Universidad Técnica Federico Santa María

Valparaíso, 3 January 2017

Introduction to Computational Reproducibility (and why we care)

Prof. Lorena A. Barba

Mechanical and Aerospace Engineering Department
The George Washington University

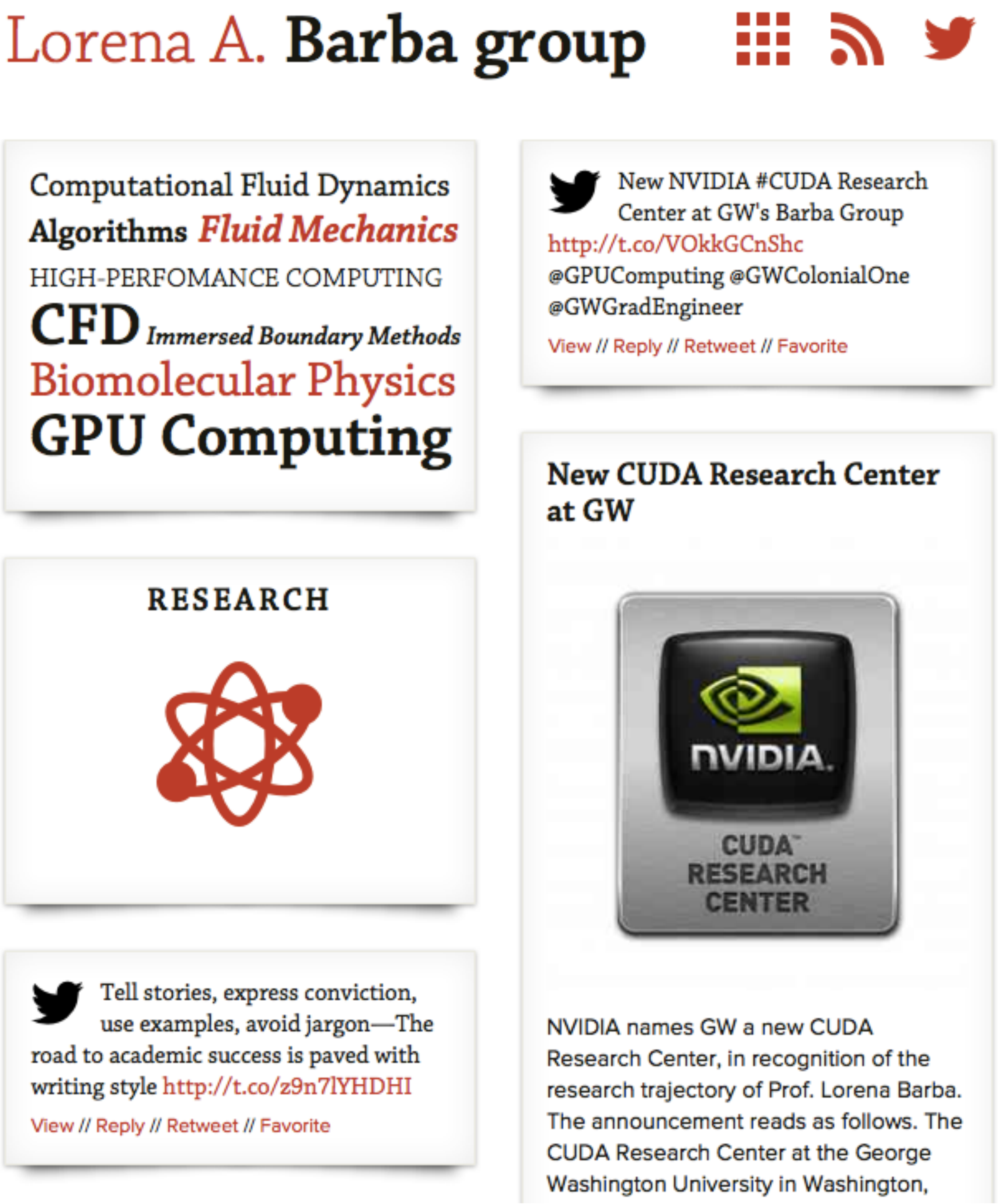
@LorenaABarba

Acknowledgements

NSF CAREER award

NVIDIA CUDA Fellows Program

About us



Lorena A. **Barba group**

Computational Fluid Dynamics
Algorithms ***Fluid Mechanics***
HIGH-PERFOMANCE COMPUTING
CFD ***Immersed Boundary Methods***
Biomolecular Physics
GPU Computing

RESEARCH

Tell stories, express conviction, use examples, avoid jargon—The road to academic success is paved with writing style http://t.co/z9n7lYHDHI

View // Reply // Retweet // Favorite

New NVIDIA #CUDA Research Center at GW's Barba Group http://t.co/VOkkGCnShc @GPUComputing @GWColonialOne @GWGradEngineer

View // Reply // Retweet // Favorite

New CUDA Research Center at GW

NVIDIA names GW a new CUDA Research Center, in recognition of the research trajectory of Prof. Lorena Barba. The announcement reads as follows. The CUDA Research Center at the George Washington University in Washington,

http://lorenabarba.com

“Essential skills for reproducible research computing”

Universidad Técnica Federico Santa María

First week of January 2017

A Barba-group workshop for graduate students

`https://barbagroup.github.io/essential_skills_RRC/`

with Barba-group members:

Gilbert Forsyth		Natalia Clementi
@gforsyth	GitHub	@ncclementi
@gilforsyth	Twitter	@ncclementi

What is Science?

- American Physical Society:
  - Ethics and Values, 1999

    "The success and credibility of science are anchored in the willingness of scientists to […] Expose their ideas and results to independent testing and replication by others. This requires the open exchange of data, procedures and materials."

`https://www.aps.org/policy/statements/99_6.cfm`

Published online 13 October 2010 | *Nature* **467**, 775-777 (2010) | doi:10.1038/467775a

News Feature

Computational science: ...Error

...why scientific programming does not compute.

