Reproducible Data Visualization in Jupyter Notebooks

Huajin Wang
Ana Van Gulick

UL-DataServices@andrew.cmu.edu

November 18, 2019

Carnegie Mellon University
Libraries

The evolution of scientific discovery

Is Einstein's theory of special relativity no longer true?

Faster than light particles found, claim scientists

Particle physicists detect neutrinos travelling faster than light, a feat forbidden by Einstein's theory of special relativity

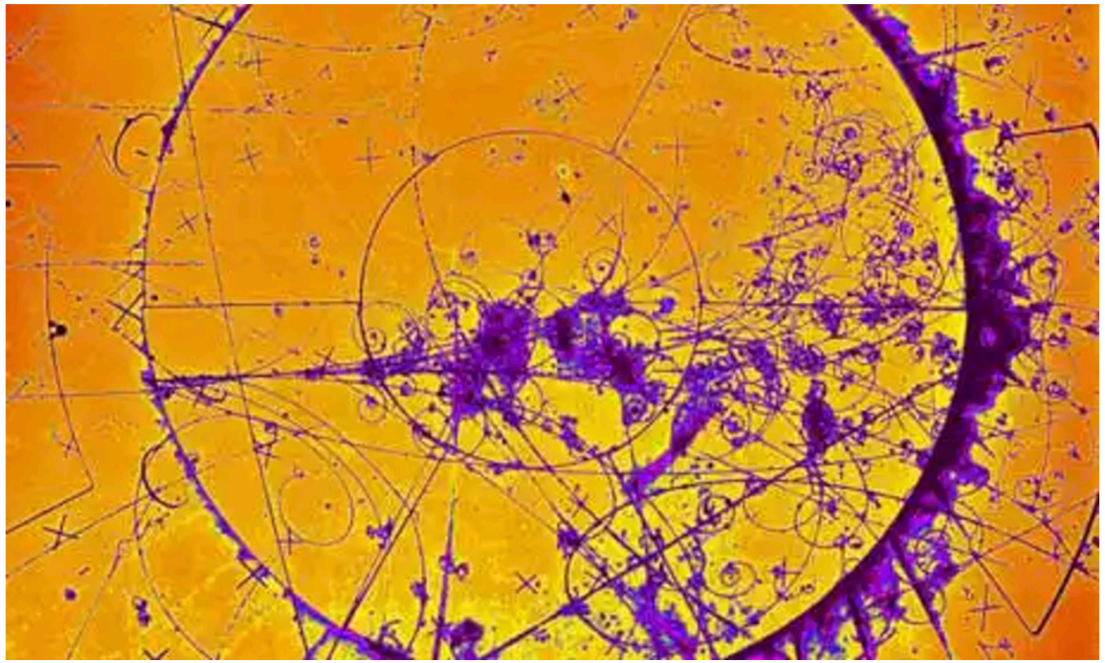
▲ Neutrinos, like the ones above, have been detected travelling faster than light, say particle physicists. Photograph: Dan Mccoy /Corbis

The Guardian. 2011/09/22

He was right after all!

Neutrino researchers admit Einstein was right

Nine months after its results caused a furore, experiment that suggested neutrinos could travel faster than light declared faulty



▲ Cern's research director, Sergio Bertolucci, told a conference in Japan that faulty wiring had led to readings suggesting neutrinos could travel faster than the speed of light. Photograph: Cern/Science Photo Library

The Guardian. 2012/06/08

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

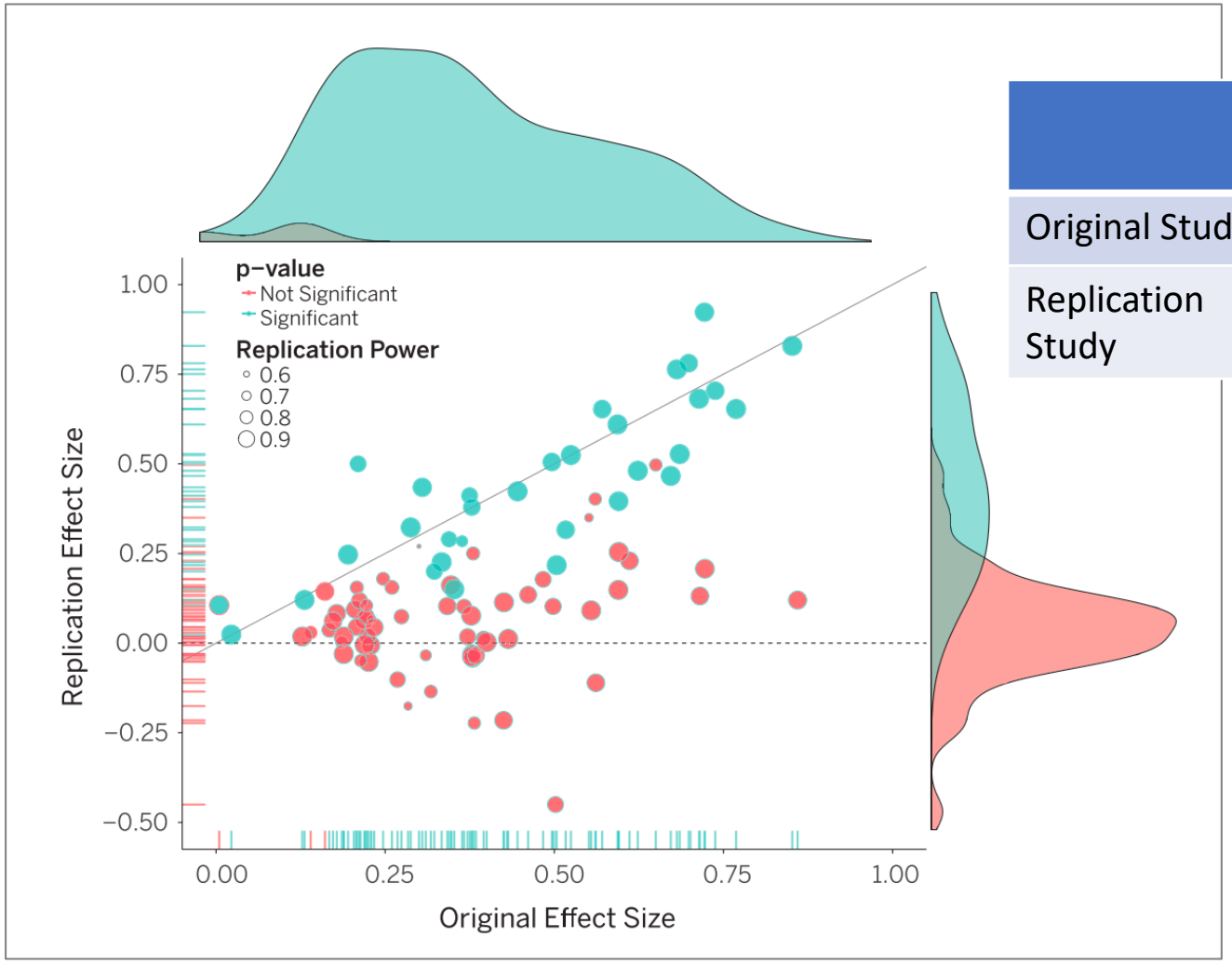
Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct

