Reproducible Research: Why and How

Part 4: Dissemination Solutions

Sam Harper

SER Pre-Conference Workshop
2020-10-30

4. Dissemination Solutions

4.1. Replication Files

4.2 Sharing

4. Dissemination Solutions

4.1. Replication Files

4.2 Sharing

Replication files provide the 'recipe' for reproducing your results.

Should:

- be complete but parsimonious. Don't over do it.
- run / reproduce results with minimal effort (1-click).
- be "literate" (human readable).
- protect confidential data.

There is no single, perfect way to organize or prepare files for replication.

Find a workflow that works well for you.

Clark (2017)

Step 1: Generate replication files

1. Create a new organized empty replication folder within your project directory (e.g., “replication_files/”)

2. Subfolders: Should be same as overall file structure:

    - code/ — scripts
    - data_clean/ — manipulated data
    - data_raw/ — original data
    - output/ — generated tables, graphs, etc.
    - extra/ — misc. extras (e.g., code book)

3. A "README" text/markdown file to document contents, sources, software/system versions, other info necessary for replication/comprehension.

Clark (2017)

Step 2: Replicate your own results

1. Copy data and code to your new replication directory.
2. After copying all of the relevant files, see if you can replicate the results in your paper.
3. May want to start with the "final" products (i.e., tables and figures from clean data), which should be "easiest" to replicate.
4. Check for errors and make sure all is well.
5. Now copy the original raw data and cleaning scripts and run the entire thing.
6. All good? If not, debug and try again.

Clark (2017)

Step 3: Final check

1. Shut down and restart software package.
2. Replicate again...all good?
3. Or have a friend / colleague try on another computer.
4. Fix any remaining bugs and try again.
5. Now ready to disseminate!

Clark (2017)

How replication can help

Facilitate reproducibility

Anyone can reproduce your tables and figures.

Detects errors

Coding is hard. We all make mistakes.

Extends work

Probes reliability of findings, answers new questions.

Hatzenbuehler et al. *Soc Sci Med* 2014:

Reported 12 year decrease in life expectancy for sexual minorities living in more prejudiced communities.

Hatzenbuehler et al. *SSM* (2014). NY Times

Replication attempt using same public data "failed".

Re-analysis commissioned by original authors.

Coding errors discovered.

Study retracted.

> *This article has been retracted at the request of the authors and the Editors-in-Chief.*
>
> *The reason for the retraction is that the authors discovered an error in the study, which, once corrected, rendered the association between structural stigma and mortality risk no longer statistically significant in the sample of 914 sexual minorities. The authors published a Corrigendum (Corrigendum to "Structural stigma and all-cause mortality in sexual minority populations" [Soc. Sci. Med. 103 (2014) 33–41], Volume 200, March 2018, p 271), pending a re-analysis of the data. Re-analysis confirmed that the original finding was erroneous and the authors wish to fully retract their original study accordingly.*

Source: Retraction Watch

4. Dissemination Solutions

4.1. Replication Files

4.2 Sharing

Why share?

Credibility

Others can reproduce or interrogate your findings.

Social Good

Resource for other questions and new ideas.

Changing norms

Professional norms are insufficient to change behavior.

What to share?

Pre publication

Preregistration/pre-analysis plan

Codebook / data documentation

Code to create / analyze data.

Replication files

Reports / preprints

Post publication

Peer-reviewed papers / postprints

Entire project?

What about pre-prints (or post-prints)?

Most publishers allow posting of a final "accepted" proof.

Consult the agreement you sign.

It is your work!

What rights do I retain as an Oxford Journal author?

- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, for their own personal use, including their own classroom teaching purposes;
- The right, after publication by Oxford Journals, to use all or part of the Article and abstract, in the preparation of derivative works, extension of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that this not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgment is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article and to Oxford University Press and/or the learned society.

Standard Oxford University Press License (e.g., AJE, IJE)

Rationale for sharing data and code

Online repositories last longer, are indexed.

Concerns:

- Can usually be embargoed, or provide only what is necessary for replication (e.g., unused survey Qs).
- Biggest risk isn't having your data/ideas stolen, it's having your research ignored! (King 1995)
- *More* difficult if research products are proprietary.

Many resources to help

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 80–88
https://doi.org/10.1080/00031305.2017.1375986

Packaging Data Analytical Work Reproducibly Using R (and Friends)

Ben Marwick[a], Carl Boettiger[b], and Lincoln Mullen[c]

[a]University of Washington, Seattle, WA; [b]University of Wollongong, Wollongong, New South Wales; [c]University of California, Berkeley, CA; [d]George Mason University, Fairfax, VA

ABSTRACT
Computers are a central tool in the research process, enabling complex and large-scale data analysis. As computer-based research has increased in complexity, so have the challenges of ensuring that this research is reproducible. To address this challenge, we review the concept of the research compendium as a solution for providing a standard and easily recognizable way for organizing the digital materials of a research project to enable other researchers to inspect, reproduce, and extend the research. We investigate how the structure and tooling of software packages of the R programming language are being used to produce research compendia in a variety of disciplines. We also describe how software engineering tools and services are being used by researchers to streamline working with research compendia. Using real-world examples, we show how researchers can improve the reproducibility of their work using research compendia based on R packages and related tools.

ARTICLE HISTORY
Received May 2017
Revised August 2017

KEYWORDS
Computational science; Data science; Open source software; Reproducible research

Same code, different environment

Code may not run somewhere else. People are working on that:

Docker containers

packages up everything needed to run an application: code, runtime, system tools, system libraries and settings in a standalone, executable package.

