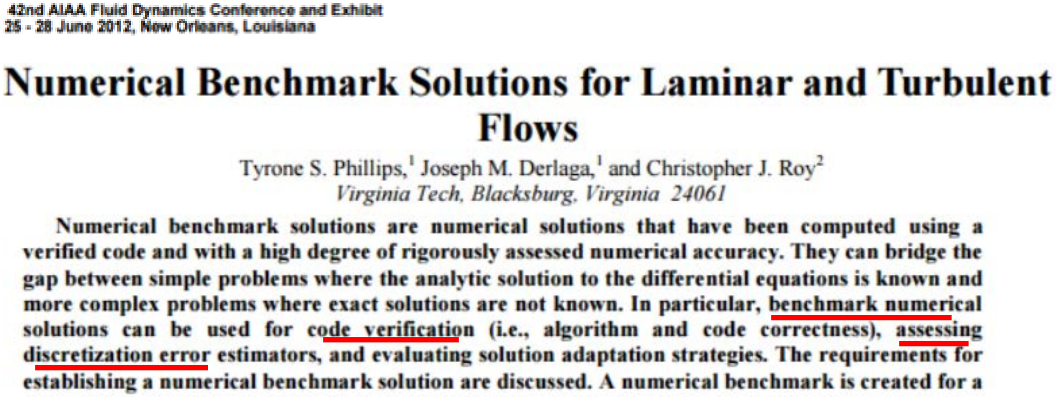


This paper focuses on one aspect of the needed improvements to software reliability and physics modeling, namely, the construction and use of highly demanding V&V benchmarks. The benchmarks of interest are those related to the accuracy and reliability of physics models and codes. We are not interested here in benchmarks that relate to computer performance issues, such as the computing speed of codes on different types of computer hardware and operating systems.

In CES:

"Poor performance" often means "large error"

Occasionally, the concept of "speed" appears



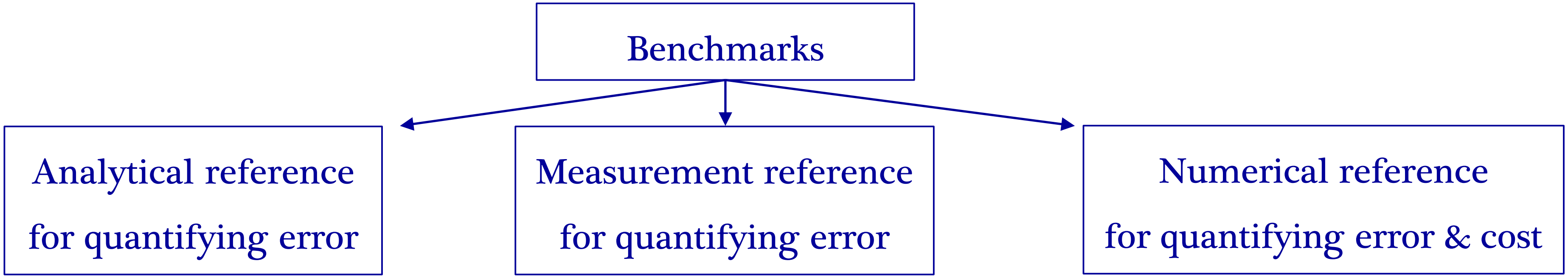
42nd AIAA Fluid Dynamics Conference and Exhibit
25 - 28 June 2012, New Orleans, Louisiana

Numerical Benchmark Solutions for Laminar and Turbulent Flows

Tyrone S. Phillips,[1] Joseph M. Derlaga,[1] and Christopher J. Roy[2]
Virginia Tech, Blacksburg, Virginia 24061

Numerical benchmark solutions are numerical solutions that have been computed using a verified code and with a high degree of rigorously assessed numerical accuracy. They can bridge the gap between simple problems where the analytic solution to the differential equations is known and more complex problems where exact solutions are not known. In particular, benchmark numerical solutions can be used for code verification (i.e., algorithm and code correctness), assessing discretization error estimators, and evaluating solution adaptation strategies. The requirements for establishing a numerical benchmark solution are discussed. A numerical benchmark is created for a

