Research Transparency and Reproducibility Training (RT2)

Session:
Reproducibility and Replicability

@ucbitss #RT2DC www.bitss.org

WORLD BANK GROUP

BITSS

Research Transparency and Reproducibility Training (RT2)

Reproducibility and Replicability

The NASEM report and the praxis of reproducible research

🐦 @LorenaABarba   http://lorenabarba.com

About me

Lorena A. Barba group

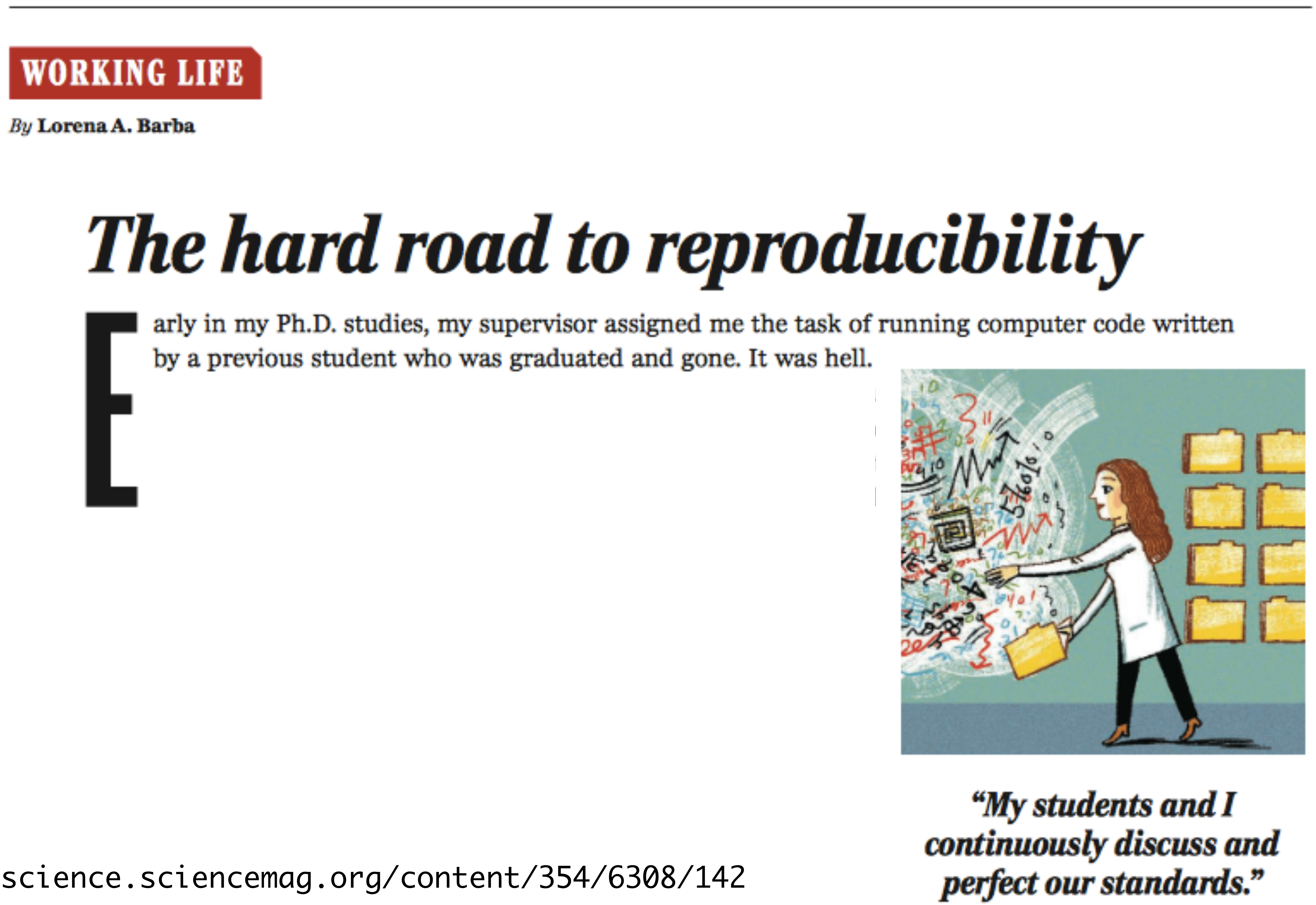
Reproducibility PI Manifesto

- Reproducibility PI Manifesto *figshare*, 2012
- "The hard road to reproducibility" *Science*, Oct. 2016
- "Repro Packs," *Nature blogs*, Apr. 2017
- CiSE editor for Reproducible Research
- SC19 Reproducibility Chair
- NASEM Committee member

http://lorenabarba.com

Reproducibility PI Manifesto (2012)

- 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



WORKING LIFE

By Lorena A. Barba

The hard road to reproducibility

Early in my Ph.D. studies, my supervisor assigned me the task of running computer code written by a previous student who was graduated and gone. It was hell.

"My students and I continuously discuss and perfect our standards."

http://science.sciencemag.org/content/354/6308/142 nature.com

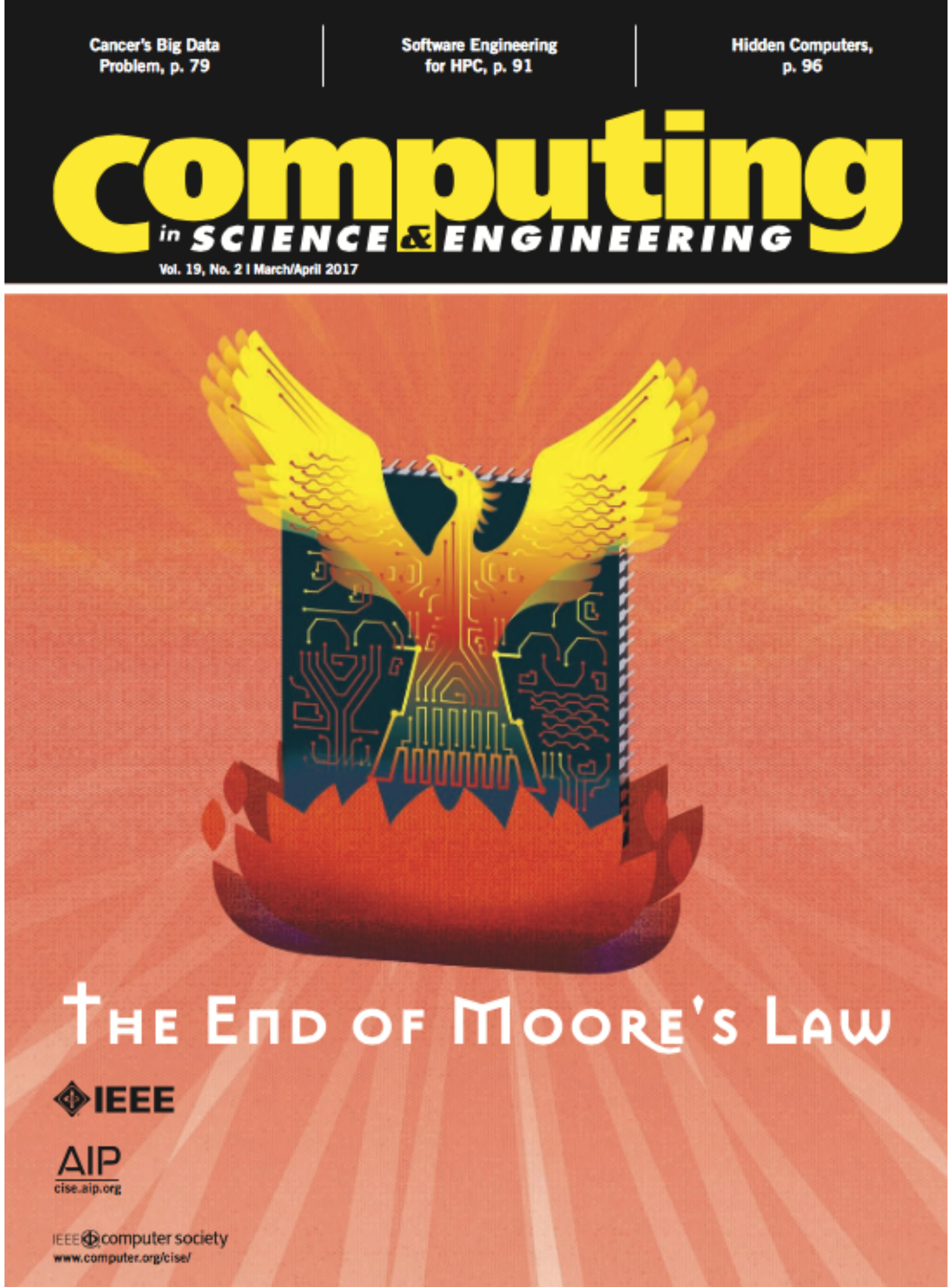
NATUREJOBS | NATUREJOBS BLOG

TechBlog: My digital toolbox: Lorena Barba

17 Apr 2017 | 12:00 BST | Posted by Jeffrey Perkel | Category: Blog, Technology

Repro-Packs: our signature open-science practice

http://blogs.nature.com/naturejobs/2017/04/17/techblog-my-digital-toolbox-lorena-barba/