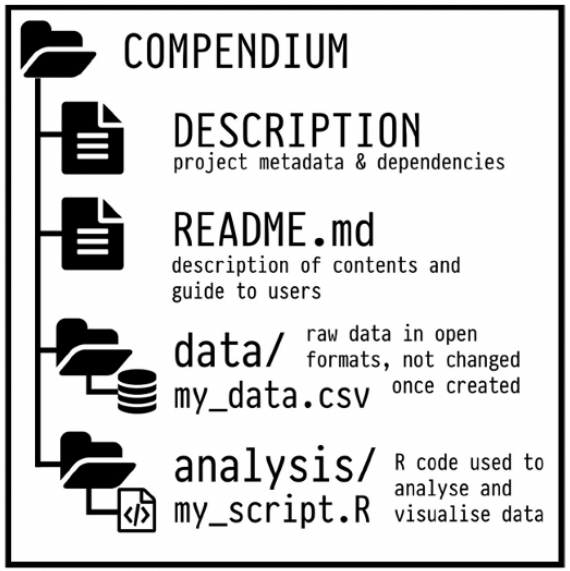
Binder

a service that provides your code and the hardware and software to execute it (mostly R, Python).

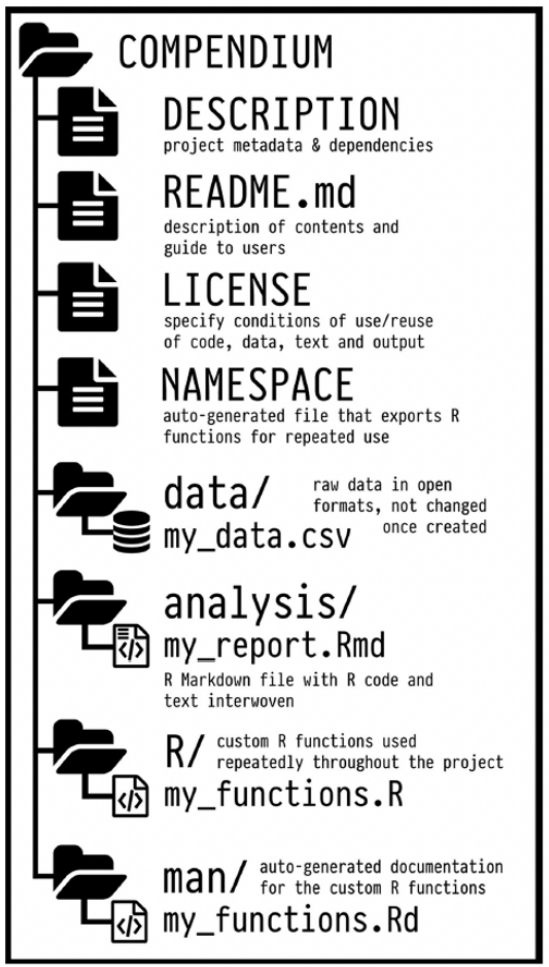
More advanced than we can cover today. See https://mybinder.org and https://hub.docker.com/ for more.

Can be done for any size project 👇

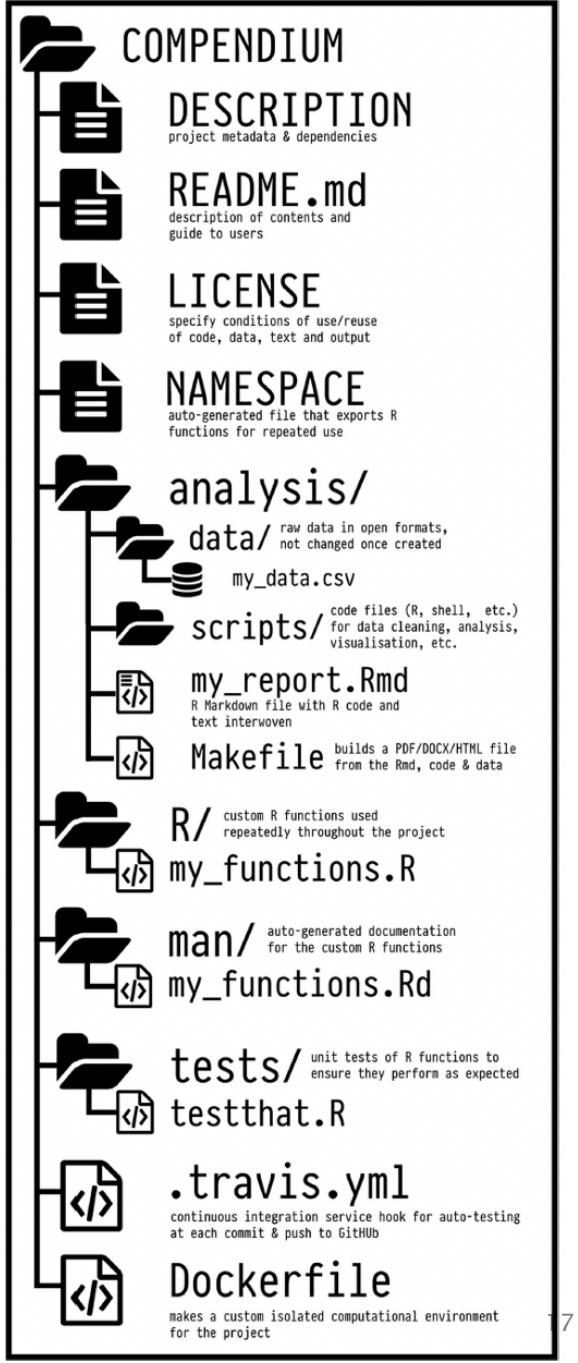
Small



Medium



Large



When to share data or code?

Many options

1. Include everything with submitted paper (public)

2. Include everything with submitted paper (for review)

3. Post-publication (recall TOP guidelines)

What if someone steals my idea?

Replication files may not matter!

Annals of Internal Medicine IDEAS AND OPINIONS

Dear Plagiarist: A Letter to a Peer Reviewer Who Stole and Published Our Manuscript as His Own

Michael Dansinger, MD

Dr. Doctor,

I am aware that you recently admitted to wrongly publishing, as your own, a scientific research paper that I had submitted to *Annals of Internal Medicine*. After serving as an external peer reviewer on our manuscript, you published that same manuscript in a different medical journal a few months later. You removed the names of the authors and the research site, replacing them with the names of your coauthors and your institution.

many research papers. It just doesn't make sense. Whether the pressure to publish is so intense, or whether the culture where you work is relatively permissive such that plagiarism is not taken as seriously, or whether getting caught seemed unlikely–it is hard to imagine why you would take this chance.

