Cyberinfrastructure in the Classroom: A Case Study Using Whole Tale

Victoria Stodden
School of Information Sciences
University of Illinois at Urbana-Champaign

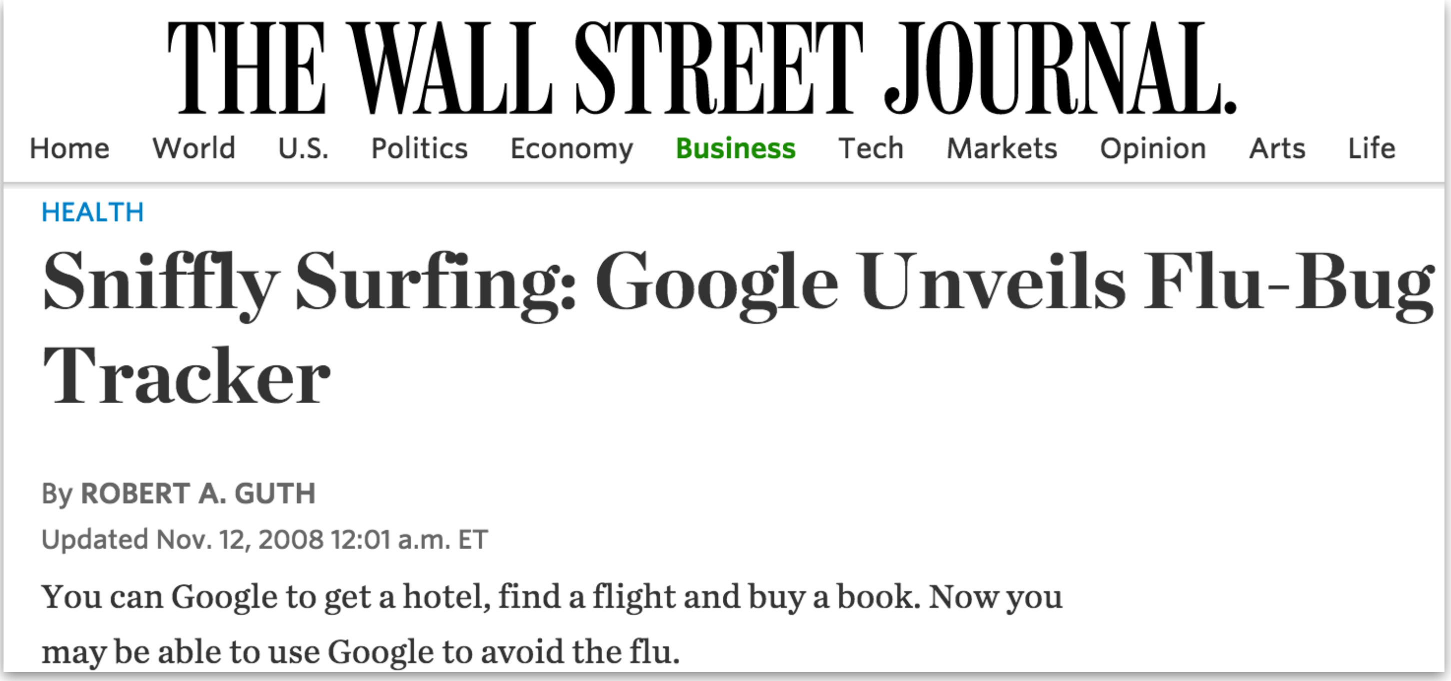
Workshop on Education and Training for Reproducible Research - An Infrastructure Perspective
National Center for Supercomputing Applications
March 7, 2019

Agenda

1. Teaching Reproducibility: Why and What

2. Two Examples of Leveraging Reproducibility Platforms in the Classroom

Remember Google Flu Trends?



HEALTH

Sniffly Surfing: Google Unveils Flu-Bug Tracker

By ROBERT A. GUTH

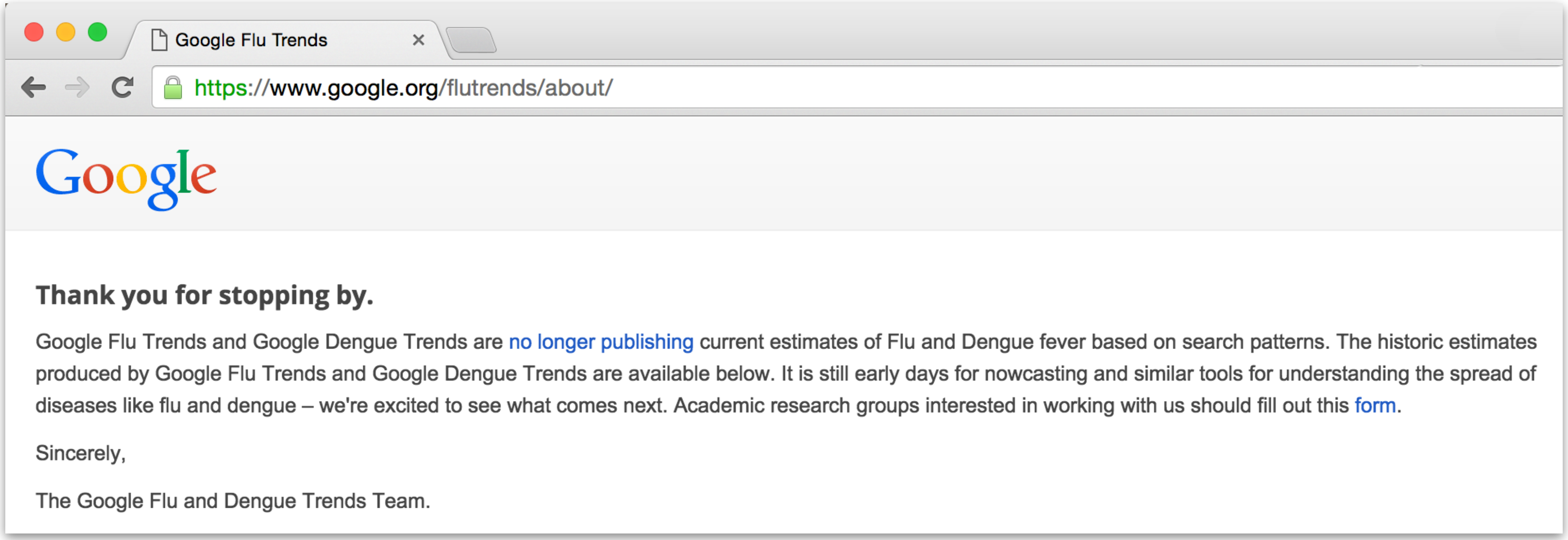
Updated Nov. 12, 2008 12:01 a.m. ET

You can Google to get a hotel, find a flight and buy a book. Now you may be able to use Google to avoid the flu.

In 2008 Google Flu Trends claimed it can tell you whether “the number of influenza cases is increasing in areas around the U.S., earlier than many existing methods”

In 2013 Google Flu Trends was predicting more than double the proportion of doctor visits for flu than the CDC.

Today:



Thank you for stopping by.

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for nowcasting and similar tools for understanding the spread of diseases like flu and dengue – we're excited to see what comes next. Academic research groups interested in working with us should fill out this form.

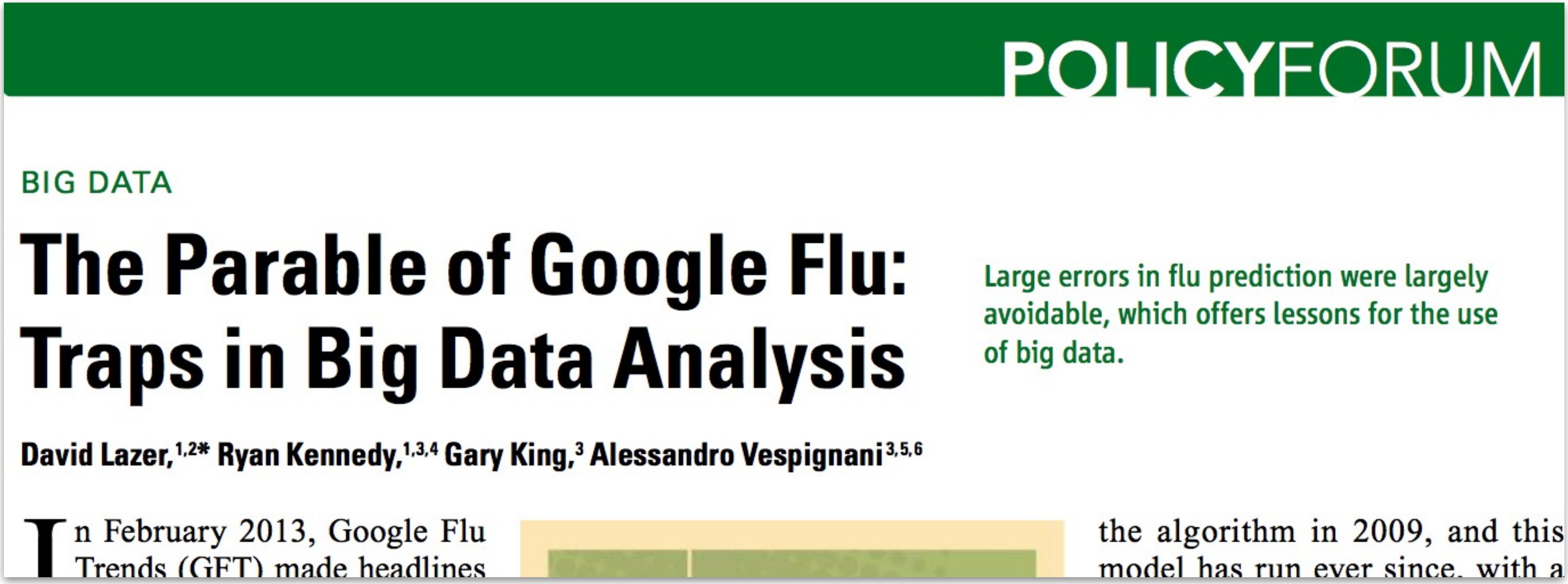
Sincerely,

The Google Flu and Dengue Trends Team.

What Happened?

- How did Google Flu Trends work? What was the data collection process? What was the algorithm?

- Why should we believe Google Flu Trends output? Many people did in 2008..



POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines

the algorithm in 2009, and this model has run ever since, with a

Technological Sources of Impact

1. Big Data / Data Driven Discovery: high dimensional data, p >> n,

2. Computational Power: simulation of the complete evolution of a physical system, systematically varying parameters,

3. Deep intellectual contributions now encoded only in software.

CSHL Keynote; Dr. Lior Pachter, UC Berkeley

Claim: *Virtually all published discoveries today have a computational component.*
Corollary: *There is a mismatch between traditional scientific dissemination practices and modern computational research processes, leading to reproducibility concerns.*

The software contains "ideas that enable biology..."
Stories from the Supplement, 2013