Zeeya Merali

```
C:\lab>
f77 -o
data.exe
>
>
...ERROR


...why scientific programming does not
compute

>
```

A dolphin's demise

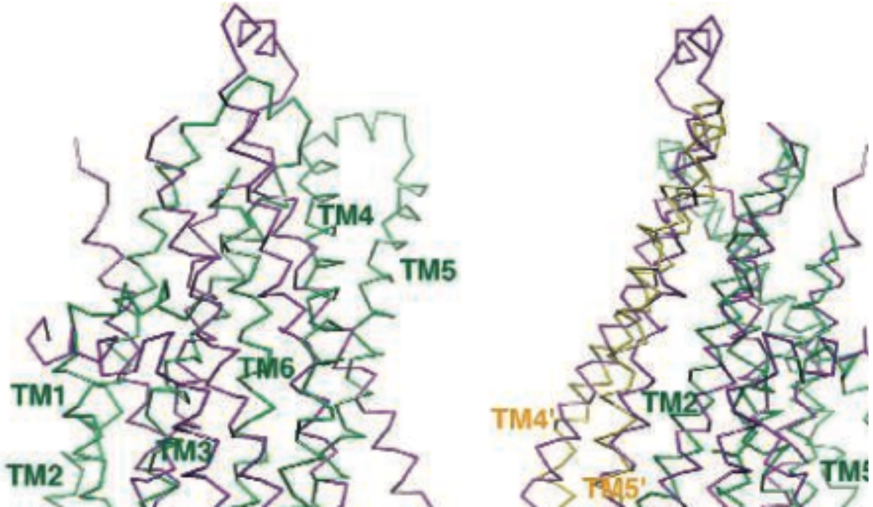
1860

SCIENTIFIC PUBLISHING

A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a Presidential Early Career Award for Scientists and Engineers, the country's highest honor for young researchers. His lab generated a stream of high-profile papers detailing the molecular structures of important proteins embedded in cell membranes.

Then the dream turned into a nightmare. In September, Swiss researchers published a paper in *Nature* that cast serious doubt on a 2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many



14 OCTOBER 2010 | VOL 467 | NATURE | 775

> As a general rule, researchers do not test or document their programs rigorously, and they rarely release their codes, making it almost impossible to reproduce and verify published results generated by scientific software …

14 OCTOBER 2010 | VOL 467 | NATURE | 775

QUOTE:
"There are terrifying statistics showing that almost all of what scientists know about coding is self-taught," says Wilson. "They just don't know how bad they are."

Retraction Watch

Error in one line of code sinks cancer study

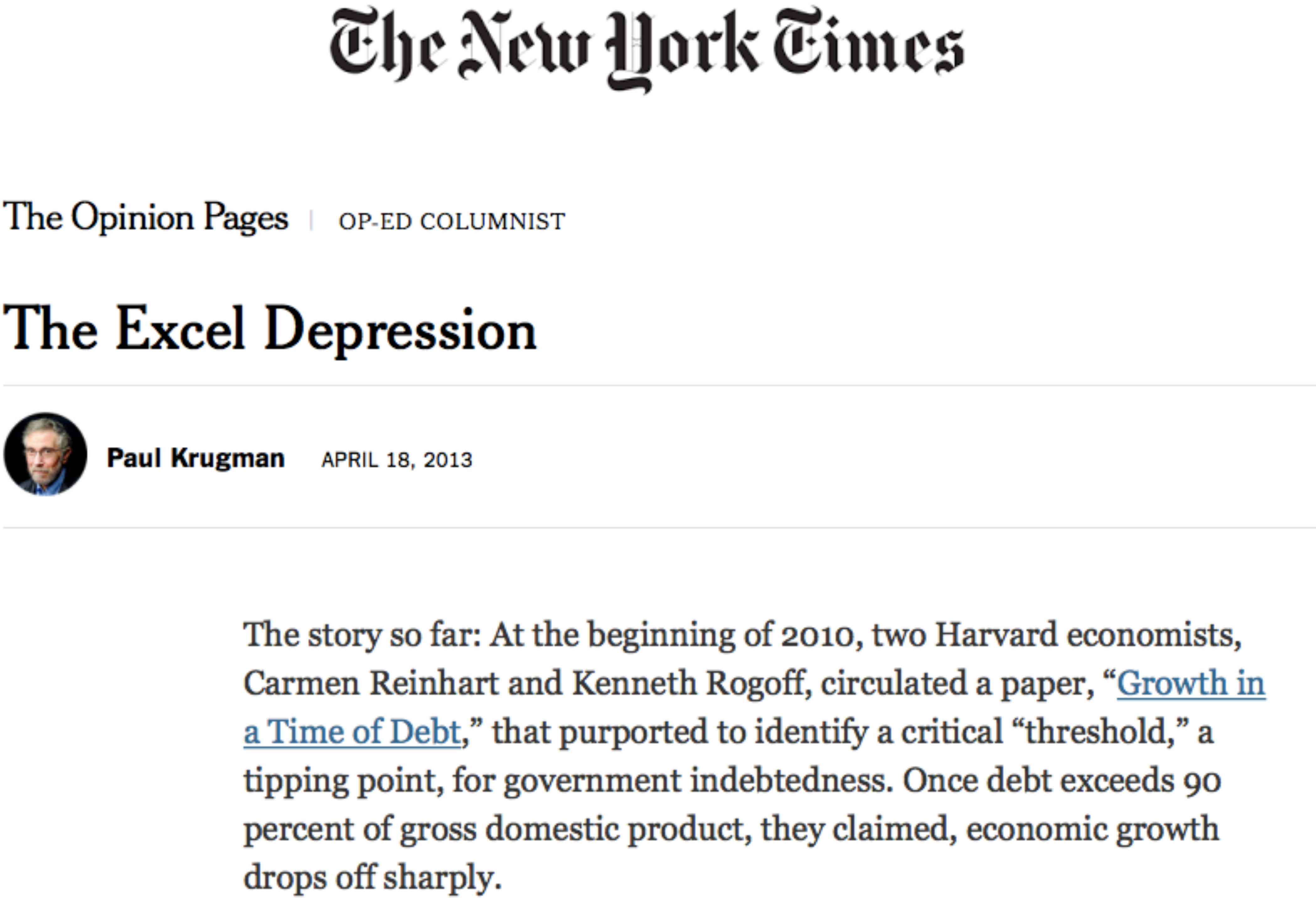
without comments

Authors of a 2016 cancer paper have retracted it after finding an error in one line of code in the program used to calculate some of the results.