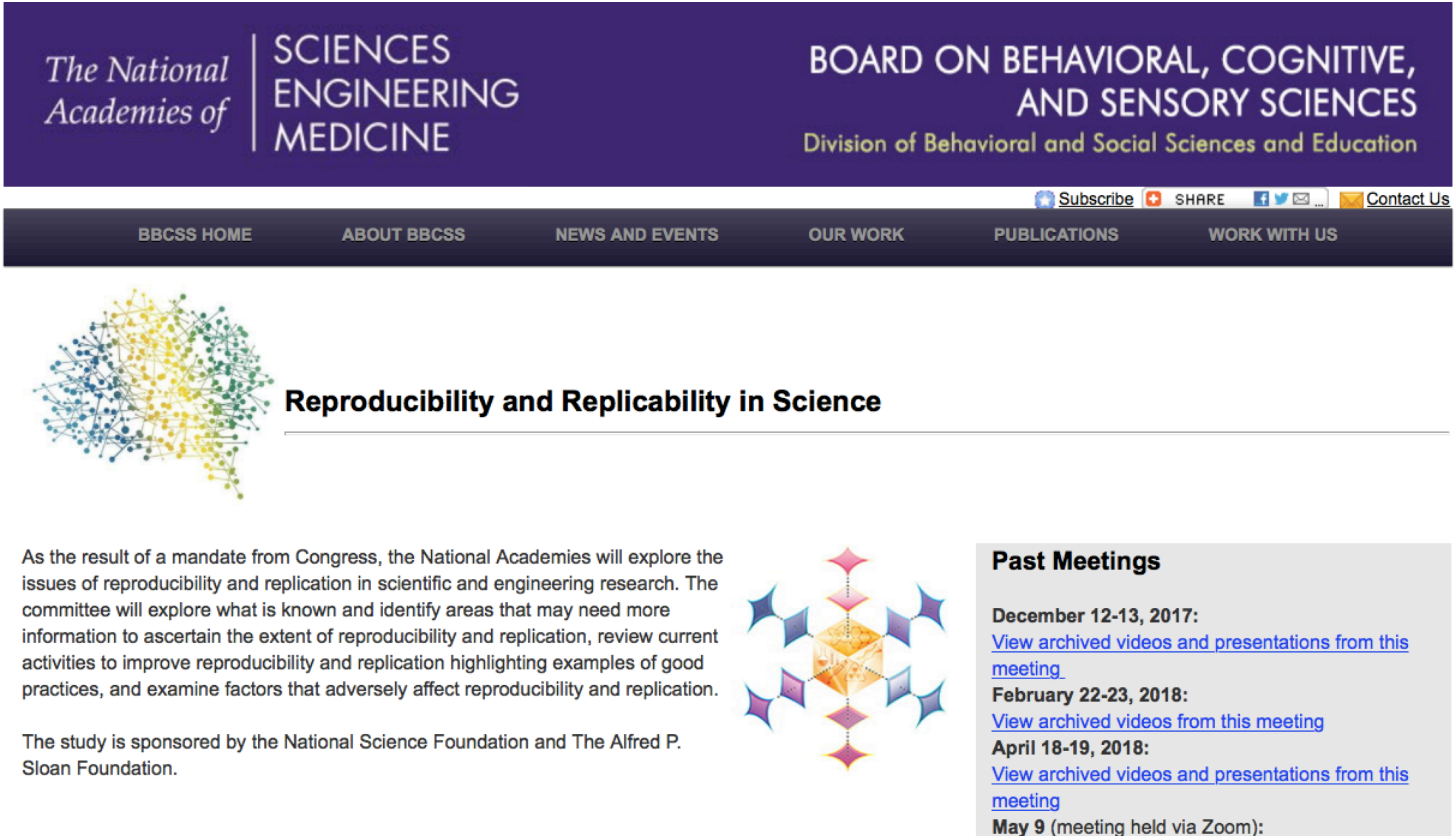
Cancer's Big Data Problem, p. 79
Software Engineering for HPC, p. 91
Hidden Computers, p. 96
computing in SCIENCE & ENGINEERING
Vol. 19, No. 2 I March/April 2017
The End of Moore's Law
IEEE
AIP
cise.aip.org
IEEE computer society
www.computer.org/cise/

Reproducible Research Track (peer reviewed)

Lorena A. Barba
George Washington University
labarba@gwu.edu

George K. Thiruvathukal
Loyola University Chicago
gkt@cs.luc.edu

`https://www.computer.org/cise/`


The National Academies of
SCIENCES
ENGINEERING
MEDICINE
BOARD ON BEHAVIORAL, COGNITIVE, AND SENSORY SCIENCES
Division of Behavioral and Social Sciences and Education
Subscribe
SHARE
Contact Us
BBCSS HOME
ABOUT BBCSS
NEWS AND EVENTS
OUR WORK
PUBLICATIONS
WORK WITH US

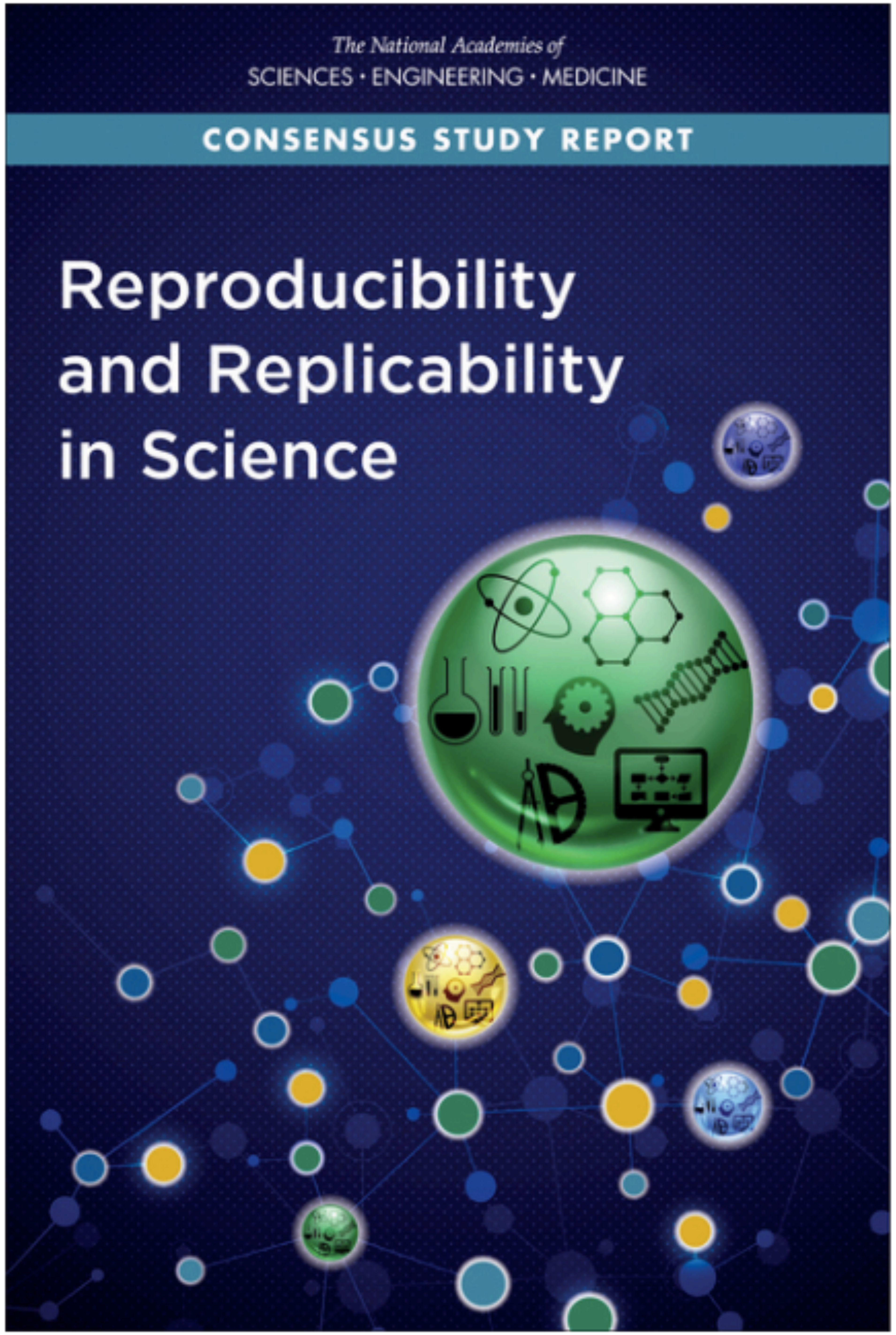
Reproducibility and Replicability in Science

As the result of a mandate from Congress, the National Academies will explore the issues of reproducibility and replication in scientific and engineering research. The committee will explore what is known and identify areas that may need more information to ascertain the extent of reproducibility and replication, review current activities to improve reproducibility and replication highlighting examples of good practices, and examine factors that adversely affect reproducibility and replication.