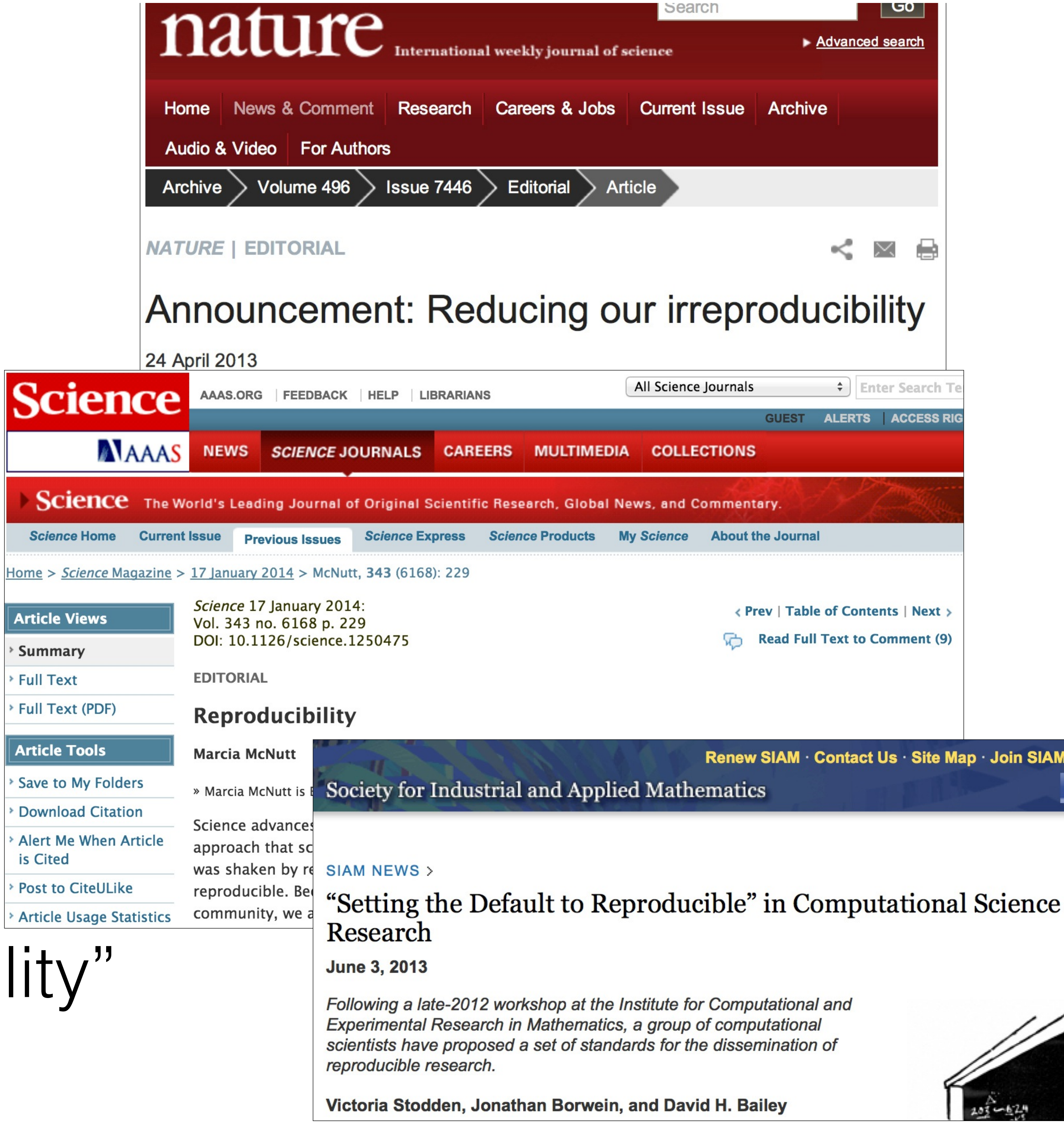
Parsing Reproducibility

“Empirical Reproducibility”

“Statistical Reproducibility”

“Computational Reproducibility”

V. Stodden, IMS Bulletin (2013)



NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

Home > *Science* Magazine > 17 January 2014 > McNutt, 343 (6168): 229

Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is E

Science advances
approach that sc
was shaken by re
reproducible. Bec
community, we a

Renew SIAM · Contact Us · Site Map · Join SIAM
Society for Industrial and Applied Mathematics

SIAM NEWS ›

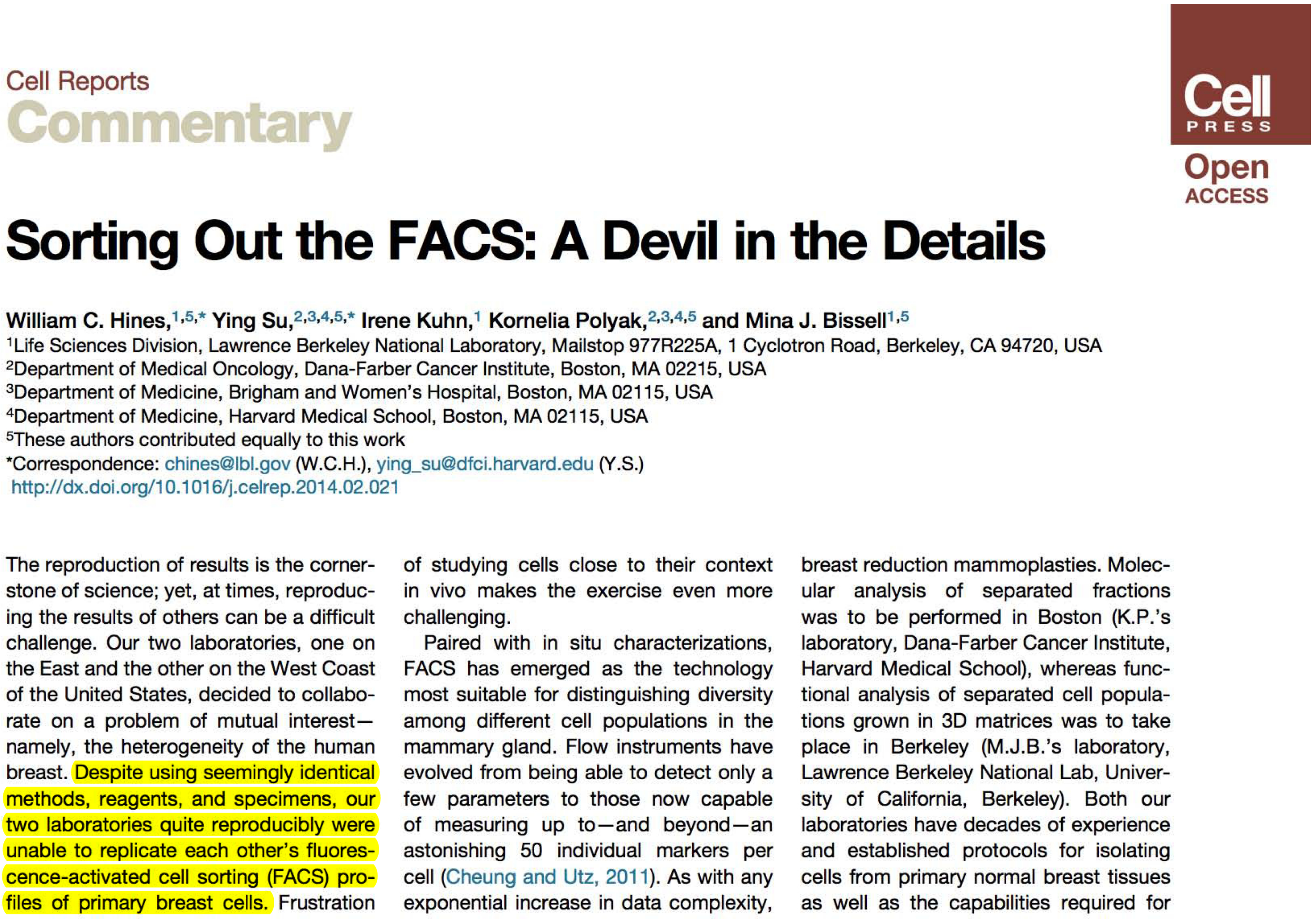
“Setting the Default to Reproducible” in Computational Science Research

June 3, 2013

Following a late-2012 workshop at the Institute for Computational and Experimental Research in Mathematics, a group of computational scientists have proposed a set of standards for the dissemination of reproducible research.

Victoria Stodden, Jonathan Borwein, and David H. Bailey

Empirical Reproducibility



Cell Reports

Commentary

Cell
PRESS
Open
ACCESS

Sorting Out the FACS: A Devil in the Details

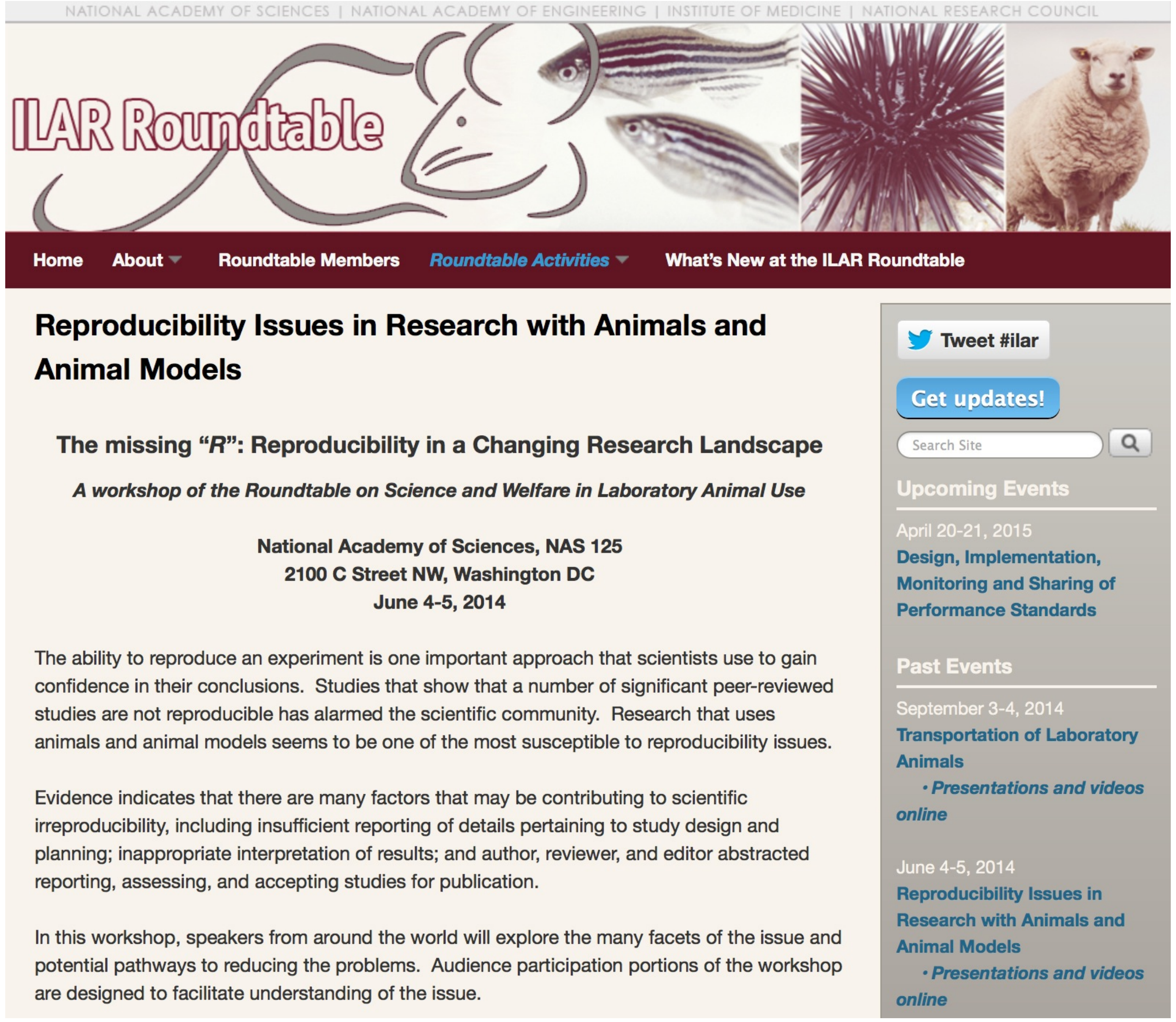
William C. Hines,[1,5,*] Ying Su,[2,3,4,5,*] Irene Kuhn,[1] Kornelia Polyak,[2,3,4,5] and Mina J. Bissell[1,5]

[1] Life Sciences Division, Lawrence Berkeley National Laboratory, Mailstop 977R225A, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3] Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[4] Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
[5] These authors contributed equally to this work
*Correspondence: chines@lbl.gov (W.C.H.), ying_su@dfci.harvard.edu (Y.S.)

http://dx.doi.org/10.1016/j.celrep.2014.02.021

The reproduction of results is the cornerstone of science; yet, at times, reproducing the results of others can be a difficult challenge. Our two laboratories, one on the East and the other on the West Coast of the United States, decided to collaborate on a problem of mutual interest—namely, the heterogeneity of the human breast. Despite using seemingly identical methods, reagents, and specimens, our two laboratories quite reproducibly were unable to replicate each other's fluorescence-activated cell sorting (FACS) profiles of primary breast cells. Frustration of studying cells close to their context in vivo makes the exercise even more challenging.

