Reproducibility in Data Science

Jessica Minnier, PhD

Assistant Professor of Biostatistics

OHSU-PSU School of Public Health

Knight Cancer Institute Biostatistics Shared Resource

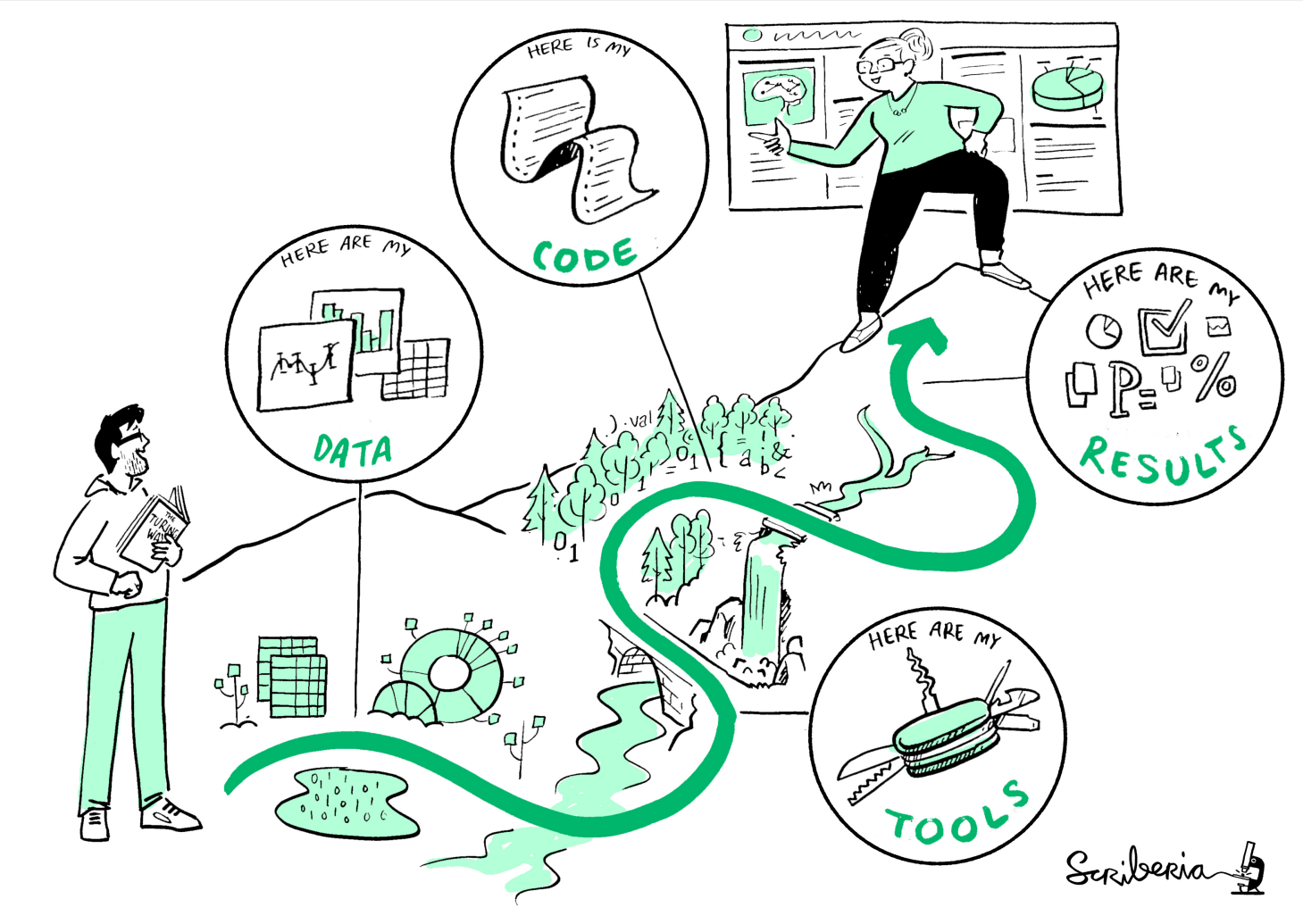
Oregon Health & Science University

HIP 523 Computerized Data Management

bit.ly/hip-repro

datapointier

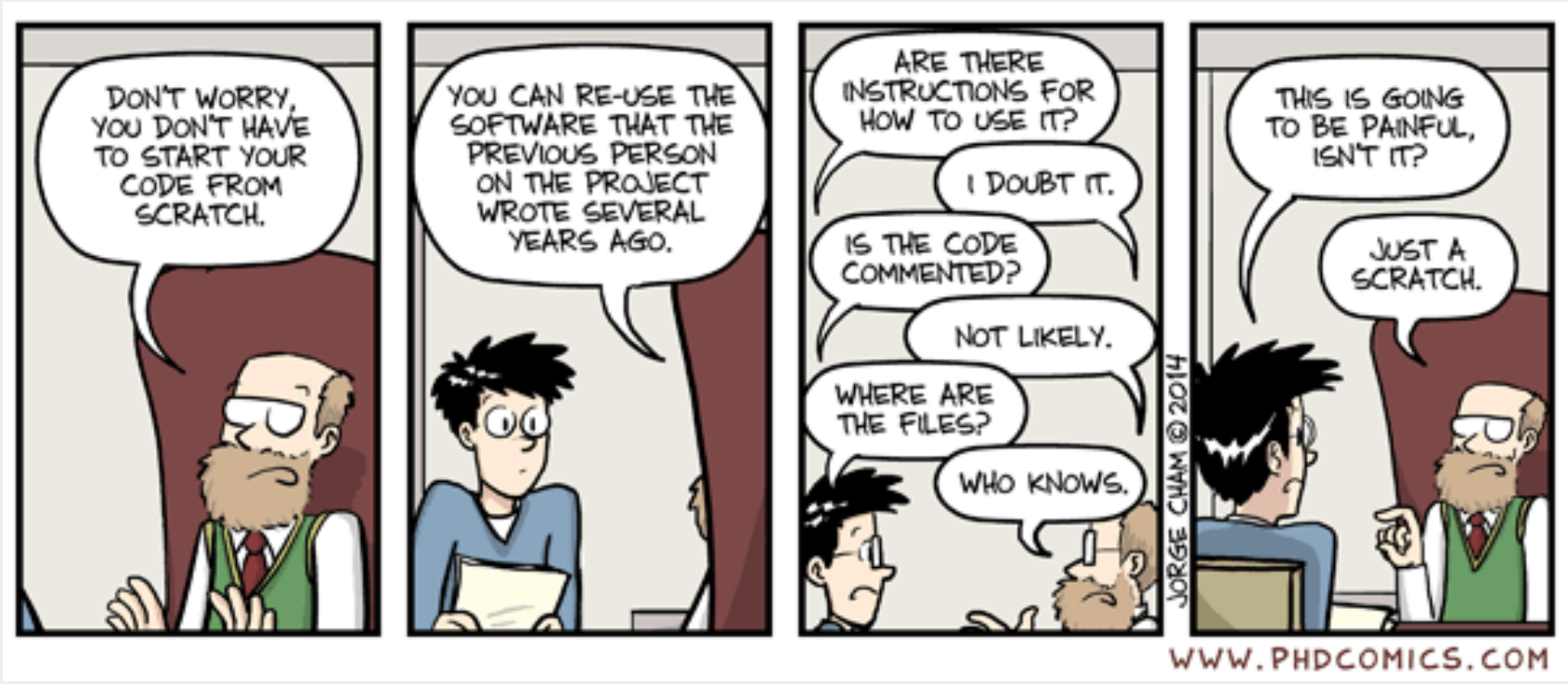
Introduction



Illustrations from the Turing Way book dashes; This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence

