Reproducible Workflow in Research Using R, LaTeX, and GitHub

Ozlem Tuncel
otuncelgurlek1@gsu.edu

19 September 2023

POLS 8830 – Georgia State University

Learning objectives

- Understand **reproducibility** and **replicability** and its impact on research

Learning objectives

- Understand **reproducibility** and **replicability** and its impact on research
- Learn practices that can be implemented in your research to enhance reproducibility

Learning objectives

- Understand **reproducibility** and **replicability** and its impact on research
- Learn practices that can be implemented in your research to enhance reproducibility
- Design a reproducible workflow using data management, coding, and dissemination tools

Learning objectives

- Understand **reproducibility** and **replicability** and its impact on research
- Learn practices that can be implemented in your research to enhance reproducibility
- Design a reproducible workflow using data management, coding, and dissemination tools

Reproduction vs Replication

- **Reproducibility** refers to instances in which the original researcher's data and computer codes are used to regenerate the results,

Reproduction vs Replication

- **Reproducibility** refers to instances in which the original researcher's data and computer codes are used to regenerate the results,

https://www.ncbi.nlm.nih.gov/books/NBK547546/#sec_ch3_2

Reproduction vs Replication

- **Reproducibility** refers to instances in which the original researcher's data and computer codes are used to regenerate the results,
- **Replicability** refers to instances in which a researcher collects new data to arrive at the same scientific findings as a previous study.

https://www.ncbi.nlm.nih.gov/books/NBK547546/#sec_ch3_2

What is workflow?

- https://www.fharrell.com/post/rflow/
- https://hbiostat.org/rflow/

So, which one?

- Today's agenda $\rightsquigarrow$ Computational reproducible research

Political Science Approach to Reproducible Research

PS: Political Science & Politics

Article contents

- Extract
- Footnotes
- References

Replication, Replication

Published online by Cambridge University Press: **02 September 2013**

Gary King

Show author details

Article Metrics

Get access Share Cite Rights & Permissions

Extract

Political science is a community enterprise; the community of empirical political scientists needs access to the body of data necessary to replicate existing studies to understand, evaluate, and especially build on this work. Unfortunately, the norms we have in place now do not encourage, or in some cases even permit, this aim. Following are suggestions that would facilitate replication and are easy to implement—by teachers, students, dissertation writers, graduate programs, authors, reviewers, funding agencies, and journal and book editors.

Type	Research Article
Information	PS: Political Science & Politics, Volume 28, Issue 3, September 1995, pp. 444 - 452 DOI: https://doi.org/10.2307/420301

Figure 1: It all started with King (2013) article!

New developments!

Political Science Replication Initiative

The Political Science Replication Initiative (PSRI) aims to promote replication research in political science. It was co-founded by Nicole Janz, Stephanie Wykstra and Seth Werfel. It is currently run by Nicole Janz, Assistant Professor at the University of Nottingham. On this page we point to resources on reproducibility and replications, and we provide an online database of replication studies in political science. As research transparency grows within the discipline, so does the ability to explore previous findings through re-analysis and robustness checks. Replication studies are an important element in scientific discovery that allow researchers to evaluate the validity and scope conditions of their proposed causal relationships. At the same time, replications tend to remain unpublished ('file drawer problem'), and even if they are published, they are not easily detectable because they rarely have the term 'replication' in the title or abstract.

If you are affiliated with a political science department and have conducted a replication study, we invite you to submit it here. Our submission process is simple:

1. Read our submission guidelines to ensure that your replication study is eligible for posting in the repository.

2. Go to the replication tab on this site, and click "create a new study" in the upper right-hand corner of the screen.

3. You will be prompted to complete a form with your replication study metadata, and upload your replication code and data to Dataverse.

If the submission complies with all requirements, it will be accepted into the Dataverse. If your submission is incomplete, you will receive an email with further instructions.

If you have already posted your replication materials to your own Dataverse, and would like to have your Dataverse linked so that it appears in the compiled list, please email us with a request.

Reproducible research matter! A lot!

Reproducible research matter! A lot!

Figure 2: https://datacolada.org/109

So, how to avoid this?

- Be honest!

So, how to avoid this?

- Be honest!
- Be organized! Keep track of **everything**!

How to keep track of things?

- Smaller projects $\rightsquigarrow$ Take notes separately or in your code!
- Larger projects $\rightsquigarrow$ Organization before you start your project!
  - Set appropriate working directory
  - Keep meaningful data file (and variable) names
  - Take notes in your code (extensively)
  - Keep research diary

Yeah, people do really take notes!

