Carnegie Mellon University

Reproducible Data Visualization in Jupyter Notebooks

University Libraries Workshop Series

Huajin Wang, Ph.D.
Liaison Librarian, Biology and Computer Science
Co-director, Open Science & Data Collaborations program

huajinw@cmu.edu

Reproducibility! ← Remember this word.

What is it?

"reproducibility refers to the ability of a researcher to **duplicate the results of a prior study** using the **same materials** as were used by the **original investigator**. That is, a second researcher might use the **same raw data** to build the **same** analysis files and implement the **same** statistical analysis in an attempt to yield the **same results**.... Reproducibility is a minimum necessary condition for a finding to be **believable** and **informative**." - National Science Foundation Subcommittee on Replicability in Science

Conversation gained momentum in psychological science

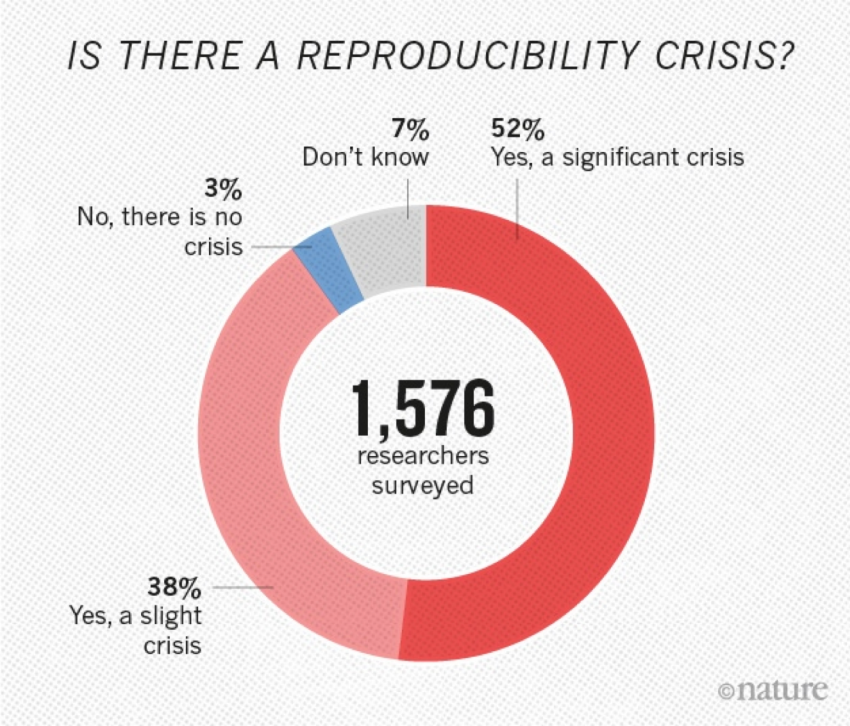
Carnegie Mellon University

Reproducibility crisis?