Types of Proto-Benchmarks in CEM



Benchmarks

Analytical reference for quantifying error

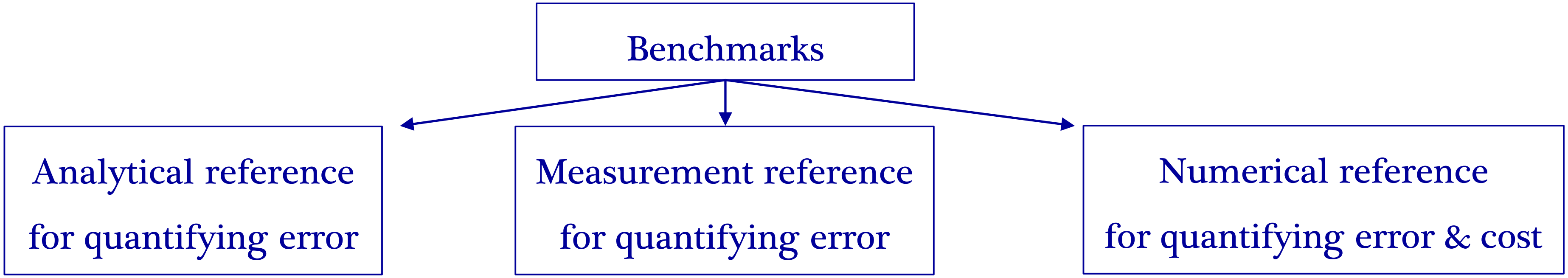
Measurement reference for quantifying error

Numerical reference for quantifying error & cost

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

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

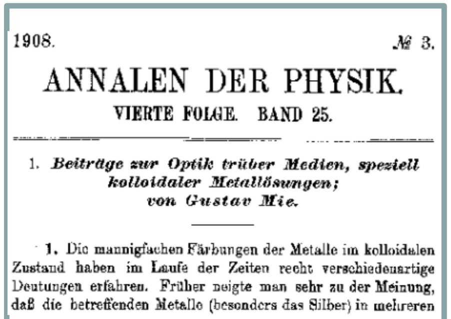
Types of Proto-Benchmarks in CEM


Benchmarks

Analytical reference for quantifying error

Measurement reference for quantifying error

Numerical reference for quantifying error & cost

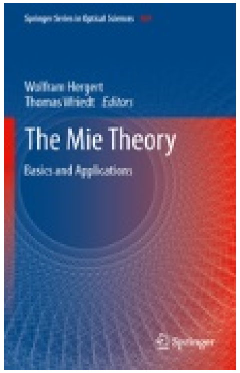

1908. № 3.

ANNALEN DER PHYSIK.

VIERTE FOLGE. BAND 25.

1. *Beiträge zur Optik trüber Medien, ſpeziell kolloidaler Metallöſungen; von Guſtav Mie.*

1. Die mannigfachen Färbungen der Metalle im kolloidalen Zuſtand haben im Laufe der Zeiten recht verſchiedenartige Deutungen erfahren. Früher neigte man ſehr zu der Meinung, daß die betreffenden Metalle (beſonders das Silber) in mehreren

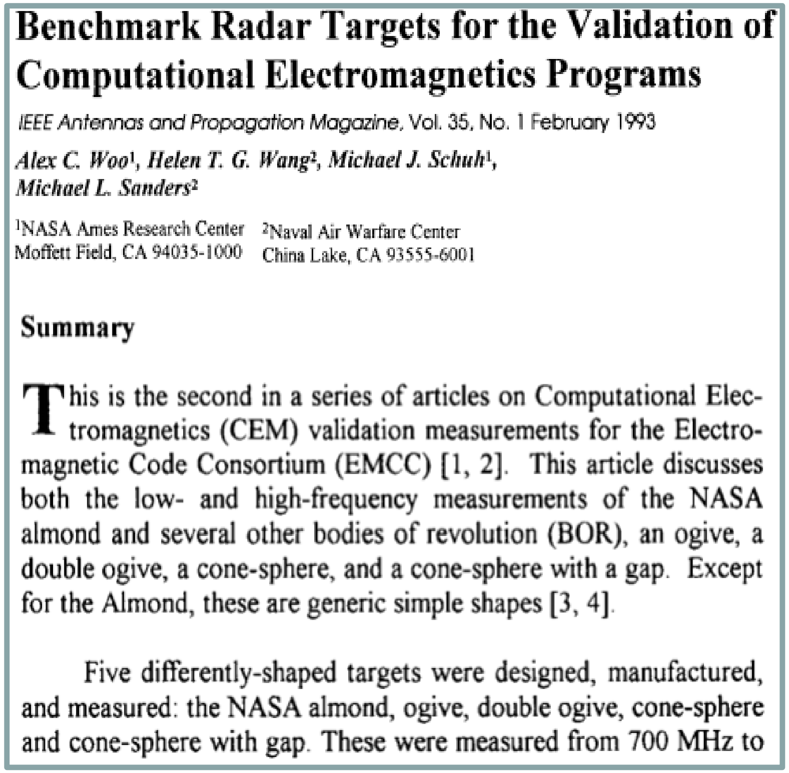

Springer Series in Optical Sciences

Wolfram Hergert
Thomas Wriedt Editors

The Mie Theory

Basics and Applications

Springer


Benchmark Radar Targets for the Validation of Computational Electromagnetics Programs

IEEE Antennas and Propagation Magazine, Vol. 35, No. 1 February 1993

Alex C. Woo[1], *Helen T. G. Wang*[2], *Michael J. Schuh*[1], *Michael L. Sanders*[2]

[1]NASA Ames Research Center
Moffett Field, CA 94035-1000

[2]Naval Air Warfare Center
China Lake, CA 93555-6001

Summary

This is the second in a series of articles on Computational Electromagnetics (CEM) validation measurements for the Electromagnetic Code Consortium (EMCC) [1, 2]. This article discusses both the low- and high-frequency measurements of the NASA almond and several other bodies of revolution (BOR), an ogive, a double ogive, a cone-sphere, and a cone-sphere with a gap. Except for the Almond, these are generic simple shapes [3, 4].