Ioannidis, 2005

Reproducibility of Psychological Science



	Statistical Significance	Mean Effect Size
Original Study	97%	0.403
Replication Study	36%	0.197

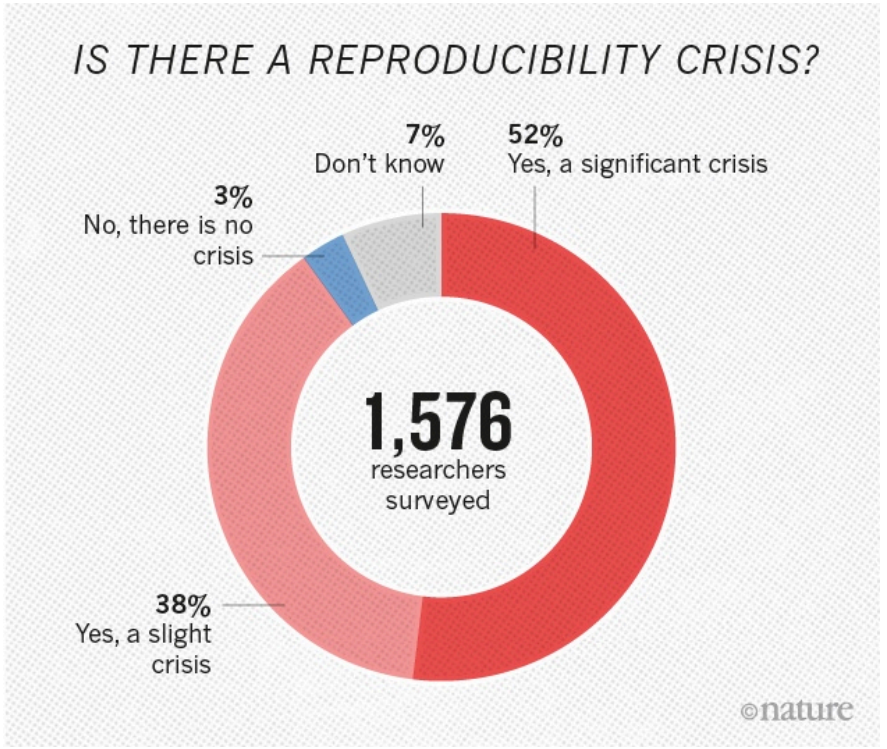
Open Science Collaboration, Nosek. *Science.* (2015)

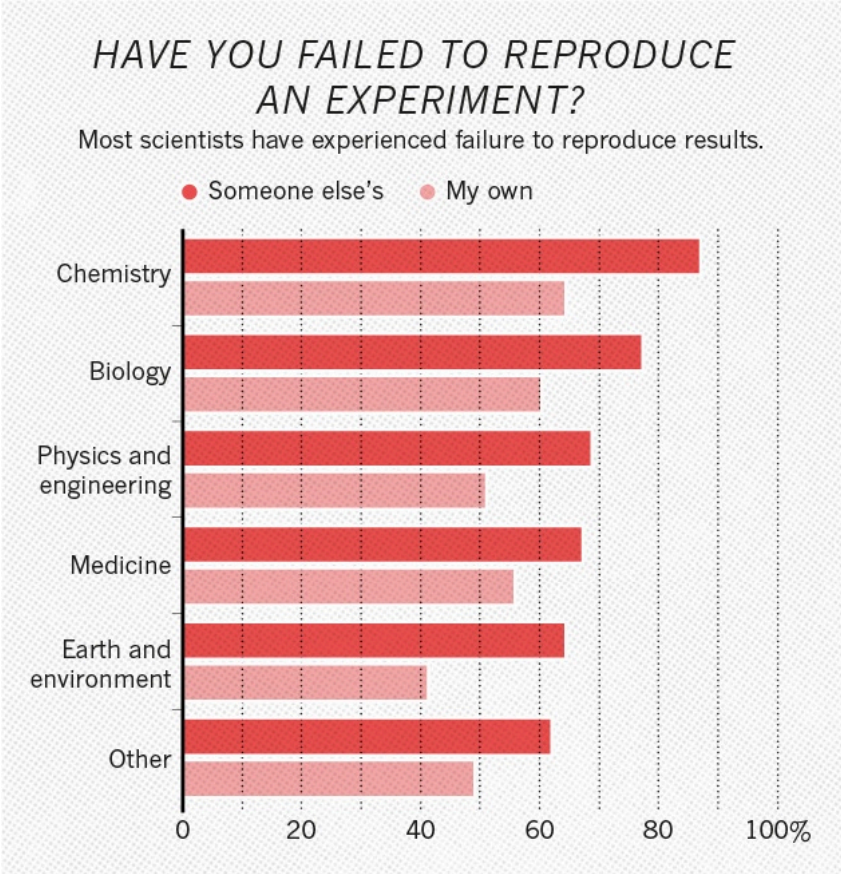
The Reproducibility Crisis (?)

Nature News | News Feature

1,500 scientists lift the lid on reproducibility

Survey sheds light on the ‘crisis’ rocking research.