The study is sponsored by the National Science Foundation and The Alfred P. Sloan Foundation.

Past Meetings

December 12-13, 2017:
View archived videos and presentations from this meeting
February 22-23, 2018:
View archived videos from this meeting
April 18-19, 2018:
View archived videos and presentations from this meeting
May 9 (meeting held via Zoom):



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility
and Replicability
in Science

- Study mandated by public law 114-329 (Jan. 2017)
- commissioned by the National Science Foundation (NSF) to The National Academies of Sciences, Engineering and Medicine (NASEM)
- 15 experts convened
- 18 months of in-person meetings, teleconferences, commissioned papers, deliberations, writing
- report released 7 May 2019

http://doi.org/c5jp

Defining Reproducibility & Replicability

Def.— Reproducibility

obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis

Def.— Replicability

obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data

Data Replication & Reproducibility

PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

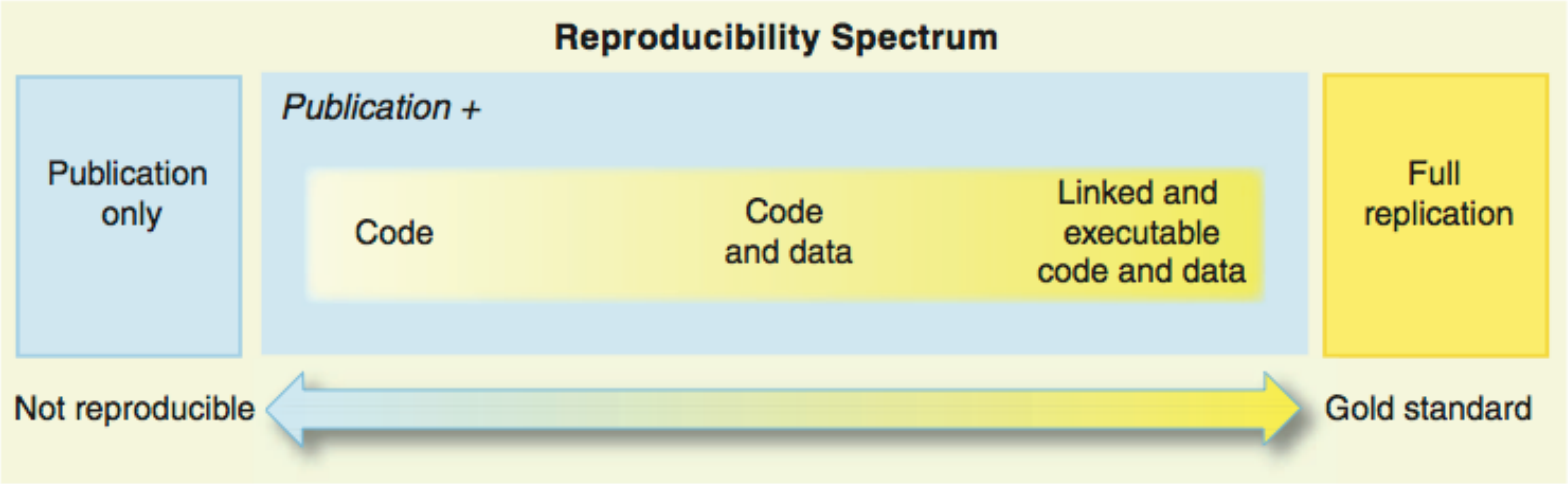


Fig. 1. The spectrum of reproducibility.

Reproducible Research

Widespread use of computation & data in science

- As important as the telescopes were the software libraries and data products needed to create the first image of a black hole

(now iconic photo of Dr. Katie Bouman)

- 92% of academics use research software
- 69% say that their research would not be practical without it
- 56% develop their own software
- 21% of those have no training in software development

S.J. Hettrick et al. (2014), *UK Research Software Survey* doi:10.5281/zenodo.14809

> “reproducibility . . . requires having the complete software environment [...] and the full source code available for inspection, modification, and application under varied parameter settings.”
>
> —Buckheit and Donoho (1995)

RECOMMENDATION 4-1: To help ensure the reproducibility of computational results, researchers should convey clear, specific, and **complete information about any computational methods and data products that support their published results** in order to enable other researchers to repeat the analysis, unless such information is restricted by non-public data policies. That information should include the data, study methods, and computational environment.

RECOMMENDATION 4-2: The National Science Foundation should consider investing in research that **explores the limits of computational reproducibility** in instances in which bitwise reproducibility is not reasonable in order to ensure that the meaning of consistent computational results remains in step with the development of new computational hardware, tools, and methods.

Sources of non-reproducibility

- Inadequate record keeping
- Nontransparent reporting
- Obsolescence of the digital artifacts
- Flawed attempts to reproduce other’s results
- Barriers in culture

Improving reproducibility

- Automatic capture of computational details; workflow management systems
- Source code and data version control
- Tools for reproducing results via virtualization, cloud computing, packaging, containers (e.g., Docker, Singularity)
- Interactive computational notebooks (e.g., Jupyter)

RECOMMENDATION 6-3: Funding agencies and organizations should consider investing in research and development of **open-source, usable tools and infrastructure that support reproducibility** for a broad range of studies across different domains in a seamless fashion. Concurrently, investments would be helpful in outreach to inform and train researchers on best practices and how to use these tools.

Step 1: Publish the software

- ***“We’re not a discipline, until we value software”***

  —L. Barba at 2015 SIAM Conference on Computational Science and Engineering (CSE) panel “The Future of CSE as a Discipline”

The Journal of Open Source Software
About
Papers
Docs
Blog
Submit
Log in with ORCID

The Journal of Open Source Software is a **developer friendly**, open access journal for research software packages.

Committed to publishing quality research software with zero article processing charges or subscription fees.

Submit a paper to JOSS
🎉 Volunteer to review
Explore Papers
Documentation
Learn More

Recently Published Papers 690

PUBLISHED Published 2 days ago

GoFish: Fishing for Line Observations in Protoplanetary Disks

Python

@richteague

DOI 10.21105/joss.01632

The Journal of Open Source Software
About
Papers
Docs
Blog
Submit
Log in with ORCID
Open Source Initiative
A Fiscally Sponsored Project of
NUMFOCUS
OPEN CODE = BETTER SCIENCE
creative commons

About the Journal of Open Source Software

The Journal of Open Source Software is a **developer friendly**, open access journal for research software packages.

Journal of Open Source Software is an affiliate of the Open Source Inititative.

Journal of Open Source Software is part of Open Journals, which is a NumFOCUS-sponsored project.

ISSN 2475-9066

http://joss.theoj.org

JOSS infrastructure

- GitHub—open-source software hosting & collaboration
- Zenodo—data repository by CERN
- ORCID—author identification
- CrossRef—DOI minting
- custom web app and Ruby bot

JOSS

Open-Source Software (OSS)

Reproducible research is vitally connected to open-source software, open data and open science.

Be aware:

… just because the source code is available on a website, doesn’t mean it is open source!

open source
initiative

Standard public licenses

It's not sufficient to make the source public to read. We must attach a **license** that allows others to modify and distribute the code.

Open-source licenses:

Anyone developing software in an academic setting should have working knowledge of software licenses.

OPEN ACCESS Freely available online

Education

A Quick Guide to Software Licensing for the Scientist-Programmer

Andrew Morin[1], **Jennifer Urban**[2], **Piotr Sliz**[1]*

1 Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, United States of America, **2** Samuelson Law, Technology & Public Policy Clinic, School of Law, University of California Berkeley, Berkeley, California, United States of America

Permissive vs. copy-left?

Permissive licenses

- Fewest restrictions
- Allow use, distribution, modification
- Only require giving credit to code authors
- Best choice for academic use!
- e.g., Berkeley Software Distribution (BSD), MIT License, Apache License

Copy-left licenses

- Guarantees perpetual access to the source code
- Requires any derivative work be under the same license
- a.k.a. “share-alike” licenses
- Are considered restrictive
- e.g., GPL license

How to choose?

For academic work: simple & permissive is best.