Five differently-shaped targets were designed, manufactured, and measured: the NASA almond, ogive, double ogive, cone-sphere and cone-sphere with gap. These were measured from 700 MHz to

?

Tiers of Benchmarks: Backyard/Party to Olympic

Original images from:
http://kidsactivitiesblog.com/9055/target-practice-game
http://sometimes-homemade.com/ultimate-nerf-battle-birthday-party-ideas/
http://supportforstudents.msu.edu/articles/2015-olympic-sports-feature-archery-shooting
https://worldarchery.org/news/143721/top-10-pictures-2016-olympics

Outline

- Motivation & Observations
  - What is Benchmarking?
    - Performance
    - Theory of benchmarking
    - Proto benchmarks vs. benchmarks
    - Types of benchmarks
  - Why?
  - Is CEM Ready as a Field?
- Conclusions

Why Benchmark?

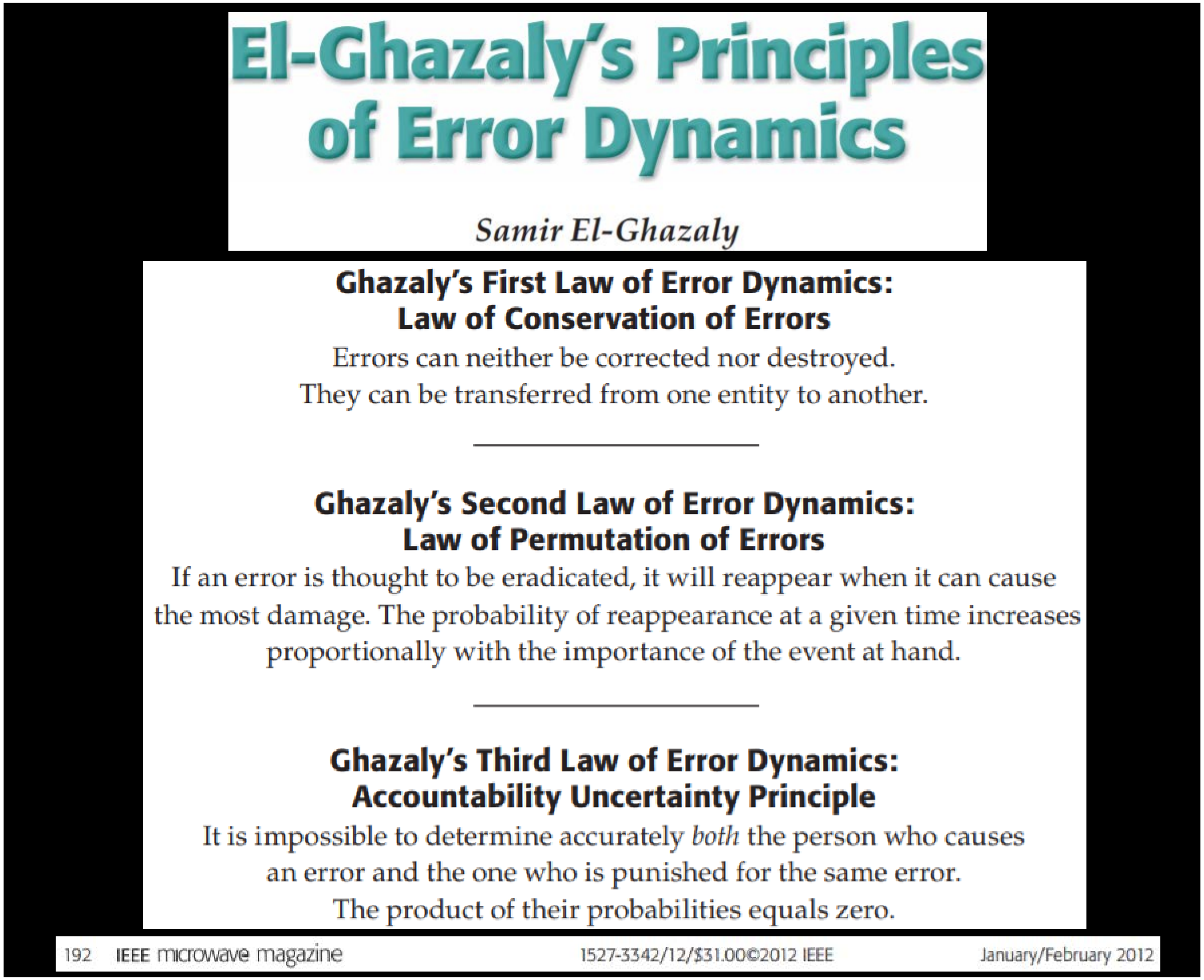
IEEE
Software
Developing Scientific Software
JULY / AUGUST 2008
8 | What Semantic Wikis Offer
89 | The Way We Program
92 | Developers' Motivation: A Study
IEEE
IEEE computer society
www.computer.org/software

"We have shown that benchmarking can have a strong positive effect on the scientific maturity of a research community. The benefits of benchmarking include a stronger consensus on the community's research goals, greater collaboration between laboratories, more rigorous examination of research results, and faster technical progress."