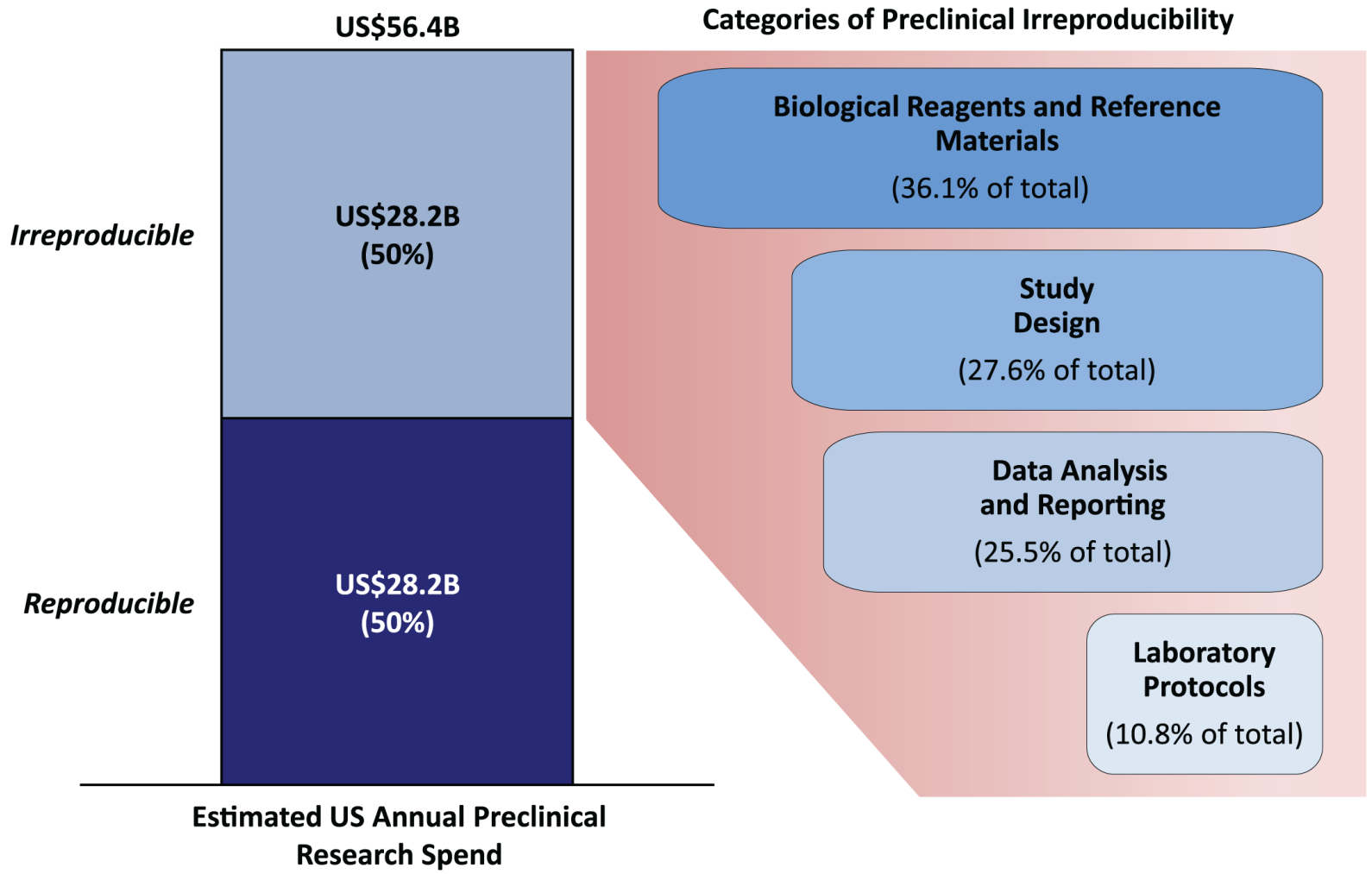




Baker, *Nature*, 2016

Why should I care?

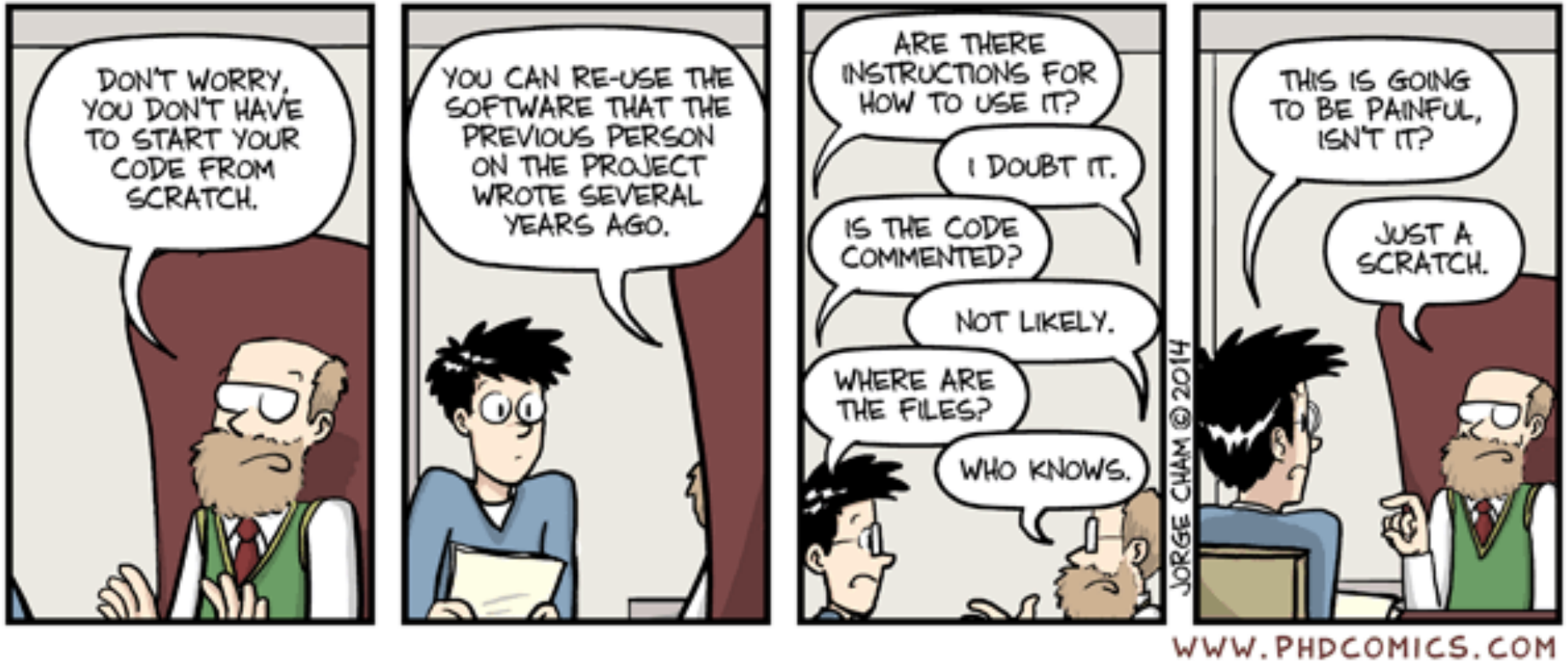
Data is expensive and valuable



Freedman et al. *PLoS Biology*. (2015)

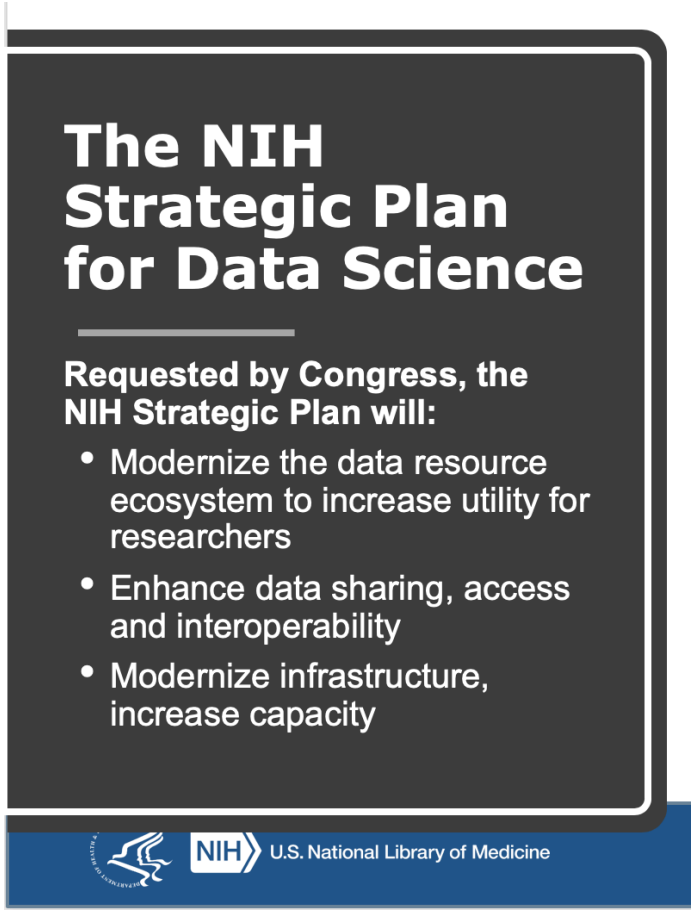
Poorly documented research is hard for collaborators, trainees, and yourself

Piled Higher and Deeper *by Jorge Cham* **www.phdcomics.com**



title: "Scratch" - originally published 3/12/2014

Data Science techniques, collaboration, and the future of science require large amounts of data that is well documented.