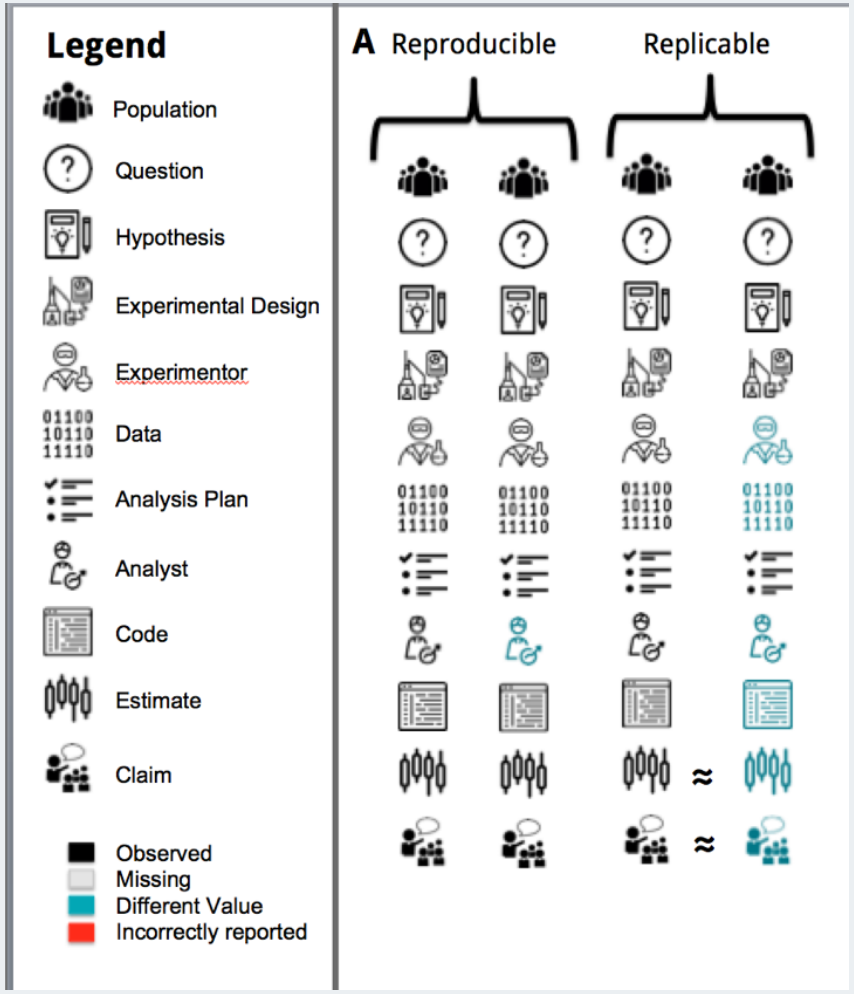
Goals

- Define reproducible research
- Discuss current issues surrounding reproducibility
- Important components of reproducibility
- Relevance to data science and analysis
- Introduction to version control
- Introduction to literate programming in R



PhD Comics by Jorge Cham

What is Reproducible Research?

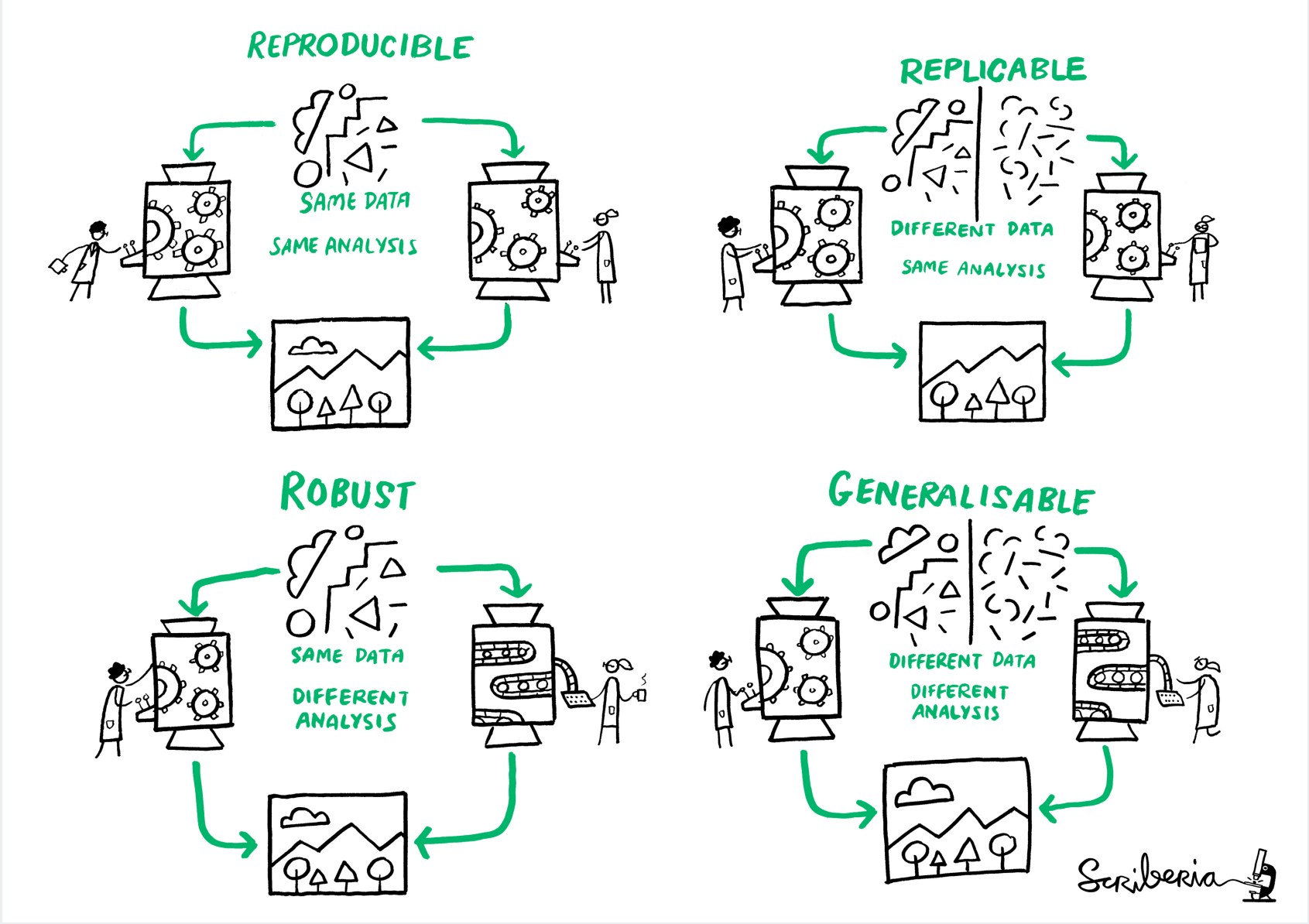


Patil, Peng, & Leek 2016

Reproducibility: ability to recompute data analytic results given the data set and knowledge of the data analysis pipeline

Replicability: the chance that an independent experiment targeting the same scientific question will produce a consistent result

-- Peng (2011) "Reproducible research in computational science" and Leek and Peng (2015) "Opinion: Reproducible research can still be wrong: Adopting a prevention approach"



Illustrations from the Turing Way book dashes; This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence

Reproducible = Replicable + Transparant

Sometimes, replicable is used instead of reproducible.

> Research results are **replicable** if there is sufficient information available for independent researchers to make the same findings using the same procedures.
>
> In **computational sciences this means**: the data and code used to make a finding are available and they are sufficient for an independent researcher to recreate the finding.
>
> In practice, research needs to be **easy for independent researchers to reproduce**.

-- Ball and Medeiros (2012); King (1995) from Gandrud (2013)

> **Replicability** has been a key part of scientific inquiry from perhaps the 1200s. It has even been called the "demarcation between science and non-science."

-- Gandrud (2013) book "Reproducible Research with R and R Studio" and references therein, including Roger Bacon's "Opera quaedam hactenus inedita Vol. 1" from 1267

What are the different kinds of reproducibile research?

> Victoria Stodden, a prominent scholar on this topic, has identified some useful distinctions in reproducible research:
>
> **Computational reproducibility**: when detailed information is provided about code, software, hardware and implementation details.
>
> **Empirical reproducibility**: when detailed information is provided about non-computational empirical scientific experiments and observations. In practice this is enabled by making data freely available, as well as details of how the data was collected.
>
> **Statistical reproducibility**: when detailed information is provided about the choice of statistical tests, model parameters, threshold values, etc. This mostly relates to pre-registration of study design to prevent p-value hacking and other manipulations.

ROpenSci Reproducibility Guide

Spectrum of Research

> Stodden et al. (2013) place computational reproducibility on a spectrum with five categories that account for many typical research contexts:
>
> - Reviewable research: The descriptions of the research methods can be independently assessed and the results judged credible. (This includes both traditional peer review and community review, and does not necessarily imply reproducibility.)
> - Replicable research: Tools are made available that would allow one to duplicate the results of the research...
> - Confirmable research: ... main conclusions of the research can be attained independently without the use of software provided by the author ...
> - Auditable research: Sufficient records (including data and software) have been archived ...
> - Open or Reproducible research: Auditable research made openly available. This comprised well-documented and fully open code and data that are publicly available that would allow one to (a) fully audit the computational procedure, (b) replicate and also independently reproduce the results of the research, and (c) extend the results or apply the method to new problems.

ROpenSci Reproducibility Guide

Reproducibility in Data Science

> "Reproducibility is important because it is the **only thing that an investigator can guarantee about a study.**"
>
> "**a study can be reproducible and still be wrong**"
>
> "These days, with the complexity of data analysis and the subtlety of many claims (particularly about complex diseases), reproducibility is pretty much the only thing we can hope for. Time will tell whether we are ultimately right or wrong about any claims, but **reproducibility is something we can know right now**."
>
> "By using the word reproducible, I mean that the **original data (and original computer code) can be analyzed (by an independent investigator) to obtain the same results of the original study**. In essence, it is the notion that the data analysis can be successfully repeated. Reproducibility is particularly important in large computational studies where the data analysis can often play an outsized role in supporting the ultimate conclusions."

-- Roger Peng's 2014 blog post on Simply Statistics "The Real Reason Reproducible Research is Important" also see Peng (2011) "Reproducible research in computational science"

"Your primary collaborator is yourself 6 months from now, and your past self doesn't answer emails."

-- Software Carprentry workshops

Early notions of reproducibility: "Claerbout's Principle"

> 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 software development environment and the complete set of instructions which generate the figures.
>
> It takes **some effort to organize your research to be reproducible**.
>
> We found that although the effort seems to be directed to helping other people stand up on your shoulders, the principal beneficiary is generally the author herself.
>
> This is because time turns each one of us into another person, and by making effort to communicate with strangers, we help ourselves to communicate with our **future selves**.