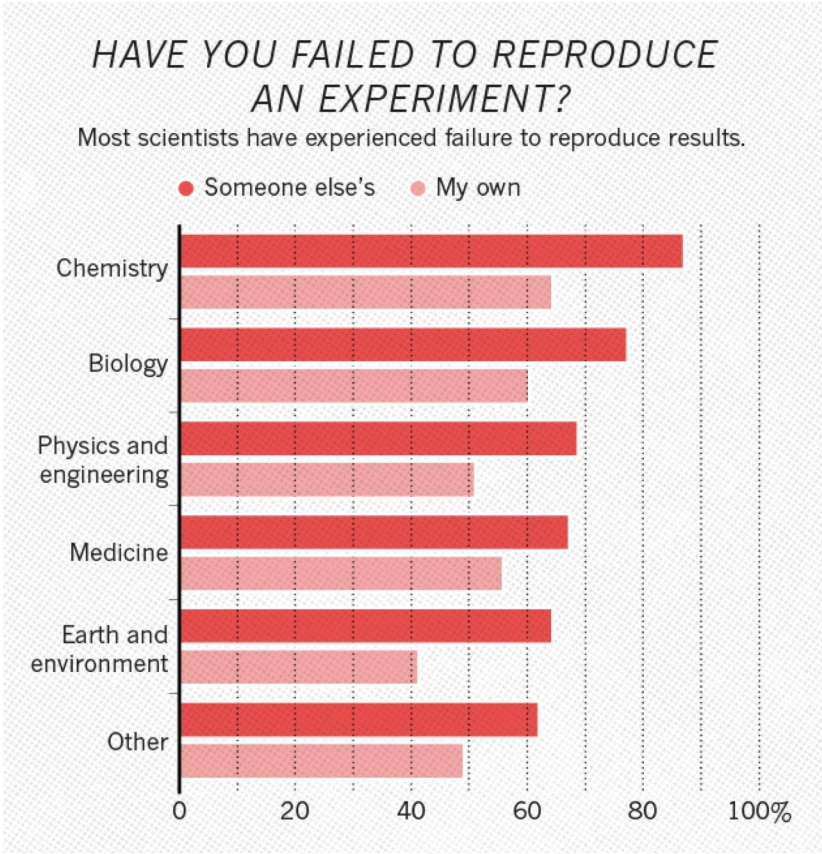
Nature News | News Feature

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.



Yes: 90%



Baker, *Nature*, 2016

Carnegie Mellon University

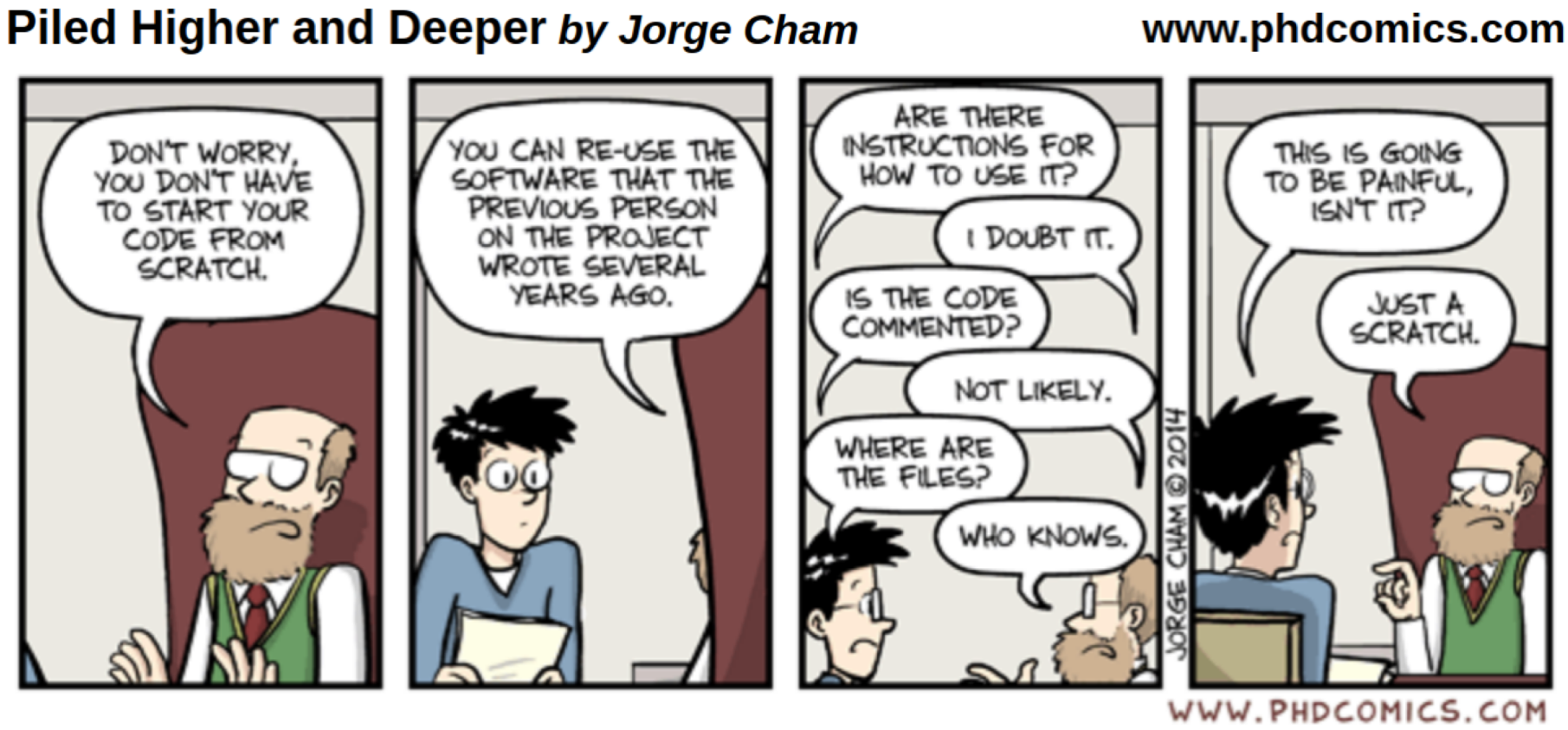
Factors contributing to low reproducibility

- Fraud (often not the case)
- Poor study design or human error (How to avoid: better planning)
  - Cherry-picking/selective reporting
  - Inconsistent reagents, mislabeling, contamination
  - Incorrect equipment calibration
  - Insufficient statistical power or wrong statistical methods
  - Insufficient training
- Incomplete documentation (How to avoid: better documentation)
- Lack of transparency in research methods and outcomes (How to avoid: better sharing)

So, why should I care?