Lisa Federer, CMU Open Science Symposium, 2019

Definitions of Reproducibility

Empirical, computational or statistical

Stodden, 2014

- For economists /social scientists:
  - Computer **code and data are available** so that someone would be able to redo the **same analysis** using the **same data**

- For bench scientists:
  - Another scientist using the **same methods** gets **similar results** and can draw the **same conclusions**

Editorial. *Nature. (2016)*

MENU

nature human behaviour

Perspective | Open Access | Published: 10 January 2017

A manifesto for reproducible science

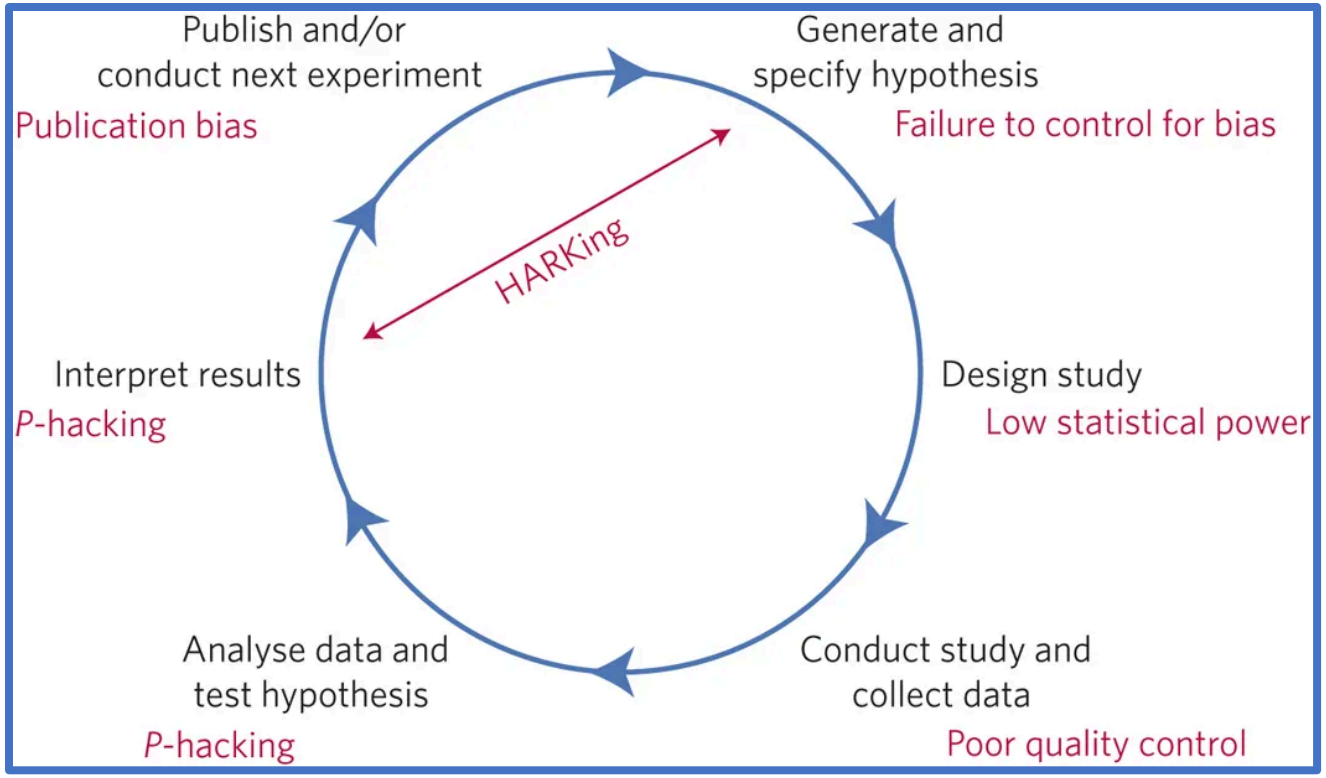
Marcus R. Munafò ✉, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

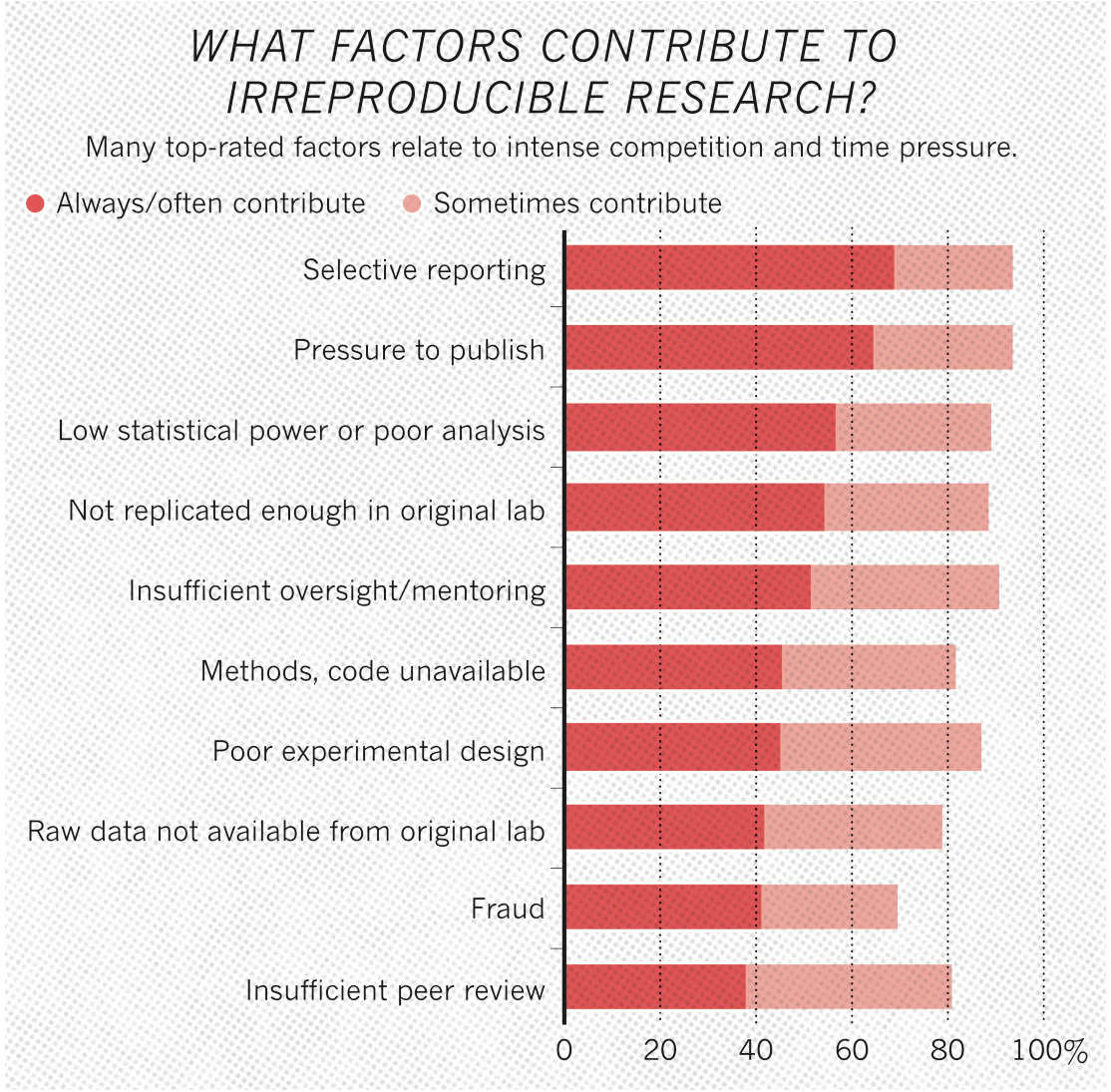
34k Accesses | **524** Citations | **2591** Altmetric | Metrics