S. E. Sim, S. Easterbrook, R. C. Holt, "Using benchmarking to advance research: A challenge to software engineering," *Proc. Int. Conf. Software Eng.,* May 2003.

Why Benchmark?

- Ubiquity of (human) error



El-Ghazaly's Principles of Error Dynamics

Samir El-Ghazaly

Ghazaly's First Law of Error Dynamics:
Law of Conservation of Errors

Errors can neither be corrected nor destroyed.
They can be transferred from one entity to another.

Ghazaly's Second Law of Error Dynamics:
Law of Permutation of Errors

If an error is thought to be eradicated, it will reappear when it can cause the most damage. The probability of reappearance at a given time increases proportionally with the importance of the event at hand.

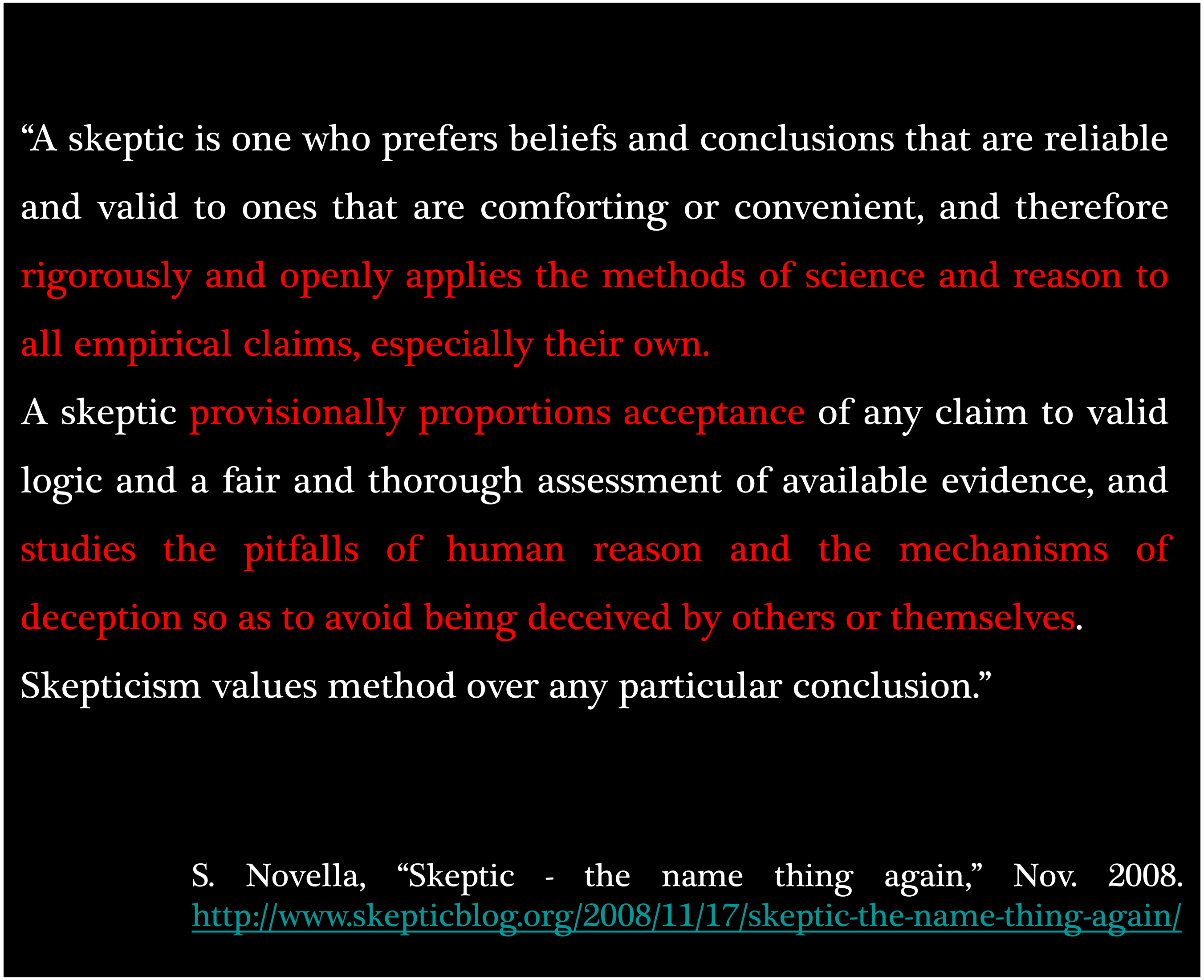
Ghazaly's Third Law of Error Dynamics:
Accountability Uncertainty Principle

It is impossible to determine accurately *both* the person who causes an error and the one who is punished for the same error.
The product of their probabilities equals zero.