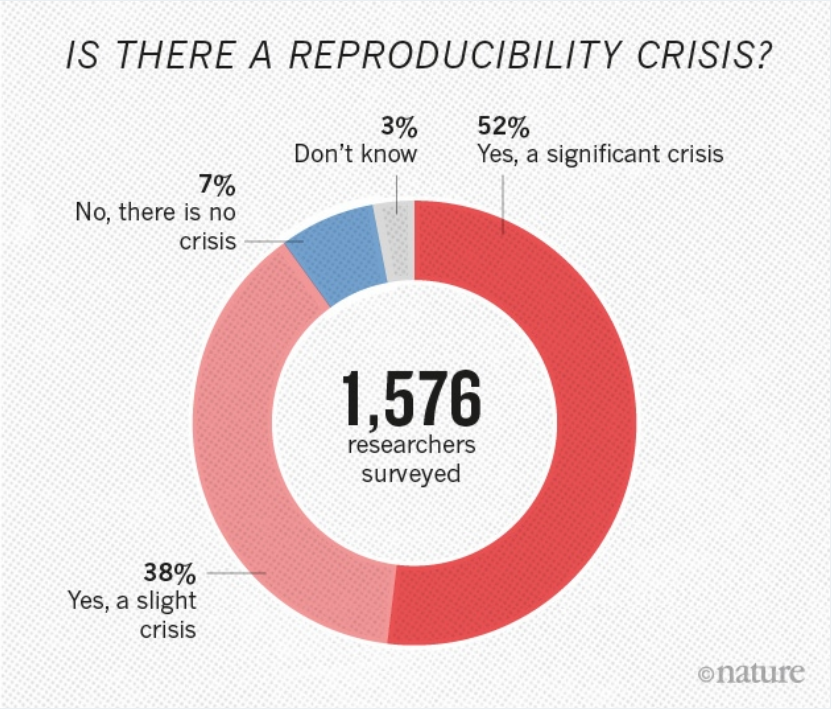
- Claerbout and Karrenbach (1992) "Electronic documents give reproducible research a new meaning"
- Buckheit and Donoho (1995) "Wavelab and reproducible research"
- Schwab, Karrenbach, and Claerbout (2000) "Making scientific computations reproducible"
- De Leeuw (2001) "Reproducible research. the bottom line"

> (Jon F. Claerbout is the Cecil Green Professor Emeritus of Geophysics at Stanford University. He was one of the first scientists to emphasize that computational methods threaten the reproducibility of research unless open access is provided to both the data and the software underlying a publication.)

Current Issues and Discussion

Nature series "Challenges in Irreproducible Research"

May 25, 2016 Editorial "Reality check on reproducibility"



> One-third of survey respondents report that they have taken the initiative to improve reproducibility. The simple presence of another person ready to **question** whether a data point or a sample should really be excluded from analysis can help to cut down on cherry-picking, conscious or not. A couple of senior scientists have set up **workflows** that avoid having a single researcher in charge of preparing images or collecting results. Dozens of respondents reported steps to make **better use of statistics, randomization or blinding**. One described an institution-level initiative to **teach scientists computer tools so they could share and analyse data collaboratively**. Key to success was making sure that their **data-management system also saved time**.

How to Make More Published Research True

Ioannidis (2014) "How to Make More Published Research True" in PLOS Medicine, the author writes a follow up to Ioannidis (2005) "Why most published research findings are false."

He suggests reproducibility as one key component to the cause:

> "To make more published research true, practices that have improved credibility and efficiency in specific fields may be transplanted to others which would benefit from them—possibilities include
>
> - the adoption of large-scale collaborative research;
> - **replication culture**;
> - registration; sharing; **reproducibility practices**;
> - better statistical methods;
> - standardization of definitions and analyses;
> - more appropriate (usually more stringent) statistical thresholds; and
> - improvement in study design standards, peer review, *reporting and dissemination of research*, and training of the scientific workforce."

Availability of code in peer-reviewed journals

Stodden, Guo, and Ma (2013) "Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals"

Table 1. Code Availability in the Journal of the American Statistical Association.

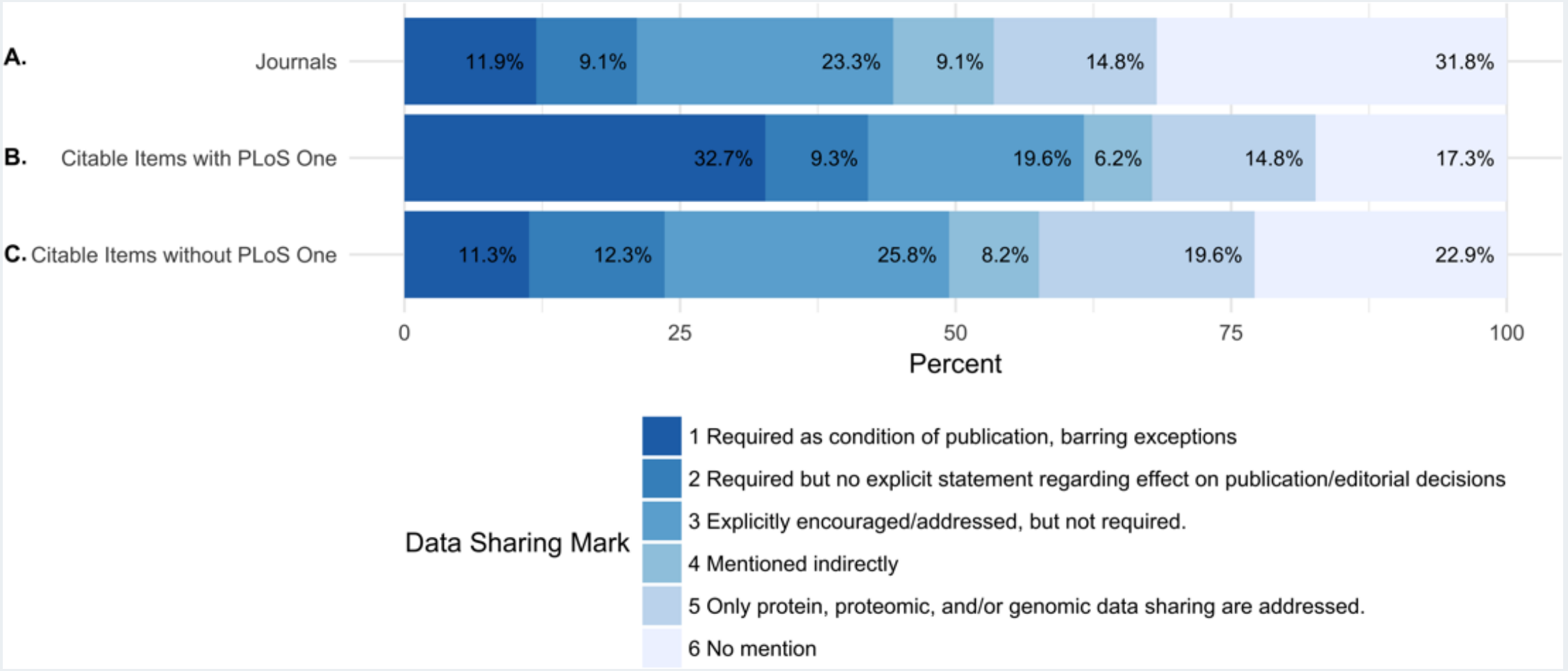
JASA June	Computational Articles	Code Publicly Available
1996	9 of 20	0%
2006	33 of 35	9%
2009	32 of 32	16%
2011	29 of 29	21%

doi:10.1371/journal.pone.0067111.t001

- Studied change in policies between 2011-2012
- Open data and code policy adoption vs.impact factor and publisher
- Higher impact journals more likely to have open data and code policies
- Scientific societies more likely to have open data and code policies than commercial publishers.