Carnegie Mellon University

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

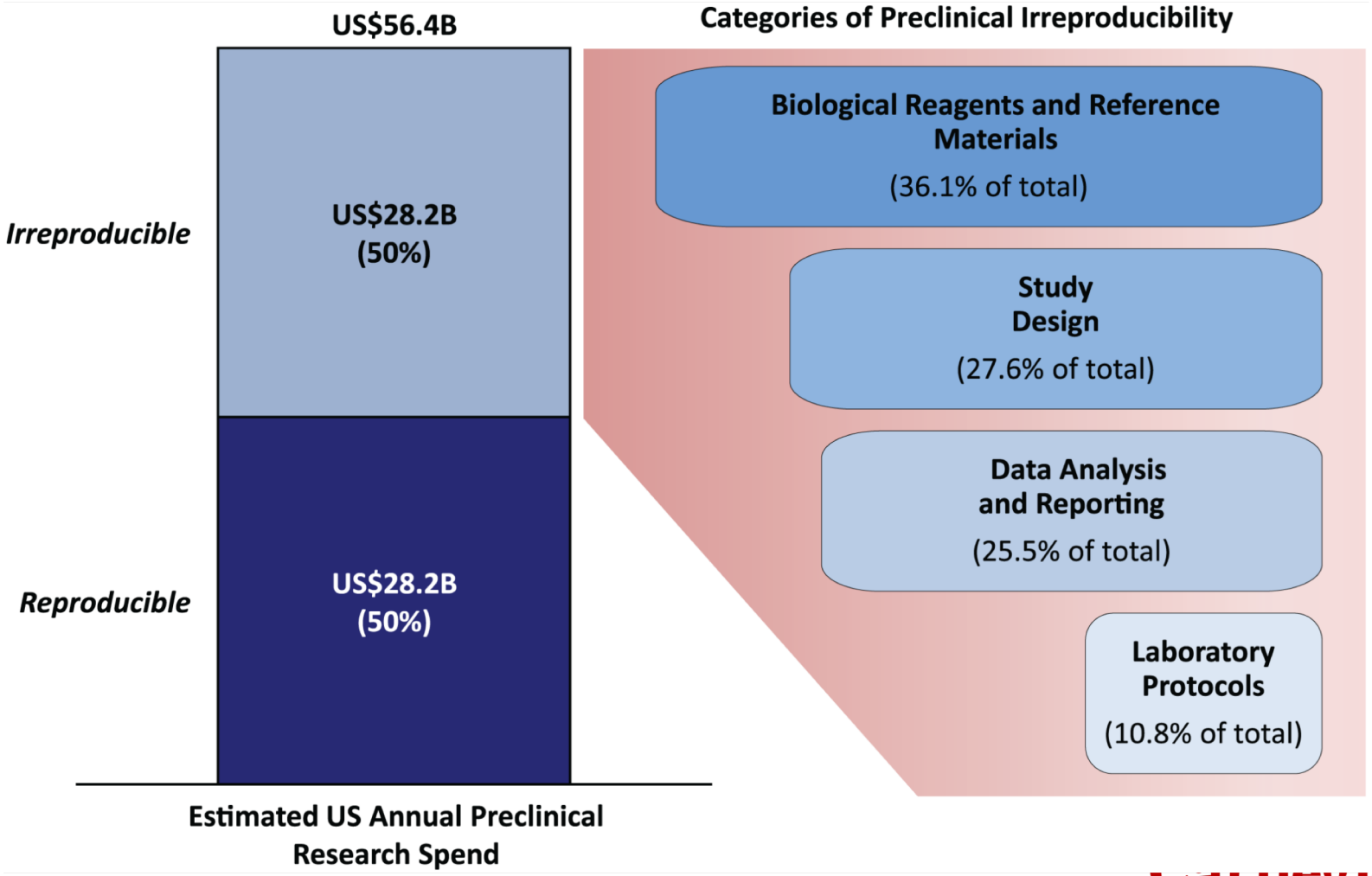


title: "Scratch" - originally published 3/12/2014

Carnegie Mellon University

Data is expensive and valuable

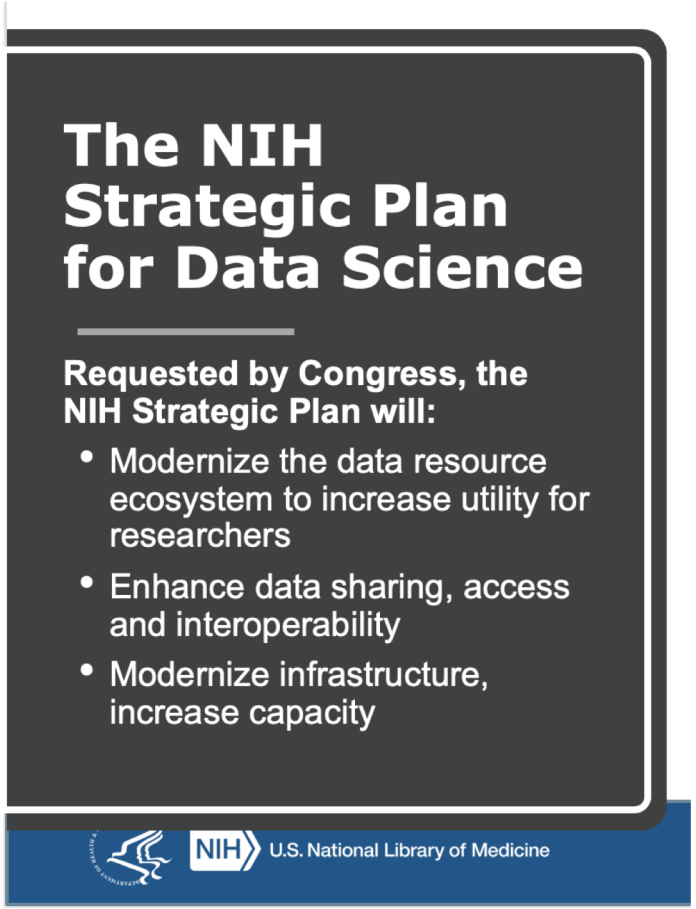
A 2015 meta-analysis of past studies regarding the cost of non-reproducible research estimated that $28 billion per year is spent on preclinical research that is not reproducible



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

Carnegie Mellon University

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

Carnegie Mellon University

Okay! I'm convinced. How do I make reproducible research?

Good Data Management and Open Science (and its principles)!