—BSD3 for code; CC-BY for content

http://choosealicense.com/

Step 2: Reproducible workflows

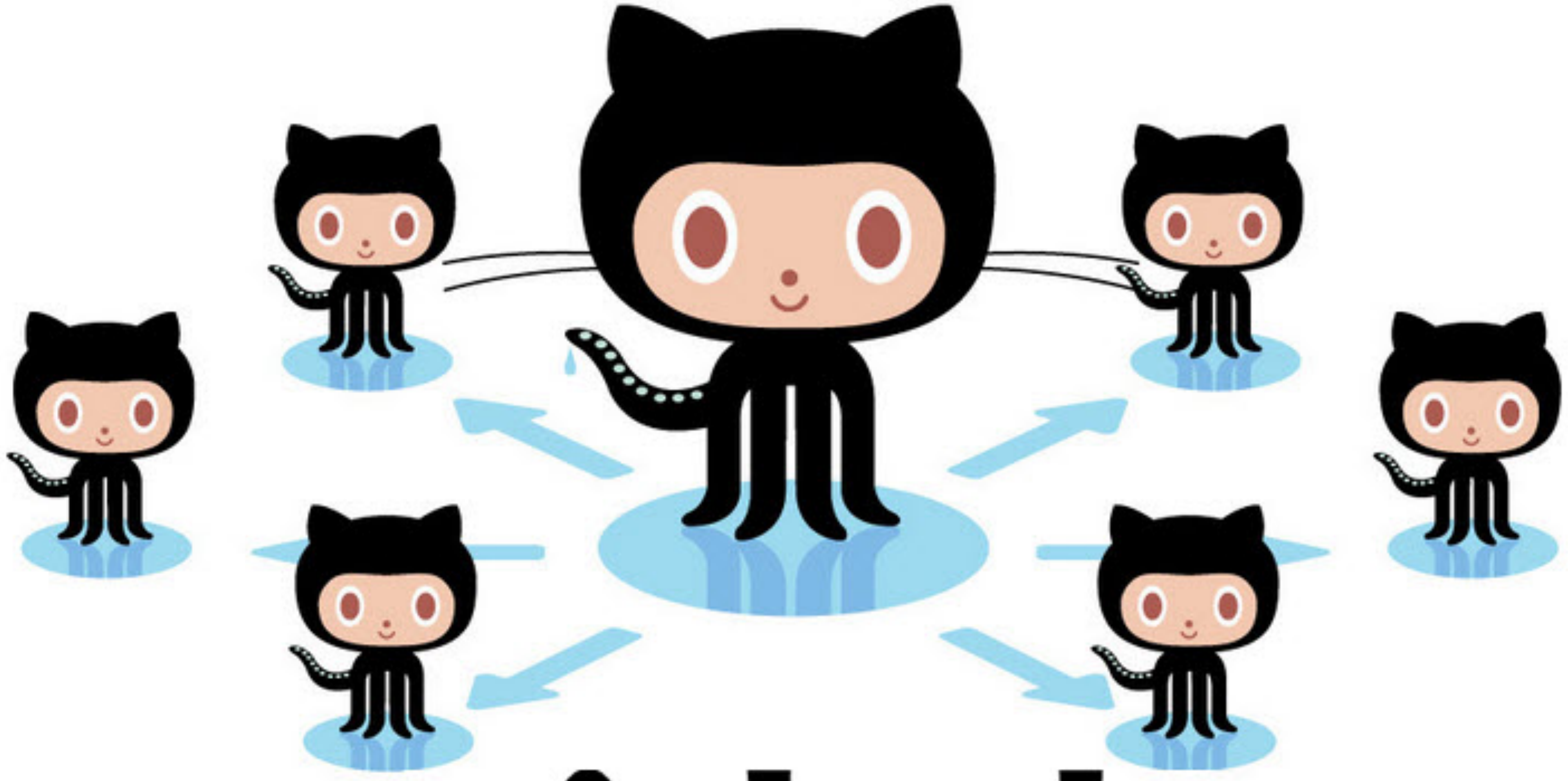
- Version control
- Script, automate, document
- Avoid GUIs for manipulating figures

Version Control

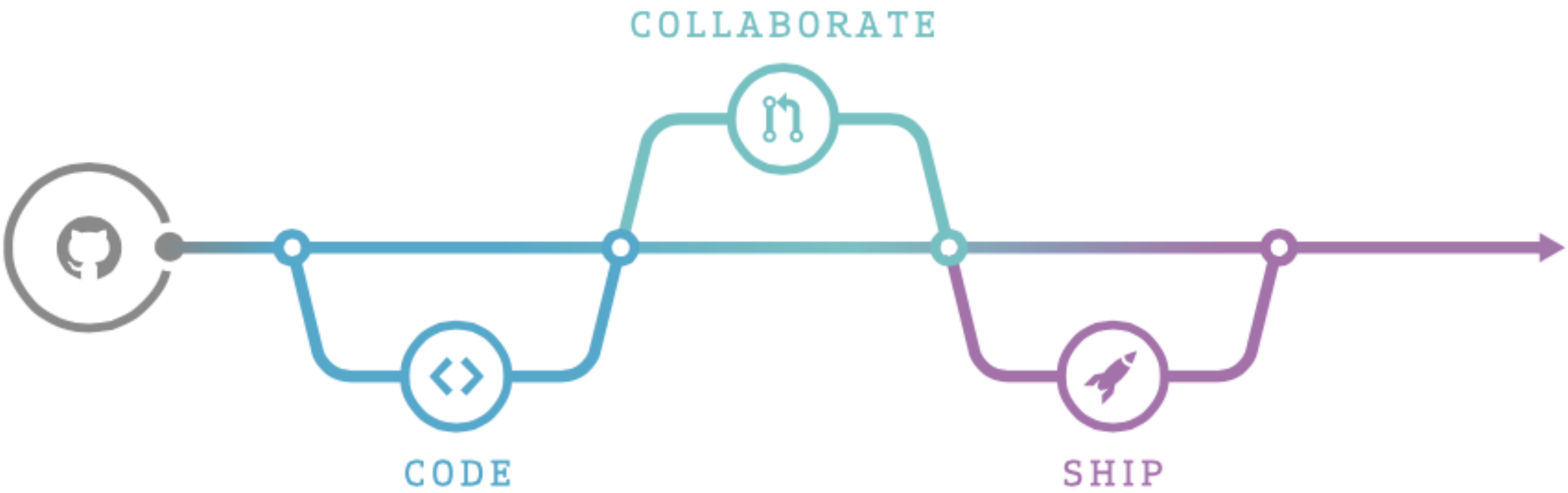
A system to automatically record and manage changing versions of source code stored as text

We use version control:

- internal reports on Markdown or Jupyter
- manuscripts in LaTeX



github
SOCIAL CODING



GitHub is a tool-of-the-trade in the open-source world that supports its workflow, and promotes a culture of collaboration.

Open source as a development model

Linus’s Law — “Given enough eyeballs, all bugs are shallow.”

jupyter

A set of open-source tools for interactive and exploratory computing.