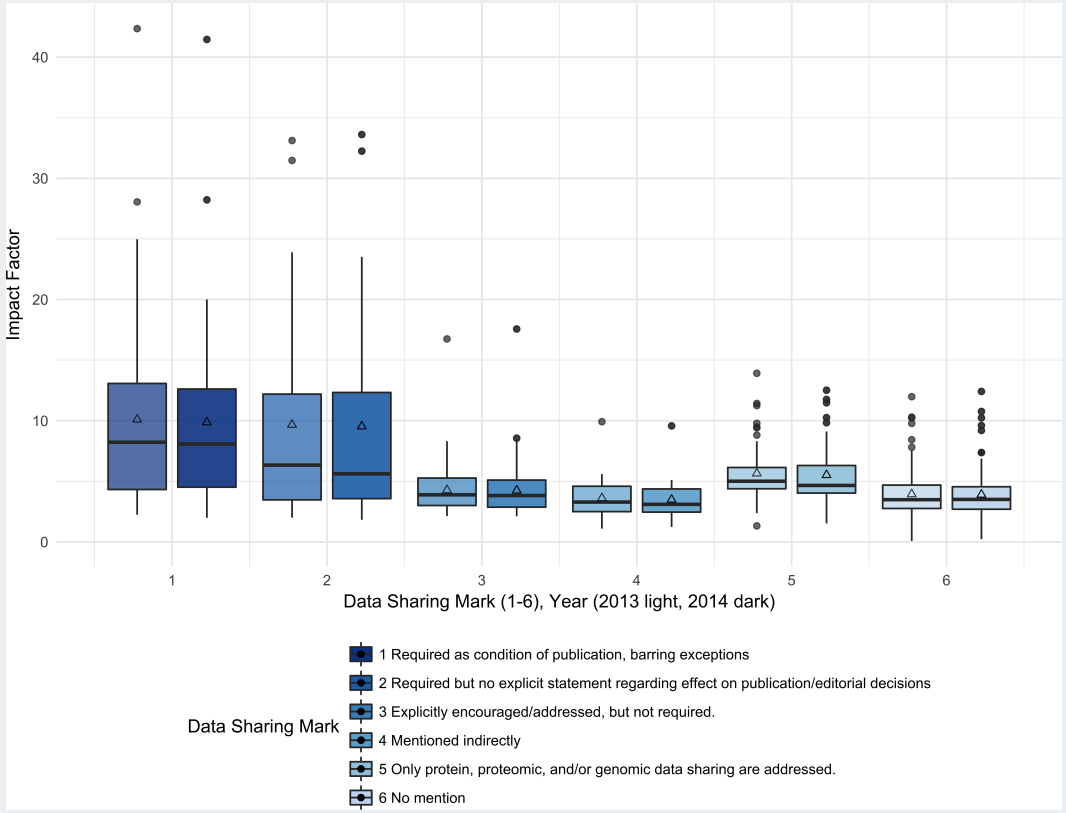
Data sharing requirements in journals

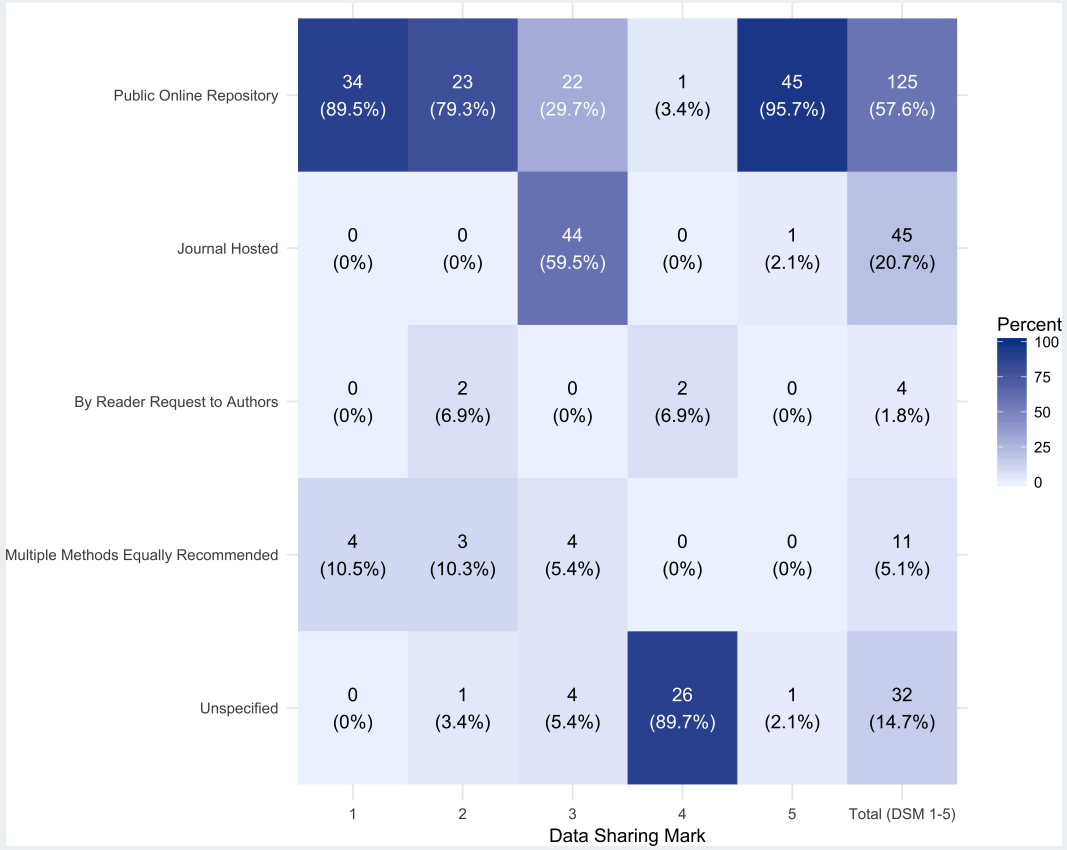
Vasilevsky, Minnier, Haendel, and Champieux (2017) "Reproducible and reusable research: are journal data sharing policies meeting the mark?"



Data sharing requirements in journals

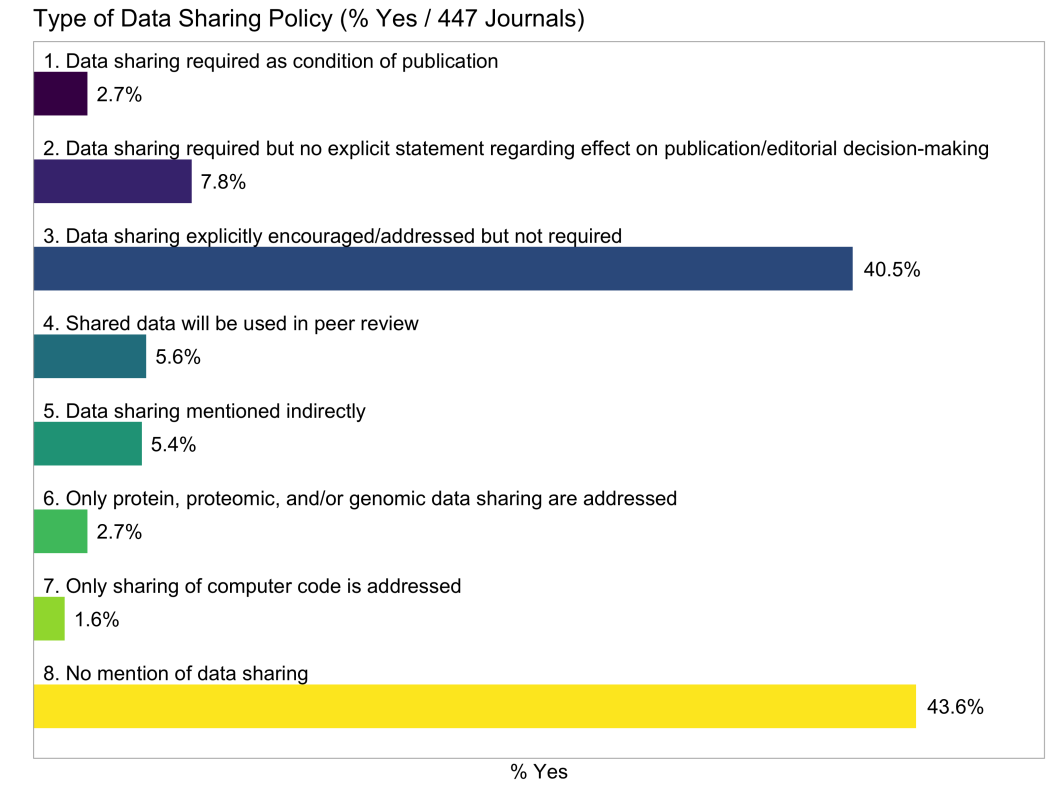
Vasilevsky, Minnier, Haendel, et al. (2017) "Reproducible and reusable research: are journal data sharing policies meeting the mark?"