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

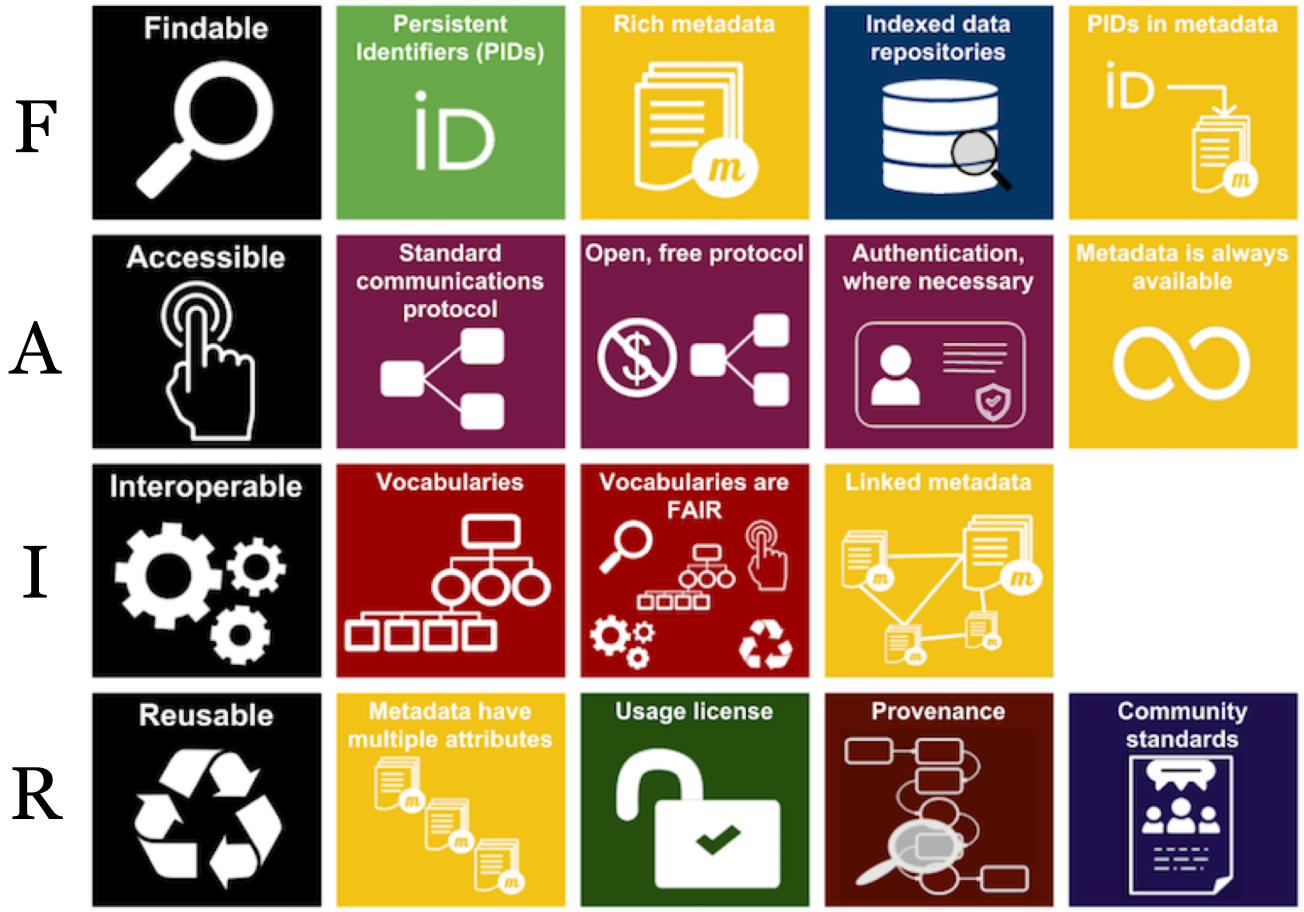
Box 1. What can I do right now?

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.

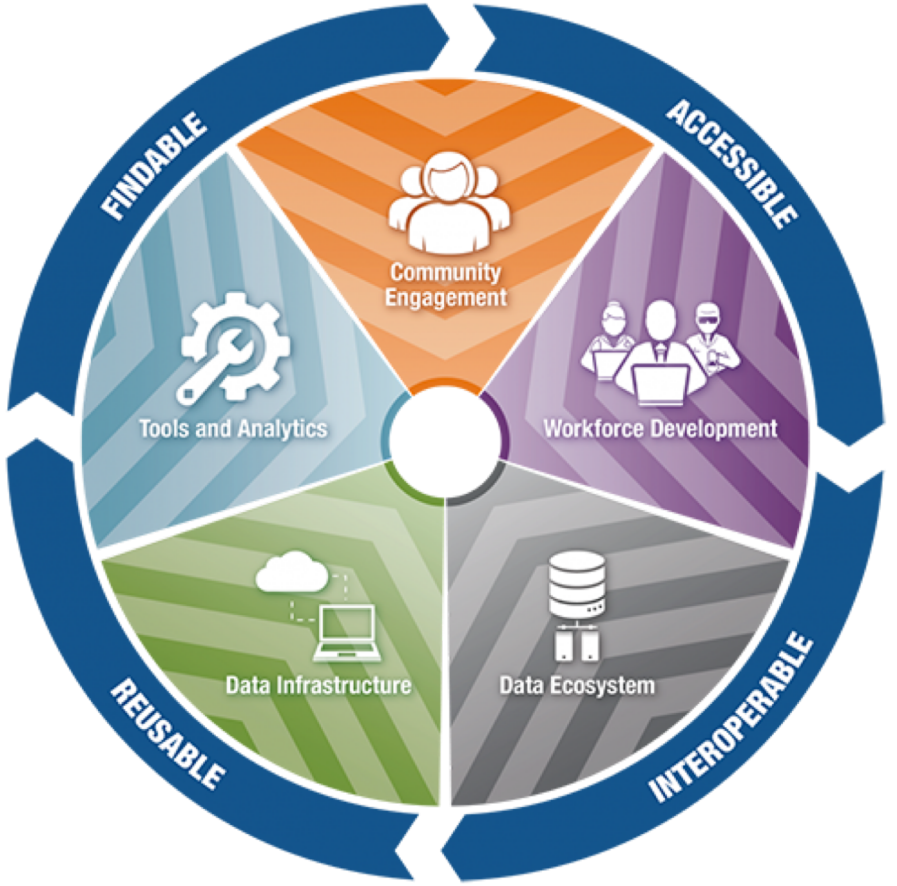
1. **Post free copies of previously published articles in a public repository**. Over 70% of publishers allow researchers to post an author version of their manuscript online, typically 6-12 months after publication (see section "Publish where you want and archive openly").
2. **Deposit preprints of 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