I hope you will not steal anyone else's research in the future. Instead, perhaps there is some way you can assist the scientific community's efforts to reverse the growing epidemic of plagiarism and scientific fraud.

Biased anecdote for benefits of sharing *with* paper submission

"Thanks for the opportunity to review this interesting paper. It is exciting to see the FARS data used in this way. **It was also exciting that you shared your code and this allowed me to review your work in a way I have not done before.**"

-Reviewer 1

"Overall verdict: This paper was both exciting and a pleasure to read. Clearly written, well argued, and with a highly commendable open science approach. As a referee, one frequently thinks 'why didn't they report the results of a model with...' - and in this case **I was able to download the data, assess their code scripts, and run my own specification to see how it changed things**."

-Reviewer 2

We probably got lucky, but the paper was accepted on the first submission.

Where to share?

Depends on discipline: find appropriate registry at http://www.re3data.org/, or check out ...

- Harvard's Dataverse
- Open Science Framework
- OpenICPSR
- figshare
- Data Dryad
- Many others...

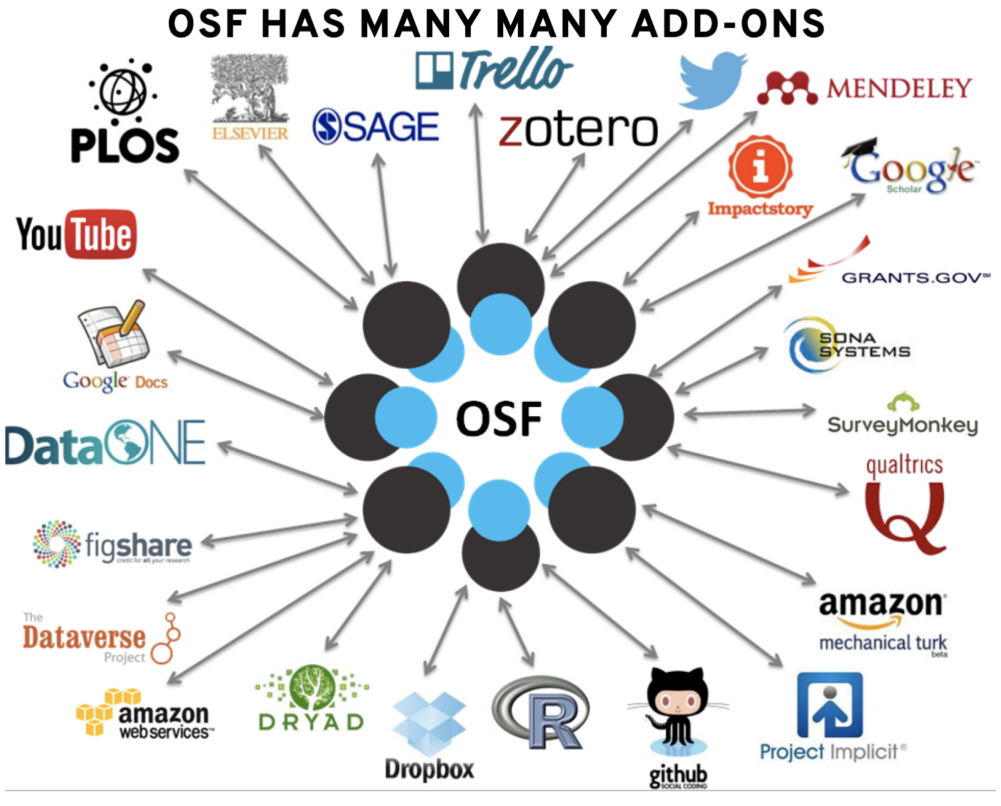
Open Science Foundation

OSF (http://osf.io) provides a central location to manage project files.

Any type of files can be uploaded (up to 5GB).

Most common file types will render to be viewed on OSF.

OSF also provides a more comprehensive system for planning, documenting, executing, and disseminating your research over the entire life cycle of a project--and beyond.



OSF workflow

1. Create a structured workspace.

2. (Possibly) pre-register study

3. Deposit / add study materials.

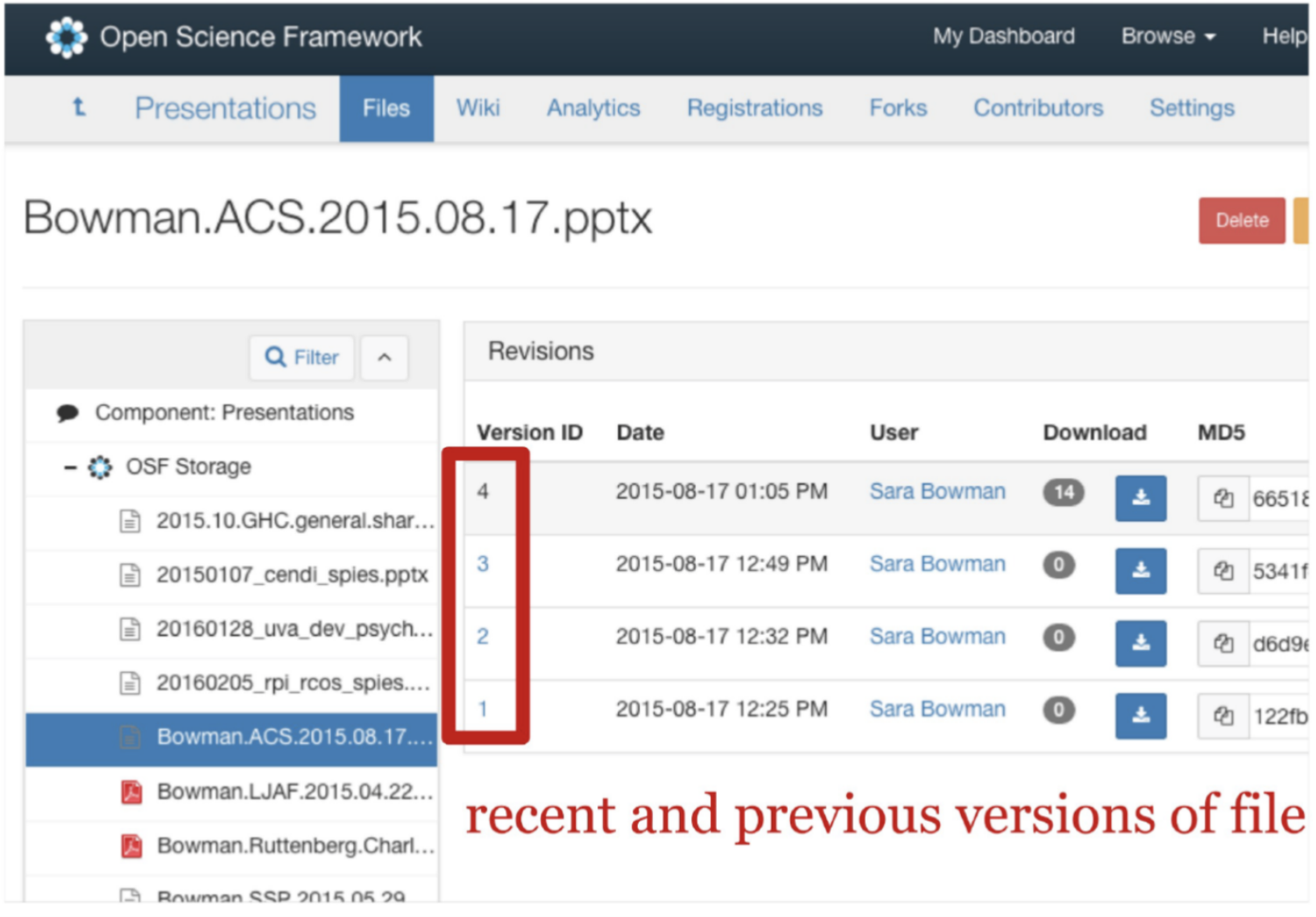
4. Add and document analyses.