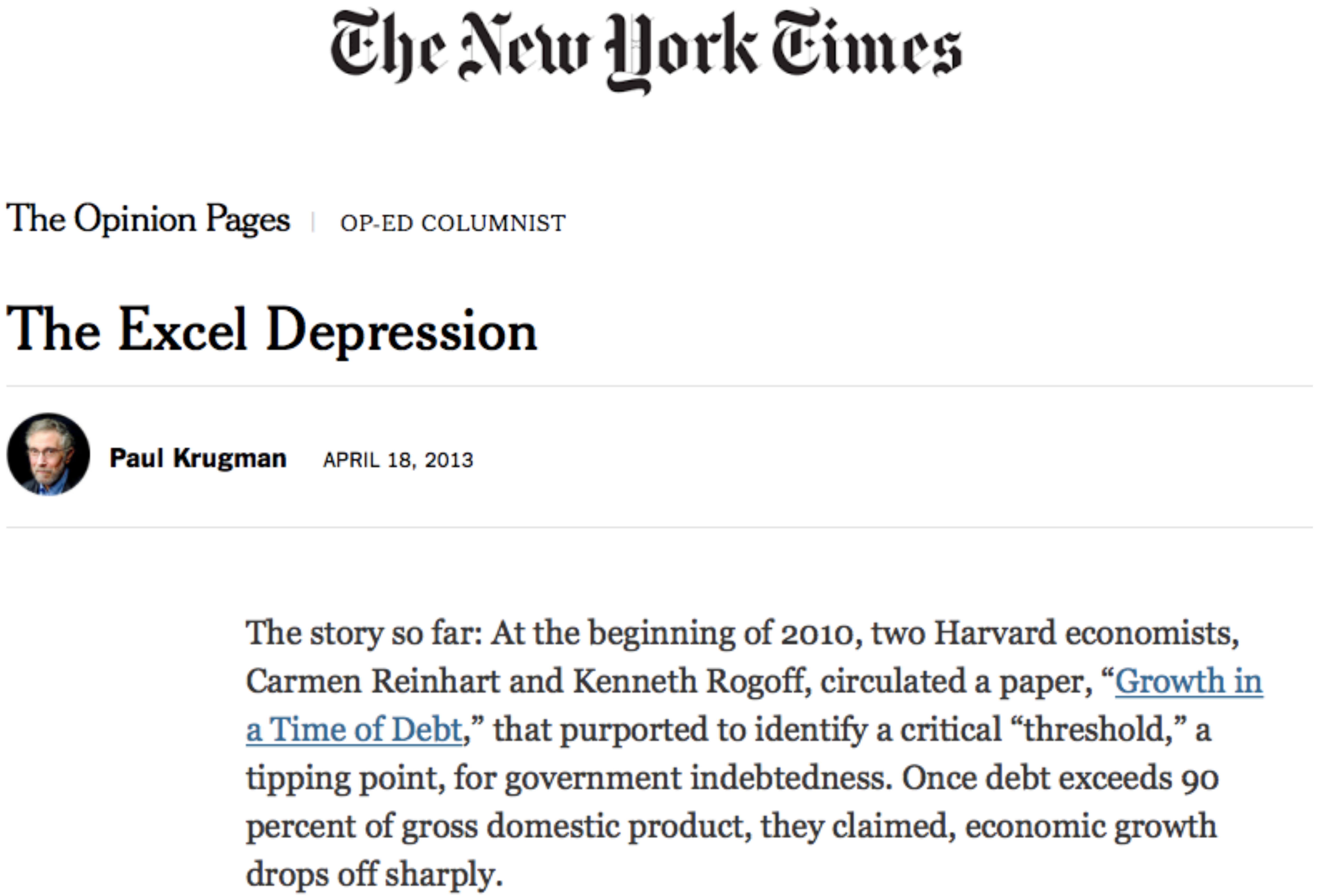
Jupyter grant proposal:

"…the core problem we are trying to solve is the collaborative creation of **reproducible** computational narratives."

Interactive →← Reproducible

> “I’ve learned that interactive programs are [tyranny] (unless they include the ability to arrive in any previous state by means of a script).”
>
> — Jon Claerbout

Interactive →← Reproducible



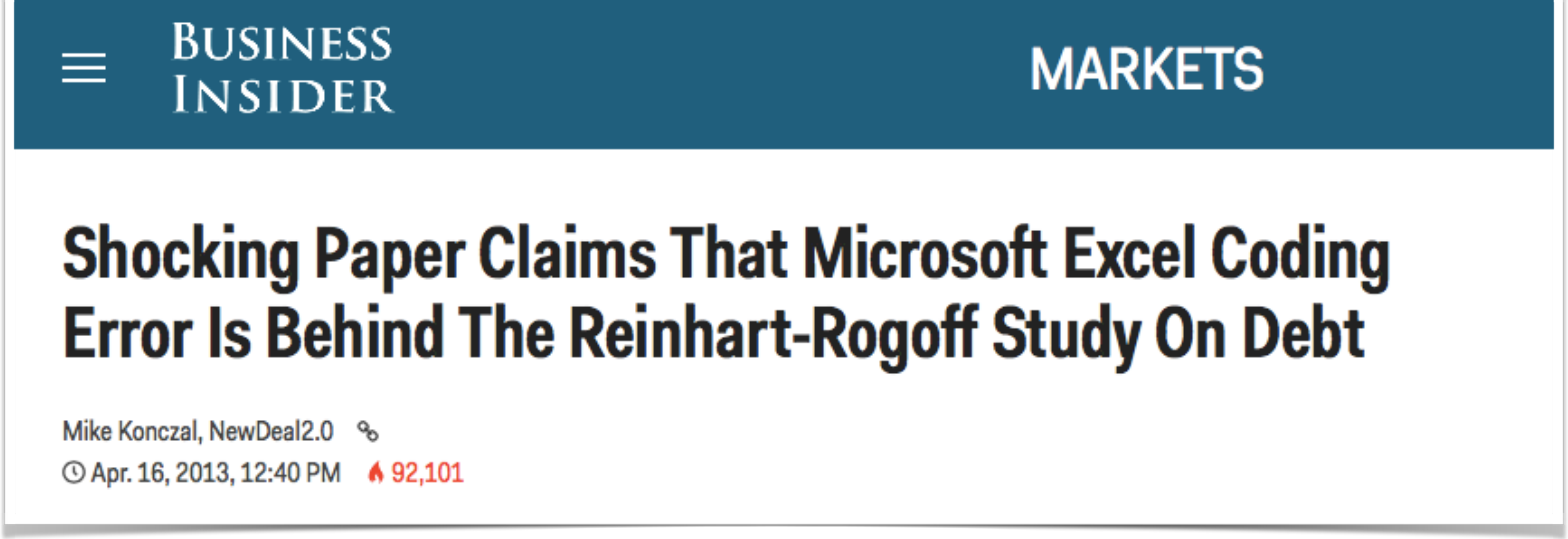
The New York Times

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.