Abstract

Improving the reliability and efficiency of scientific research will increase the credibility of the published scientific literature and accelerate discovery. Here we argue for the adoption of measures to optimize key elements of the scientific process: methods, reporting and dissemination, reproducibility, evaluation and incentives. There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures, but their broad adoption by researchers, institutions, funders and journals will require iterative evaluation and improvement. We discuss the goals of these measures, and how they can be implemented, in the hope that this will facilitate action toward improving the transparency, reproducibility and efficiency of scientific research.



Factors contributing to poor reproducibility



- Pressure to publish
- Poor methodological training
- Poor laboratory record keeping
- Poor statistical knowledge
- Lack of resources to appropriately executive the experiments

Baker M. Nature. (2016)

ASCB Member Survey on Reproducibility (2015)

Factors contributing to poor reproducibility

- Fraud (often not the case)
- Poor study design or human error – better planning
  - Cherry-picking / selective reporting
  - Inconsistent reagents, mis-labeling, or contamination
  - Incorrect equipment calibration
  - Insufficient statistical powder or wrong statistical methods
  - Insufficient training
- Incomplete documentation – better documentation
- Lack of transparency in research methods and outcome – better sharing

Practicing Open Science

eLIFE elifesciences.org

FEATURE ARTICLE

POINT OF VIEW

How open science helps researchers succeed

Abstract Open access, open data, open source and other open scholarship practices are growing in popularity and necessity. However, widespread adoption of these practices has not yet been achieved. One reason is that researchers are uncertain about how sharing their work will affect their careers. We review literature demonstrating that open research is associated with increases in citations, media attention, potential collaborators, job opportunities and funding opportunities. These findings are evidence that open research practices bring significant benefits to researchers relative to more traditional closed practices.

DOI: 10.7554/eLife.16800.001

ERIN C MCKIERNAN*, PHILIP E BOURNE, C TITUS BROWN, STUART BUCK, AMYE KENALL, JENNIFER LIN, DAMON MCDOUGALL, BRIAN A NOSEK, KARTHIK RAM, COURTNEY K SODERBERG, JEFFREY R SPIES, KAITLIN THANEY, ANDREW UPDEGROVE, KARA H WOO AND TAL YARKONI

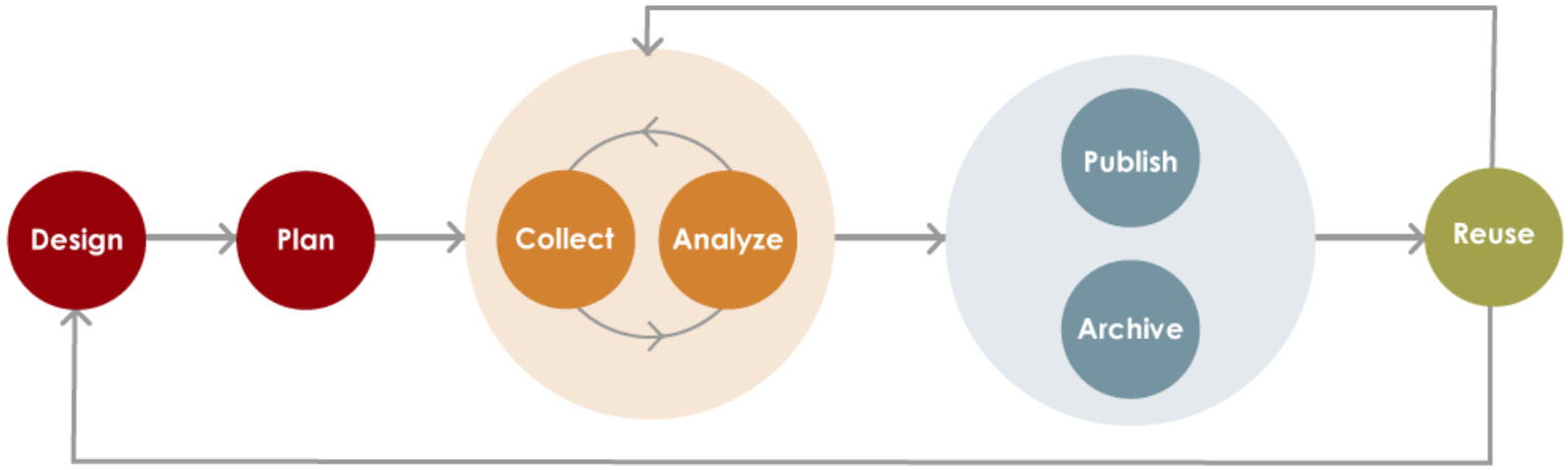
Box 1. What can I do right now?

Engaging in open science need not require a long-term commitment or intensive effort. There are a number of practices and resolutions that researchers can adopt with very little effort that can help advance the overall open science cause while simultaneously benefiting the individual researcher.

1. **Post free copies of previously published articles in a public repository**. Over 70% of publishers allow researchers to post an author version of their manuscript online, typically 6-12 months after publication (see section "Publish where you want and archive openly").
2. **Deposit preprints of all manuscripts in publicly accessible repositories** as soon as possible – ideally prior to, and no later than, the initial journal submission (see section "Postprints").
3. **Publish in open access venues** whenever possible. As discussed in Prestige and journal impact factor, this need not mean forgoing traditional subscription-based journals, as many traditional journals offer the option to pay an additional charge to make one's article openly accessible.
4. **Publicly share data and materials via a trusted repository**. Whenever it is feasible, the data, materials, and analysis code used to generate the findings reported in one's manuscripts should be shared. Many journals already require authors to share data upon request as a condition of publication; pro-actively sharing data can be significantly more efficient, and offers a variety of other benefits (see section "Resource management and sharing").
5. **Preregister studies**. Publicly preregistering one's experimental design and analysis plan in advance of data collection is an effective means of minimizing bias and enhancing credibility (see section "Open questions"). Since the preregistration document(s) can be written in a form similar to a Methods section, the additional effort required for preregistration is often minimal.

DOI: 10.7554/eLife.16800.006

Solution: Good data management throughout the research life cycle!



Literature
Hardware
Software
Notebooks
Statistics
Data Management Plan

Data Backup
File Naming Systems
Collaboration
Documentation

Reference Management
Writing Platform
Sharing
Open Access

Preservation
Visibility
Research Profile

Library RDM Website http://www.library.cmu.edu/RDM

RDM Best Practices

Including how to

- Organize
- Document
- Back-up
- Secure your work!