5. Share study data, materials, and code.

OSF project landing page

Manage access/permissions

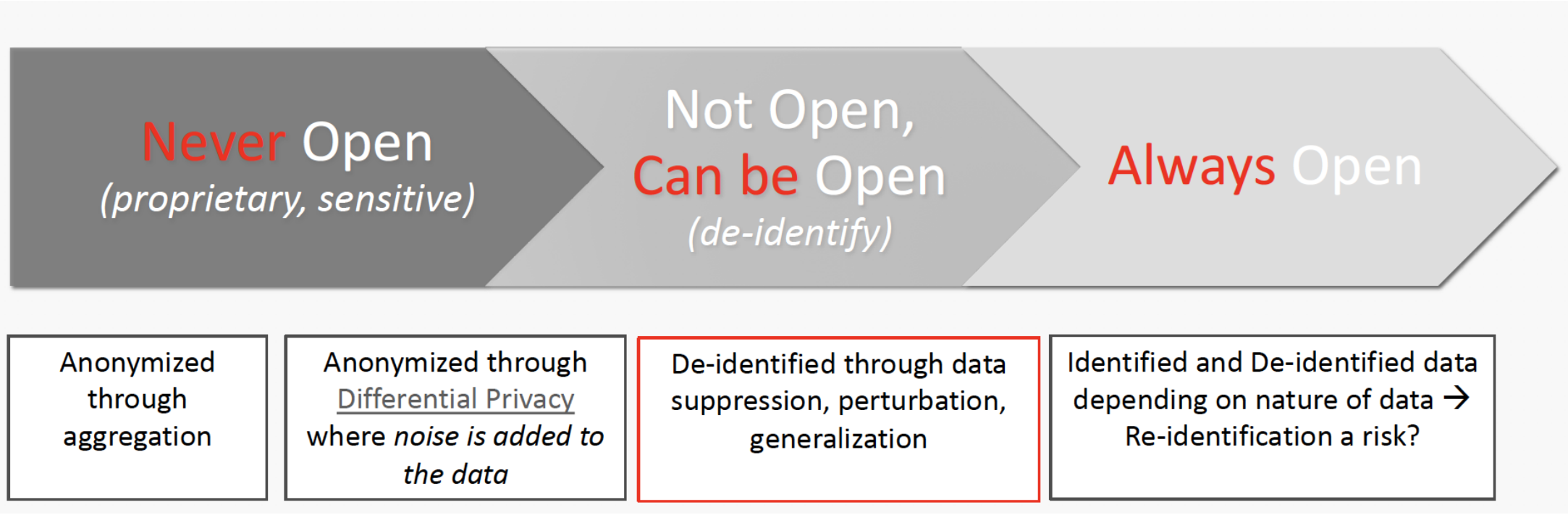
Automate version control



Persistent, unique identifiers for published work

Not everything can (or should) be shared

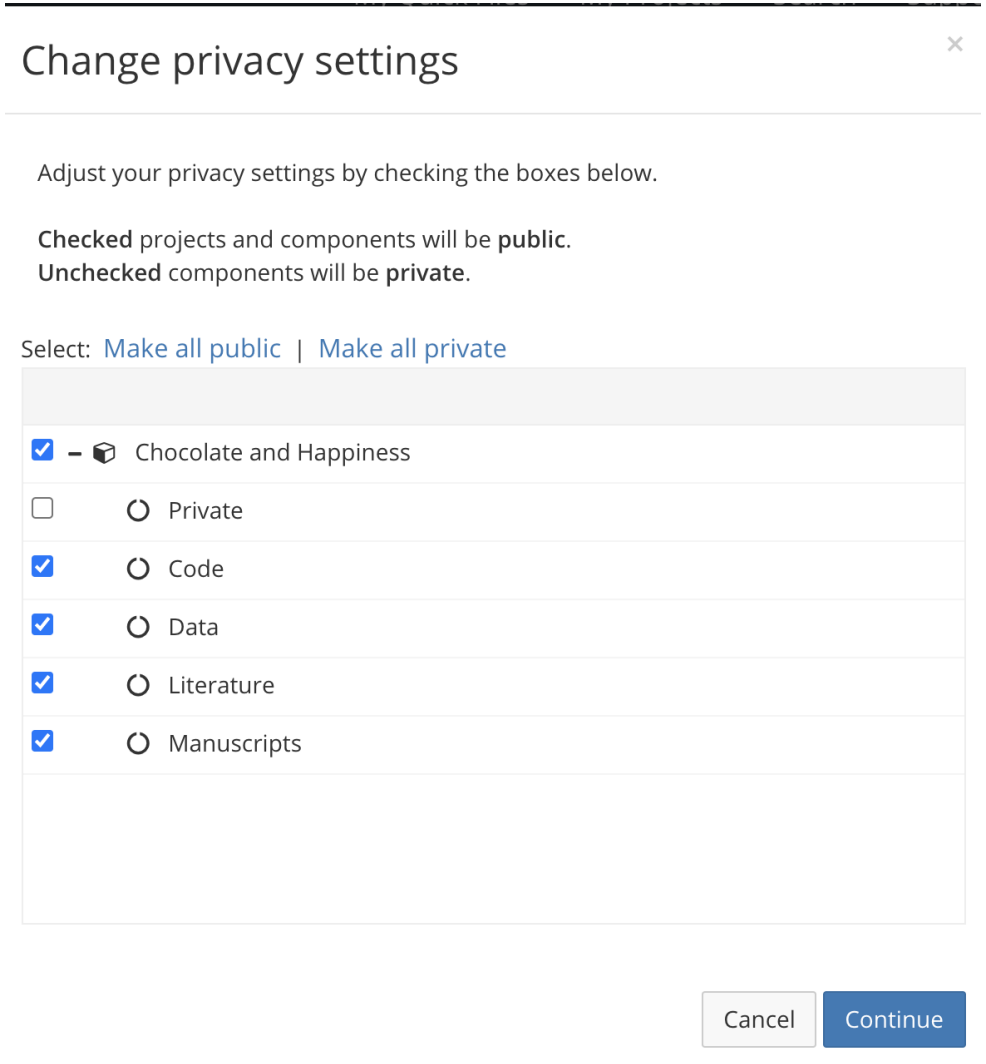
Spectrum for sharing sensitive material:



Source: Jennifer Sturdy (https://osf.io/5yq4u/)

OSF allows both public and private components

Allows for maximum sharing without sacrificing privacy concerns.

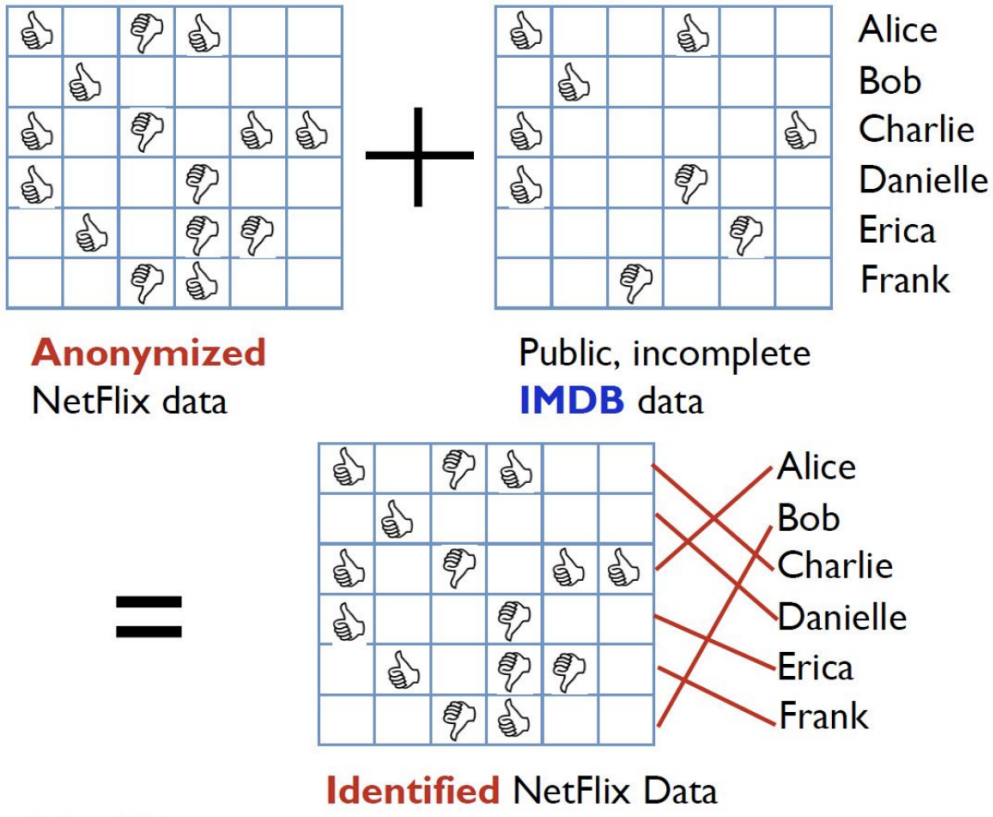


Why we worry (re-identification)

Netflix "contest" data:

- unique "random" ID
- film rated
- date
- ⭐⭐⭐

But people rate in many places, some of which do contain personal data.



Narayanan (2008)

Concerns about privacy

Trade-offs

Can't pretend there aren't social costs to open data.

Dangers of re-identification

Wealth of data after social media revolution.

Limits

Solutions may be too severe to justify utility of sharing.

Synthetic data may also be possible

- mimics an original dataset, preserving its statistical properties and relationships between variables
- classification and regression tree approach
- 0% disclosure risk
- `synthpop` R package 📦
- code and materials on GitHub

Quintana (2020)

Synthetic datasets: A non-technical primer for the biobehavioural sciences to promote reproducibility and hypothesis-generation

Daniel S. Quintana

Norwegian Centre for Mental Disorders Research (NORMENT), Division of Mental Health and Addiction, University of Oslo, and Oslo University Hospital, Oslo, Norway.