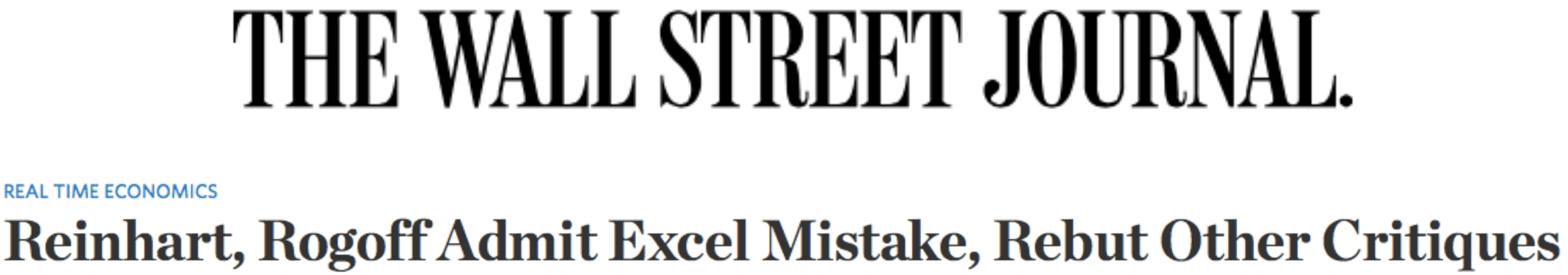
BUSINESS INSIDER

MARKETS

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

Mike Konczal, NewDeal2.0

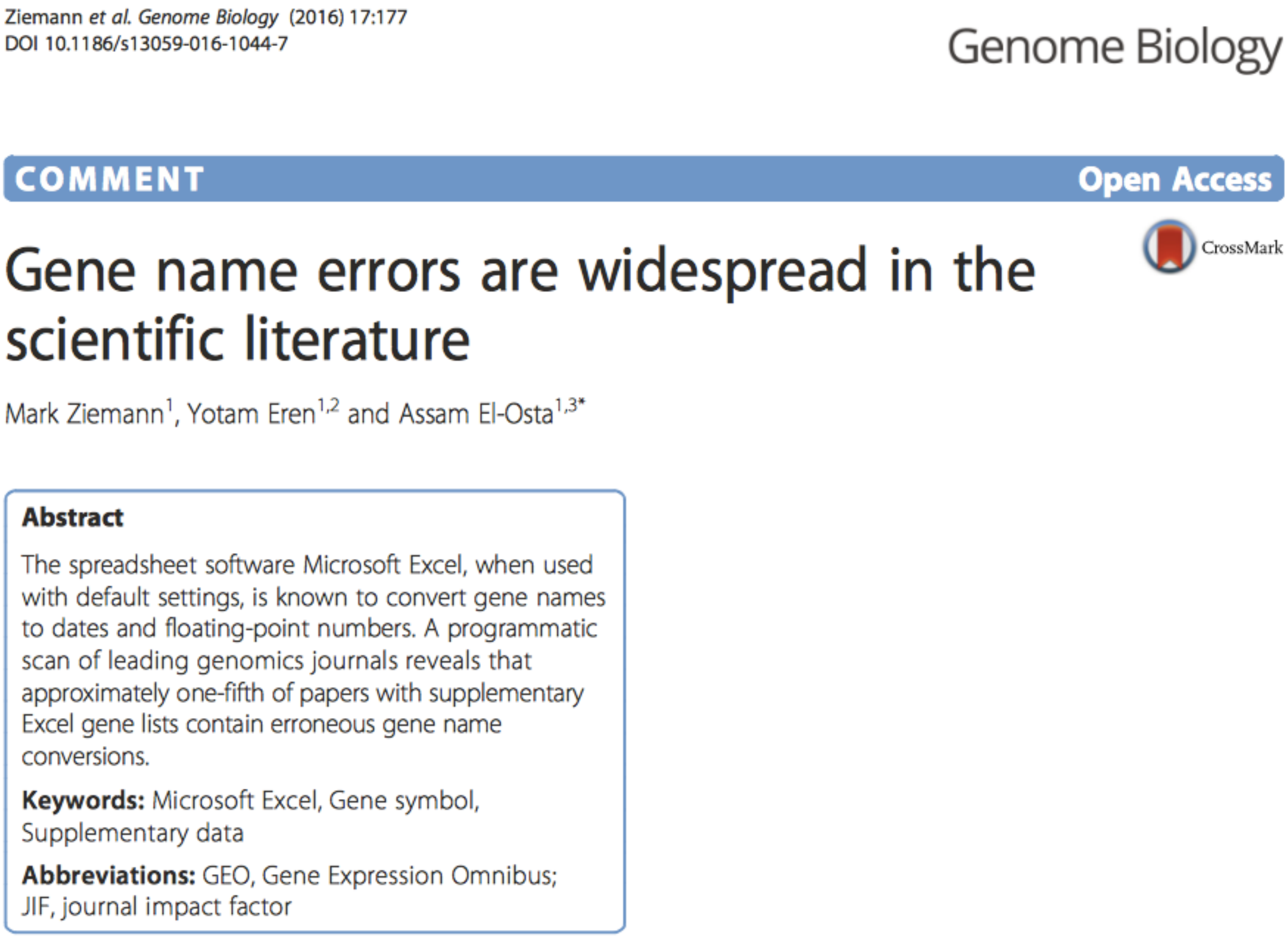
Apr. 16, 2013, 12:40 PM 🔥 92,101



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

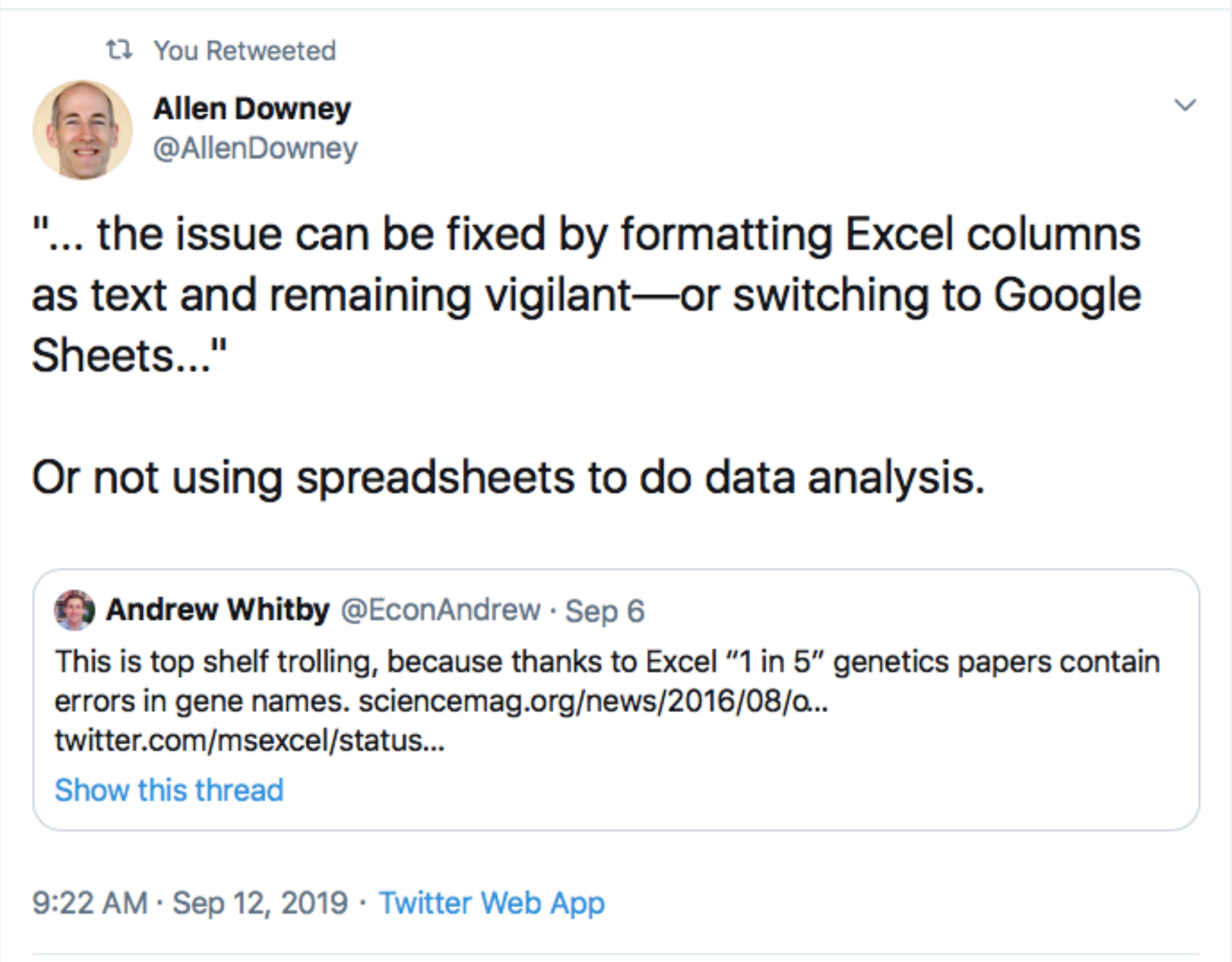
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

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



You Retweeted

Allen Downey
@AllenDowney

"... the issue can be fixed by formatting Excel columns as text and remaining vigilant—or switching to Google Sheets..."

Or not using spreadsheets to do data analysis.

Andrew Whitby @EconAndrew · Sep 6

This is top shelf trolling, because thanks to Excel “1 in 5” genetics papers contain errors in gene names. sciencemag.org/news/2016/08/o... twitter.com/msexcel/status...

