An introduction to missing data analysis

Introduction

Waylon Howard

Webinar, November 12, 2024

About me

Waylon Howard

- Senior researcher at *GESIS - Leibniz Institute for the Social Sciences* (Department Survey Data Curation) & (co-) leader of the team *Research Data & Methods* at the *Center for Advanced Internet Studies* (CAIS)
  - Uses and effects of digital media
  - Computational methods
  - Data management & Open science

About you

- What's your name?
- What is your research area?
- What are your experiences with reproducible research (and the tools we cover in this course)?
- What are your expectations for this course?

Preliminaries

Slides and material are available at

http://frederikaust.com/reproducible-research-practices-workshop

- The workshop consists of a combination of lectures and hands-on exercises
- Feel free to ask questions anytime
- We will have frequent breaks (please remind us if necessary)

Preliminaries

Did you have any trouble with the setup for this workshop?

Installing...

- `git`
- `R` & *RStudio*
- a `TeX` distribution (e.g., TinyTeX)
- `papaja`

Preliminaries

Did you have any trouble with the setup for this workshop?

Creating...

- a *GitHub* account
- an account for the KU Leuven *GitLab*
- access credentials for *GitHub* (HTTPS, PAT) and *GitLab* (SSH)

Course schedule

Wednesday, April 27th, 2022

When?		What?
10:00 - 12:00		Introduction: Reproducible Workflows
12:00 - 13:00		*Lunch break*
13:00 - 15:00		Introduction to R Markdown
15:00 - 15:30		*Coffee break*
15:30 - 17:30		papaja

Course schedule

Thursday, April 28th, 2022

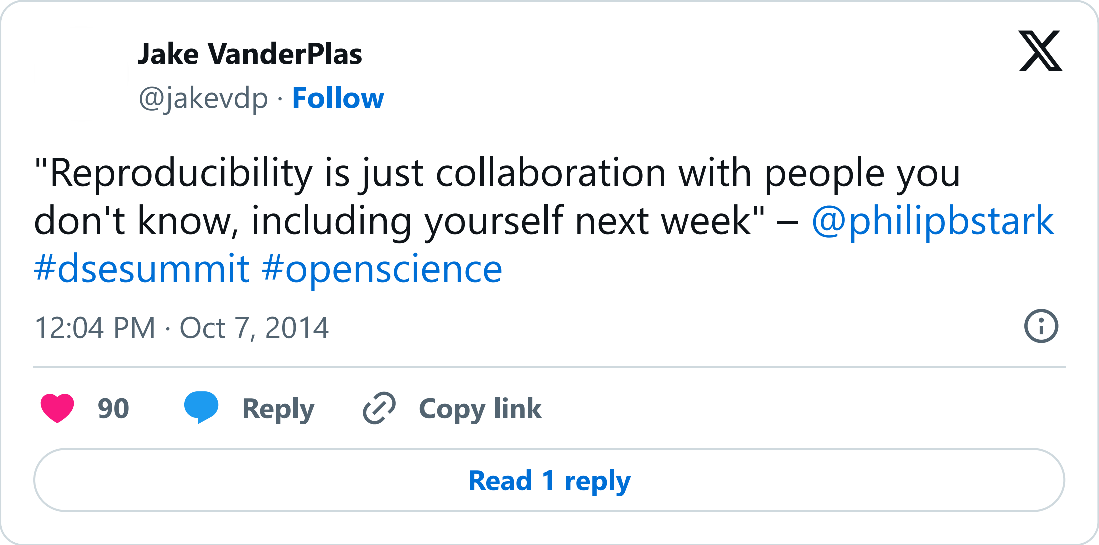
When?		What?
09:00 - 10:15		Introduction to Git & GitHub
10:15 - 10:30		*Coffee break*
10:30 - 11:30		Git in RStudio
11:30 - 12:30		Git & GitHub for collaboration - Part 1

Course schedule

Thursday, April 28th, 2022

When?	What?
12:30 - 13:30	*Lunch break*
13:30 - 14:00	Git & GitHub for collaboration - Part 2
14:00 - 14:15	*Coffee break*
14:15 - 15:30	Other tools & workflows
15:30 - 16:00	Wrap-Up

What is reproducibility?