Abstract

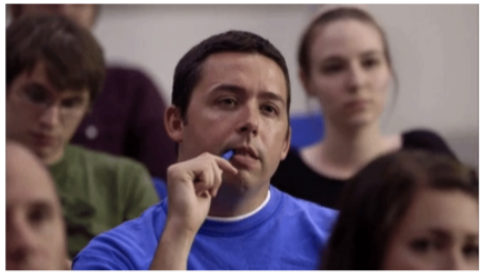
Open research data provides considerable scientific, societal, and economic benefits. However, disclosure risks can sometimes limit the sharing of open data, especially in datasets that include sensitive details or information from individuals with rare disorders. This article introduces the concept of synthetic datasets, which is an emerging method originally developed to permit the sharing of confidential census data. Synthetic datasets mimic real datasets by preserving their statistical properties and the relationships between variables. Importantly, this method also reduces disclosure risk to essentially nil as no record in the synthetic dataset represents a real individual. This practical guide with accompanying R script enables biobehavioural researchers to create synthetic datasets and assess their utility via the *synthpop* R package. By sharing synthetic datasets that mimic original datasets that could not otherwise be made open, researchers can ensure the reproducibility of their results and facilitate data exploration while maintaining participant privacy.

5. Reproducible Example

Today's Research Project

Does chocolate increase graduate student happiness?

Subjects



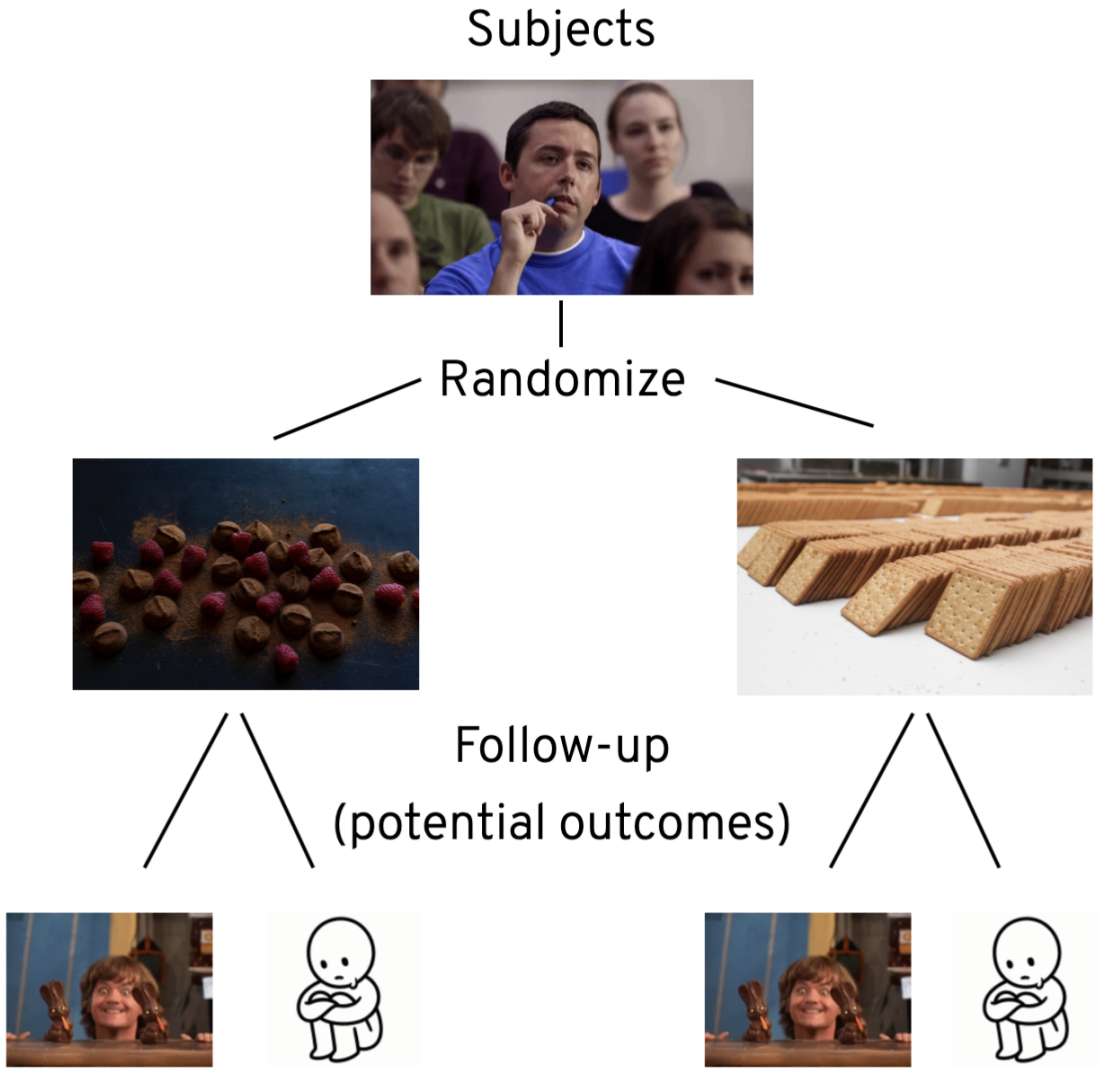
Randomize

Follow-up

(potential outcomes)