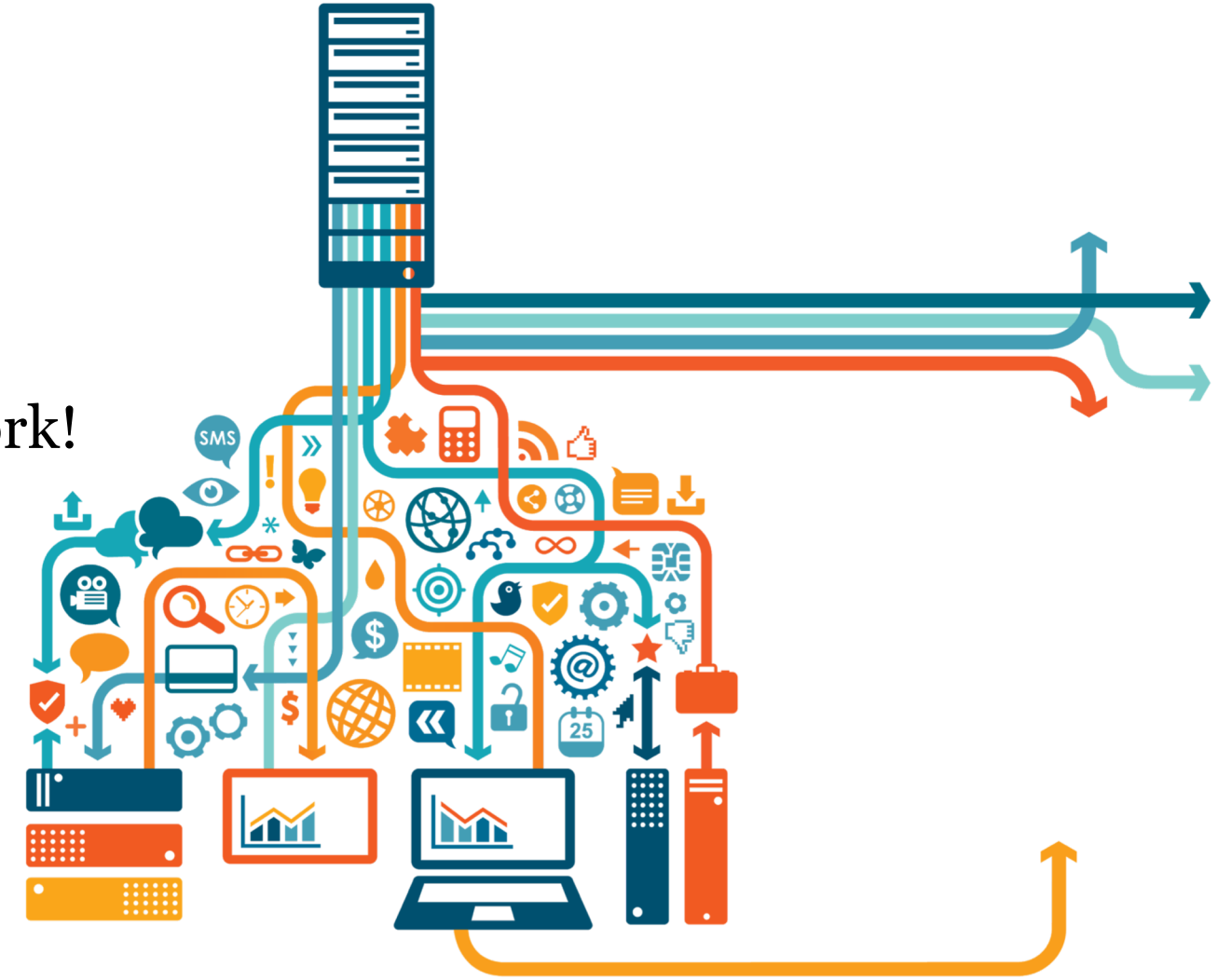


Image: Compute Canada

FAIR Principles for all research products

F

A

I

R

Australian Research Data Commons

Data Science at NIH

Plan Ahead:
What Should be in a Data Management Plan?

- What data will be generated in this project?
- Who will be responsible for the data at each stage?
- How will datasets need to be connected?
- What formats will be used (Excel, MySQL, jpg, etc.)?
- What information about the data will need to be captured so that others can understand it?
- Where should the data be stored and who should have access to it?
- How should the data be organized and named?
- How will the data be published or archived at the end of the project?

3-2-1 Backup!

- 3 copies of your data
- 2 different formats (e.g. laptop, external hard)
- 1 off-site back-up (e.g. PSC storage) or in the cloud (e.g.CMU Box account)

File Formats

- The ability to share and re-use your data.
- Plan for future hardware and software obsolescence.
- Save the dataset in multiple open documented formats, when possible, to ensure long term preservation.

Some preferred file formats

- **Containers**: TAR, GZIP, ZIP
- **Databases**: XML, CSV
- **Geospatial**: SHP, DBF, GeoTIFF, NetCDF
- **Moving images**: MOV, MPEG, AVI, MXF
- **Sounds**: WAVE, AIFF, MP3, MXF
- **Statistics**: ASCII, DTA, POR, SAS, SAV
- **Still images**: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- **Tabular data**: CSV
- **Text**: XML, PDF/A, HTML, ASCII, UTF-8

Slide courtesy of Sarah Young

File Naming Conventions

- Create your FNC by identifying key elements of the project, e.g. date of creation, author's name, project name, or section

- Have a code book or data dictionary

- Have a readme file that lists all files and any file hierarchy

File Naming Conventions

- Avoid special characters
- Use underscores instead of periods or spaces
- Err on the side of brevity (<25 characters)
- Include all necessary descriptive information independent of where it is stored
- Include dates, format consistently
- Include a version number
- **Be consistent**

Slide courtesy of Sarah Young

Files without employing a naming convention:

- Test_data_2013
- Project_Data
- Design for project.doc
- Lab_work_Eric
- Second_test
- Meeting Notes Oct 23