192 IEEE microwave magazine 1527-3342/12/$31.00©2012 IEEE January/February 2012

Why Benchmark?

- Ubiquity of (human) error



“A skeptic is one who prefers beliefs and conclusions that are reliable and valid to ones that are comforting or convenient, and therefore rigorously and openly applies the methods of science and reason to all empirical claims, especially their own.

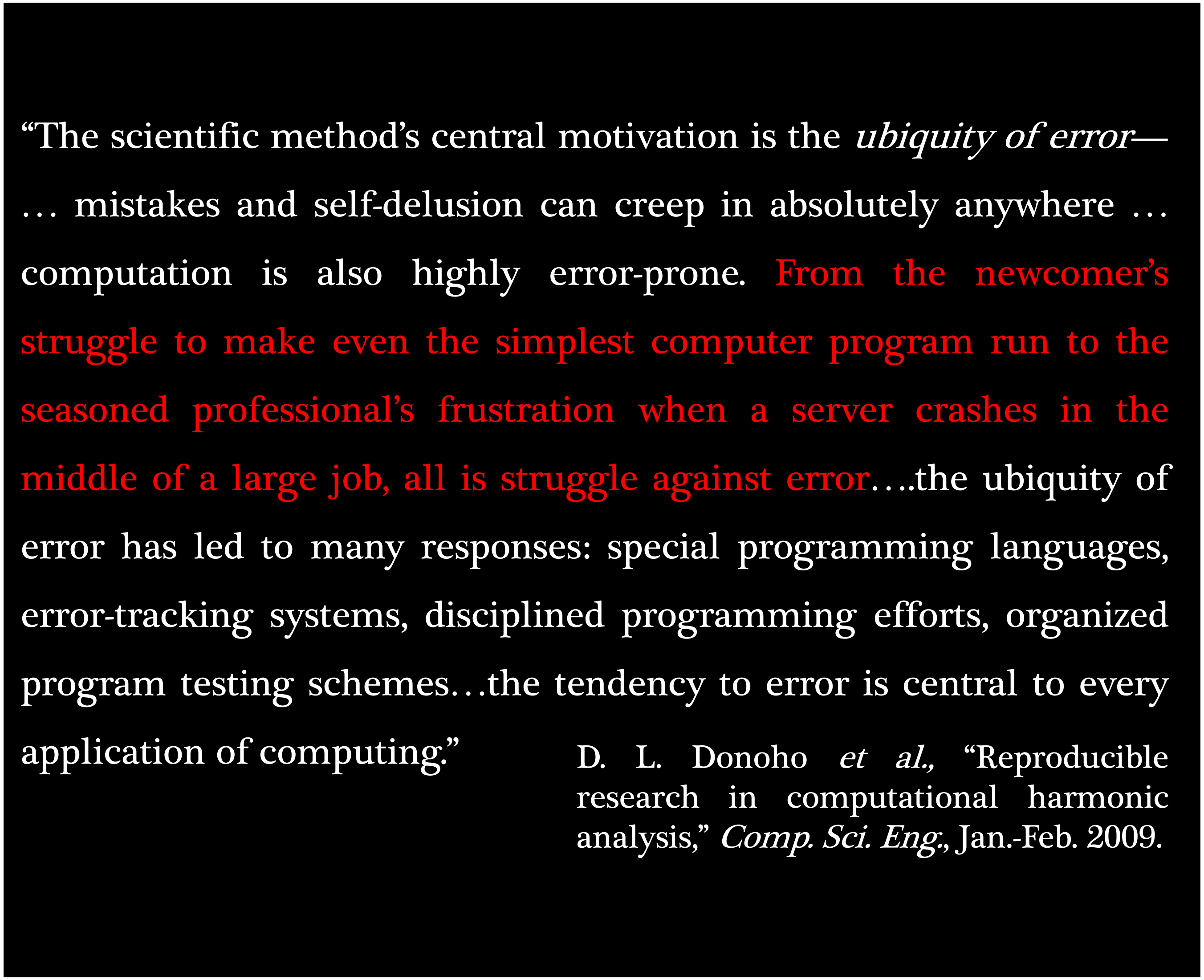
A skeptic provisionally proportions acceptance of any claim to valid logic and a fair and thorough assessment of available evidence, and studies the pitfalls of human reason and the mechanisms of deception so as to avoid being deceived by others or themselves.

Skepticism values method over any particular conclusion.”