Paired with in situ characterizations, FACS has emerged as the technology most suitable for distinguishing diversity among different cell populations in the mammary gland. Flow instruments have evolved from being able to detect only a few parameters to those now capable of measuring up to—and beyond—an astonishing 50 individual markers per cell (Cheung and Utz, 2011). As with any exponential increase in data complexity, breast reduction mammoplasties. Molecular analysis of separated fractions was to be performed in Boston (K.P.'s laboratory, Dana-Farber Cancer Institute, Harvard Medical School), whereas functional analysis of separated cell populations grown in 3D matrices was to take place in Berkeley (M.J.B.'s laboratory, Lawrence Berkeley National Lab, University of California, Berkeley). Both our laboratories have decades of experience and established protocols for isolating cells from primary normal breast tissues as well as the capabilities required for



NATIONAL ACADEMY OF SCIENCES | NATIONAL ACADEMY OF ENGINEERING | INSTITUTE OF MEDICINE | NATIONAL RESEARCH COUNCIL
ILAR Roundtable
Home
About
Roundtable Members
Roundtable Activities
What's New at the ILAR Roundtable
Tweet #ilar
Get updates!
Search Site

Reproducibility Issues in Research with Animals and Animal Models

The missing "*R*": Reproducibility in a Changing Research Landscape

A workshop of the Roundtable on Science and Welfare in Laboratory Animal Use

National Academy of Sciences, NAS 125
2100 C Street NW, Washington DC
June 4-5, 2014

The ability to reproduce an experiment is one important approach that scientists use to gain confidence in their conclusions. Studies that show that a number of significant peer-reviewed studies are not reproducible has alarmed the scientific community. Research that uses animals and animal models seems to be one of the most susceptible to reproducibility issues.

Evidence indicates that there are many factors that may be contributing to scientific irreproducibility, including insufficient reporting of details pertaining to study design and planning; inappropriate interpretation of results; and author, reviewer, and editor abstracted reporting, assessing, and accepting studies for publication.

In this workshop, speakers from around the world will explore the many facets of the issue and potential pathways to reducing the problems. Audience participation portions of the workshop are designed to facilitate understanding of the issue.

Upcoming Events

April 20-21, 2015
Design, Implementation, Monitoring and Sharing of Performance Standards

Past Events

September 3-4, 2014
Transportation of Laboratory Animals
- ***Presentations and videos online***

June 4-5, 2014
Reproducibility Issues in Research with Animals and Animal Models
- ***Presentations and videos online***

Statistical Reproducibility

- False discovery, p-hacking (Simonsohn 2012), file drawer problem, overuse and mis-use of p-values, lack of multiple testing adjustments,
- Low power, poor experimental design, nonrandom sampling, insufficient sample size,
- Data preparation, treatment of outliers and missing values, re-combination of datasets,
- Inappropriate tests or models, model misspecification, poor parameter estimation techniques,
- Model robustness to parameter changes and data perturbations,
- …

Computational Reproducibility

Traditionally two branches to the scientific method:

- Branch 1 (deductive): mathematics, formal logic,
- Branch 2 (empirical): statistical analysis of controlled experiments.

Now, new branches due to technological changes?

- Branch 3,4? (computational): large scale simulations / data driven computational science.

The Ubiquity of Error

The central motivation for the scientific method is to root out error:

- Deductive branch: the well-defined concept of the proof,

- Empirical branch: the machinery of hypothesis testing, appropriate statistical methods, structured communication of methods and protocols.

Claim: Computation and Data Science present *potential* third/fourth branches of the scientific method (Donoho et al. 2009), until the development of comparable standards.

Teaching and Reproducibility: Why

1. Reproducibility as a *research practice*: e.g. workflow information and documentation; artifact sharing.

2. Reproducibility as a *skill*: e.g. tool and platform use; generalizability of research findings (appropriate statistical techniques).

3. Reproducibility as a *tool*: e.g. leveraging teaching practices; providing teaching platforms; critical analysis of findings.

Teaching and Reproducibility: What

My experience in the classroom with a reproducibility platform:

- What problems were solved, what happened, and what was the reaction.

- Larger impact: understanding the ideas and reasons for reproducible research.

The Whole Tale Platform

- A *Quantitative Programming Environment* designed to capture the end to end computational research workflow.

- Implements well-known interfaces to R and python, RStudio and the Jupyter Notebook, and a unix terminal window.

What problems were solved?

- I teach “Introduction to Data Science”
- Students come from different backgrounds and different degree programs.
- Illinois has no centralized, openly available, computing resources, so students use what their unit provides, which is highly variable.
- Students were able to access Whole Tale and were presented with a uniform computing environment.

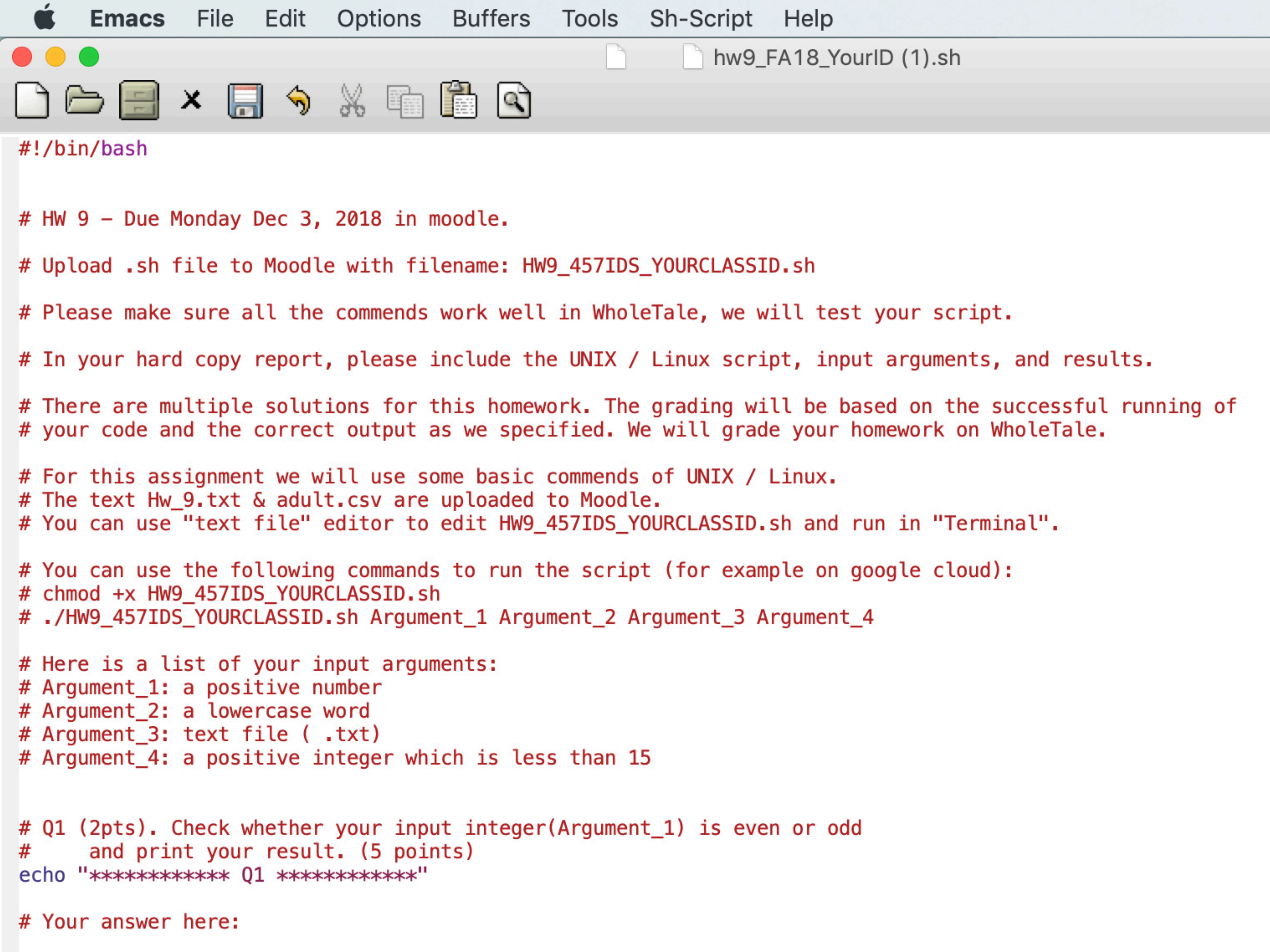
What problems were solved?

- This made it possible to teach unix and shell scripting skills to a diverse group of students.
- It also made it possible for students to create reproducible research compendia as part of their homework assignments.

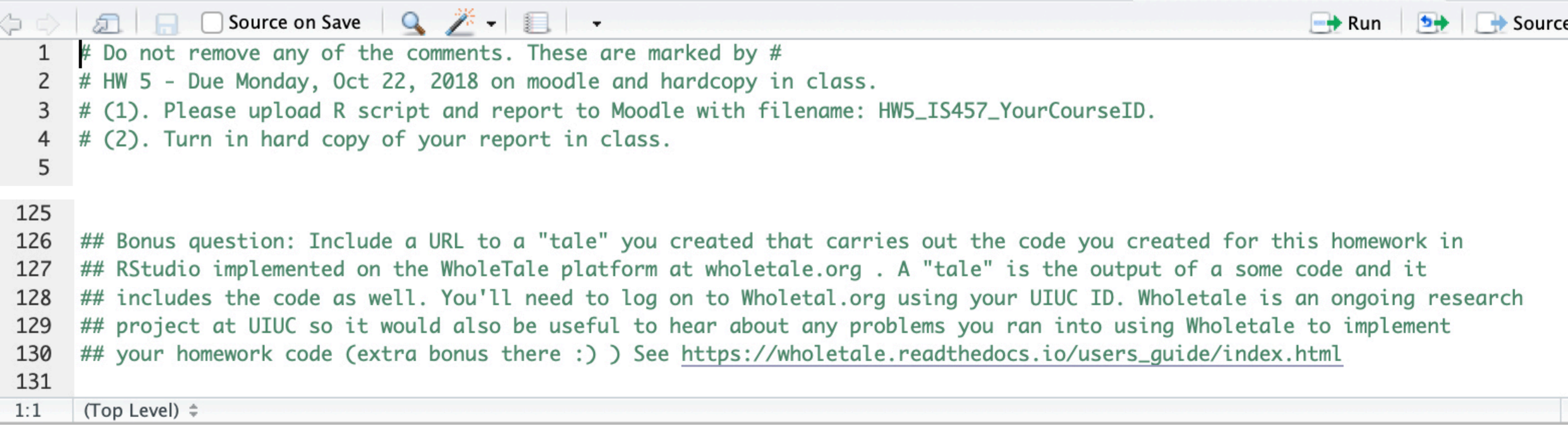
What Happened?

1. Students used Whole Tale for their shell scripting homework as there are no uniform shell environments available to all students. The students created “tales,” and then we (me and TAs) ran their scripts on the Whole Tale Platform.