FAIR Principles for research products



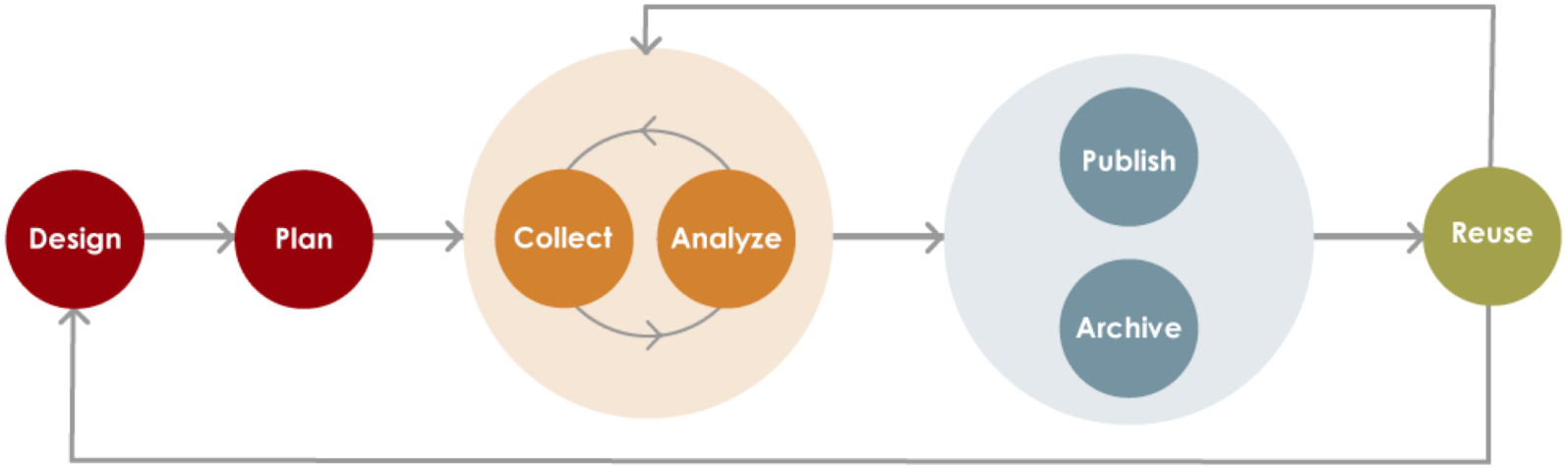
Australian Research Data Commons



Data Science at NIH

Carnegie Mellon University

Open science is built on good data management throughout the research lifecycle



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

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

Carnegie Mellon University

Plan ahead with a data management plan (DMP)

What data will be generated in this project?

Who will be responsible for the data at each stage?

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?

Carnegie Mellon University

3-2-1 backup

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

Open (when possible), sustainable 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

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 codebook or data dictionary

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

File naming conventions, cont.

Avoid special characters

Use underscores instead of periods or spaces

No more than 35 characters, ideally

Include all necessary descriptive information independent of where it is stored

Include dates, format consistently

Include a version number

Be consistent! and write it down

Files without employing a naming convention:

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

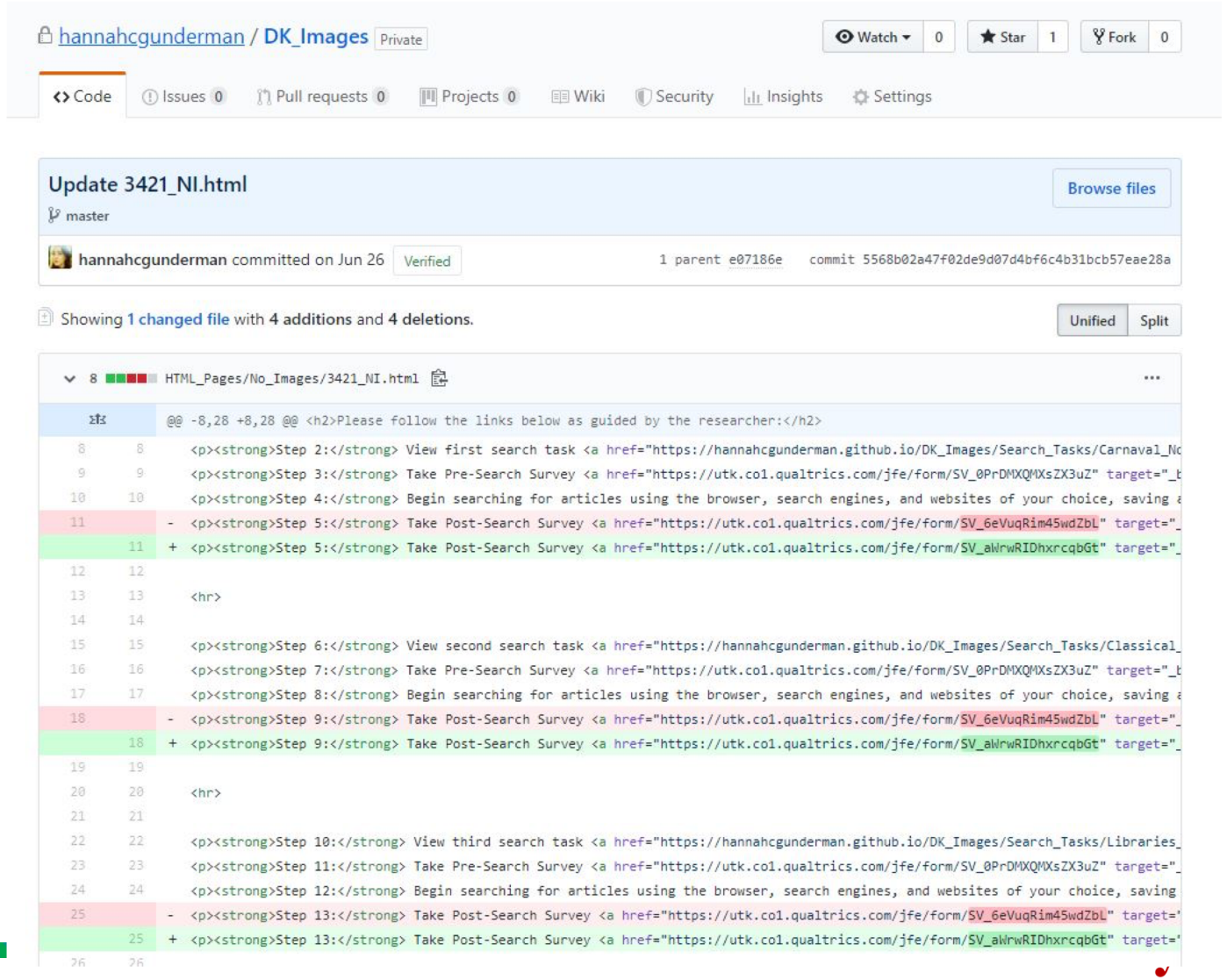
Files with a naming convention:

- 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

Version control (don’t assume you’ll remember what you did)

GitHub is a platform for version control

Use your andrew.cmu.edu email to get a premium account for free



Metadata

Metadata is data that describes a dataset:

What is the data?

Who created it?

How may it be used?

What generated it?