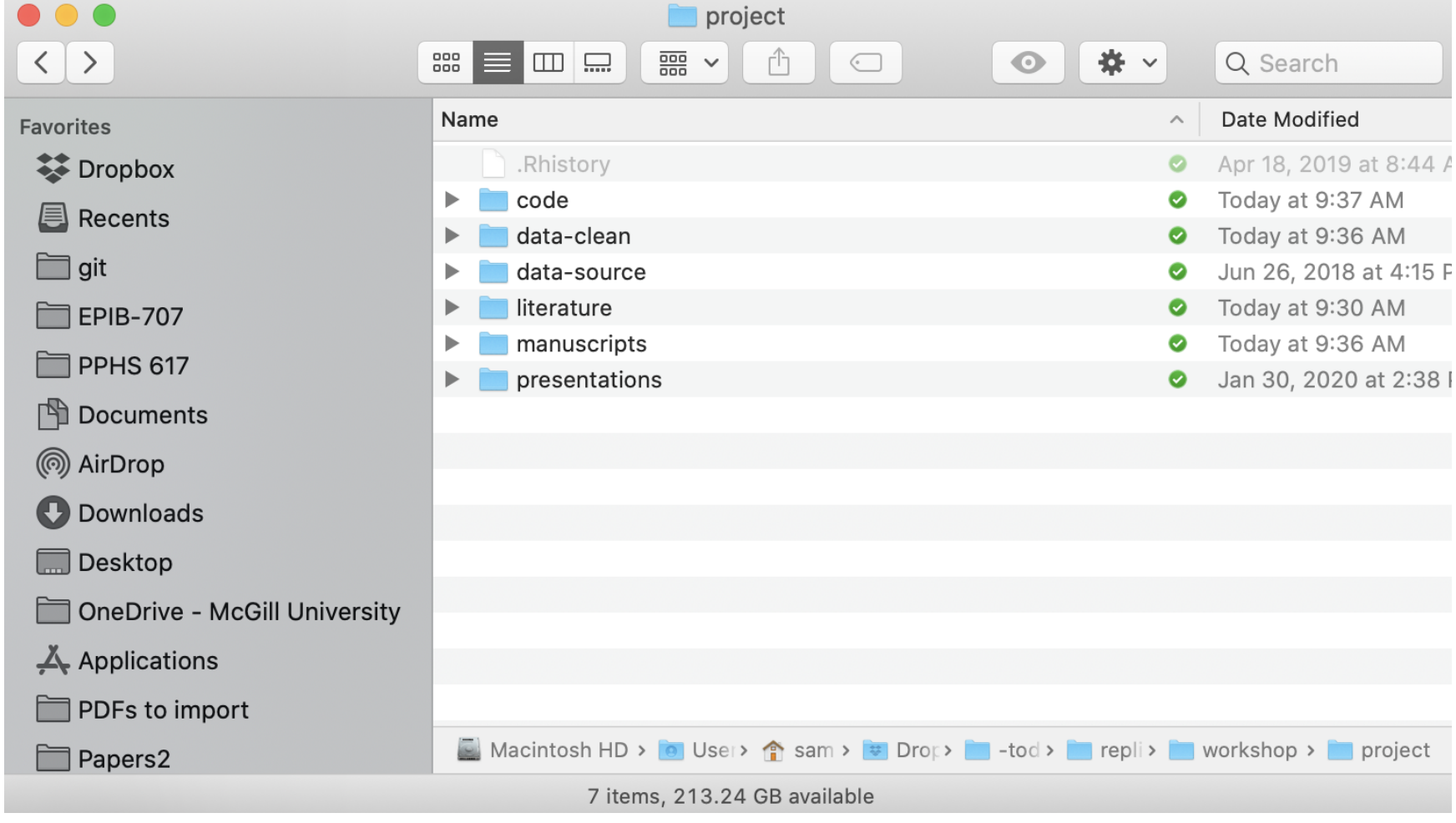
Example

- Folder structure
- Organizing via OSF
- Cleaning
- Analysis
  - Descriptives
  - Tables
  - Figures
- Write-up



Folder structure

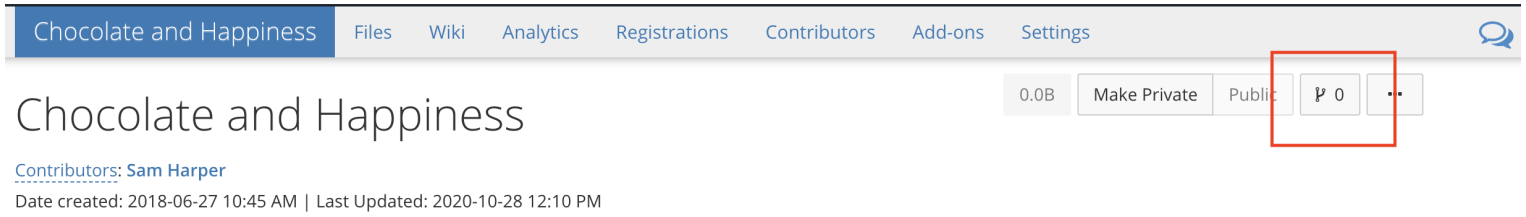
Separate code from data and manuscript:



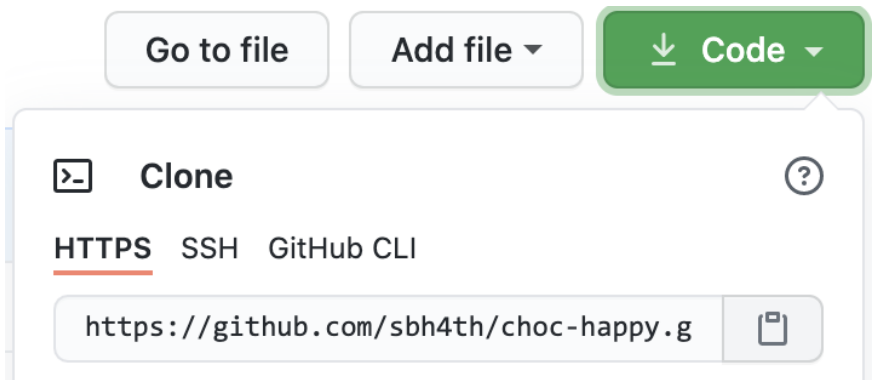
Live Demo

Follow along

Create a copy of the OSF repository (called 'forking')

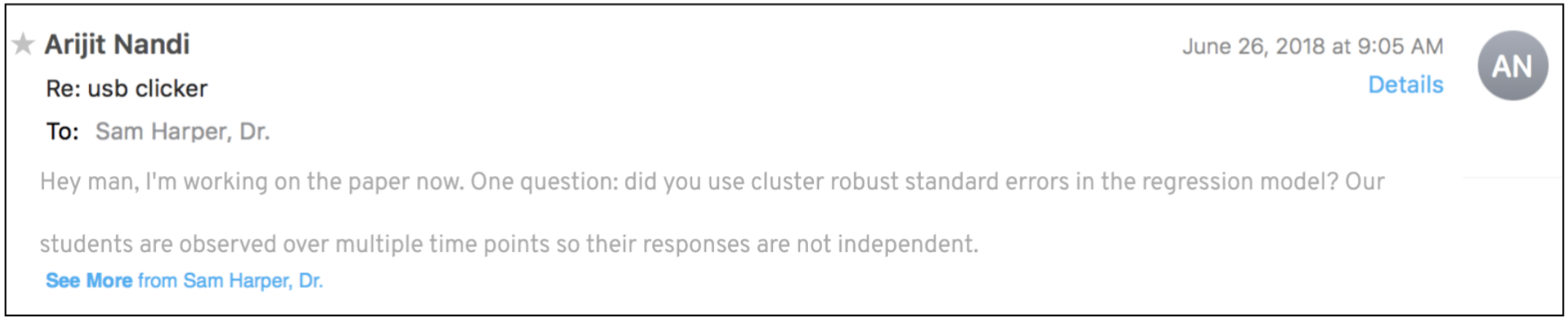


Clone the GitHub repository (download to your machine)