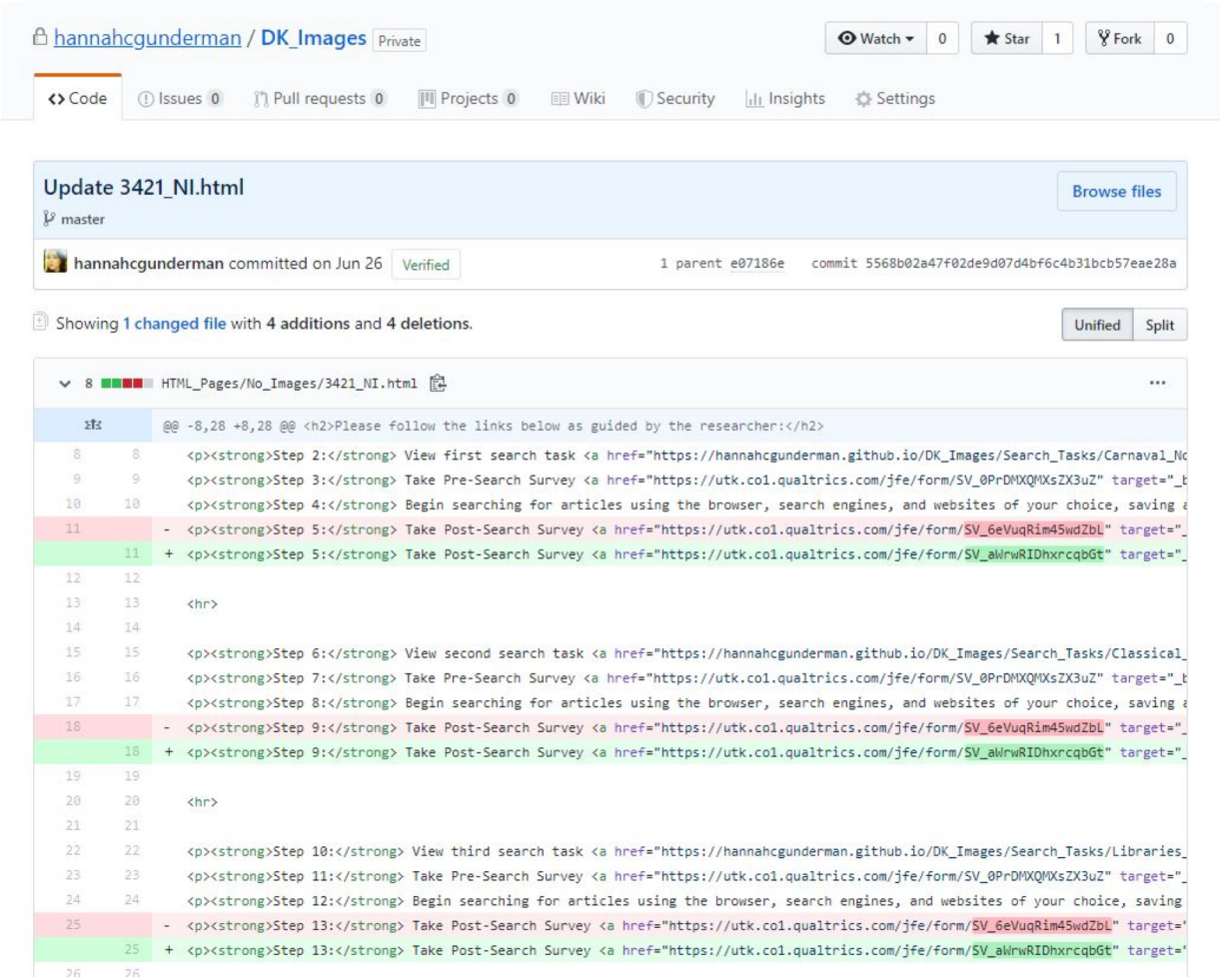
Files with a naming convention:

- 20130503_DOEProject_DesignDocument_Smith_v2-01.docx
- 20130709_DOEProject_MasterData_Jones_v1-00.xlsx
- 20130825_DOEProject_Ex1Test1_Data_Gonzalez_v3-03.xlsx
- 20130825_DOEProject_Ex1Test1_Documentation_Gonzalez_v3-03.xlsx
- 20131002_DOEProject_Ex1Test2_Data_Gonzalez_v1-01.xlsx
- 20141023_DOEProject_ProjectMeetingNotes_Kramer_v1-00.docx

Slide courtesy of Sarah Young

Version Control is Key! (Don’t assume you’ll remember what you did)

- GitHub for version control
- Students can get advanced account (private repositories!)



Metadata: Makes Your Work Reusable

Metadata is data that describes a dataset:

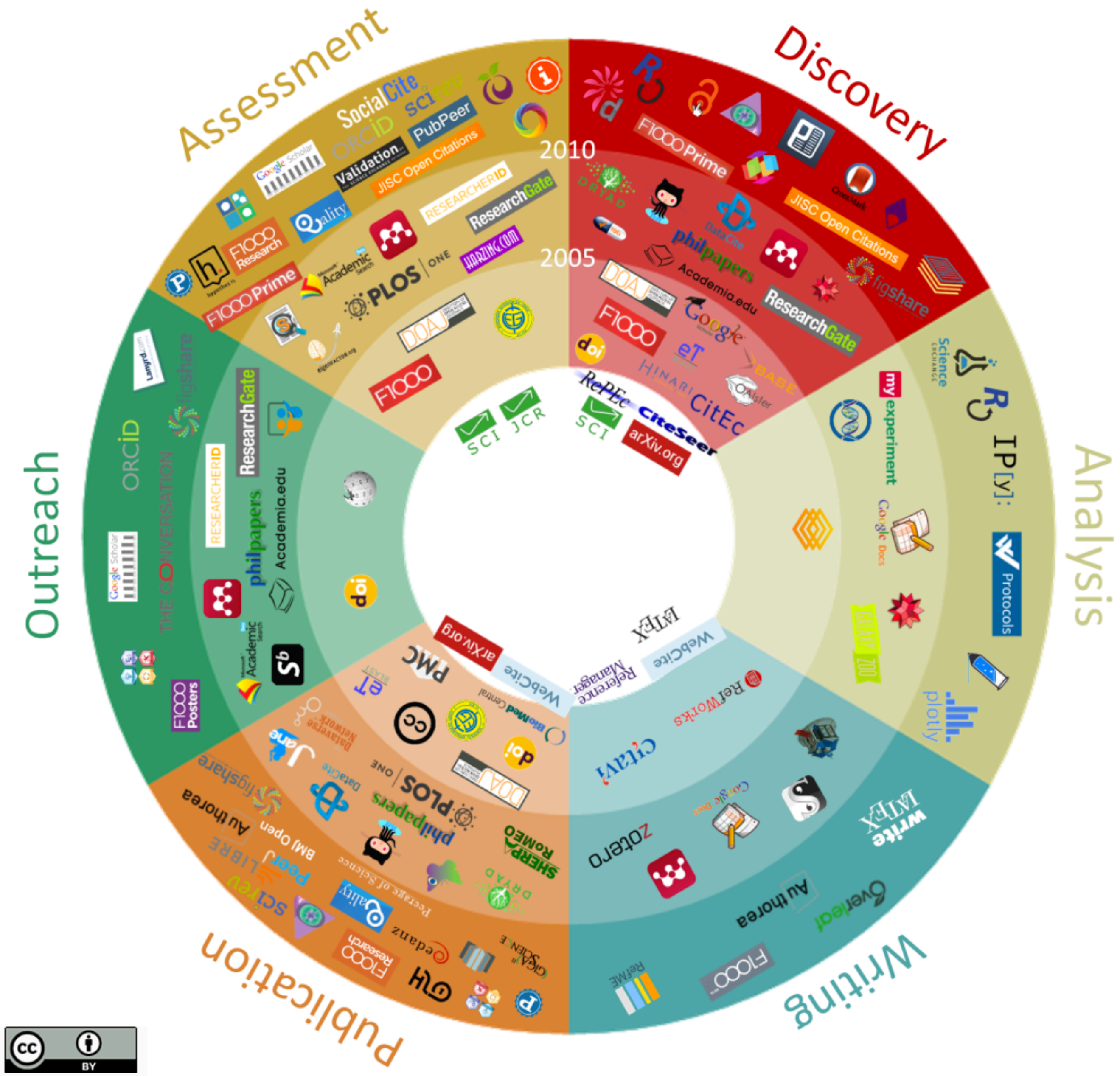
- What is the data?
- Who created it?
- How may it be used?
- What generated it?