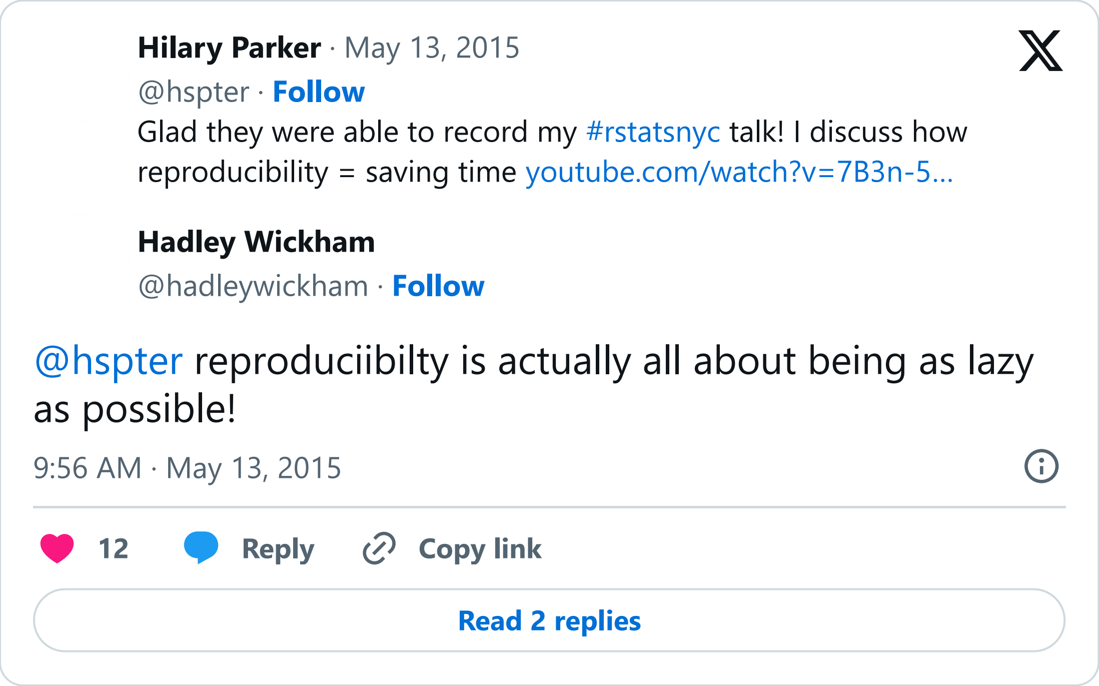


Jake VanderPlas
@jakevdp · **Follow**

"Reproducibility is just collaboration with people you don't know, including yourself next week" – @philipbstark #dsesummit #openscience

12:04 PM · Oct 7, 2014

Why reproducibility matters



Hilary Parker · May 13, 2015

@hspter · Follow

Glad they were able to record my #rstatsnyc talk! I discuss how reproducibility = saving time youtube.com/watch?v=7B3n-5...

Hadley Wickham

@hadleywickham · Follow

@hspter reproduciibilty is actually all about being as lazy as possible!

9:56 AM · May 13, 2015

Defining reproducibility

As with (almost) everything in science, there are different definitions of reproducibility. We will discuss some of them in the following.

Defining dimensions

3-dimensional concept space

By Christof Schöch. Source: https://dh-trier.github.io/trr/#/2/1

Defining dimensions

The Turing Way Project illustration by Scriberia. DOI: 10.5281/zenodo.3332807

The Turing Way definition

Source: https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

Replication or reproduction?

By Christof Schöch. Source: https://dh-trier.github.io/trr/#/2/2

Reproducible research workflows

> being an open scientist means adopting a few straightforward research management practices, which lead to less error-prone, reproducible research workflows (Klein et al., 2018, p. 11)

Research management practices

There are quite a few practices that researchers can adopt to increase the reproducibility of their work.

- Project-oriented workflow
- Folder structures
- Naming things
- …

Exercise: Folder structures

If you feel comfortable with that, feel free to share screenshots of some of your project folders. You can choose examples that you think are particularly well-organized as well as negative examples from your "dark past" as someone whose research might have been difficult to reproduce.

Sharing is caring

One prerequisite for research being reproducible (by others) is sharing research materials. There are many parts of their work that researchers can share to increase the reproducibility as well as the (potential) replicability of their work (see Klein et al., 2018). Four main types of output are:

1. Data
2. Code & scripts (for data collection, processing, and analysis)
3. Other study materials (e.g., questionnaires or stimulus materials)
4. (Detailed) Information about the study procedure

Notably, all of these outputs should be well-documented (e.g., via README files, metadata or comments in code).

Sharing for reproducibility

While sharing study materials and information about the procedure are important for replicability, for reproducibility, the most important things to share are the data and code.

Sharing data & code

There are many different ways in which researchers can share their data and code/scripts (see Klein et al., 2018). Keeping only local copies of things and sharing upon personal request is not a very sustainable or scalable solution. The better option is sharing via institutional or public archives and repositories.

Fantastic repositories and where to find them 🐉

The paper by Klein et al. (2018) provides an overview of public repositories that hold psychological data.[1] A good tool for finding suitable repositories is the *Registry of Research Data Repositories*.

[1] However, parts of this overview have inevitably become somewhat outdated since the paper was published.

How to choose a repository

In general, research data (and code) that are publicly archived should follow the so-called FAIR principles (Wilkinson et al., 2016), meaning that the shared materials should be…

- Findable: Persistent identifiers; metadata; indexed
- Accessible: Retrievable by identifier; controlled access where necessary
- Interoperable: Standardized metadata; open, lightweight, and interoperable file formats (e.g., CSV, TSV, JSON, ODS)
- Reusable: Documented; clear usage license

How to choose a repository

Some more specific key criteria for choosing a repository are that it should...

- use persistent and unique identifiers (such as DOIs)
- accommodate licensing
- feature access controls (e.g., allowing the restriction of access to a particular set of users)
- have persistence guarantees for long-term access
- store data in accordance with local legislation (e.g. the GDPR in Europe)

See Klein et al., (2018) for further details

Public archiving options

Two archiving options that are quite popular among researchers (esp. also in psychology) are the *Open Science Framework* (OSF) and *Zenodo*.

While these two archives are not curated, which somewhat reduces findability, they are quite flexible and easy to use. They can be used to share different types of content, including data and code, and also offer some degree of access control. A nice feature of the *OSF* and *Zenodo*, especially for sharing code, is that they offer integration with *GitHub* (which facilitates version control).

Tools & tool stacks

As you probably already know, there are lots of different tools and workflows that can be used to increase the reproducibility of research. These tools can then be combined into different tool stacks. We will introduce you to some of those in this workshop, but there are many more, and, in the end, it depends on your personal preferences and needs what tools/tool stacks and workflows you (should) employ.[1]

[1] As you will see, the two of us also have different preferences in our workflows and tool use.

Tools in this workshop

As stated before, we will focus on the following tools for reproducible research in this workshop:

- Git & *GitHub*/ *GitLab*
- R Markdown

We will focus on using them via *RStudio*.

Note: In the outlook part, we will also briefly introduce some other/additional tools.

Any questions so far?