JOURNAL OF CLINICAL ONCOLOGY
Official Journal of the American Society of Clinical Oncology
www.jco.org
ASCO
EDICIÓN ESPAÑOLA

Sarah Darby, last author of the now-retracted paper from the University of Oxford, UK, told *Retraction Watch* that the mistake was made by a doctoral student. When the error was realized, Darby said, she contacted the *Journal of Clinical Oncology* (*JCO*), explained the issue, and asked whether they would prefer a retraction or a correction. *JCO* wanted a retraction, and she complied.

The journal allowed the authors to publish a correspondence article outlining their new results.



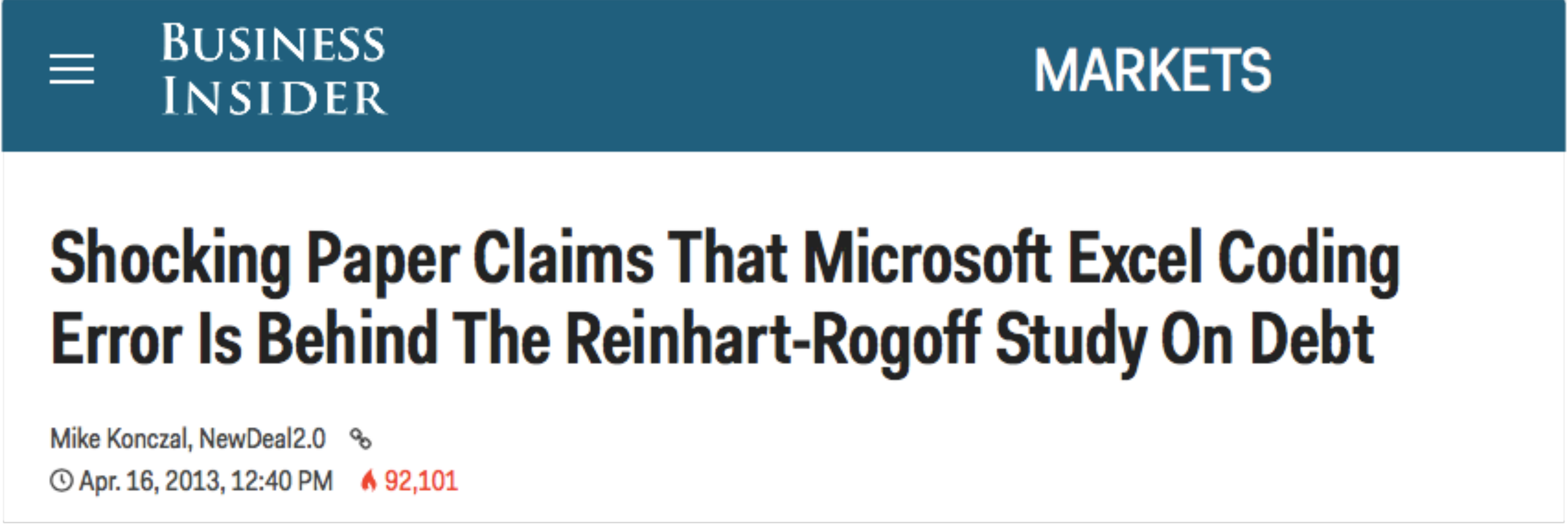
The New York Times

The Opinion Pages | OP-ED COLUMNIST

The Excel Depression

Paul Krugman APRIL 18, 2013

The story so far: At the beginning of 2010, two Harvard economists, Carmen Reinhart and Kenneth Rogoff, circulated a paper, "Growth in a Time of Debt," that purported to identify a critical "threshold," a tipping point, for government indebtedness. Once debt exceeds 90 percent of gross domestic product, they claimed, economic growth drops off sharply.



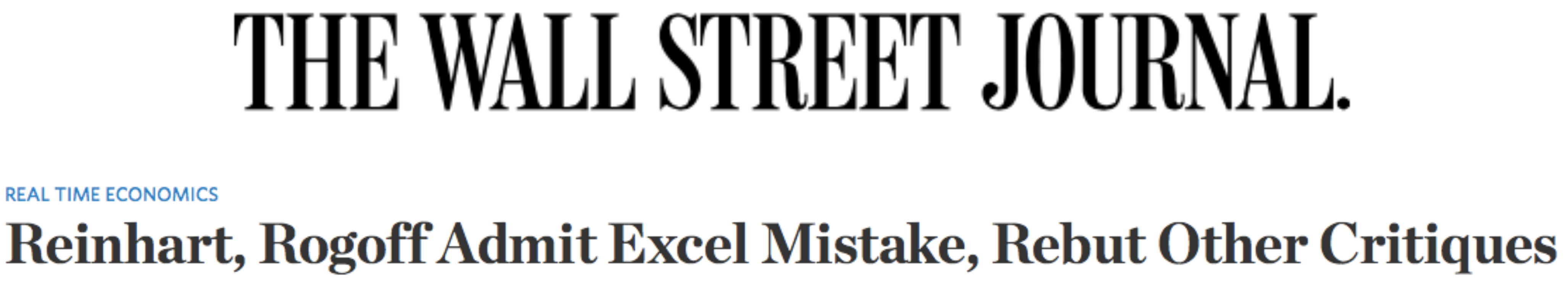
BUSINESS INSIDER

MARKETS

Shocking Paper Claims That Microsoft Excel Coding Error Is Behind The Reinhart-Rogoff Study On Debt

Mike Konczal, NewDeal2.0

Apr. 16, 2013, 12:40 PM 92,101



THE WALL STREET JOURNAL.