2. Students created a “tale” of an R homework assignment on Whole Tale and again we ran their scripts on Whole Tale.



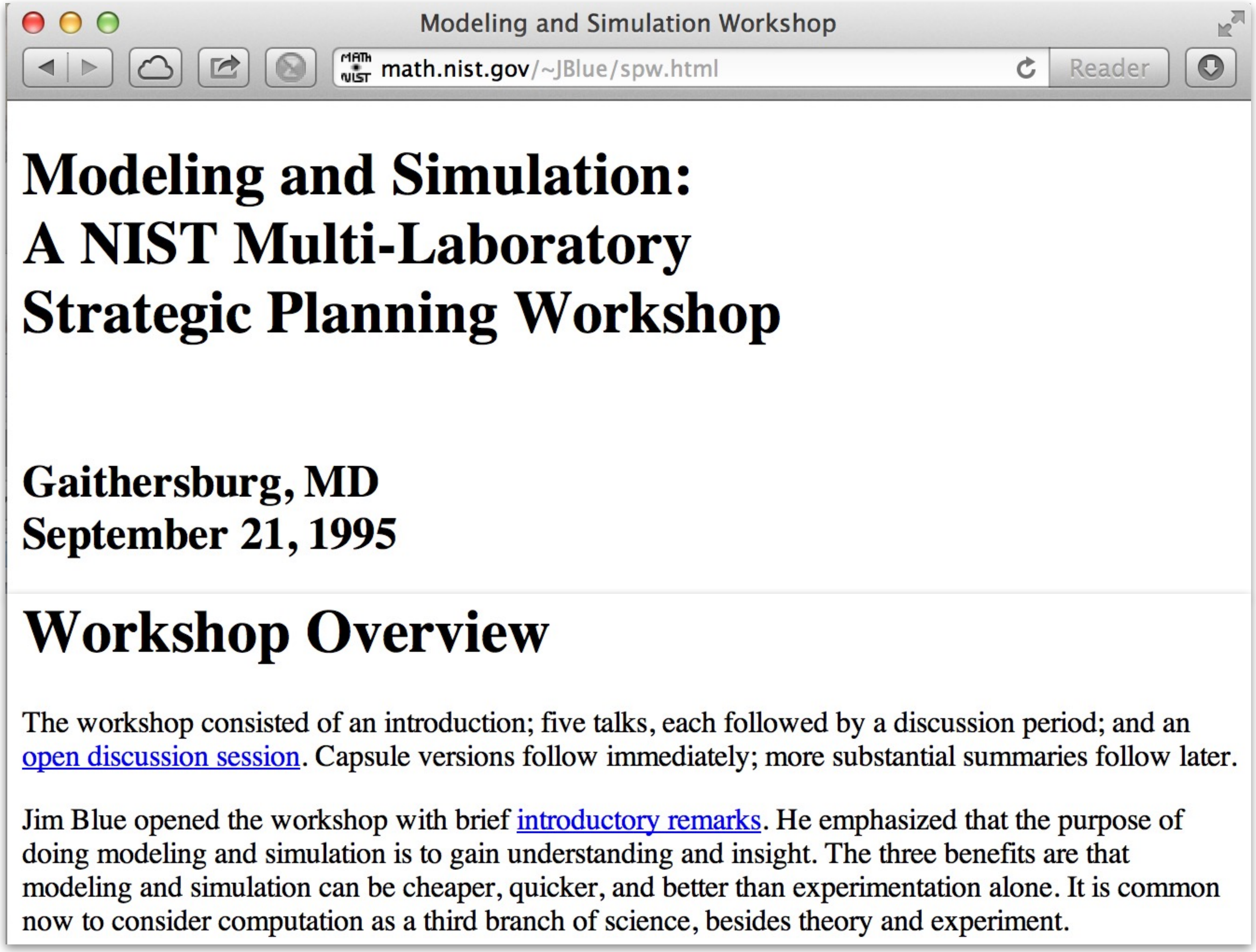
```
#!/bin/bash


# HW 9 - Due Monday Dec 3, 2018 in moodle.

# Upload .sh file to Moodle with filename: HW9_457IDS_YOURCLASSID.sh

# Please make sure all the commends work well in WholeTale, we will test your script.

# In your hard copy report, please include the UNIX / Linux script, input arguments, and results.

# There are multiple solutions for this homework. The grading will be based on the successful running of
# your code and the correct output as we specified. We will grade your homework on WholeTale.

# For this assignment we will use some basic commends of UNIX / Linux.
# The text Hw_9.txt & adult.csv are uploaded to Moodle.
# You can use "text file" editor to edit HW9_457IDS_YOURCLASSID.sh and run in "Terminal".

# You can use the following commands to run the script (for example on google cloud):
# chmod +x HW9_457IDS_YOURCLASSID.sh
# ./HW9_457IDS_YOURCLASSID.sh Argument_1 Argument_2 Argument_3 Argument_4

# Here is a list of your input arguments:
# Argument_1: a positive number
# Argument_2: a lowercase word
# Argument_3: text file ( .txt)
# Argument_4: a positive integer which is less than 15


# Q1 (2pts). Check whether your input integer(Argument_1) is even or odd
#     and print your result. (5 points)
echo "************ Q1 ************"

# Your answer here:
```



```
# Do not remove any of the comments. These are marked by #
# HW 5 - Due Monday, Oct 22, 2018 on moodle and hardcopy in class.
# (1). Please upload R script and report to Moodle with filename: HW5_IS457_YourCourseID.
# (2). Turn in hard copy of your report in class.


## Bonus question: Include a URL to a "tale" you created that carries out the code you created for this homework in
## RStudio implemented on the WholeTale platform at wholetale.org . A "tale" is the output of a some code and it
## includes the code as well. You'll need to log on to Wholetal.org using your UIUC ID. Wholetale is an ongoing research
## project at UIUC so it would also be useful to hear about any problems you ran into using Wholetale to implement
## your homework code (extra bonus there :) ) See https://wholetale.readthedocs.io/users_guide/index.html

```

Reaction and Larger Impact

- Students generally liked the interaction with Whole Tale.
- A little more than half provided bonus “tales” to us.
- Many provided constructive and valuable criticism regarding their experience with Whole Tale.


Modeling and Simulation Workshop
math.nist.gov/~JBlue/spw.html
Reader

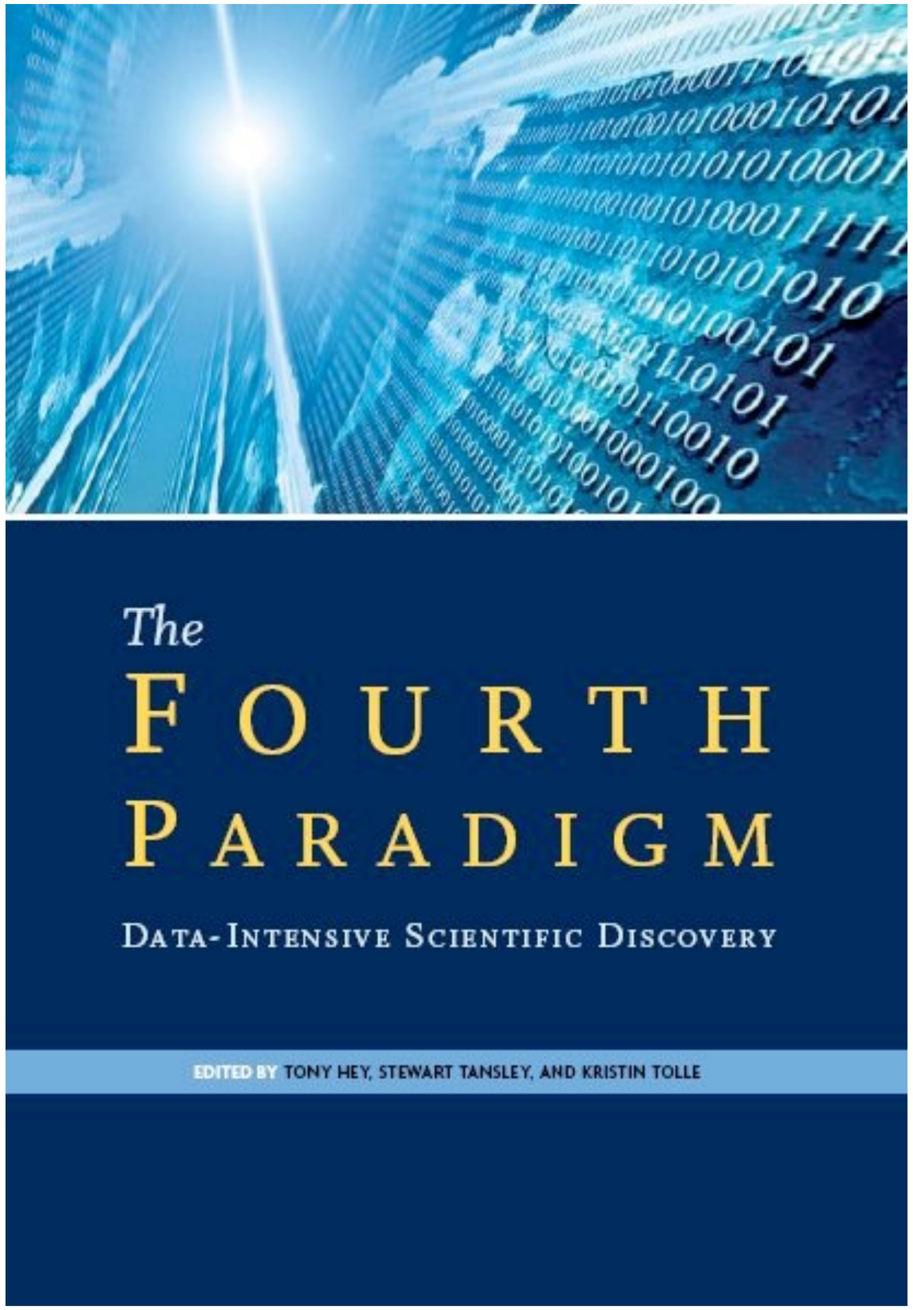
Modeling and Simulation: A NIST Multi-Laboratory Strategic Planning Workshop

Gaithersburg, MD
September 21, 1995

Workshop Overview

The workshop consisted of an introduction; five talks, each followed by a discussion period; and an open discussion session. Capsule versions follow immediately; more substantial summaries follow later.

Jim Blue opened the workshop with brief introductory remarks. He emphasized that the purpose of doing modeling and simulation is to gain understanding and insight. The three benefits are that modeling and simulation can be cheaper, quicker, and better than experimentation alone. It is common now to consider computation as a third branch of science, besides theory and experiment.



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

"It is common now to consider computation as a third branch of science, besides theory and experiment."

"This book is about a new, fourth paradigm for science based on data-intensive computing."

Really Reproducible Research

“Really Reproducible Research” (1992) inspired by Stanford Professor Jon Claerbout:

> “**The idea is: An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete ... set of instructions [and data] which generated the figures.**” David Donoho, 1998

A Convergence of Trends

- ➡ Scientific projects will become massively more computing intensive, and
- ➡ Scientific computing will become dramatically more transparent

Simultaneity: better transparency allows much more ambitious computational experiments. *And* better computational experiment infrastructure allows greater transparency.

Such a system is used not out of ethics or hygiene, but because this is a corollary of managing massive amounts of computational work, enabling *efficiency* and *productivity*, and *discovery*.

“Quantitative Programming Environments”

Define and create “**Quantitative Programming Environments**” to (easily) manage the conduct of massive computational experiments and expose the resulting data for analysis and structure the subsequent data analysis

Address the two trends simultaneously: better transparency will allow people to run **much more ambitious computational experiments**. *And* better computational experiment infrastructure will allow researchers to be **more transparent**.

Whole Tale: Merging **Science** and **Cyberinfrastructure** Pathways

Bertram Ludaescher, Kyle Chard, Niall Gaffney, Matthew B. Jones, Jaroslaw Nabrzyski, Victoria Stodden, Matt Turk

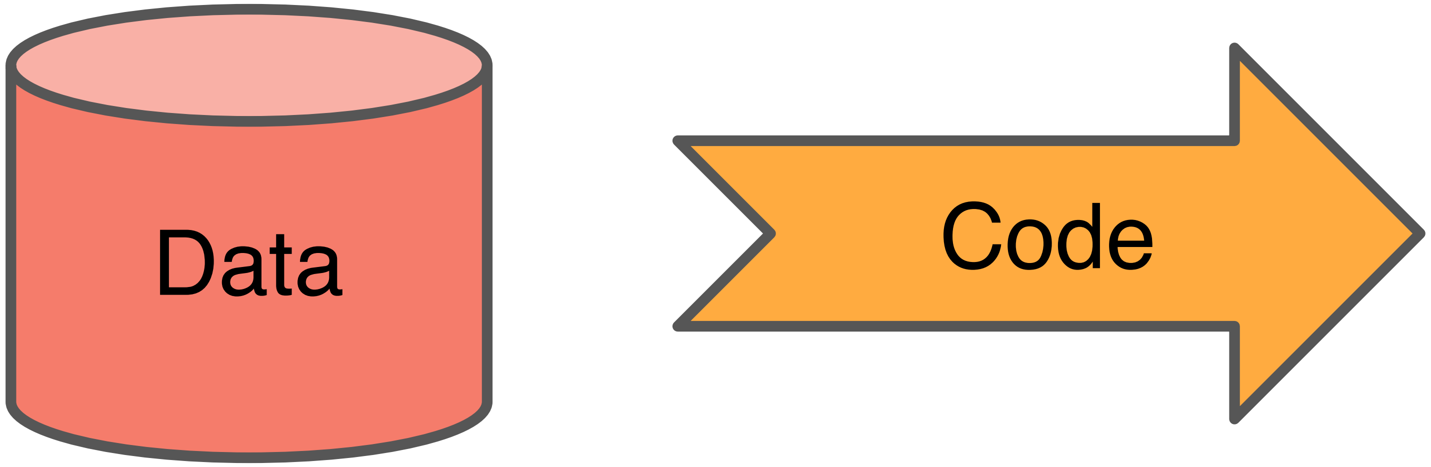
wholetale.org

ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · THE UNIVERSITY OF CHICAGO · UCSB UNIVERSITY OF CALIFORNIA SANTA BARBARA · TEXAS The University of Texas at Austin · UNIVERSITY OF NOTRE DAME