S. Novella, “Skeptic - the name thing again,” Nov. 2008. http://www.skepticblog.org/2008/11/17/skeptic-the-name-thing-again/

Why Benchmark?

- Ubiquity of (human) error



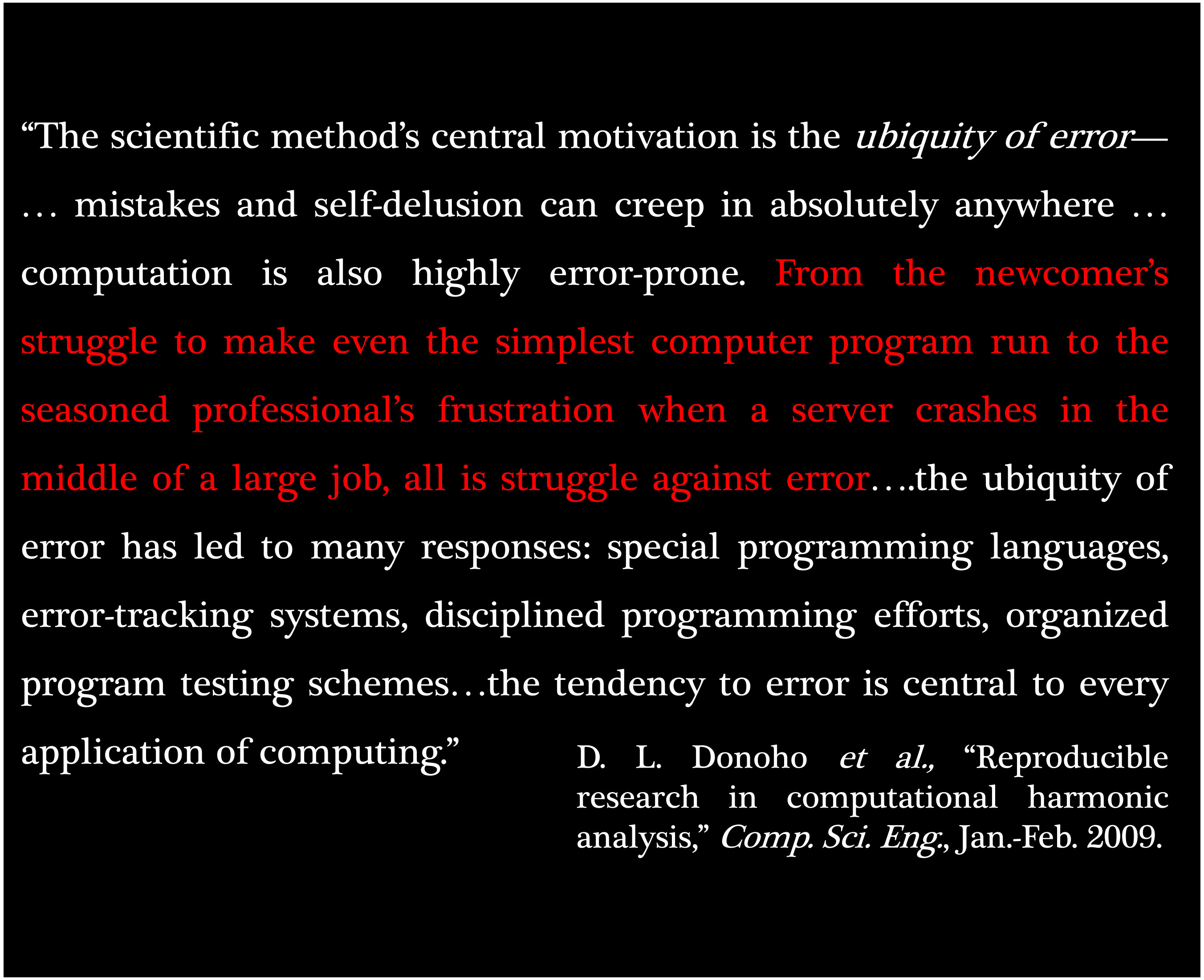
“The scientific method’s central motivation is the *ubiquity of error*—… mistakes and self-delusion can creep in absolutely anywhere … computation is also highly error-prone. From the newcomer’s struggle to make even the simplest computer program run to the seasoned professional’s frustration when a server crashes in the middle of a large job, all is struggle against error.…the ubiquity of error has led to many responses: special programming languages, error-tracking systems, disciplined programming efforts, organized program testing schemes…the tendency to error is central to every application of computing.”

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

Why Benchmark?