Show this thread

9:22 AM · Sep 12, 2019 · Twitter Web App

Following
Tweet your reply

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

1:14 AM - 11 Aug 2014

41 Retweets 38 Likes

4 41 38

Philip Stark @philipbstark · 11 Aug 2014
Replying to @philipbstark
2\

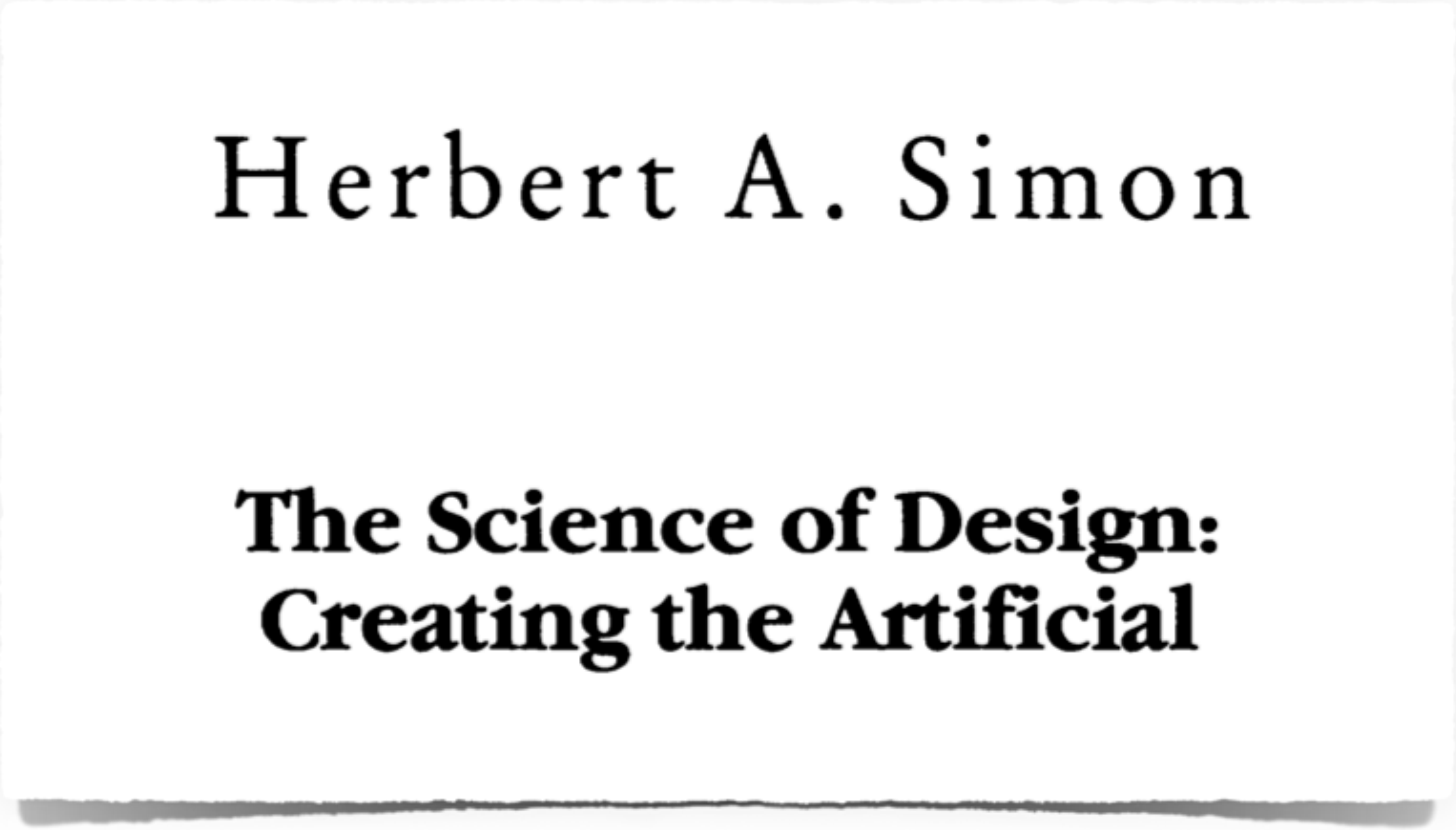
Philip Stark
@philipbstark Follows you

On spreadsheets:

"...the user interface conflates input, output, code, and presentation, making testing code and discovering bugs difficult."

— Philip Stark, *Science is 'show me,' not 'trust me'* (2015)

How do we design our tools for reproducibility?



Herbert A. Simon

The Science of Design: Creating the Artificial

“Designing the User Interface”

—Ben Shneiderman, 6th ed.

Tools that succeed are:

- comprehensible,
- predictable, and
- controllable

Those who have authority and responsibility must have adequate levels of control.

Responsibility should guide design.

Human control ↓↑ Automation

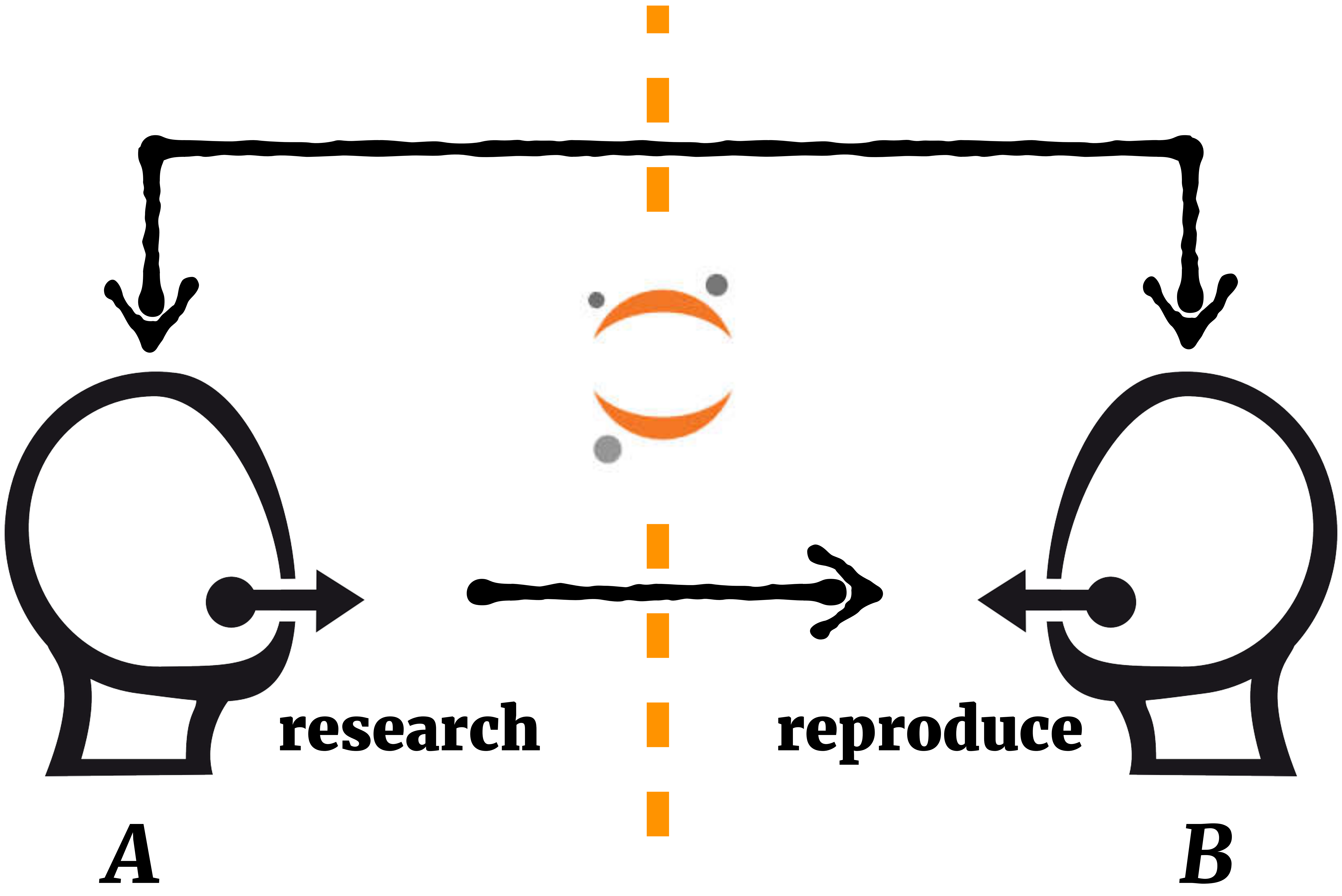
“Ensuring human control while increasing automation.”

On 21st-century design:

“…design has expanded from giving form to creating systems that support human interactions.”

— Hugh Dubberly & Paul Pangaro,
Cybernetics and Design: Conversations for action (2015)

Conversation builds *trust*



“***I have a button here. I push the button.*** **That’s not a conversation.**”

— Paul Pangaro,
Rethinking Design Thinking, PICNIC Festival Amsterdam (2010)

Reproducibility:
not a one-click solution

Step 3: Open data / Open science

- Archive interim data products (e.g., meshes)
- Share input files, configuration, parameter lists, runtime options
- Archive secondary data, figures, and plotting scripts (“repro-packs”)

Good data management

FAIR Principles: digital artifacts of research should be Findable, Accessible, Interoperable and Reusable for machines and for people

—Wilkinson et al., 2016.

Data repositories

- must provide a unique global identifier for your data (typically a digital object identifier, DOI)
- must offer long-term preservation guarantees (at least 10 years)

Free data repositories:

figshare

zenodo

DRYAD

Open Science Framework

The Dataverse Project

figshare

- general-purpose repository for all kinds of digital artifacts of research
- any file format, up to 5GB in size
- free and unlimited for public items

zenodo

- created by CERN and OpenAIRE
- free and non-commercial
- log in with your ORCID
- deposit large files: up to 50GB by default

`https://zenodo.org/communities/barbagroup/`

Open-access publishing

- Yale Law School Roundtable on Data and Code Sharing (2009) recommended publishing under open-access conditions (or post pre-prints).