Final Thoughts

Working with "friends"



1. Ugh.
2. He's right.
3. Need to regenerate the model and table results.

- Do you know where your materials are?
- Is it 6 months later?
- Could someone else take over and figure out what you did?

Other resources for reproducibility in R

Version Control

- Happy Git and GitHub for the useR

RMarkdown

- R Markdown: The Definitive guide
- RMarkdown Driven Development (RmdDD): Blog post by Emily Riederer

R Packages

- R packages by Hadley Wickham and Jenny Bryan

Research Compendia

- Karthik Ram: *rstudio::conf 2019 talk*

Docker & Binder

- Getting started with binder docs
- rOpenSci Docker tutorial

Tutorials

- Rstudio Essentials Webinar series
- rrresearch: ACCE DTP course on Research Data & Project Management

Source: Anna Krystalli

This seems like a lot of extra work.

What's in it for me?

Selfish reasons to work reproducibly:

- Helps to avoid disaster.
- Makes it easier to write papers.
- Can help reviewers
- Enables continuity of your work.
- Builds your scholarly reputation.

Markowetz (2015)

Take home messages

Reproducible ≠ rigid

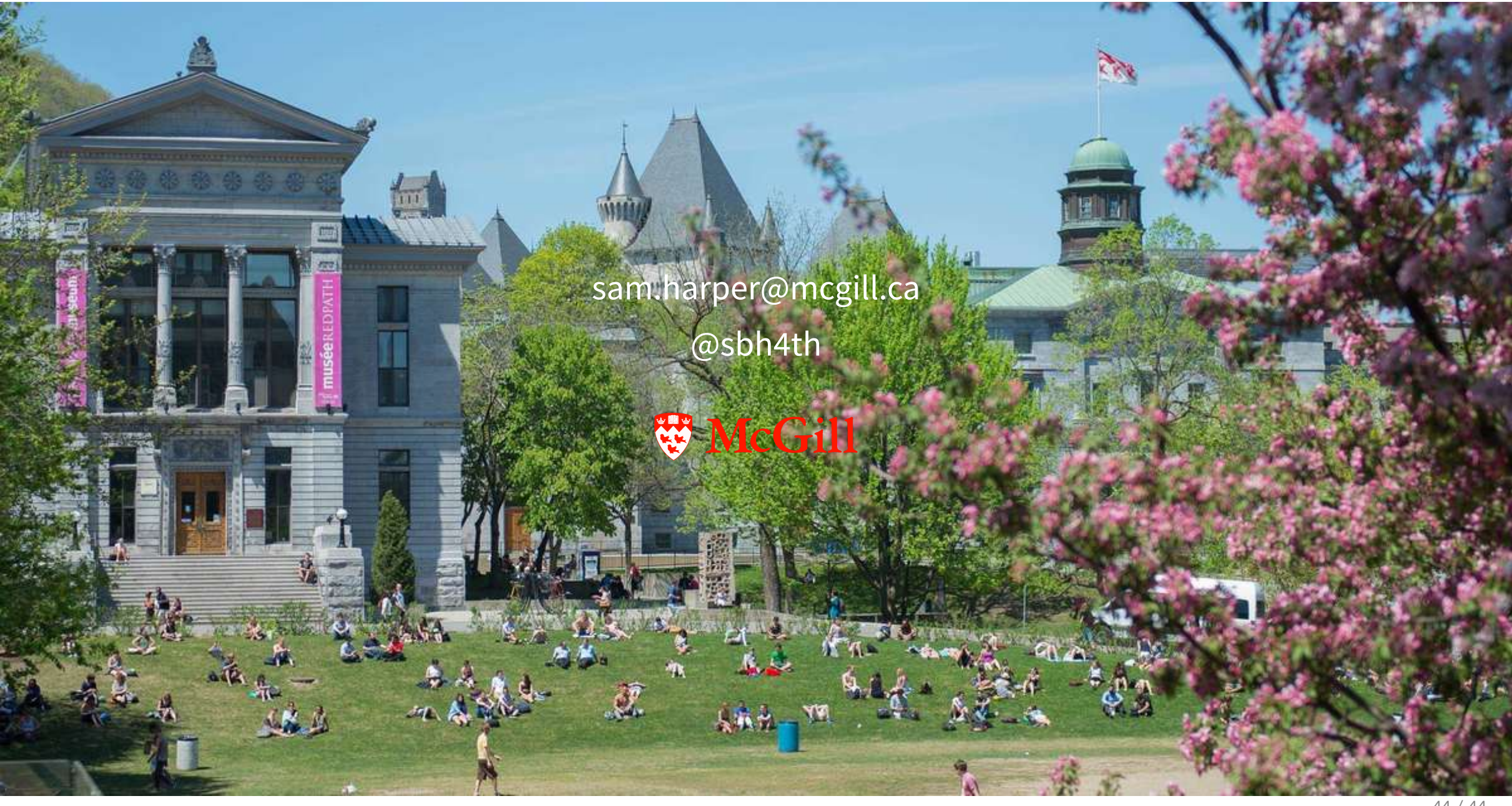
Don't like *R*? Need to use SAS? There is no single set of tools--more important to find a workflow that works for **you** and your team.

Security still matters

Some things can't be shared. Providing a rationale for what is/not shared means being transparent.

Lead by example

Transparency and openness are positive norms. We need help.



sam.harper@mcgill.ca

@sbh4th

McGill