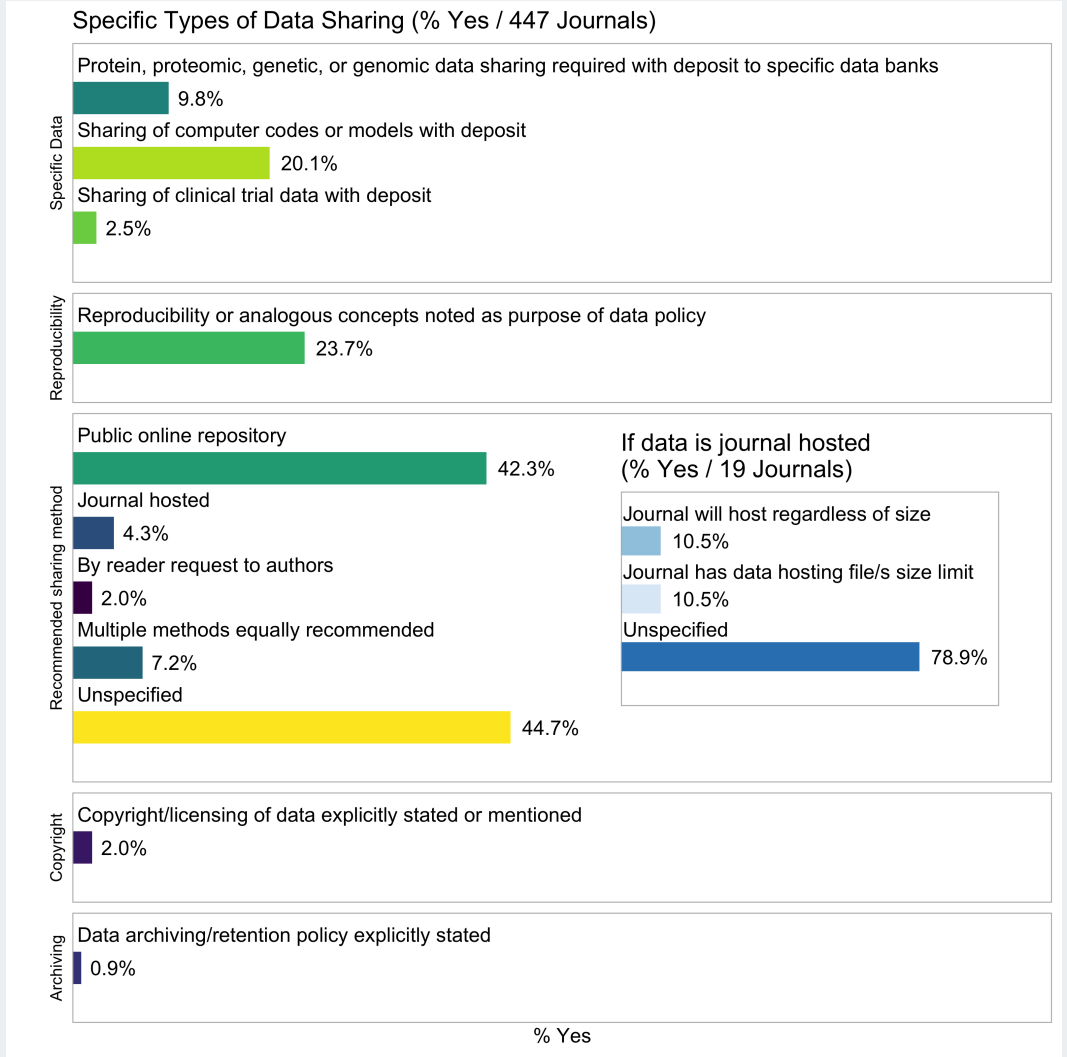




Data sharing requirements in journals

Resnik, Morales, Landrum, Shi, Minnier, Vasilevsky, and Champieux (2019) "Effect of Impact Factor and Discipline on Journal Data Sharing Policies": Stratified sampling of scientific disciplines



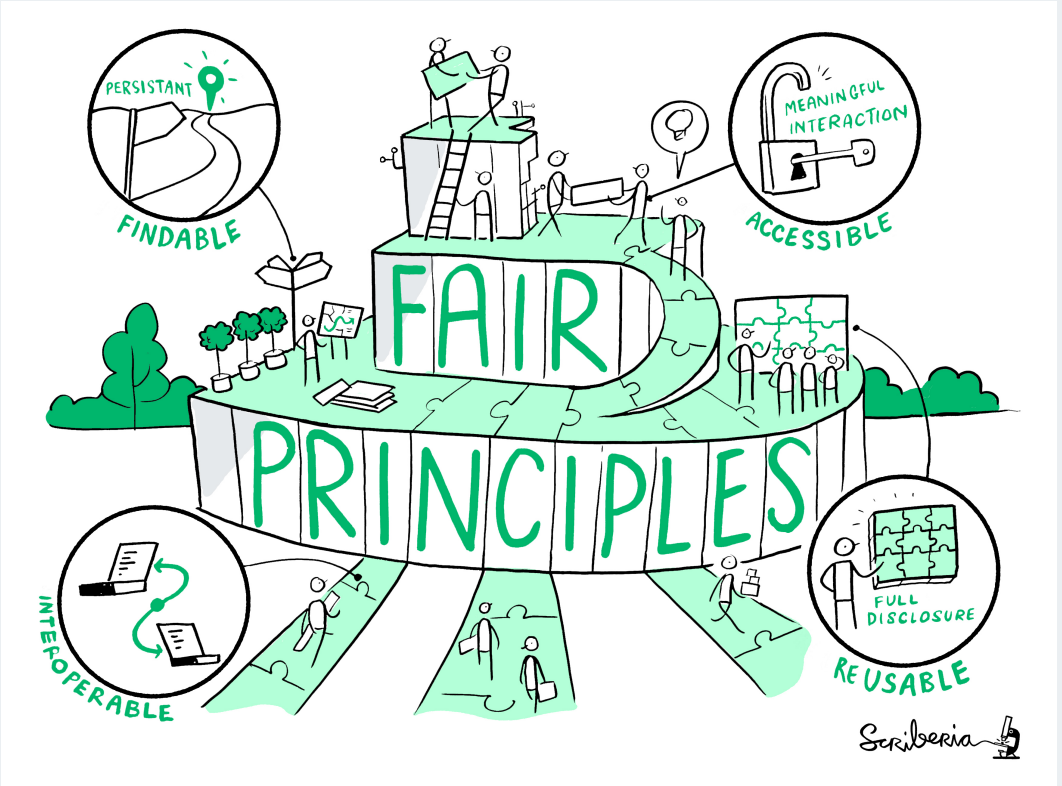


Data management: FAIR principles (Findable, Accessible, Interoperable, Reusable)

- Make data available in trusted data repository
- Include comprehensive metadata
- Store in open (non-proprietary) formats
- Attach a permanent identifier (i.e. DOI)

In Biomedical research, open access can be difficult/unethical.

Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, Blomberg, Boiten, da Silva Santos, Bourne, and et al. (2016)



NIH requirements for grants (beginning Jan 2016, updated Jan 2019)

"Enhancing Reproducibility through Rigor and Transparency"

1. *Rigor of the Prior Research*
    - "describe the general strengths and weaknesses of the prior research being cited by the investigator as crucial to support the application."
    - experimental design/power of prior studies used for hypothesis generation, weaknesses include different populations/species, unblinded, not adjusting for confounders
2. *Rigorous Experimental Design*
3. *Consideration of Sex and Other Relevant Biological Variables*
    - "sex is a biological variable that is frequently ignored in animal study designs and analyses"
4. *Authentication of Key Biological and/or Chemical Resources*
5. *Implementation*

NIH "Rigor and Reproducibility" Policy

Note: Most of this is in regards to the science, design of experiment, chemical and biological methods. **Essentially no language describing reproducibility of analyses or data management for data or results generated by the grant.**

Journals unite with NIH to encourage reproducibility

- Principles and Guidelines for Reporting Preclinical Research
- NIH held a joint workshop in June 2014 with the Nature Publishing Group and Science on the issue of reproducibility and rigor of research findings
- A video/slide presentation about this topic and how it applies to grant applications and peer review can be found here: NIH Policy Rigor for Reviewers Presentation

NIH Principles and Guidelines for Reporting Preclinical Research

Journals should aim to facilitate the interpretation and repetition of experiments as they have been conducted in the published study.

- include policies for statistical reporting in information to authors
- no limits or generous limits for methods sections
- should use a **checklist** during editorial processing to ensure the reporting of key methodological and analytical information to reviewers and readers
- Data and material sharing
  - at the minimum, data sets on which the conclusions of the paper rely must be made available upon request (where ethically appropriate) during consideration of the manuscript (by editors and reviewers) and upon reasonable request immediately upon publication.
  - Recommend deposition of data sets in public repositories, where available
  - Encourage presentation of all other data values in machine readable format
  - Encourage sharing of software and require at the minimum a statement in the manuscript describing if software is available and how it can be obtained.
- journal assumes responsibility to consider publication of refutations of that paper
- best practice guidelines for image based data and a description of biological material with enough information to uniquely identify the reagents
- do not obviate need for biological replication/validation

Checklist: authors required to report

from NIH Guidelines & Landis, Amara, Asadullah, Austin, Blumenstein, Bradley, Crystal, Darnell, Ferrante, Fillit, and et al. (2012) "A call for transparent reporting to optimize the predictive value of preclinical research"

- *Standards:* community-based standards (nomenclature etc) where applicable
- *Replicates:* report how often each experiment was performed, whether results were substantiated by repetition under a range of conditions. Sufficient information about sample collection must be provided to distinguish between independent biological data points & technical replicates.
- *Statistics:* Require statistics to be fully reported in the paper, including statistical test used, exact value of N, definition of center, dispersion & precision measures
- *Randomization:* (yes/no) & method, at a minimum for all animal experiments
- *Blinding:* were experimenters blind to group assignment & outcome assessment, at a minimum for all animal experiments.
- *Sample-size (SS) estimation:* was an appropriate SS computed during study design & include method; if no power analysis, how was SS determined?
- *Inclusion and exclusion criteria:* criteria used for exclusion of any data or subjects. Include any similar experimental results that were omitted from reporting for any reason, esp. if results don't support main findings of study; describe any outcomes/conditions that are measured/used & not reported in results section.