REAL TIME ECONOMICS

Reinhart, Rogoff Admit Excel Mistake, Rebut Other Critiques

Ziemann *et al. Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT **Open Access**

Gene name errors are widespread in the scientific literature

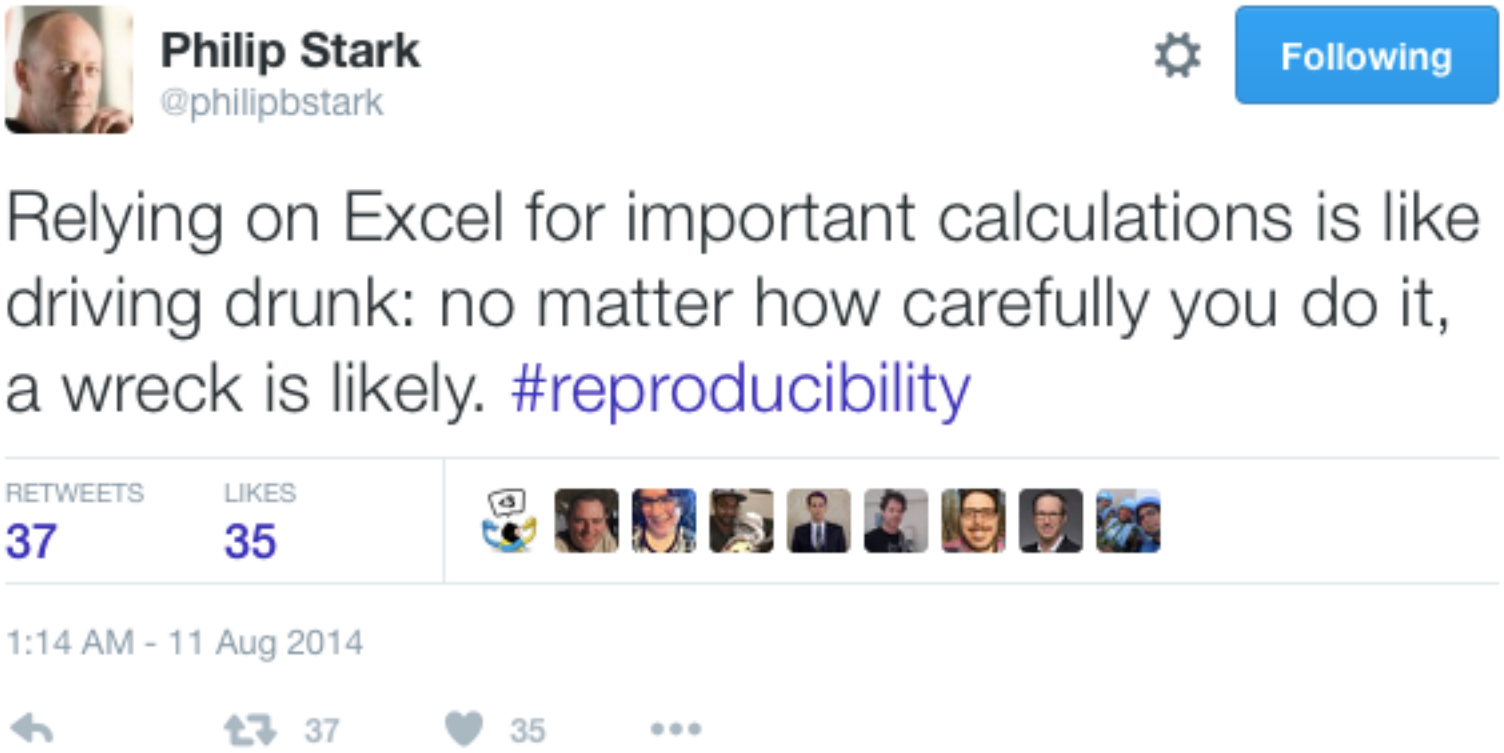
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

Keywords: Microsoft Excel, Gene symbol, Supplementary data

Abbreviations: GEO, Gene Expression Omnibus; JIF, journal impact factor



Philip Stark
@philipbstark

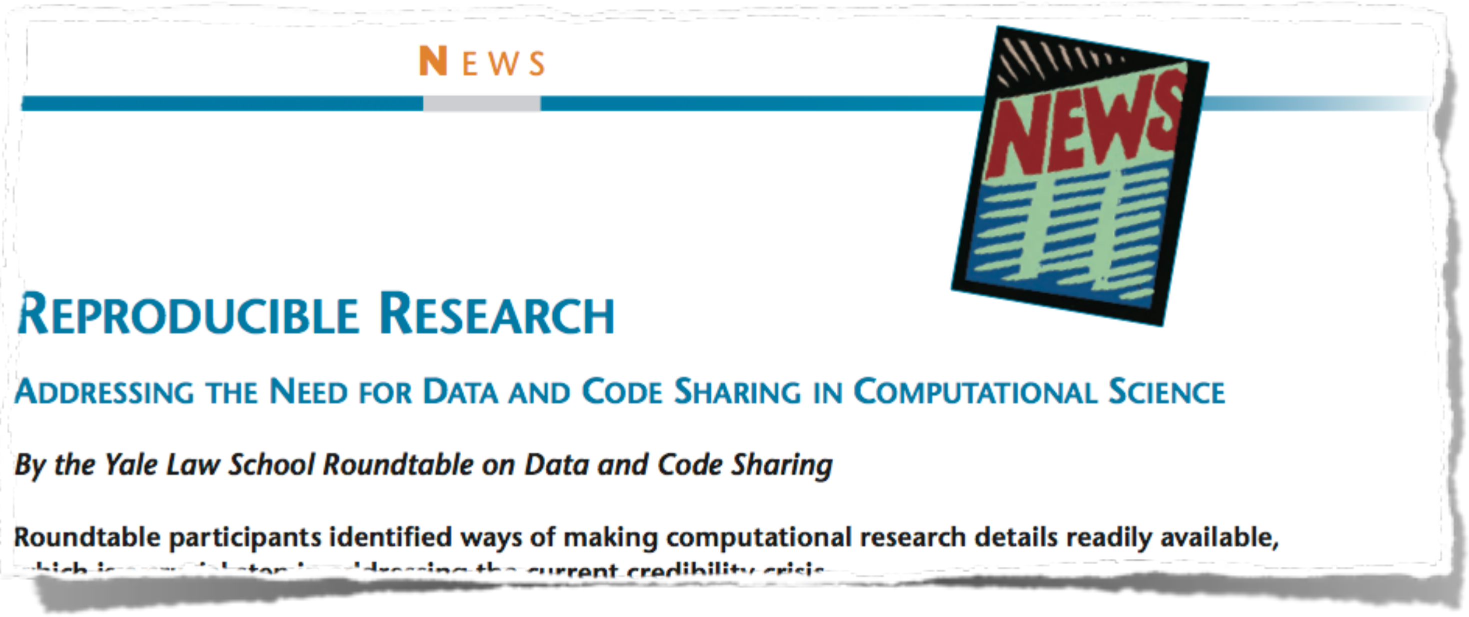
Relying on Excel for important calculations is like driving drunk: no matter how carefully you do it, a wreck is likely. #reproducibility