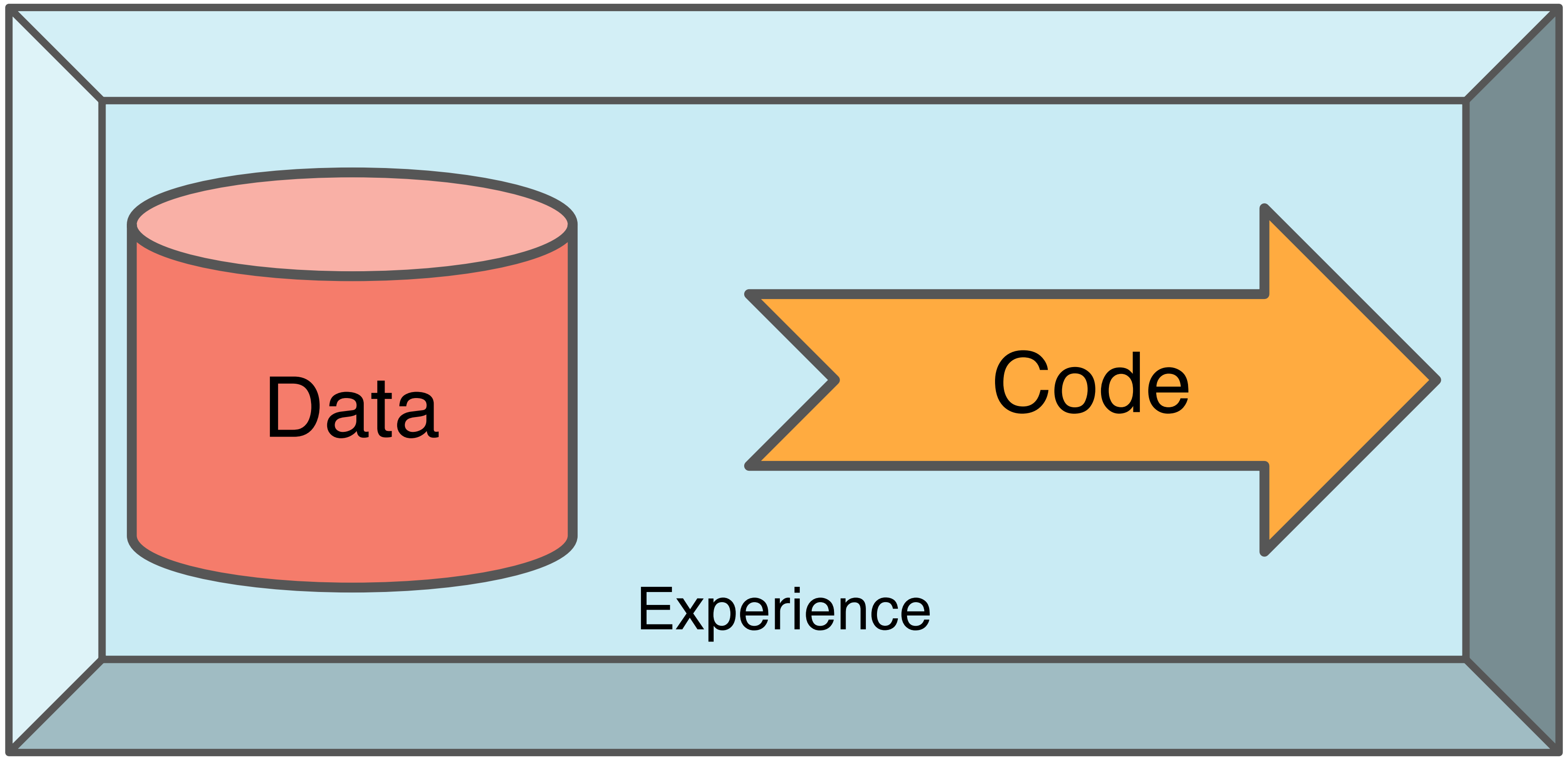
Whole Tale Vision

“It used to be, you’d publish a paper…”

Whole Tale Vision



Whole Tale Vision



Core to our mission is *active*, *meaningful* engagement with open source and research communities.

jupyter

RStudio®

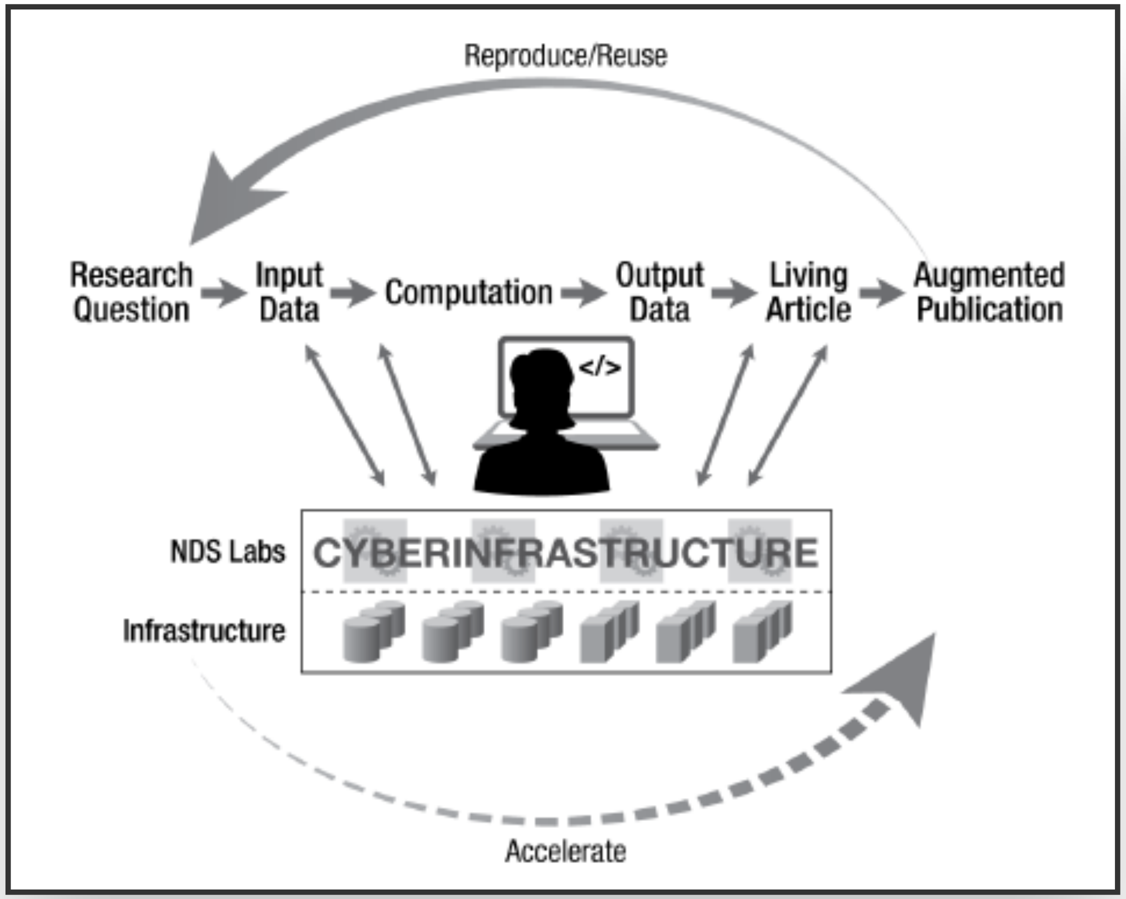
nteract

Girder

Data management platform

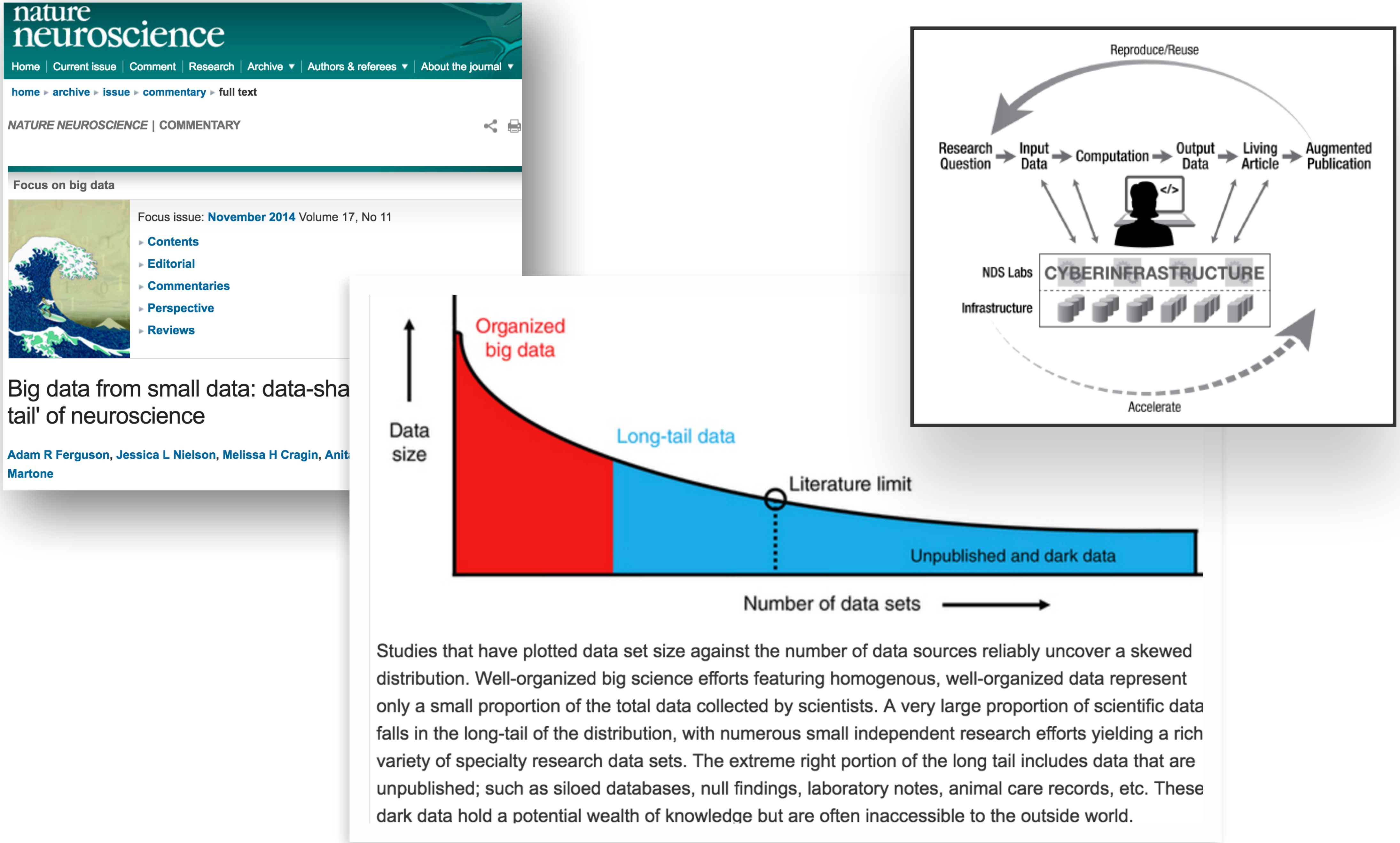
Whole Tale: What's in a name?

- (1) Whole Tale ⇔ Whole **Story**:
  - **Support** (computational & data) **scientists**
  - … along the **complete research lifecycle**
  - ... from **experiment** to (new kind of) **publication**
    - ... and back!