Preprints

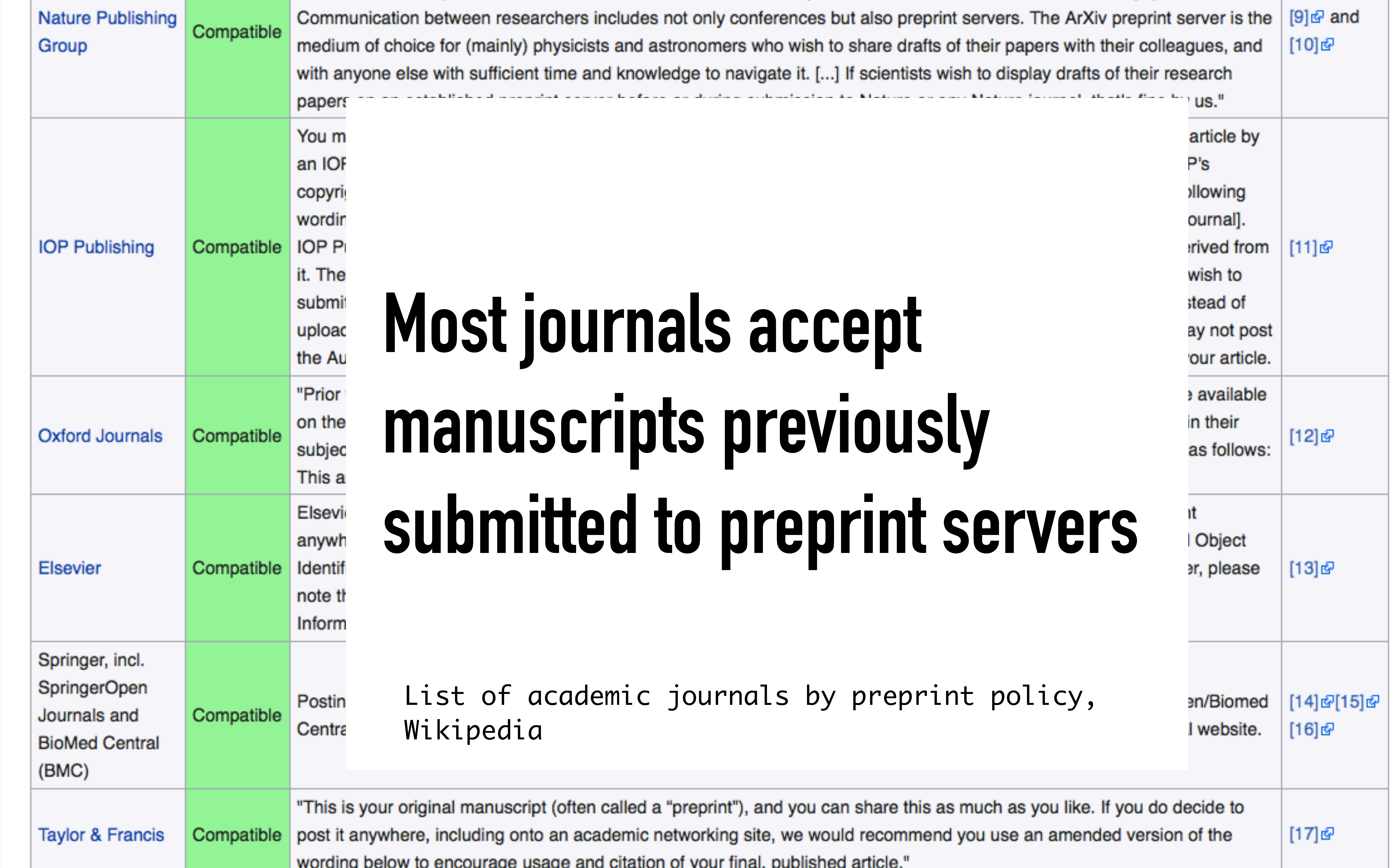
- In physics, math, CS… arXiv is a way of life
- Preprints growing by all metrics
- Explosion of ’Xiv sites

engrχiv

bioRχiv
beta
THE PREPRINT SERVER FOR BIOLOGY

SOC
ARXIV

OSFPREPRINTS



Nature Publishing Group	Compatible	Communication between researchers includes not only conferences but also preprint servers. The ArXiv preprint server is the medium of choice for (mainly) physicists and astronomers who wish to share drafts of their papers with their colleagues, and with anyone else with sufficient time and knowledge to navigate it. [...] If scientists wish to display drafts of their research papers on an established preprint server before or during submission to Nature or any Nature journal, that's fine by us."	[9] and [10]
IOP Publishing	Compatible	You m article by an IOF P's copyri llowing wordir ournal]. IOP Pu rived from it. The wish to submi stead of uploac ay not post the Au our article.	[11]
Oxford Journals	Compatible	"Prior available on the in their subjec as follows: This a	[12]
Elsevier	Compatible	Elsevi it anywh Object Identif er, please note th Inform	[13]
Springer, incl. SpringerOpen Journals and BioMed Central (BMC)	Compatible	Postin en/Biomed Centra l website.	[14][15][16]
Taylor & Francis	Compatible	"This is your original manuscript (often called a "preprint"), and you can share this as much as you like. If you do decide to post it anywhere, including onto an academic networking site, we would recommend you use an amended version of the wording below to encourage usage and citation of your final, published article."	[17]

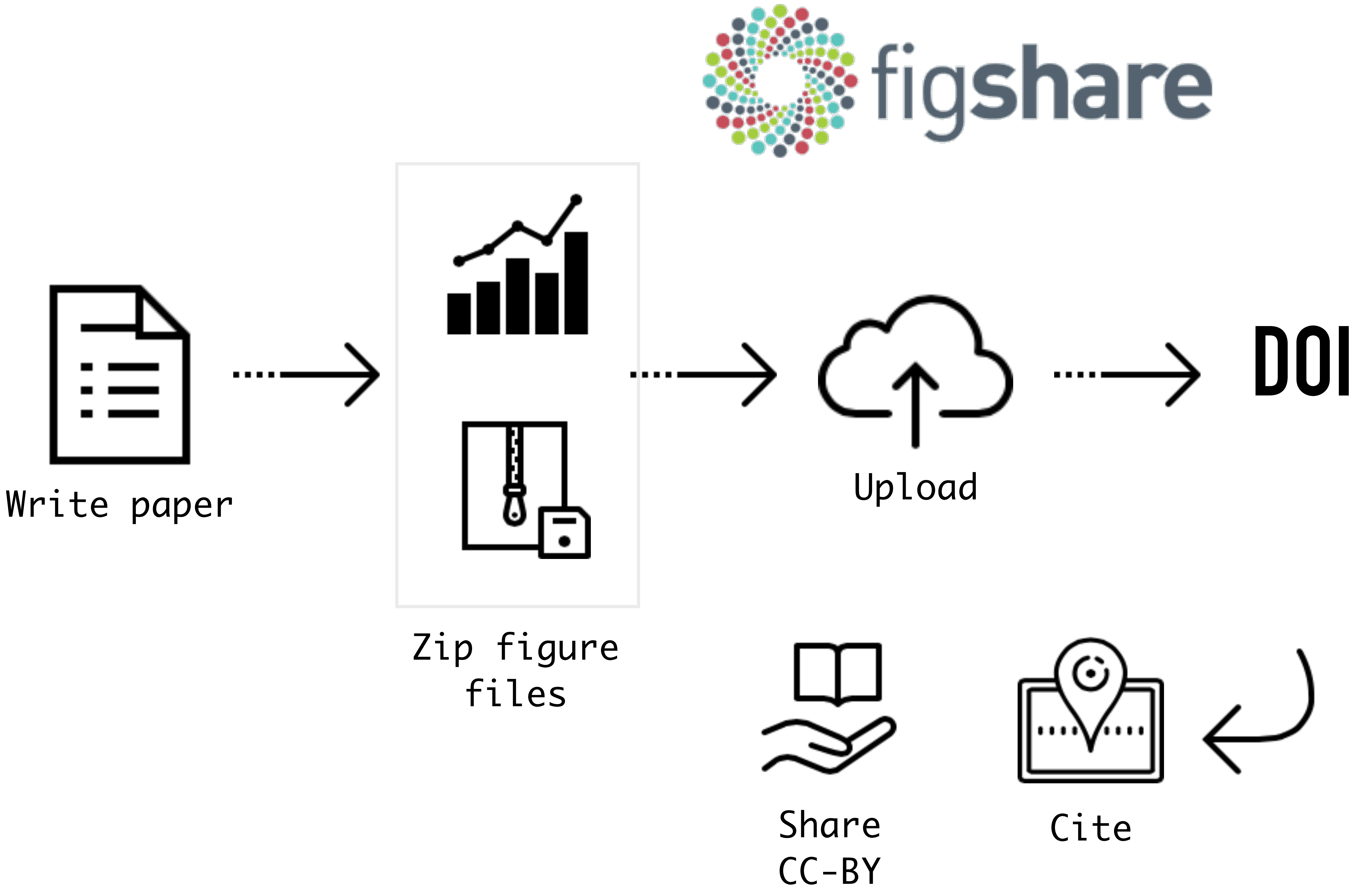
Most journals accept manuscripts previously submitted to preprint servers

List of academic journals by preprint policy, Wikipedia

ReproPacks

- For main results in a paper, we share data, plotting script & figure under CC-BY.
- Deposit the file bundle as a Figshare object and get a DOI
- We cite this DOI in the figure caption!

Our workflow:



figshare

Write paper

Zip figure files

Upload

DOI

Share CC-BY

Cite

Icons from Icons8.

Top challenges of reproducible research

- creation, curation, usage and publication of research software
- acceptance, adoption and standardization of open-science practices;
- misalignment with academic incentive structures and institutional processes for career progression

Research Transparency and Reproducibility Training (RT2)

Reproducibility and Replicability

The NASEM report and the praxis of reproducible research

@LorenaABarba http://lorenabarba.com