RETWEETS 37
LIKES 35

1:14 AM - 11 Aug 2014

2009 Yale Data and Code Sharing Roundtable

- 14 contributed thought pieces
- “Data and Code Sharing Declaration”
  … demanding a resolution to the **credibility crisis** from the lack of reproducible research in computational science.



REPRODUCIBLE RESEARCH

ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE

By the Yale Law School Roundtable on Data and Code Sharing

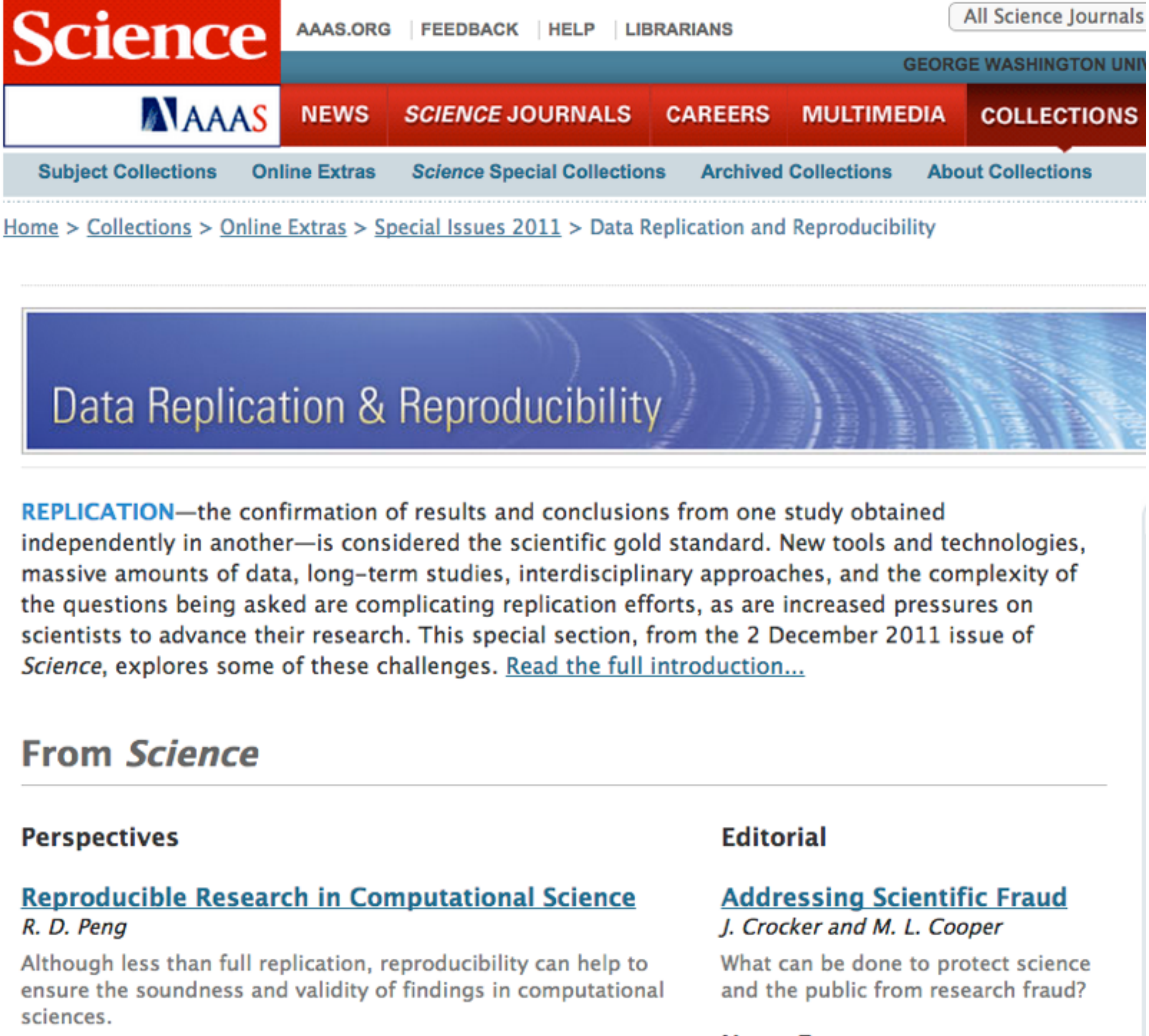
Roundtable participants identified ways of making computational research details readily available,

SEPT/OCT 2010 | COMPUTING IN SCIENCE AND ENGINEERING

Practicing safe software …

- Use a version-control system
- Track your materials
- Write testable software
- Test the software
- Encourage sharing of software

14 OCTOBER 2010 | VOL 467 | NATURE | 775



Data Replication & Reproducibility

REPLICATION—the confirmation of results and conclusions from one study obtained independently in another—is considered the scientific gold standard. New tools and technologies, massive amounts of data, long-term studies, interdisciplinary approaches, and the complexity of the questions being asked are complicating replication efforts, as are increased pressures on scientists to advance their research. This special section, from the 2 December 2011 issue of *Science*, explores some of these challenges. Read the full introduction...

From *Science*

Perspectives

Reproducible Research in Computational Science
R. D. Peng
Although less than full replication, reproducibility can help to ensure the soundness and validity of findings in computational sciences.

Editorial

Addressing Scientific Fraud
J. Crocker and M. L. Cooper
What can be done to protect science and the public from research fraud?