Whole Tale: What's in a name?

- (2) Whole Tale ⇔ For the **Long Tail of Science**
  - *"Big data & compute for mere mortals"*



nature neuroscience
Home | Current issue | Comment | Research | Archive ▾ | Authors & referees ▾ | About the journal ▾
home ▸ archive ▸ issue ▸ commentary ▸ full text
NATURE NEUROSCIENCE | COMMENTARY
Focus on big data

Focus issue: November 2014 Volume 17, No 11

▸ Contents
▸ Editorial
▸ Commentaries
▸ Perspective
▸ Reviews

Big data from small data: data-sha
tail' of neuroscience

Adam R Ferguson, Jessica L Nielson, Melissa H Cragin, Anit
Martone

Data size
Organized big data
Long-tail data
Literature limit
Unpublished and dark data
Number of data sets

Studies that have plotted data set size against the number of data sources reliably uncover a skewed distribution. Well-organized big science efforts featuring homogenous, well-organized data represent only a small proportion of the total data collected by scientists. A very large proportion of scientific data falls in the long-tail of the distribution, with numerous small independent research efforts yielding a rich variety of specialty research data sets. The extreme right portion of the long tail includes data that are unpublished; such as siloed databases, null findings, laboratory notes, animal care records, etc. These dark data hold a potential wealth of knowledge but are often inaccessible to the outside world.

Whole Tale Vision

- The Old Way:
  - Scholarly **Publication** .. || .. Data .. || .. Code

- The Emerging Way:
  - Scholarly **Publication** ⇔ **Data** .. | .. Code

- The New Way:
  - “Living” **Publication** ⇔ **Data** ⇔ **Code**

  = *Computational Narrative*
  - (more easily) *Reproducible Science*

.. participate in and share the *experience of inquiry*

Problems Facing Researchers

Workflow for data research is **fragmented:**

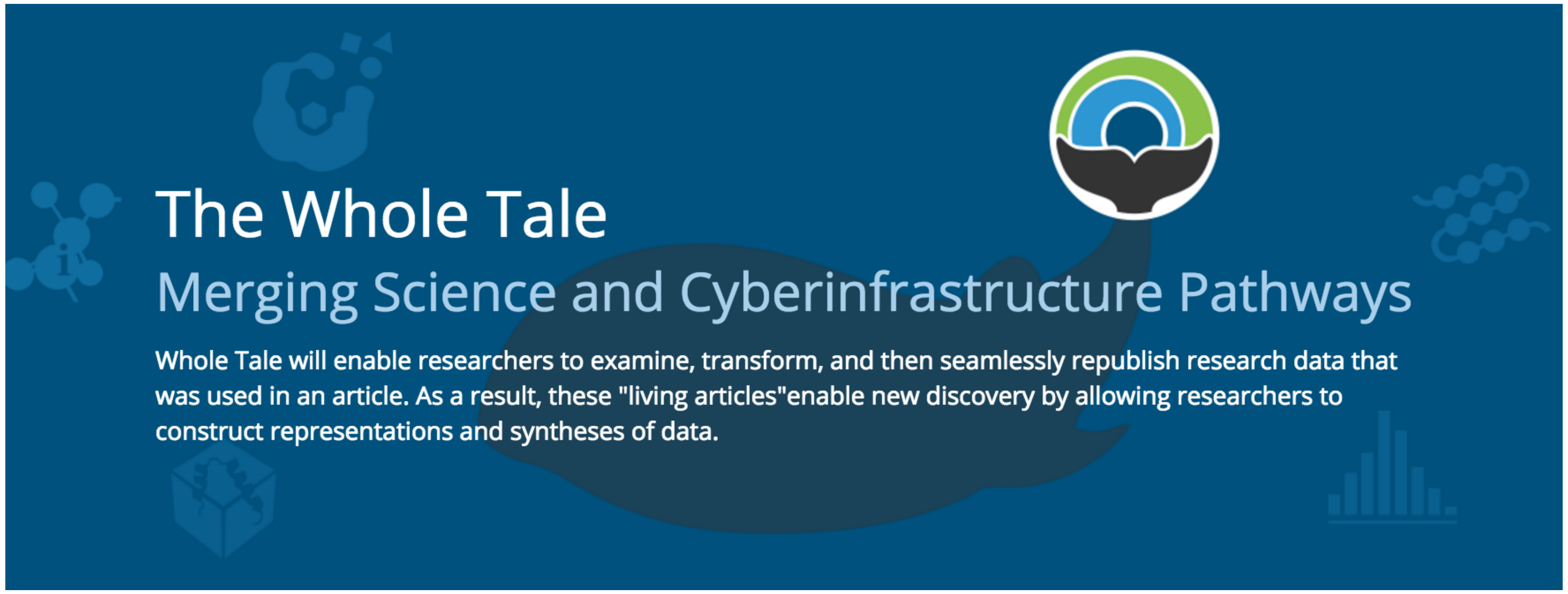
- Data comes from many sources and is **“integrated the old fashioned way”** (*email, Excel,* …)
- Use cloud services **copying data** from *(Drop)Box*, *Google-Drive*, … to local storage with a distributed directory structures to organize (and provide discovery) to data
- Data provenance is **not captured** (custom scripts, some version of a community developed and supported codebase)
- Publication of data with link to publication (never mind DOIs, DMP) is **not sufficient for reproducibility**

So what do we do about this?

- WT will leverage & contribute to **existing CI and tools** to support the **whole science story** (= run-to-pub-cycle), and providing access to big data via CI and compute for **long tail** researchers.

➡ ***Integrate** tools to **simplify usage** and promote **best practices***

- NSF CC*DNI DIBBS:
  - 5 Institutions, 5 Years ($5M total)
  - Cooperative Agreement



The Whole Tale

Merging Science and Cyberinfrastructure Pathways

Whole Tale will enable researchers to examine, transform, and then seamlessly republish research data that was used in an article. As a result, these "living articles"enable new discovery by allowing researchers to construct representations and syntheses of data.

Specific Goals of Whole Tale

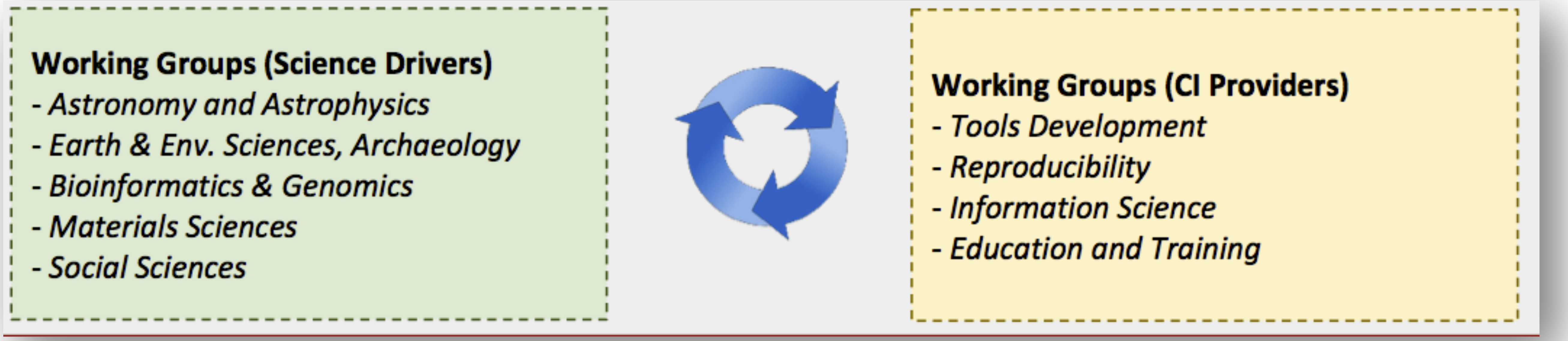
- **Expose existing CI**
  - … through popular frontends (Jupyter, RStudio, ..)

- **Develop necessary “software glue”**
  - … for seamless access to different CI-backend capabilities

- **Enhance data-to publication lifecycle**
  - … by empowering scientists to create computational narratives in their usual programming environments

Iterative Design through **Working Groups**

Merging Science & CI Pathways
… through Working Groups



Working Groups (Science Drivers)
- *Astronomy and Astrophysics*
- *Earth & Env. Sciences, Archaeology*
- *Bioinformatics & Genomics*
- *Materials Sciences*
- *Social Sciences*

Working Groups (CI Providers)
- *Tools Development*
- *Reproducibility*
- *Information Science*
- *Education and Training*

Working Groups Driving Use Cases and Adoption

Iterative Design

Working Groups to Provide Key Components

Try it!

http://wholetale.readthedocs.io/users_guide/index.html

Feedback is very welcome at feedback@wholetale.org and/or at https://github.com/whole-tale/whole-tale/issues

Conclusion

We see the convergence of two (ordinarily antagonistic) trends:

- ➡ Scientific projects will become massively more computing intensive
- ➡ Research computing will become dramatically more transparent

These are reinforcing trends, resolution essential for verifying and comparing findings.

AAAS / Arnold Foundation Reproducibility Workshop III: Code and Modeling

- *This workshop will consider ways to make code and modeling information more readily available, and include a variety of stakeholders.*

- *The computational steps that produce scientific findings are increasingly considered a crucial part of the scholarly record, permitting transparency, reproducibility, and re-use. Important information about data preparation and model implementation, such as parameter settings or the treatment of outliers and missing values, is often expressed only in code. Such decisions can have substantial impacts on research outcomes, yet such details are rarely available with scientific findings.*

- http://www.aaas.org/event/iii-arnold-workshop-modeling-and-code
  Feb 16-17, 2016

REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

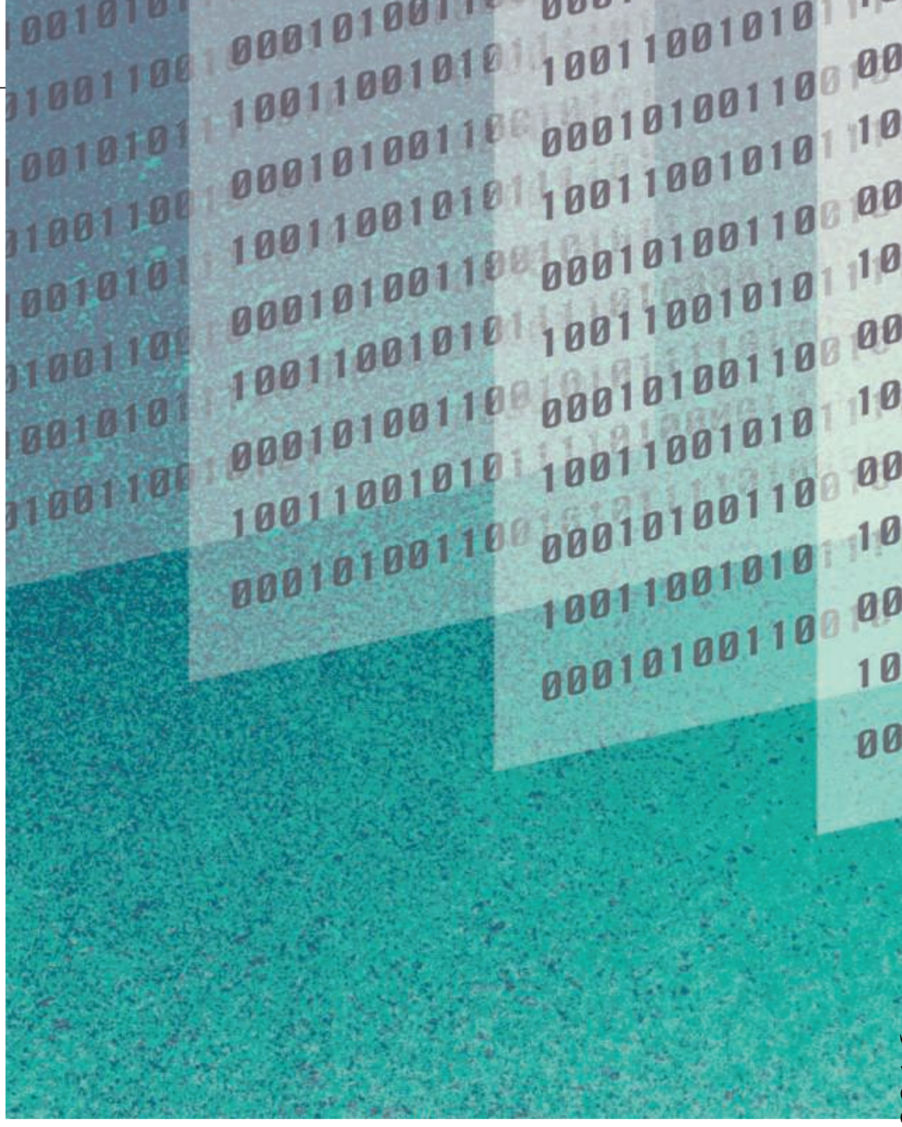
By **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[4] Brooks Hanson,[5] Michael A. Heroux,[6] John P.A. Ioannidis,[7] Michela Taufer[8]**

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general proposals from the Transparency and Openness Promotion (TOP) guidelines (*1*) and recommendations for field data (*2*), emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers repre-

to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable independent regeneration of computational results are the data, the computational steps that produced the findings, and the workflow describing how to generate the results using the data and code, including parameter settings, random number seeds, make files, or



Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fVwjPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly objects, citation should be standard practice. All data, code, and workflows, including software written by the authors, should be cited in the references section (*10*). We suggest that software citation include software version information and its unique identifier in addi-

Workshop Recommendations: “Reproducibility Enhancement Principles”

1. Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.

2. Persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3. To enable credit for shared digital scholarly objects, citation should be standard practice.

4. To facilitate reuse, adequately document digital scholarly artifacts.

Workshop Recommendations: “Reproducibility Enhancement Principles”

5. Use Open Licensing when publishing digital scholarly objects.

6. Journals should conduct a reproducibility check as part of the publication process and should enact the TOP standards at level 2 or 3.

7. To better enable reproducibility across the scientific enterprise, funding agencies should instigate new research programs and pilot studies.

Legal Issues in Software

Intellectual property is associated with software (and all digital scholarly objects) e.g the U.S. Constitution and subsequent Acts:

> "*To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries.*" (U.S. Const. art. I, §8, cl. 8)

Copyright

- Original expression of ideas falls under copyright by default (papers, code, figures, tables..)
- Copyright secures exclusive rights vested in the author to:
  - reproduce the work
  - prepare derivative works based upon the original
- limited time: generally life of the author +70 years
- Exceptions and Limitations: e.g. Fair Use.

Patents

Patentable subject matter: "*new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof*" (35 U.S.C. §101) that is

1. *Novel*, in at least one aspect,
2. *Non-obvious*,
3. *Useful*.

USPTO Final Computer Related Examination Guidelines (1996) "A practical application of a computer-related invention is statutory subject matter. This requirement can be discerned from the variously phrased prohibitions against the patenting of abstract ideas, laws of nature or natural phenomena" (see e.g. Bilski v. Kappos, 561 U.S. 593 (2010)).

Bayh-Dole Act (1980)

- Promote the transfer of academic discoveries for commercial development, via licensing of patents (ie. Technology Transfer Offices), and harmonize federal funding agency grant intellectual property regs.

- Bayh-Dole gave federal agency grantees and contractors title to government-funded inventions and charged them with using the patent system to aid disclosure and commercialization of the inventions.

- Hence, institutions such as universities charged with utilizing the patent system for technology transfer.

Legal Issues in Data

- In the US raw facts are not copyrightable, but the original “selection and arrangement” of these facts is copyrightable. (Feist Publns Inc. v. Rural Tel. Serv. Co., 499 U.S. 340 (1991)).

- Copyright adheres to raw facts in Europe.

- the possibility of a residual copyright in data (attribution licensing or public domain certification).

- Legal mismatch: What constitutes a “raw” fact anyway?

The Reproducible Research Standard

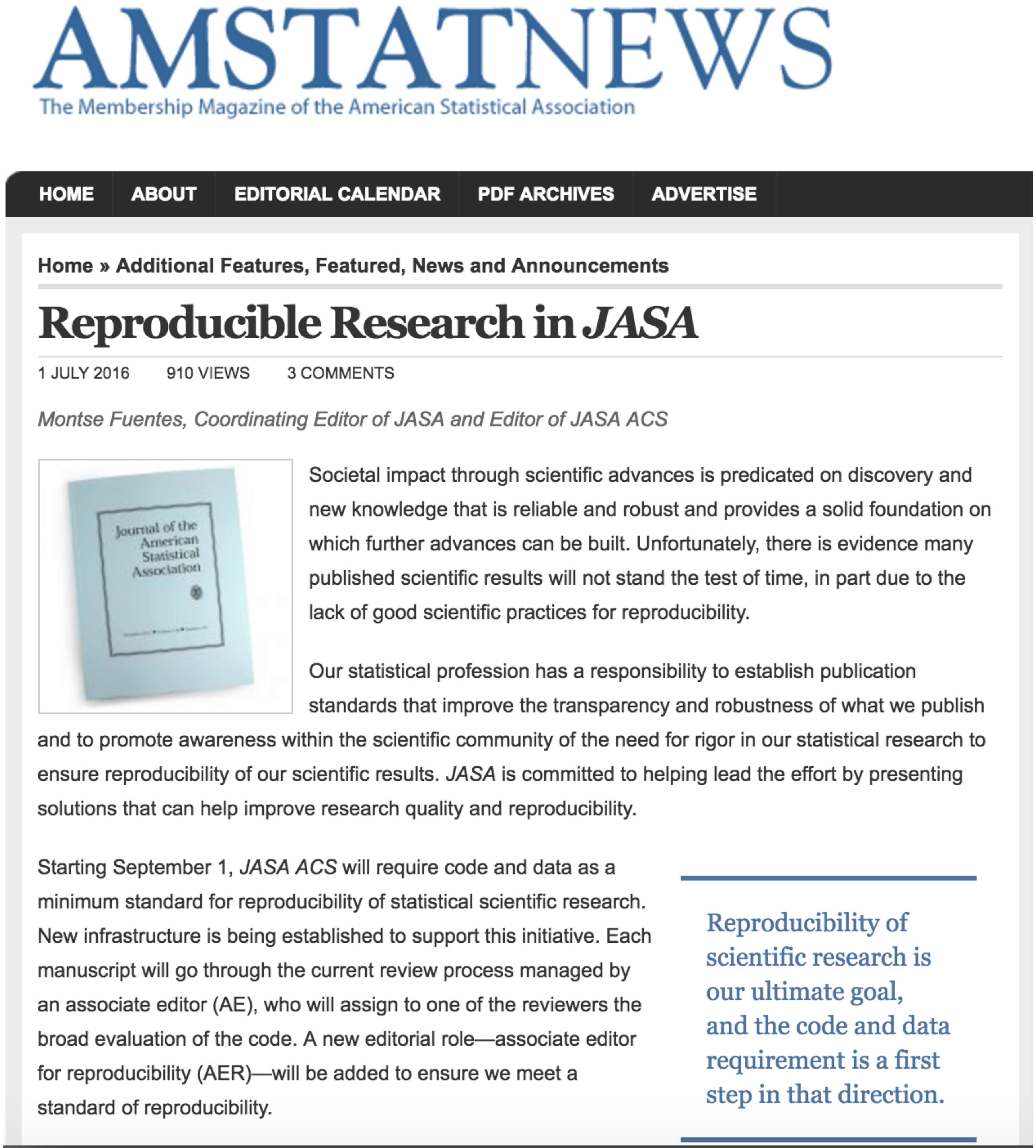
The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

➡ *Remove copyright's barrier to reproducible research and,*

➡ *Realign the IP framework with longstanding scientific norms.*



AMSTATNEWS

The Membership Magazine of the American Statistical Association

Home » Additional Features, Featured, News and Announcements

Reproducible Research in *JASA*

1 JULY 2016 910 VIEWS 3 COMMENTS

Montse Fuentes, Coordinating Editor of JASA and Editor of JASA ACS

Societal impact through scientific advances is predicated on discovery and new knowledge that is reliable and robust and provides a solid foundation on which further advances can be built. Unfortunately, there is evidence many published scientific results will not stand the test of time, in part due to the lack of good scientific practices for reproducibility.

Our statistical profession has a responsibility to establish publication standards that improve the transparency and robustness of what we publish and to promote awareness within the scientific community of the need for rigor in our statistical research to ensure reproducibility of our scientific results. *JASA* is committed to helping lead the effort by presenting solutions that can help improve research quality and reproducibility.

Starting September 1, *JASA ACS* will require code and data as a minimum standard for reproducibility of statistical scientific research. New infrastructure is being established to support this initiative. Each manuscript will go through the current review process managed by an associate editor (AE), who will assign to one of the reviewers the broad evaluation of the code. A new editorial role—associate editor for reproducibility (AER)—will be added to ensure we meet a standard of reproducibility.

Reproducibility of scientific research is our ultimate goal, and the code and data requirement is a first step in that direction.

Privacy and Data

- (U.S.) HIPAA, FERPA, Institutional Review Boards create legally binding restrictions on the sharing human subjects data (see e.g. http://www.dataprivacybook.org/)
- Potential privacy implications for industry generated data.
- Solutions: access restrictions, technological e.g. encryption, restricted querying, simulation..

Ownership: What Defines Contribution?

- Issue for producers: credit and citation.
- What is the role of peer-review?
- Repositories adding meta-data and discoverability make a contribution.
- Data repositories may be inadequate: velocity of contributions
- Future coders may contribute in part to new software, other software components may already be in the scholarly record. Attribution vs sharealike.

  ➡ (at least) 2 aspects: legal ownership vs scholarly credit.

- Redefining plagiarism for software contributions.

Licensing in Research
Background: Open Source Software

Innovation: Open Licensing

- ➡ Software with licenses that communicate alternative terms of use to code developers, rather than the copyright default.

Hundreds of open source software licenses:

- GNU Public License (GPL)
- (Modified) BSD License
- MIT License
- Apache 2.0 License
- ... see http://www.opensource.org/licenses/alphabetical

The Reproducible Research Standard

The *Reproducible Research Standard* (*RRS*) (Stodden, 2009)

A suite of license recommendations for computational science:

- Release media components (text, figures) under **CC BY**,
- Release code components under **MIT License** or similar,
- Release data to public domain (**CC0**) or attach attribution license.

  - ➡ Remove copyright's barrier to reproducible research and,
  - ➡ Realign the IP framework with longstanding scientific norms.

Computational Barriers

Barriers to Replication in Computational Science:

- rerunning same code, same parameter settings, same system can produce different results (?),
- same code (Reprozip, containerization/Docker), but updated libraries, compiler, operating system..
- software customization to underlying architectures; portability, modularity, re-usability,
- numerical stability of the underlying software architecture,
- unique hardware, scarce allocations, long runtimes..

Encouraging Reproducibility While Expanding Access to Massive Computation

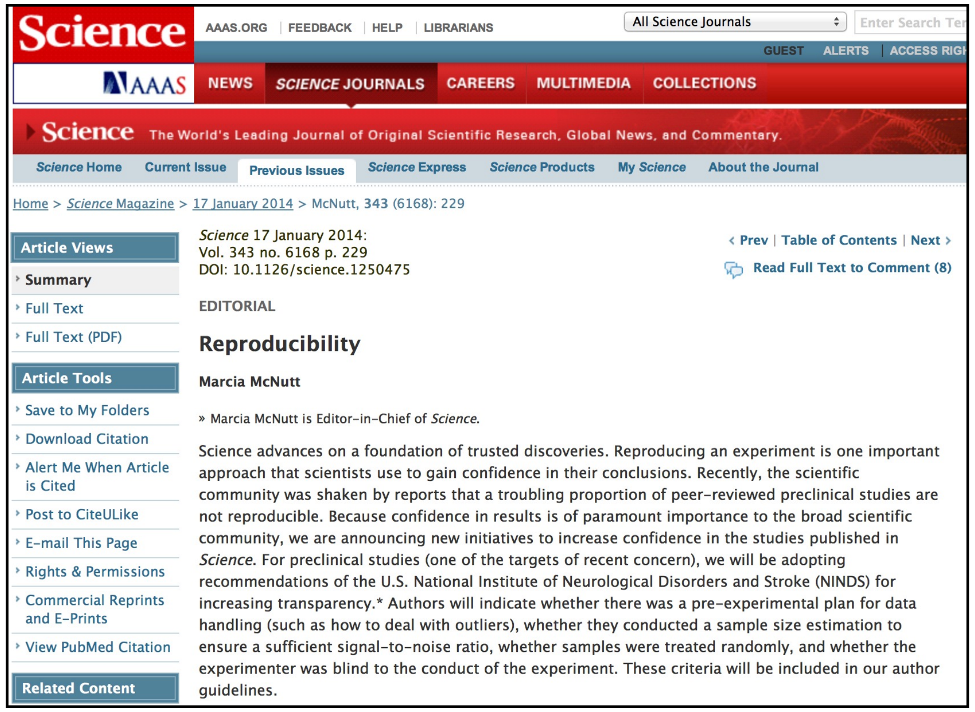
We are at the convergence of two (ordinarily antagonistic) trends:

1. Scientific projects will become massively more computing intensive,

2. Scientific computing dramatically more transparent.

These two trends can reinforce each other: better transparency will allow people to run much more ambitious computational experiments. *And* better computational experiment infrastructure will allow researchers to be more transparent.