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

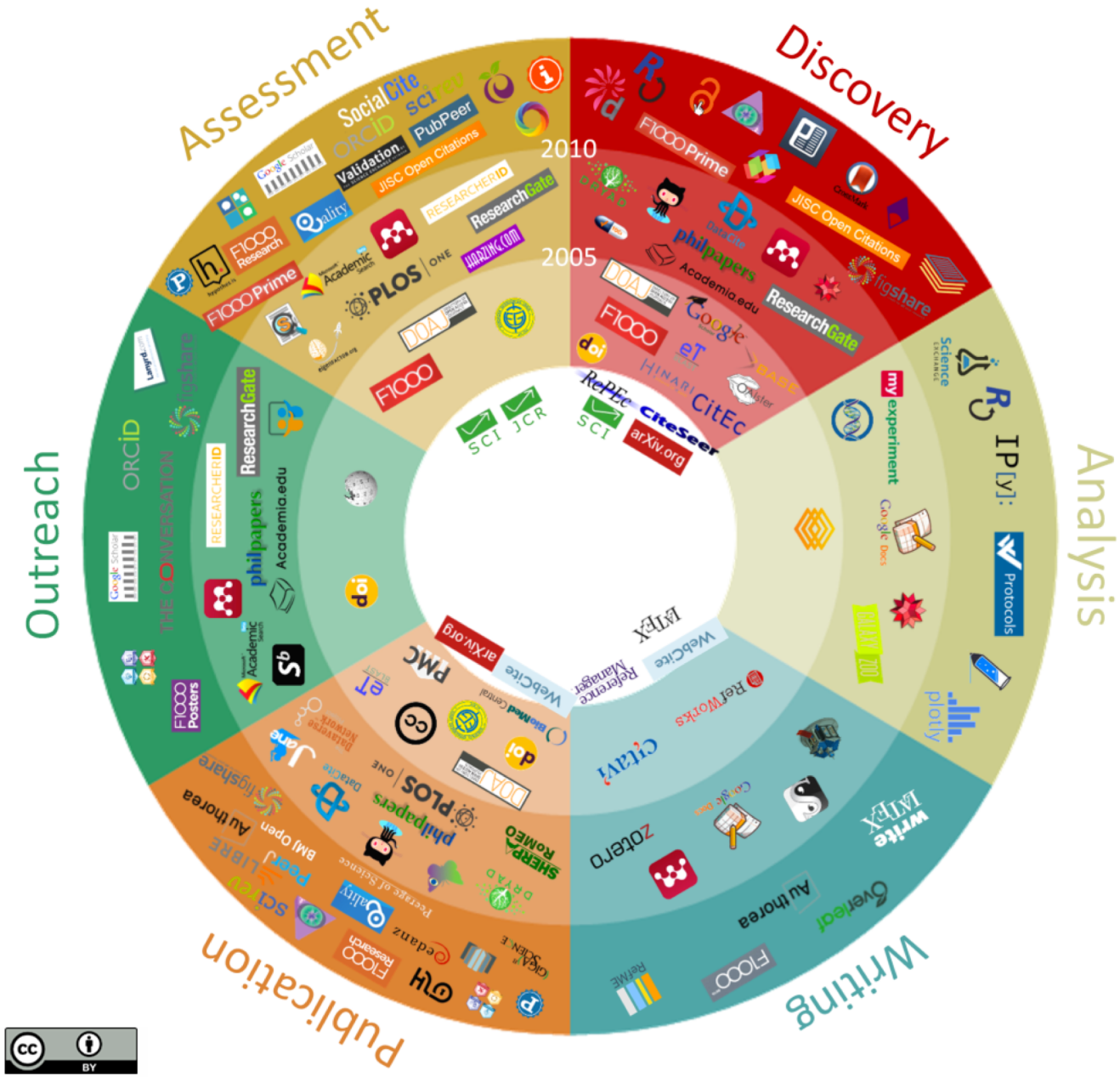
Tools for:

- Documentation
- Collaboration
- Data storage and backup
- Sharing

+ Tools that work together

+ Tools that are used in your discipline

Carnegie Mellon University



(logo's excluded)

Crowdsourced list of tools, now 400+

http://bit.ly/innoscholcomm-list

Bosman & Kramer. 2015

Tools at CMU (many of which provided by the Libraries)

Carnegie Mellon University

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

Carnegie Mellon University

Computational Reproducibility, cont.

Document all steps

Comment with textual statements

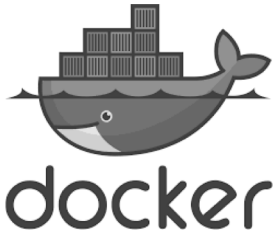
Save both raw and intermediate data

Document all dependencies

Use relative paths

- ../picture.jpg
- /Users/Huajin/Pictures/picture.jpg

Use version control to track changes



Carnegie Mellon University

Computational Reproducibility, cont.

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

Documentation of data analysis code, results, and visualization all in one place

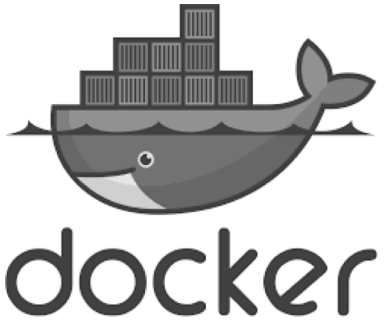
Enable others to understand and reproduce results

Example: code for LIGO project - discovery of gravitational waves
https://losc.ligo.org/s/events/GW150914/GW150914_tutorial.html

Jupyter Notebook and R Markdown

Carnegie Mellon University

Run notebooks interactively online



https://mybinder.org

Carnegie Mellon University

Contact Us

Open Science & Data Collaborations
openscience@andrew.cmu.edu

Data Services
data@cmu.libanswers.com

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

@cmulibraries

@CMULibraries

@CMULibraries