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

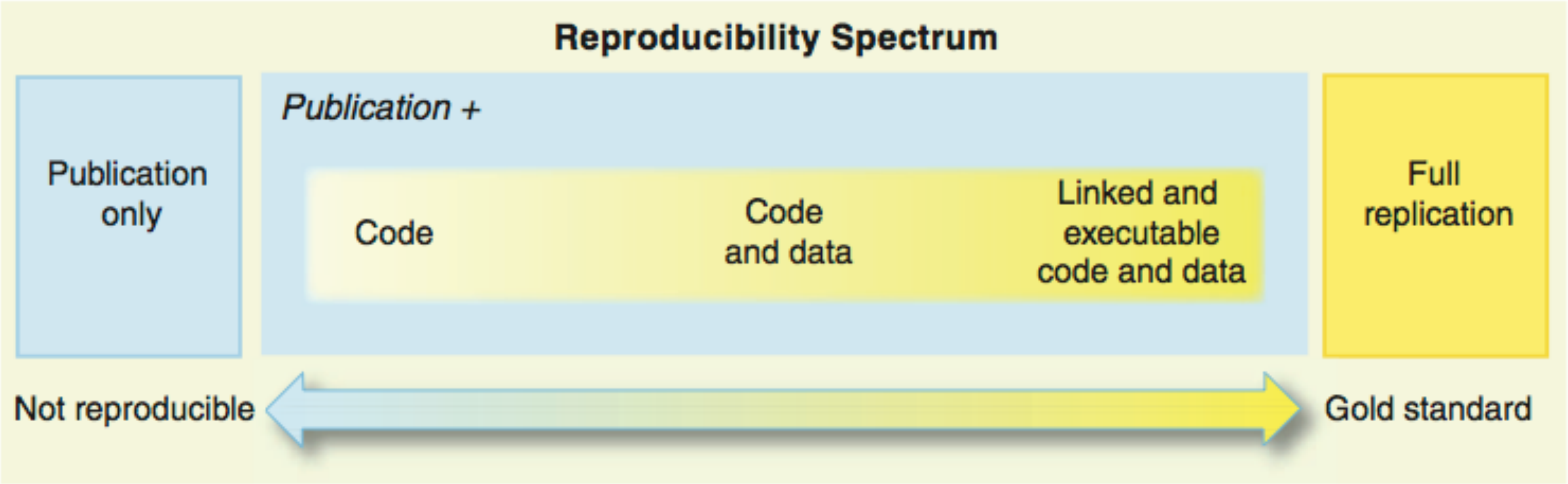
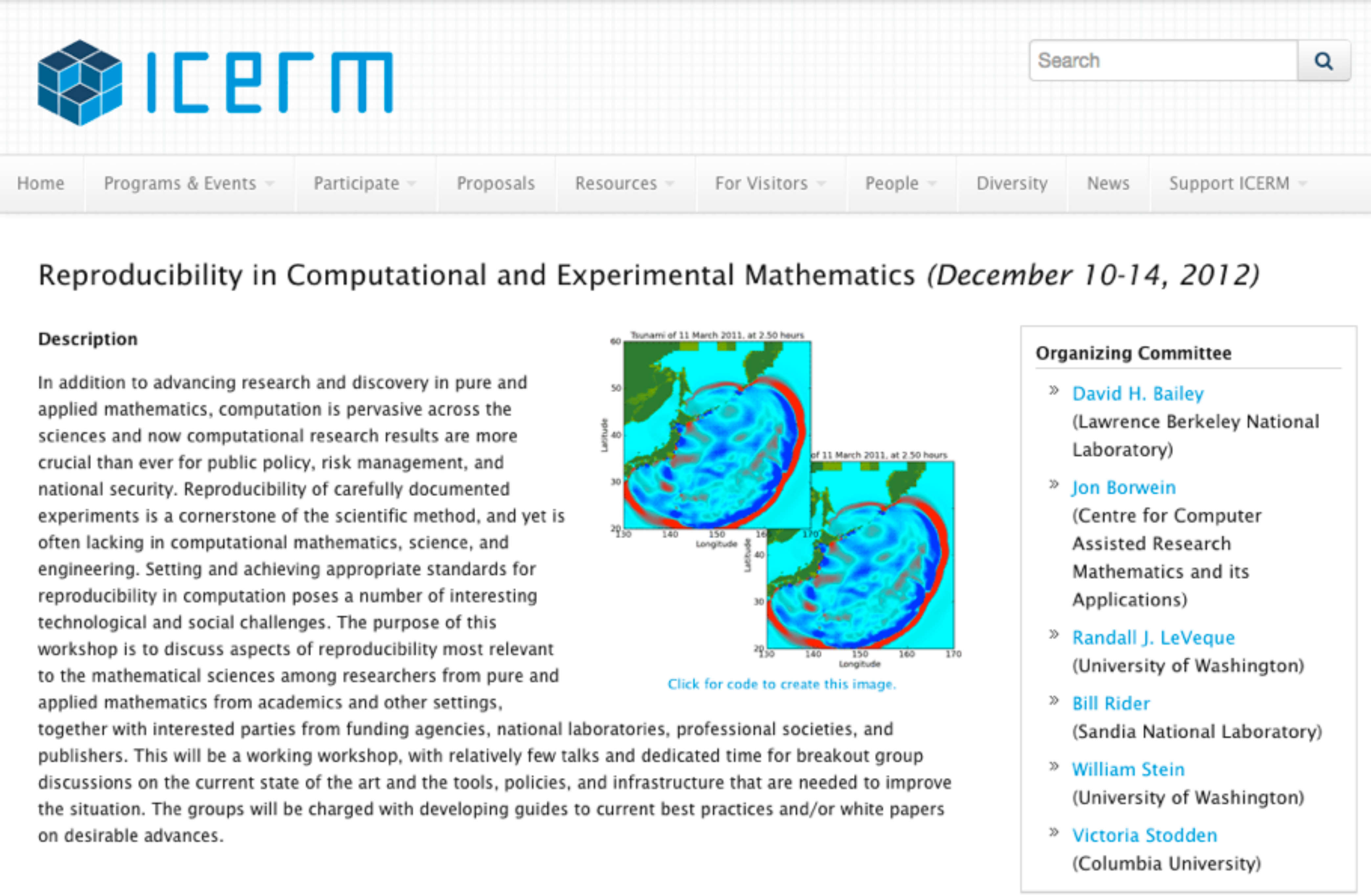


Fig. 1. The spectrum of reproducibility.