- Ubiquity of (human) error

Benchmarking …
+ a systematic method to combat error
+ does not place undue burdens of (perfect) replication

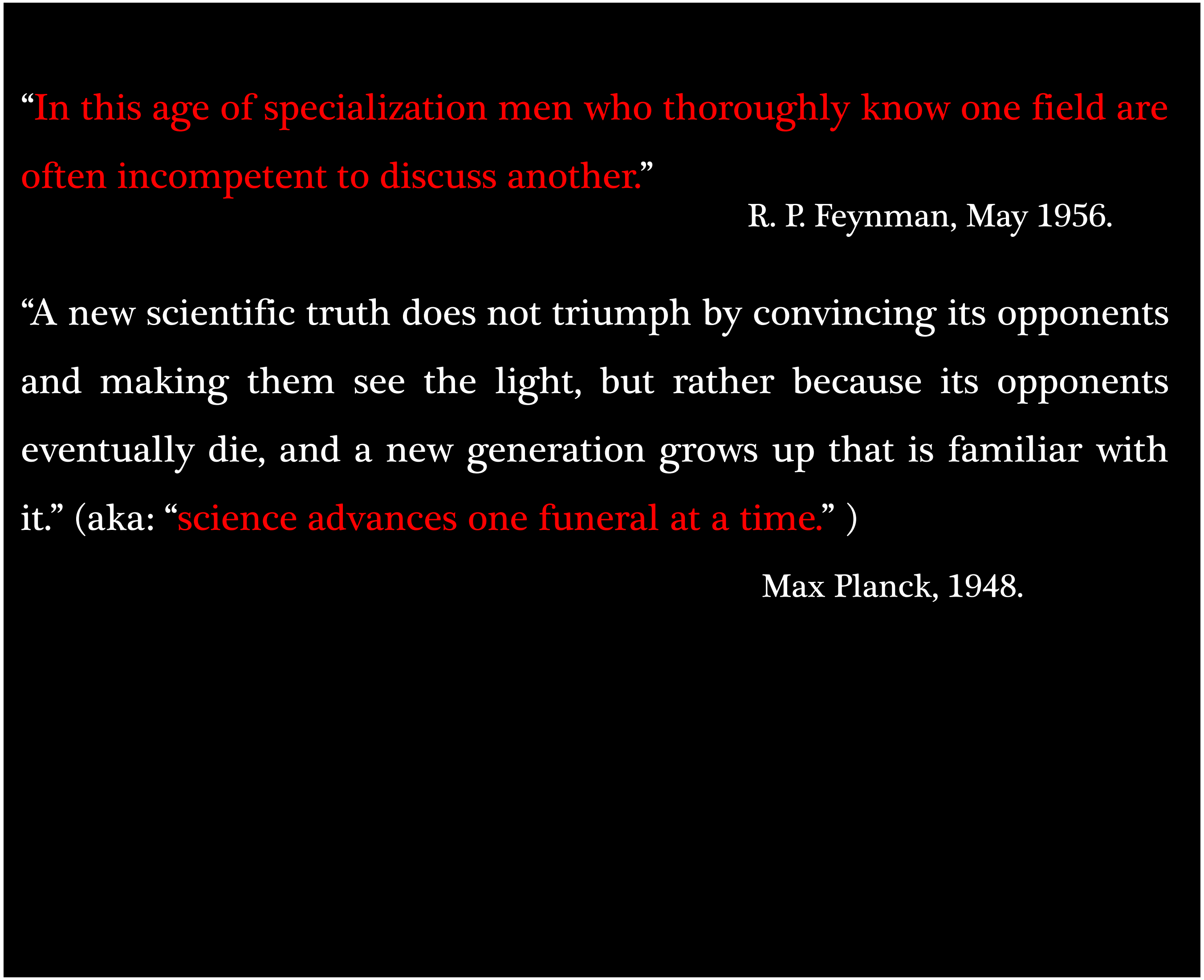


> "The scientific method's central motivation is the *ubiquity of error*— … mistakes and self-delusion can creep in absolutely anywhere … computation is also highly error-prone. From the newcomer's struggle to make even the simplest computer program run to the seasoned professional's frustration when a server crashes in the middle of a large job, all is struggle against error….the ubiquity of error has led to many responses: special programming languages, error-tracking systems, disciplined programming efforts, organized program testing schemes…the tendency to error is central to every application of computing."

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

Why Benchmark?

- Ubiquity of (human) error
- Specialization



"In this age of specialization men who thoroughly know one field are often incompetent to discuss another."

R. P. Feynman, May 1956.

"A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." (aka: "science advances one funeral at a time.")

Max Planck, 1948.

Why Benchmark?

- Ubiquity of (human) error

- Specialization

Benchmarking can…
+ inform others about important problems
+ inform others about the current state of computational systems for solving these problems
+ help us keep up with advances
+help us keep an open mind
+ lower barriers to entry of new researchers/ideas/ systems

“In this age of specialization men who thoroughly know one field are often incompetent to discuss another.”

R. P. Feynman, May 1956.

“A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it.” (aka: “science advances one funeral at a time.”)

Max Planck, 1948.

Why Benchmark?