Reproducible research and Biostatistics (the journal)

> Authors should submit the following:
>
> 1. A “main” script which directs the overall analysis. This script may load data, other software, and call the necessary functions for conducting the analysis described in the article.
> 2. Other required code files, presumably called from the “main” script file.
> 3. External data or auxiliary files containing the analytic data sets or other required information.
> 4. A “target” file (or files) containing the results which are to be reproduced. Such a file could consist of an ASCII text file containing numerical results or a PDF file containing a figure. This will aid in the comparison of computed results with published results.
>
> Although not required, authors are encouraged to use literate programming tools [...]

-- Peng (2009) "Reproducible research and *Biostatistics*"

> Our reproducible research policy is for papers in the journal to be kite-marked D if the data on which they are based are freely available, C if the authors' code is freely available, and R if both data and code are available, and our Associate Editor for Reproducibility is able to use these to reproduce the results in the paper. Data and code are published electronically on the journal's website as Supplementary Materials.

Nature series on "Challenges in Irreproducible Research"

Nature has a website containing editorials, features, news, and articles on various topics related to reproducible research: Nature special: Challenges in Irreproducible Research

Including

- a checklist for authors of Nature papers described in the 2013 announcement "Announcement: Reducing our irreproducibility"
- R Nuzzo (2014) Nature News Feature "Scientific method: Statistical errors" on "P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume."

"Enhancing Reproducibility for Computational Methods"

Stodden, McNutt, Bailey, Deelman, Gil, Hanson, Heroux, Ioannidis, and Taufer (2016) in *Science* Policy Forum

- Share data, software, workflows, and details of computational environment in open trusted repositories
- Persistent links, permanent identifiers for data, code, digital artifacts upon which the results depend
- Enable credit for shared digital scholarly objects with citations
- Adequately document to facilitate reuse
- Use Open Licensing
- Journals should conduct a reproducibility check
- Funding agencies should instigate new research programs and pilot studies

Barriers:

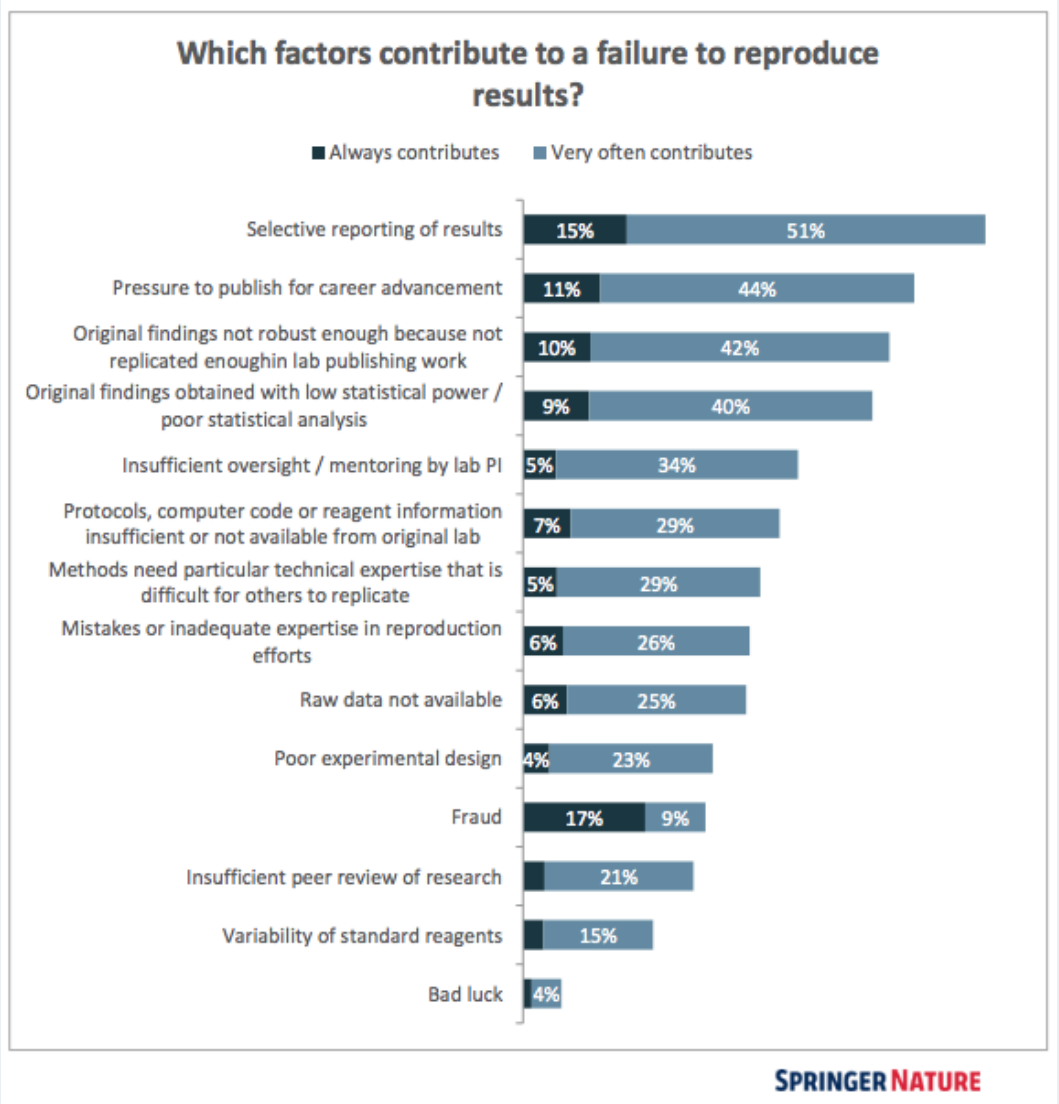
- Human subject data or proprietary code, but creative solutions should be implemented
- Journals and funders should reward reproducibility
- Students need to be taught reproducible methodology
- Need to make these efforts common place

Barriers to Reproducibility/Replicability

- Selective reporting of results
- Pressure to publish
- Not robust results
- Protocols, computer code, or reagent info insufficient or not available

Barriers to implementing R&R

- insufficient funding/time
- insufficient training (and funding for training)
- journals/funders do not have clear guidelines or priorities
- open access not encouraged or funded



Nature Survey 2017 results, n=480 authors

Reproducibility in Practice

Data management

- Write a data management plan prior to research, should cover entire project lifestyle
- Follow the FAIR principles when possible
- Document, document, document! (metadata, create Readme file)
- For open data, attach a licence or rights waiver
- DOI or Zenodo identifiers to capture open data at specific time points
- For private/patient data, access should be granted based on data sharing agreement; version control and data storage principles should still be followed for internal reproducibility and traceability; code can still be shared

> "The basics of RDM [Research Data Management] that should be applied to every research project include: i) storing data carefully and securely (according to the appropriate standards in the case of sensitive data), ii) backing up frequently and in at least two separate locations, and iii) using a file naming convention so that others within and outside a project can understand a file's content."

Kunzmann et al (PeerJ as preprint soon) "Realistic and Robust Reproducible Research for Biostatistics"

Resources:

- DMPonline by the Digital Curation Centre for writing data management plans
- FAQs for Enabling FAIR Data
- LabKey, open-source specimen and data management platform

Software, workflow, and dependency management

Problems:

- Software changes, new versions are released and older code breaks.
- Files are removed or moved and all the code breaks.
- One file is updated but the rest of the code/files are not updated.
- You forgot which files depend on which other files, or what has been changed.

Solutions for Workflow:

- **Automation** streamlined automation and documentation of the research process, e.g. editing files, moving input and output between different parts of your workflow, and compiling documents for publication (shell programs, `make`); in python: snakemake, in R: drake

Solutions for Dependency:

- **Preservation** of the environment (code, data, software) with virtualisation, i.e. Docker, VMware, VirtualBox, packrat for R package management

ROpenSci Reproducibility Guide

Archiving and citability

- Long-term availability of code and data
- Popular repositories for data: Zenodo, Dryad, Genbank, GEO
- Popular repositories for code: GitHub, bitbucket, Bioconductor, CRAN PyPI
- Citability: Digital Object Identifier (DOI) for code and data

Type	Archiving location	Pros	Cons
Code/software	Git repositories (github, gitlab, bitbucket)	continuous, version-tracked improvements	no DOI (but see below), no commitment to long-term storage
	General purpose repositories (Zenodo, Dryad, etc.)	DOI, stable version, some have integration with git repositories	Specific code is less findable, and may not be up-to-date
Data	Discipline-specific repos (ENA, ArrayExpress, etc.)	DOI, standardised, often APIs in place to push/pull data, minimal metadata must be provided	Only specific data types or formats may be supported
	General purpose repositories (Zenodo, Dryad, etc.)	DOI, wide range of data types/formats supported	Specific data content is less findable

Table 3. Pros and cons for code and data archiving in different archiving locations. DOI = digital object identifier.