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.

Click for code to create this image.

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)

`http://icerm.brown.edu/tw12-5-rcem/`

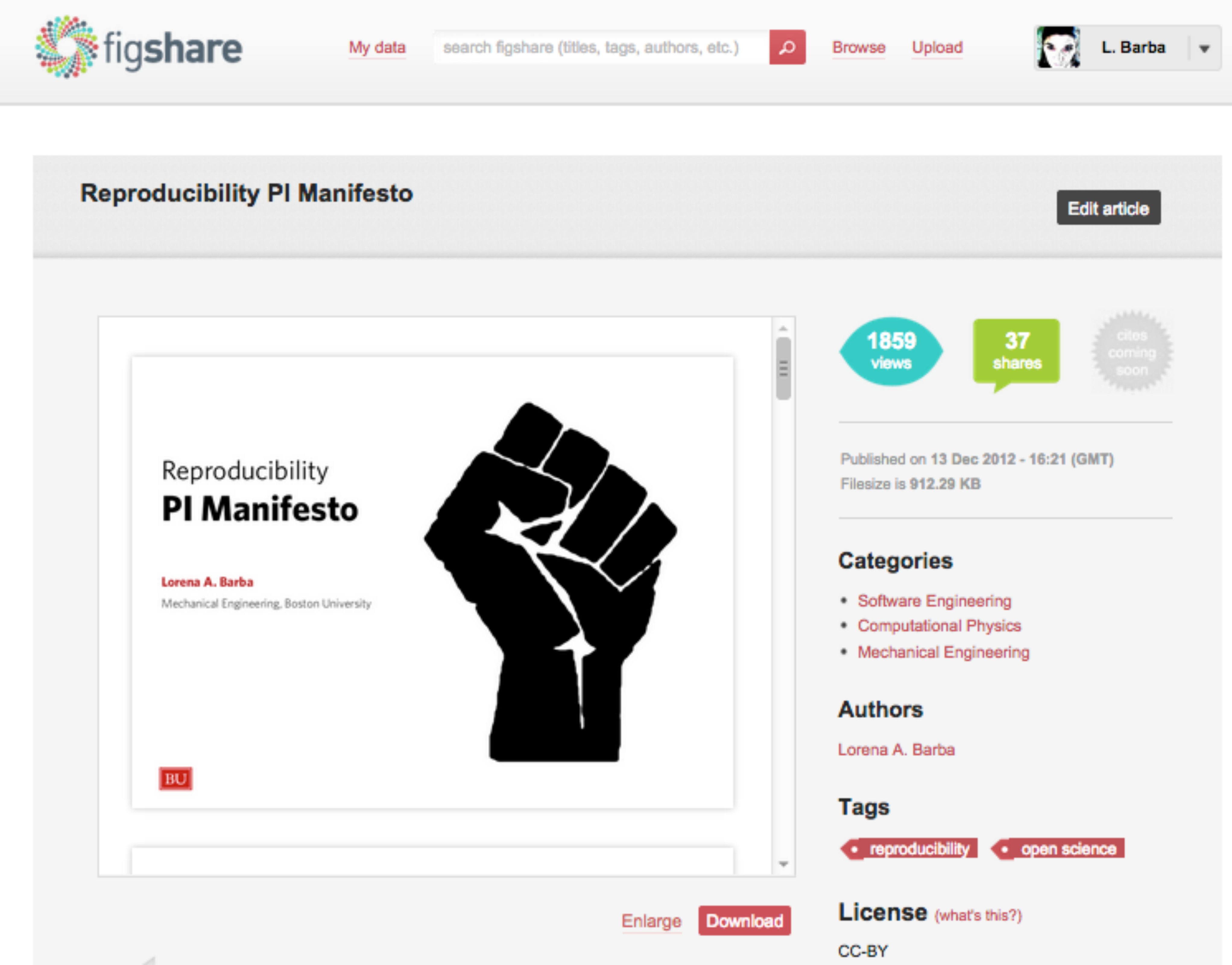
Lorena A. **Barba group**

Reproducibility PI Manifesto

http://lorenabarba.com/gallery/reproducibility-pi-manifesto/

Reproducibility **PI Manifesto**

- I teach my graduate students about reproducibility
- All our research code (and writing) is under version control
- We always carry out verification & validation (and make them public)
- For main results, we share data, plotting script & figure under CC-BY
- We upload preprint to arXiv at the time of submission to a journal
- We release code at the time of submission of a paper to a journal
- We add a "Reproducibility" declaration at the end of each paper
- I develop a consistent open-science policy & keep an up-to-date web presence