Most repositories require some basic metadata record.

It is a good practice to build metadata into your collection and analysis workflow!

Select the right tools for each research phase

- Make a plan from the start of a project
- Tools for
  - Documentation
  - Collaboration
  - Data storage and backup
  - Sharing
  - + Tools that work together
  - + Tools that are used in your discipline



Crowdsourced list of tools, now 400+ http://bit.ly/innoscholcomm-list

Bosman & Kramer. 2015

Tools at CMU

DMPTool
Build your Data Management Plan

box

Google Drive

MENDELEY

zotero

labarchives

protocols.io

OSF

figshare

Overleaf

Pilot coming

CODE OCEAN

Bench Sciences Reproducibility

- Protocols
  - Document details and variations of every step

- Reagents:
  - Antibodies, cell lines, serum, plasmids
  - Document everything: Vendor, cat #, lot #, date, recipe, …
  - Obtain from reliable sources

(Antibodies)

- Equipment:
  - Calibration and metadata

(cell lines)

Computational Reproducibility

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2]*, Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

1 Department of Informatics, University of Oslo, Blindern, Oslo, Norway, **2** Centre for Cancer Biomedicine, University of Oslo, Blindern, Oslo, Norway, **3** Department of Biochemistry and Molecular Biology and The Huck Institutes for the Life Sciences, Penn State University, University Park, Pennsylvania, United States of America, **4** Department of Biology and Department of Mathematics and Computer Science, Emory University, Atlanta, Georgia, United States of America, **5** Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway, **6** Institute for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Montebello, Oslo, Norway

PLOS | COMPUTATIONAL BIOLOGY

PERSPECTIVE

Good enough practices in scientific computing

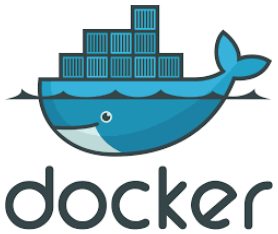
Greg Wilson[1☯]*, Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

1 Software Carpentry Foundation, Austin, Texas, United States of America, **2** RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, **3** Department of Biology, Duke University, Durham, North Carolina, United States of America, **4** Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, **5** Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, **6** Data Carpentry, Davis, California, United States of America

☯ These authors contributed equally to this work.
* gvwilson@software-carpentry.org

Computational Reproducibility

- Document all steps
- Comment with textual statements
- Save both raw and intermediate data
- Document all dependencies
- Use relative paths
  - ../picture.jpg
  - /Users/Huajin/Pictures/picture.jpg
- Use version control to track changes



Computational Reproducibility

- For analyses that include randomness, note underlying random seeds
- Avoid manual manipulation
- Use collaboration platforms for ease discovery and understanding
- Share script, analysis, and results

Literate Programming Platforms

Jupyter Notebook and R Markdown

- Documentation of data analysis code, results, and visualization all in one place
- Enable others to understand and reproduce results
- Example: code for LIGO project - discovery of gravitational waves https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html

Run notebooks online

= binder (beta)

https://mybinder.org

Sharing your work - General Repositories

Discipline-specific repositories

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

OpenNEURO
A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data

Recommended Data Repositories from *Scientific Data*:
https://www.nature.com/sdata/policies/repositories

Depending on your discipline, your data may have greater visibility in discipline-specific repositories! Always check with experts in your field.

Discipline-specific repositories

SCIENTIFIC DATA

Search | E-alert | Submit | Login

Policies

- Editorial & Publishing Policies
- For Referees
- Data Policies
- Recommended Data Repositories

Recommended Data Repositories

Scientific Data mandates the release of datasets accompanying our Data Descriptors, but we do not ourselves host data. Instead, we ask authors to submit datasets to an appropriate public data repository. Data should be submitted to discipline-specific, community-recognized repositories where possible, or to generalist repositories if no suitable

Software and Data Artifacts in the ACM Digital Library

ACM encourages authors to submit software and data sets with their papers. For years, ACM has provided mechanisms for authors to submit software, data sets, videos and other supplemental artifacts with their research papers. We have recently made these artifacts more discoverable through search and made them more prominent on abstract pages and Tables of Contents.

ACM's Reproducibility Task Force has been working with SIG conferences and journal EICs to understand and articulate common Best Practices in preparing and reviewing software and data artifacts, how they can be integrated with the ACM Digital Library, and how to reflect them in publication and enable their re-use. Many of ACM's technical communities are evolving standardized documentation and review processes to improve the chances for successful experiment re-runs and artifact re-use.

A number of ACM conferences and journals have already instituted formal processes and are implementing Best Practices for artifact review. ACM provides standard terms and definitions for labeling successful artifact reviews, and iconic badging for their associated articles, thereby establishing uniformity across ACM publications and any choosing to adopt its Best Practices.

ACM Badges
ACM Reproducibility Task Force
DL Pilot Integrations

CMU Repository: KiltHub

1TB deposit limit (accounts start with 20GB)

All grad students and faculty have accounts

GitHub Ingest, API

Self-deposit with libraries check point

DOIs and citations, metrics on reuse

Powered by

figshare
credit for all your research

Integrated with Figshare public repository = good discovery in google and trusted by publishers and funders

https://kilthub.cmu.edu/

Licensing and Privacy

- Establish ownership and access to data at the start of a research project.
- Understand implications of ownership for sharing, preservation, or publication.
- Licensing varies depending on scholarship type (data vs code vs papers)
- Considerations:
  - Human subjects (IRB, HIPPA, FERPA)
  - Animal research (IACUC)
  - Proprietary data (purchased or licensed data)

Creative Commons Licenses: Good for Text, Posters

Seven regularly used licenses [edit]

Icon	Description	Acronym	Attribution Required	Allows Remix culture	Allows commercial use	Allows Free Cultural Works	Meets 'Open Definition'
	Freeing content globally without restrictions	CC0	No	Yes	Yes	Yes	Yes
	Attribution alone	BY	Yes	Yes	Yes	Yes	Yes
	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes	Yes	Yes
	Attribution + Noncommercial	BY-NC	Yes	Yes	No	No	No
	Attribution + Noncommercial + ShareAlike	BY-NC-SA	Yes	Yes	No	No	No
	Attribution + NoDerivatives	BY-ND	Yes	No	Yes	No	No
	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	Yes	No	No	No	No

https://en.wikipedia.org/wiki/Creative_Commons_license

Code is Trickier!

- Free, open source licenses
  - Few requirements for redistribution
    - BSD 3, MIT, Apache, Python (PSFL) licenses
    - Option to reuse in closed or commercial software
  - Share alike licenses (copy left)
    - E.g. GNU General Public License

Libraries Research Data Services Team

Email us at
UL-DataServices@Andrew.cmu.edu

Contact Us

www.library.cmu.edu

Huajin Wang
huajinw@cmu.edu

Data Services Team
UL-dataservices@Andrew.cmu.edu

email, phone, chat, text
www.library.cmu.edu/help/ask

@cmulibraries

@CMULibraries

@CMULibraries