Kunzmann et al (PeerJ as preprint soon) "Realistic and Robust Reproducible Research for Biostatistics"

Literate Programming

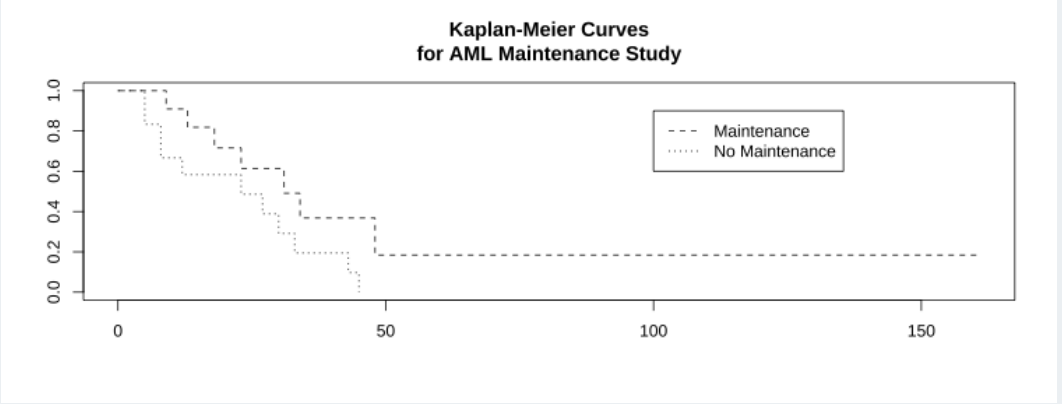
> Literate programming is an approach to programming introduced by Donald Knuth in which a program is given as an explanation of the program logic in a natural language, such as English, interspersed with snippets of macros and traditional source code, from which a compilable source code can be generated. (Knuth, 1984)

Current implementations weave code, documentation, results, and other output in the same document.

Examples: knitr (for R), Sweave; Jupyter notebooks (for python) SASweave, Statrep (for SAS); StatWeave (for STATA)

This is knitr (presentation made with knitr+RStudio):

```
library(survival)
leukemia.surv <- survfit(Surv(time, status) ~ x,
plot(leukemia.surv, lty = 2:3)
legend(100, .9, c("Maintenance", "No Maintenance
title("Kaplan-Meier Curves\nfor AML Maintenance
```



Version Control

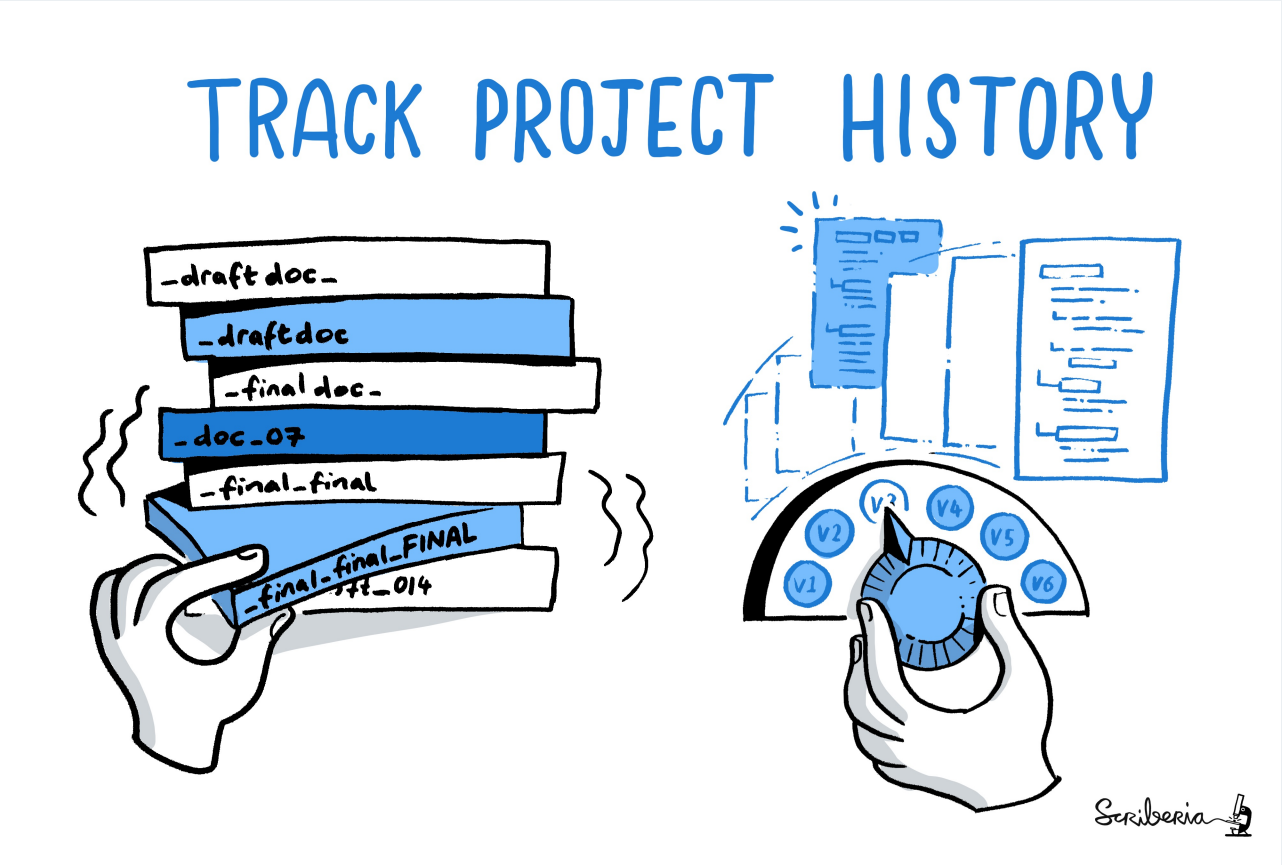
- A system that tracks and records changes to file(s)
- Allows for collaborative work on code
- i.e. **git**: "a lightweight yet robust framework that is ideal for managing the full suite of research outputs such as data sets, statistical code, figures, lab notes, and manuscripts." Ram (2013) "Git can facilitate greater reproducibility and increased transparency in science."
- online git repositories include: GitHub, bitbucket, GitLab
- more sophisticated/lightweight than track changes in Word/google docs

Resources:

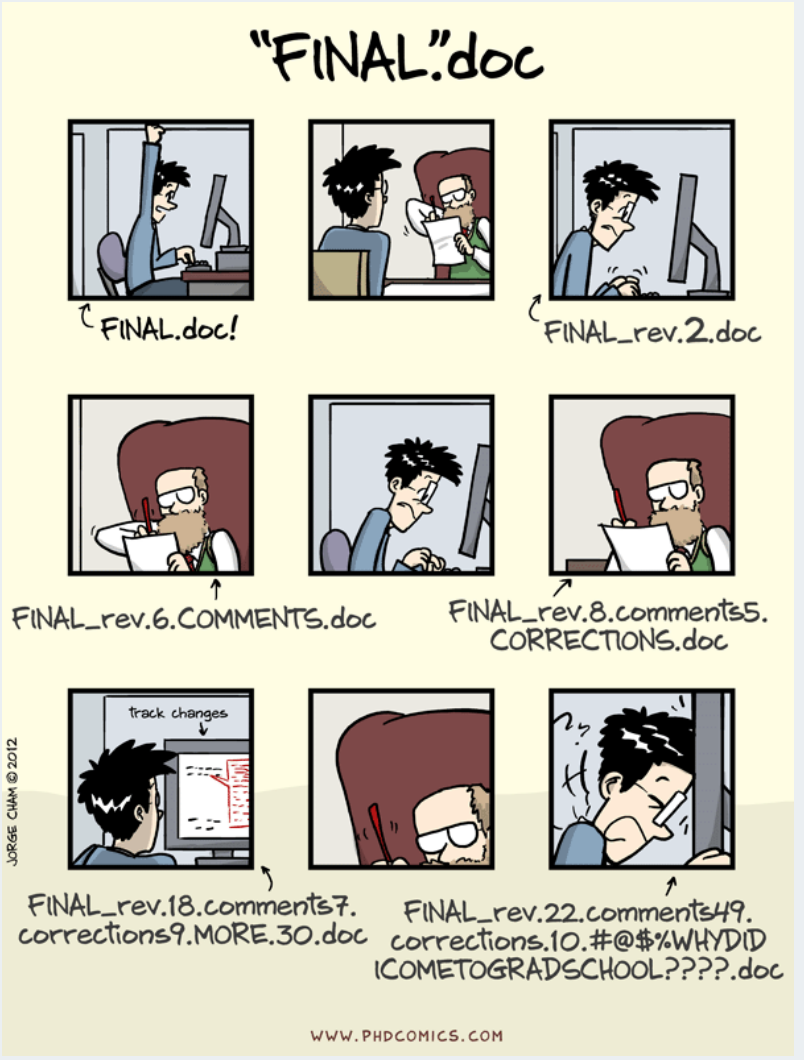
- Software Carpentry "Version Control w/Git"
- Github tutorials
- Happy Git and GitHub for the useR - Jenny Bryan, Jim Hester

Illustrations from the Turing Way book dashes; This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence

Why Use Version Control?



Illustrations from the Turing Way book dashes; This image was created by Scriberia for The Turing Way community and is used under a CC-BY licence



PhD Comics by Jorge Cham

Why Use Version Control?