- Ubiquity of (human) error
- Specialization
- Scientific integrity

"The idea is to try to give *all* of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction …learning how to not fool ourselves—of having utter scientific integrity—is, I'm sorry to say, something that we…just hope you've caught on by osmosis. The first principle is that you must not fool yourself—and you are the easiest person to fool…After… it's easy not to fool other scientists. You just have to be honest in a conventional way after that."

R. P. Feynman, 1974.

"I mean by intellectual integrity the habit of deciding vexed questions in accordance with the evidence or of leaving them undecided where the evidence is inconclusive."

Bertrand Russell, 1954.

Why Benchmark?

- Ubiquity of (human) error

- Specialization

- Scientific integrity

Benchmarking can…
+ reduce importance of subjective factors when judging simulation tools
+ increase credibility of claims made by computational scientists and engineers
+ fortify intellectual/ scientific integrity

"The idea is to try to give *all* of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction …learning how to not fool ourselves—of having utter scientific integrity—is, I'm sorry to say, something that we…just hope you've caught on by osmosis. The first principle is that you must not fool yourself—and you are the easiest person to fool…After… it's easy not to fool other scientists. You just have to be honest in a conventional way after that."

R. P. Feynman, 1974.

"I mean by intellectual integrity the habit of deciding vexed questions in accordance with the evidence or of leaving them undecided where the evidence is inconclusive."

Bertrand Russell, 1954.

Why Benchmark?

- Ubiquity of (human) error
- Specialization
- Scientific integrity
- Incentivize research advances

Benchmarking can…
+highlight open problems
+ identify weaknesses in existing computational systems
+ inspire R&D to address these

“The idea is to try to give *all* of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction …learning how to not fool ourselves—of having utter scientific integrity—is, I’m sorry to say, something that we…just hope you’ve caught on by osmosis. The first principle is that you must not fool yourself—and you are the easiest person to fool…After… it’s easy not to fool other scientists. You just have to be honest in a conventional way after that.”

R. P. Feynman, 1974.

“I mean by intellectual integrity the habit of deciding vexed questions in accordance with the evidence or of leaving them undecided where the evidence is inconclusive.”

Bertrand Russell, 1954.

Outline

- Motivation & Observations
  - What is Benchmarking?
    - Performance
    - Theory of benchmarking
    - Proto benchmarks vs. benchmarks
    - Types of benchmarks
  - Why?
  - Is CEM Ready as a Field?
- Conclusions

A Theory of (Community) Benchmarking

IEEE
Software
Developing Scientific Software
JULY / AUGUST 2008
8 | What Semantic Wikis Offer
89 | The Way We Program
92 | Developers' Motivation: A Study
IEEE
IEEE computer society
www.computer.org/software

"…theory suggests…conditions…must…exist within a discipline before construction of a benchmark can be fruitfully attempted… …a minimum level of maturity in the discipline. During the early days, when a research area is becoming established, it is necessary and appropriate to go through a stage where diverse approaches and solutions proliferate… Evidence…community…reached…required level of maturity and is ready to move to a more rigorous scientific basis comes in many forms. Typical symptoms include an increasing concern with validation of research results and with comparison between solutions developed at different labs; attempted replication of results; use of proto-benchmarks; … an increasing resistance to accept speculative papers for publication."

S. E. Sim, S. Easterbrook, R. C. Holt, "Using benchmarking to advance research: A challenge to software engineering," *Proc. Int. Conf. Software Eng.,* May 2003.

A Theory of (Community) Benchmarking

IEEE
Software
Developing Scientific Software
JULY/AUGUST 2008
8 | What Semantic Wikis Offer
89 | The Way We Program
92 | Developers' Motivation: A Study
IEEE
IEEE computer society
www.computer.org/software

"…theory suggests…conditions…must…exist within a discipline before construction of a benchmark can be fruitfully attempted…

…an ethos of collaboration within the community."

S. E. Sim, S. Easterbrook, R. C. Holt, "Using benchmarking to advance research: A challenge to software engineering," *Proc. Int. Conf. Software Eng.,* May 2003.

" Evidence of this ethos can be found in:

multi-site collaborative projects

papers with authors from disparate geographic locations and sectors of the economy

exchange visits between laboratories…

standards for terminology and publication."

S. E. Sim, "A theory of benchmarking with applications to software reverse engineering," PhD Thesis, University of Toronto, 2003.

Conclusions

- Current state of benchmarking in CEM

  + verification & validation (proto-)benchmarks exist/common in CEM

  + numerical benchmarks (with error vs. cost trade-off) underutilized

  + papers full of unreproducible numerical results

- Next-generation benchmarks can

  + become important tools for advancing CEM

  + increase credibility of computational scientists & engineers without placing undue burdens of (perfect) replication (unlike ‘really reproducible research’)

  + reduce importance of subjective factors when judging computational systems

- Meaningful benchmarking of computational systems non-trivial

  + error measures, cost metrics must be carefully chosen to reward/incentivize advances

  + even extremely different systems can be compared with precise measurement/ normalization