A Credibility Crisis



Science | AAAS.ORG | FEEDBACK | HELP | LIBRARIANS | All Science Journals | Enter Search Te
GUEST | ALERTS | ACCESS RIGH
AAAS | NEWS | SCIENCE JOURNALS | CAREERS | MULTIMEDIA | COLLECTIONS

Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home | Current Issue | Previous Issues | Science Express | Science Products | My Science | About the Journal

Home > Science Magazine > 17 January 2014 > McNutt, 343 (6168): 229

Article Views
- Summary
- Full Text
- Full Text (PDF)

Article Tools
- Save to My Folders
- Download Citation
- Alert Me When Article is Cited
- Post to CiteULike
- E-mail This Page
- Rights & Permissions
- Commercial Reprints and E-Prints
- View PubMed Citation

Related Content

Science 17 January 2014:
Vol. 343 no. 6168 p. 229
DOI: 10.1126/science.1250475

‹ Prev | Table of Contents | Next ›
Read Full Text to Comment (8)

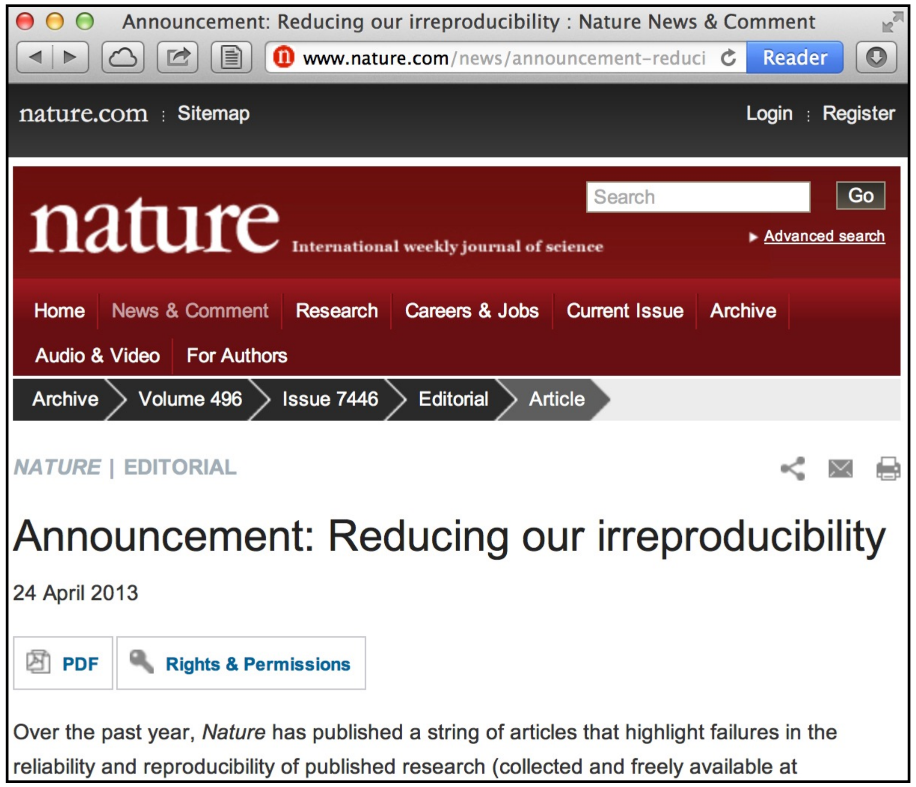
EDITORIAL

Reproducibility

Marcia McNutt

» Marcia McNutt is Editor-in-Chief of *Science*.

Science advances on a foundation of trusted discoveries. Reproducing an experiment is one important approach that scientists use to gain confidence in their conclusions. Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible. Because confidence in results is of paramount importance to the broad scientific community, we are announcing new initiatives to increase confidence in the studies published in *Science*. For preclinical studies (one of the targets of recent concern), we will be adopting recommendations of the U.S. National Institute of Neurological Disorders and Stroke (NINDS) for increasing transparency.* Authors will indicate whether there was a pre-experimental plan for data handling (such as how to deal with outliers), whether they conducted a sample size estimation to ensure a sufficient signal-to-noise ratio, whether samples were treated randomly, and whether the experimenter was blind to the conduct of the experiment. These criteria will be included in our author guidelines.



Announcement: Reducing our irreproducibility : Nature News & Comment
www.nature.com/news/announcement-reduci | Reader

nature.com | Sitemap | Login | Register

nature International weekly journal of science | Search | Go | Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 496 > Issue 7446 > Editorial > Article

NATURE | EDITORIAL

Announcement: Reducing our irreproducibility

24 April 2013

PDF | Rights & Permissions

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at

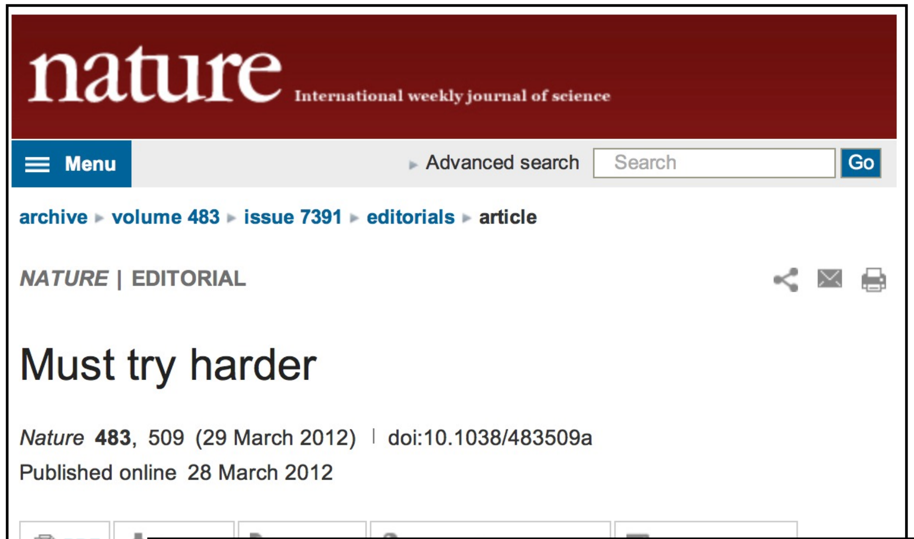
The Economist

OCTOBER 19TH–25TH 2013 Economist.com

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG

99
Einsteinium



nature International weekly journal of science

Menu | Advanced search | Search | Go

archive ▸ volume 483 ▸ issue 7391 ▸ editorials ▸ article

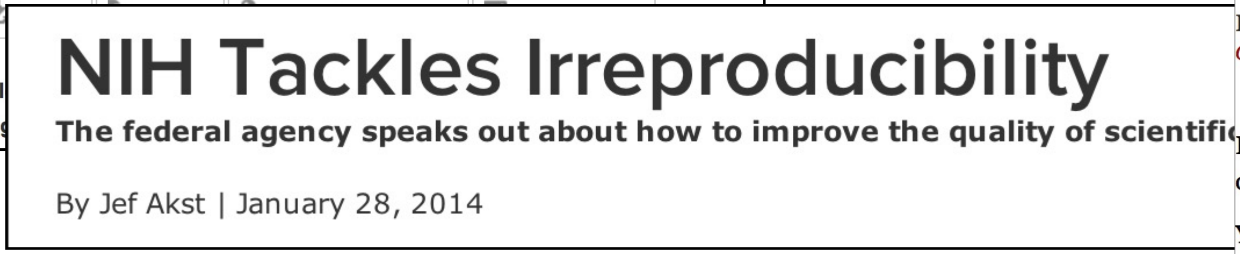
NATURE | EDITORIAL

Must try harder

Nature **483**, 509 (29 March 2012) | doi:10.1038/483509a
Published online 28 March 2012

PDF

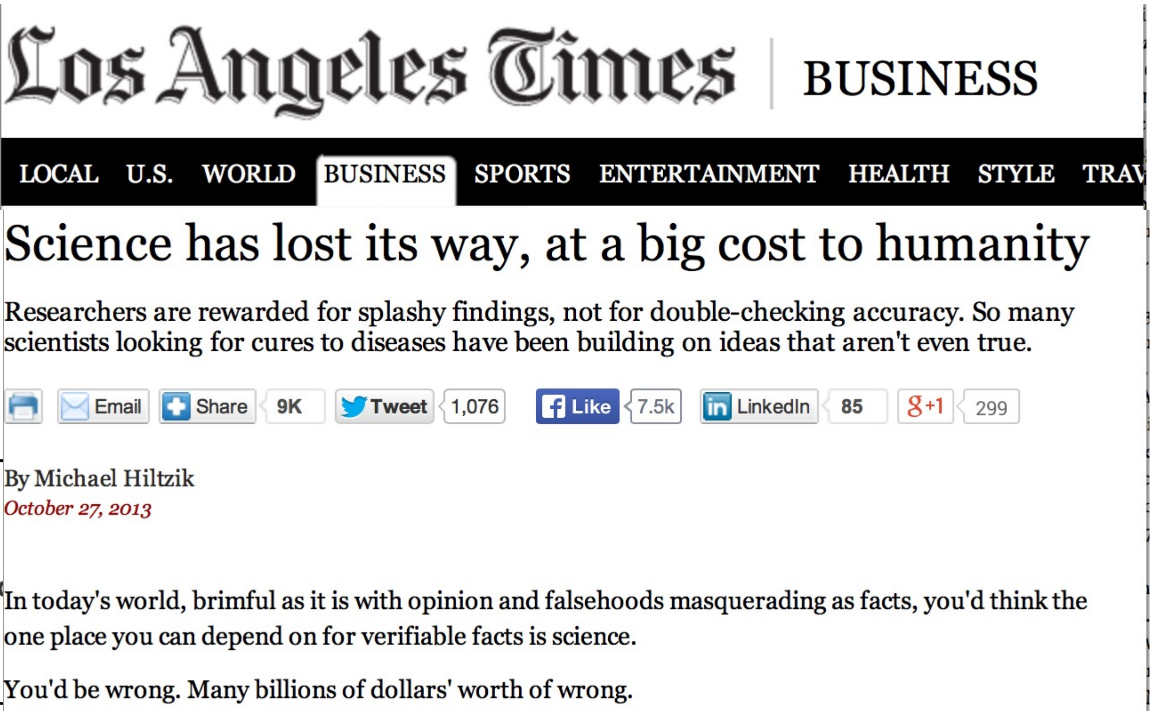
Too many sl
look more ri



NIH Tackles Irreproducibility

The federal agency speaks out about how to improve the quality of scientific

By Jef Akst | January 28, 2014



Los Angeles Times | BUSINESS

LOCAL | U.S. | WORLD | BUSINESS | SPORTS | ENTERTAINMENT | HEALTH | STYLE | TRAV

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true.

Email | Share 9K | Tweet 1,076 | Like 7.5k | LinkedIn 85 | +1 299

By Michael Hiltzik
October 27, 2013

In today's world, brimful as it is with opinion and falsehoods masquerading as facts, you'd think the one place you can depend on for verifiable facts is science.

You'd be wrong. Many billions of dollars' worth of wrong.