> Have you ever:
>
> - Made a change to code, realised it was a mistake and wanted to revert back?
> - Lost code or had a backup that was too old?
> - Had to maintain multiple versions of a product?
> - Wanted to see the difference between two (or more) versions of your code?
> - Wanted to prove that a particular change broke or fixed a piece of code?
> - Wanted to review the history of some code?
> - Wanted to submit a change to someone else's code?
> - Wanted to share your code, or let other people work on your code?
> - Wanted to see how much work is being done, and where, when and by whom?
> - Wanted to experiment with a new feature without interfering with working code?
>
> In these cases, and no doubt others, **a version control system should make your life easier.**

Stack Overflow question: Why should I use version control?

Make a Plan

Checklists for Reproducibility

- "Checklists work to improve science" - Nature editorial, April 2018
- In 2013, *Nature* announced that authors submitting manuscripts to *Nature* journals would need to complete a checklist addressing key factors underlying irreproducibility for reviewers and editors to assess during peer review
- Survey of researchers who had published in a Nature journal between July 2016 and march 2017

> Of the 480 who responded, 49% thought that the checklist had improved the quality of research published in Nature (15% disagreed); 37% thought the checklist had improved quality in their field overall (20% disagreed).

Checklists can be useful for computational research in general.

An excellent example: ROpenSci's Reproducibility Checklist

"Ten Simple Rules for Reproducible Computational Research"

- Rule 1: For Every Result, Keep Track of How It Was Produced
- Rule 2: Avoid Manual Data Manipulation Steps
- Rule 3: Archive the Exact Versions of All External Programs Used
- Rule 4: Version Control All Custom Scripts
- Rule 5: Record All Intermediate Results, When Possible in Standardized Formats
- Rule 6: For Analyses That Include Randomness, Note Underlying Random Seeds
- Rule 7: Always Store Raw Data behind Plots
- Rule 8: Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
- Rule 9: Connect Textual Statements to Underlying Results
- Rule 10: Provide Public Access to Scripts, Runs, and Results

Sandve, Nekrutenko, Taylor, and Hovig (2013)

Example ideals for a computational group

- Literate programming (use R Markdown)
- Agree on best practices for writing code (i.e. ROpenSci's Reproducibility & Writing Code Guide and "Best Practices for Scientific Computing" Wilson, Aruliah, Brown, Hong, Davis, Guy, Haddock, Huff, Mitchell, Plumbley, and others (2014)
- Agree on folder structure, naming conventions
- Use html web-based output
  - see Matthew Shotwell's slides
  - nearly universal compatibility
  - persistent
  - images handled more naturally
  - avoid changes (i.e. in Word)
- Use `make` files to rerun analyses when certain files change
- Version/revision control systems such as git for all files
- Version control of data, stored with backup in persistent location
- Store metadata, readme files
- Software/package versions need to be maintained (i.e. packrat for R)

Brief Tutorials

R Markdown/Knitr (literate programming)

- BERD workshop "Reproducible Reports with R Markdown" slides: bit.ly/berd_rmd and recording link and other info is found on github
- Example, new git project in Rstudio
- Example, cloning a github repo

Git/github (version control)

- Using git in Rstudio (for more instruction, highly recommend happygitwithr.com)
- Explore an example GitHub repo OHSU-Library/Biomedical_Journal_Data_Sharing_Policies

Resources

Recommended Books

- Stodden, Victoria, Friedrich Leisch, and Roger D. Peng, eds. Implementing reproducible research. CRC Press, 2014.
- Gandrud, Christopher. Reproducible Research with R and R Studio. CRC Press, 2013.
- Xie, Yihui. Dynamic Documents with R and knitr. Vol. 29. CRC Press, 2013.

Online classes

- Karl Broman's class "Tools for Reproducible Research" at UWisconsin-Madison http://kbroman.org/Tools4RR/
- "Reproducible Research" by Johns Hopkins on Coursera (Peng, Leek, Caffo) https://www.coursera.org/learn/reproducible-research
- Learn git: https://try.github.io/levels/1/challenges/1

NIH Rigor & Reproducibility Resources

- Website: http://grants.nih.gov/reproducibility/index.htm
- FAQs: http://grants.nih.gov/reproducibility/faqs.htm
- NIH Training Module: https://grants.nih.gov/reproducibility/module_1/presentation.html

Websites/slides/blogs

- ROpenSci's "Reproducibility in Science" guide including the reproducibility checklist
- Victoria Stodden's list of talks on various topics from "Reproducibility: Breakin' it Down" to "Legal Issues in Reproducible Research"
- Matthew Shotwell's slides (2011) "Approaches and Barriers to Reproducible Practices in Biostatistics"
- M Shotwell and JM Álvarez' slides "Approaches and Barriers to Reproducible Practices in Biostatistics" and "Barriers to Reproducible Research and a Web-Based Solution"
- ROpenSci's blog post "Reproducible research is still a challenge" by R. FitzJohn, M. Pennell, A. Zanne, W. Cornwell, June 9, 2014, describes the experience of running an example analysis
- Stodden (2014) "What scientific idea is ready for retirement?"
- StackOverflow question "Why should I use version control?"
- Karl Broman's class "Tools for Reproducible Research" resource page and "Why Reproducibility is Hard"
- CRAN's task view on Reproducible Research
- Frank Harrell's wiki on statistical reporting

References 1

Ball, R. and N. Medeiros (2012). "Teaching Integrity in Empirical Research: A Protocol for Documenting Data Management and Analysis". In: *The Journal of Economic Education* 43.2, pp. 182-189.

Buckheit, J. B. and D. L. Donoho (1995). *Wavelab and reproducible research*. Springer.

Claerbout, J. and M. Karrenbach (1992). "Electronic documents give reproducible research a new meaning". In: *Proc. 62nd Ann. Int. Meeting of the Soc. of Exploration Geophysics.* , pp. 601-604.

De Leeuw, J. (2001). "Reproducible research. the bottom line". In: *Department of Statistics, UCLA*.

Gandrud, C. (2013). *Reproducible Research with R and R Studio*. CRC Press.

Ioannidis, J. (2005). "Why most published research findings are false". In: *PLoS Med* 2.8, p. e124.

Ioannidis, J. P. (2014). "How to make more published research true".

King, G. (1995). "Replication, replication". In: *PS: Political Science & Politics* 28.03, pp. 444-452.

Knuth, D. E. (1984). "Literate programming". In: *The Computer Journal* 27.2, pp. 97-111.

References 2

Landis, S. C, S. G. Amara, K. Asadullah, et al. (2012). "A call for transparent reporting to optimize the predictive value of preclinical research". In: *Nature* 490.7419, pp. 187-191. ISSN: 1476-4687. DOI: 10.1038/nature11556. URL: http://dx.doi.org/10.1038/nature11556.

Leek, J. T. and R. D. Peng (2015). "Opinion: Reproducible research can still be wrong: Adopting a prevention approach". In: *PNAS; Proceedings of the National Academy of Sciences* 112.6, pp. 1645-1646.

Peng, R. D. (2009). "Reproducible research and Biostatistics". In: *Biostatistics* 10.3, pp. 405-408.

Peng, R. D. (2011). "Reproducible research in computational science". In: *Science (New York, Ny)* 334.6060, p. 1226.

Ram, K. (2013). "Git can facilitate greater reproducibility and increased transparency in science." In: *Source code for biology and medicine* 8.1, p. 7.

Resnik, D. B, M. Morales, R. Landrum, et al. (2019). "Effect of impact factor and discipline on journal data sharing policies". In: *Accountability in Research* 26.3, pp. 139-156. ISSN: 1545-5815. DOI: 10.1080/08989621.2019.1591277. URL: http://dx.doi.org/10.1080/08989621.2019.1591277.

Sandve, G. K, A. Nekrutenko, J. Taylor, et al. (2013). "Ten simple rules for reproducible computational research".

References 3

Schwab, M, M. Karrenbach, and J. Claerbout (2000). "Making scientific computations reproducible". In: *Computing in Science & Engineering* 2.6, pp. 61-67.

Stodden, V, P. Guo, and Z. Ma (2013). "Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals". In: *PloS one* 8.6, p. e67111.

Stodden, V, M. McNutt, D. H. Bailey, et al. (2016). "Enhancing reproducibility for computational methods". In: *Science* 354.6317, pp. 1240-1241. ISSN: 1095-9203. DOI: 10.1126/science.aah6168. URL: http://dx.doi.org/10.1126/science.aah6168.

Vasilevsky, N. A, J. Minnier, M. A. Haendel, et al. (2017). "Reproducible and reusable research: are journal data sharing policies meeting the mark?" In: *PeerJ* 5, p. e3208. ISSN: 2167-8359. DOI: 10.7717/peerj.3208. URL: http://dx.doi.org/10.7717/peerj.3208.

Wilkinson, M. D, M. Dumontier, I. J. Aalbersberg, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: http://dx.doi.org/10.1038/sdata.2016.18.

Wilson, G, D. Aruliah, C. T. Brown, et al. (2014). "Best practices for scientific computing". In: *PLoS Biol* 12.1, p. e1001745.

Thank you!

Contact info:

- Jessica Minnier: *minnier@ohsu.edu*

This workshop info:

- Code for these slides are on github, with links to other talks and course materials: jminnier/talks_etc
- The `.Rmd` file that generated the slides is on github, though you need to download the whole R project to knit the file.

This presentation is made with Knitr + RStudio

This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

This is a document written in plain text (.Rmd file) with text and R code embedded with the special syntax. Within RStudio when you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.