Reproducibility PI Manifesto

Reproducibility
PI Manifesto

Lorena A. Barba
Mechanical Engineering, Boston University

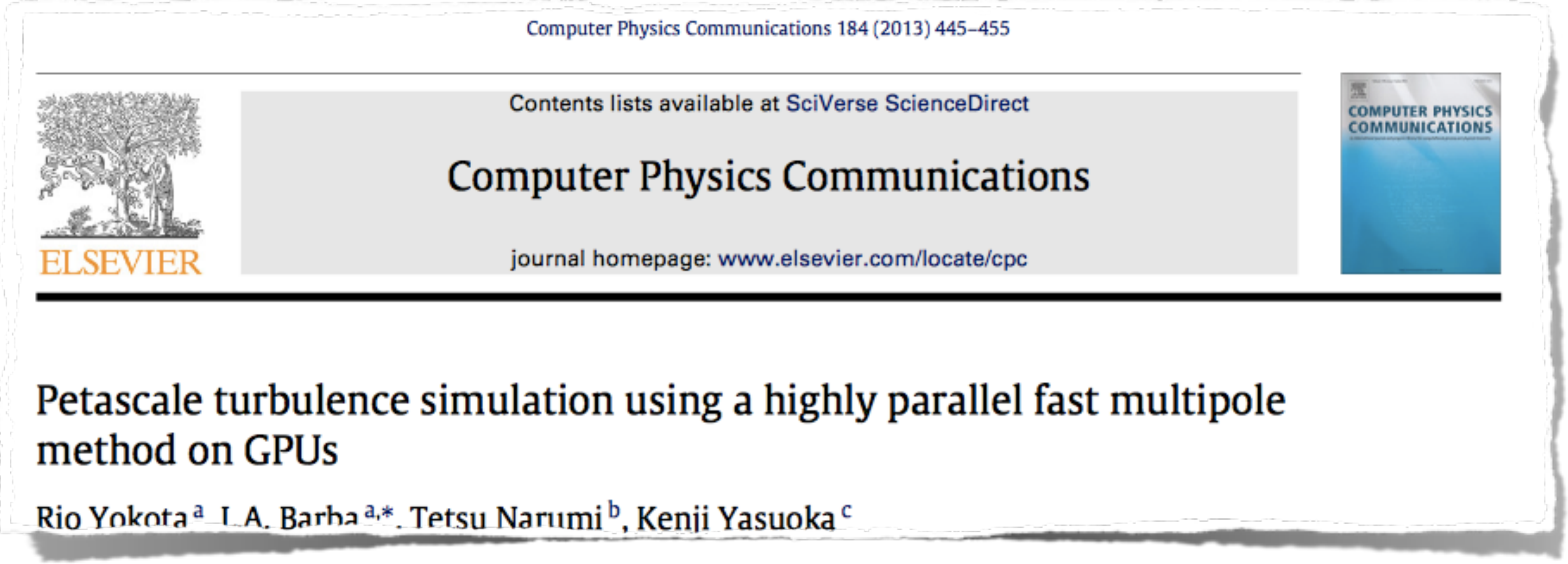
Published on 13 Dec 2012 - 16:21 (GMT)
Filesize is 912.29 KB

Categories
- Software Engineering
- Computational Physics
- Mechanical Engineering

Authors
Lorena A. Barba

Tags
reproducibility
open science

License (what's this?)
CC-BY



Computer Physics Communications 184 (2013) 445–455

Contents lists available at SciVerse ScienceDirect

Computer Physics Communications

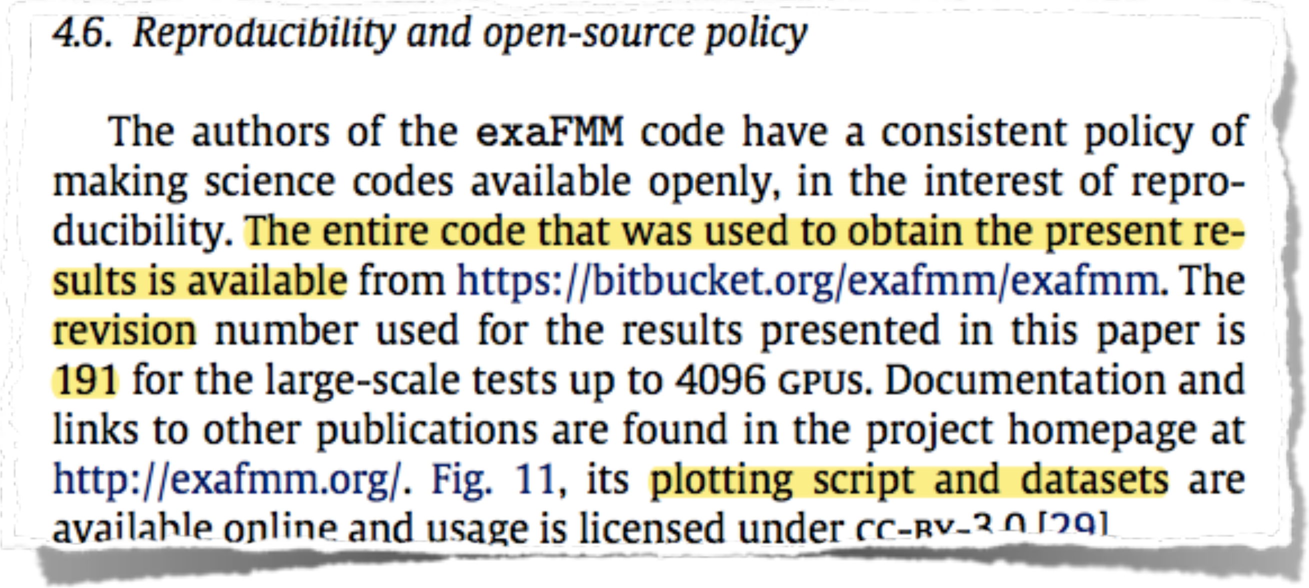
journal homepage: www.elsevier.com/locate/cpc

ELSEVIER

COMPUTER PHYSICS COMMUNICATIONS

Petascale turbulence simulation using a highly parallel fast multipole method on GPUs

Rio Yokota [a], L.A. Barba [a,*], Tetsu Narumi [b], Kenji Yasuoka [c]



4.6. Reproducibility and open-source policy

The authors of the `exaFMM` code have a consistent policy of making science codes available openly, in the interest of reproducibility. The entire code that was used to obtain the present results is available from https://bitbucket.org/exafmm/exafmm. The revision number used for the results presented in this paper is 191 for the large-scale tests up to 4096 GPUs. Documentation and links to other publications are found in the project homepage at http://exafmm.org/. Fig. 11, its plotting script and datasets are available online and usage is licensed under CC-BY-3.0 [29]

Why does it matter?

We use computers to create scientific knowledge.

“Essential skills for reproducible research computing”

A syllabus for research computing

1. command line utilities in Unix/Linux
2. an open-source scientific software ecosystem (our favorite is Python's)
3. software version control (we advocate the distributed kind: our favorite is git)
4. good practices for scientific software development: code hygiene and testing
5. knowledge of licensing options for sharing software

`https://barbagroup.github.io/essential_skills_RRC/`