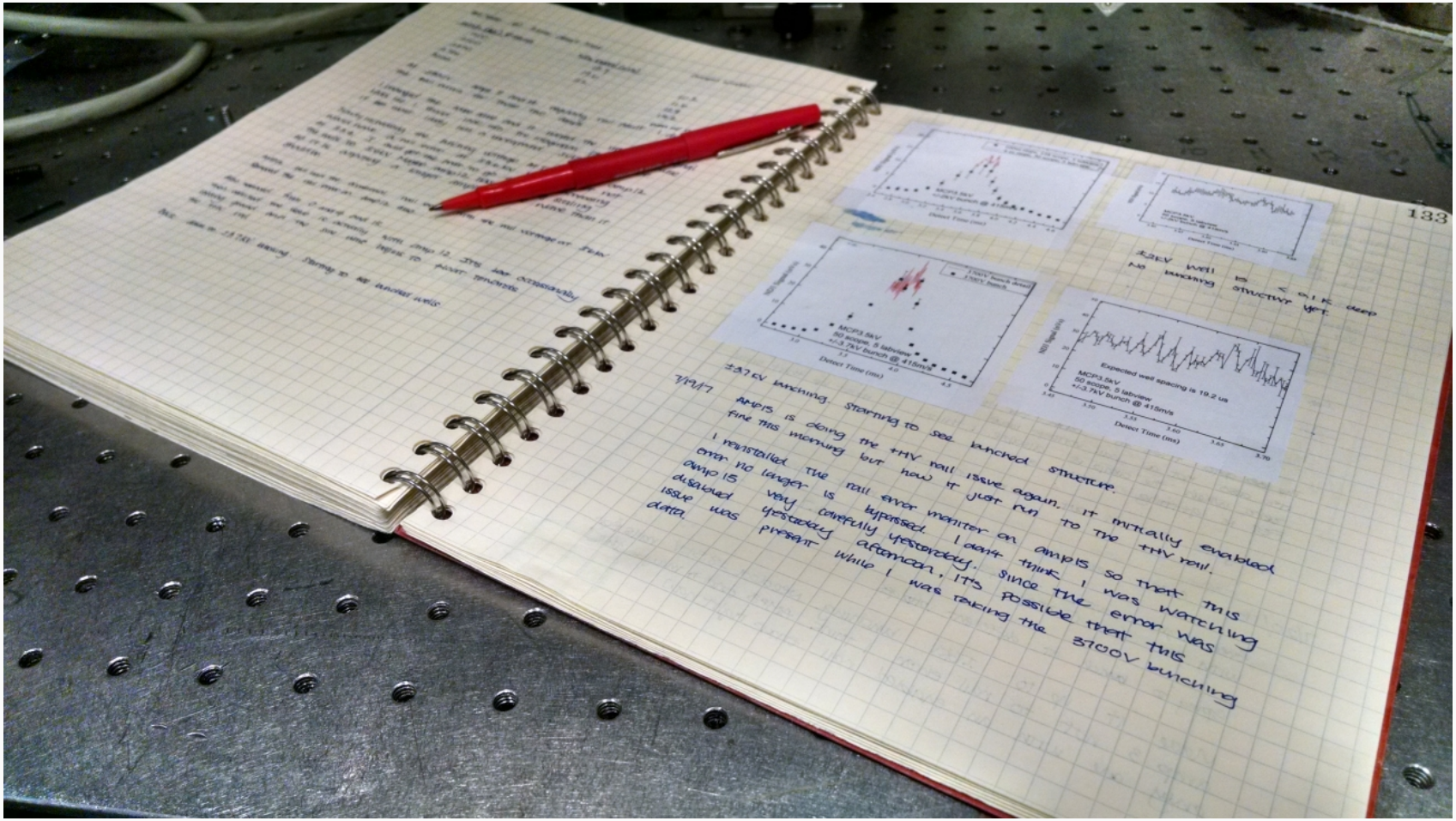


Figure 3: Lab notebooks are necessary part of research in hard sciences

How to keep track of things?

- Smaller projects $\rightsquigarrow$ Take notes separately or in your code!
- Larger projects $\rightsquigarrow$ Organization before you start your project!
  - Set appropriate working directory
  - Keep meaningful data file (and variable) names
  - Take notes in your code (extensively)
  - Keep research diary
  - Use R, LaTeX, and other tools efficiently!

Reproducible research overview

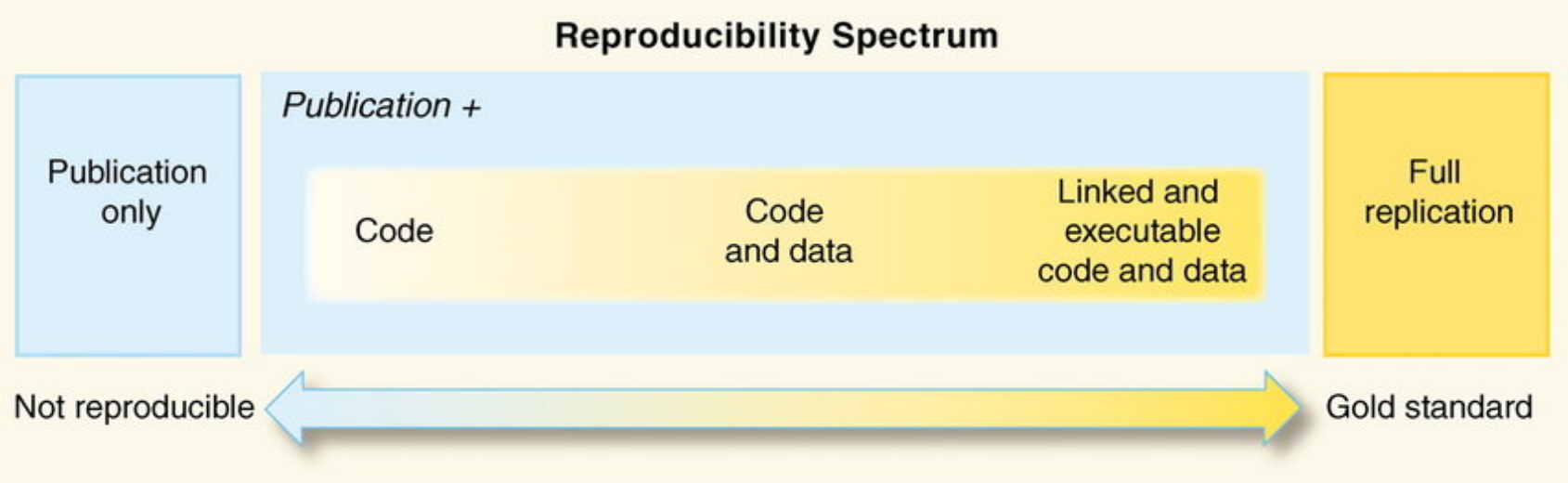
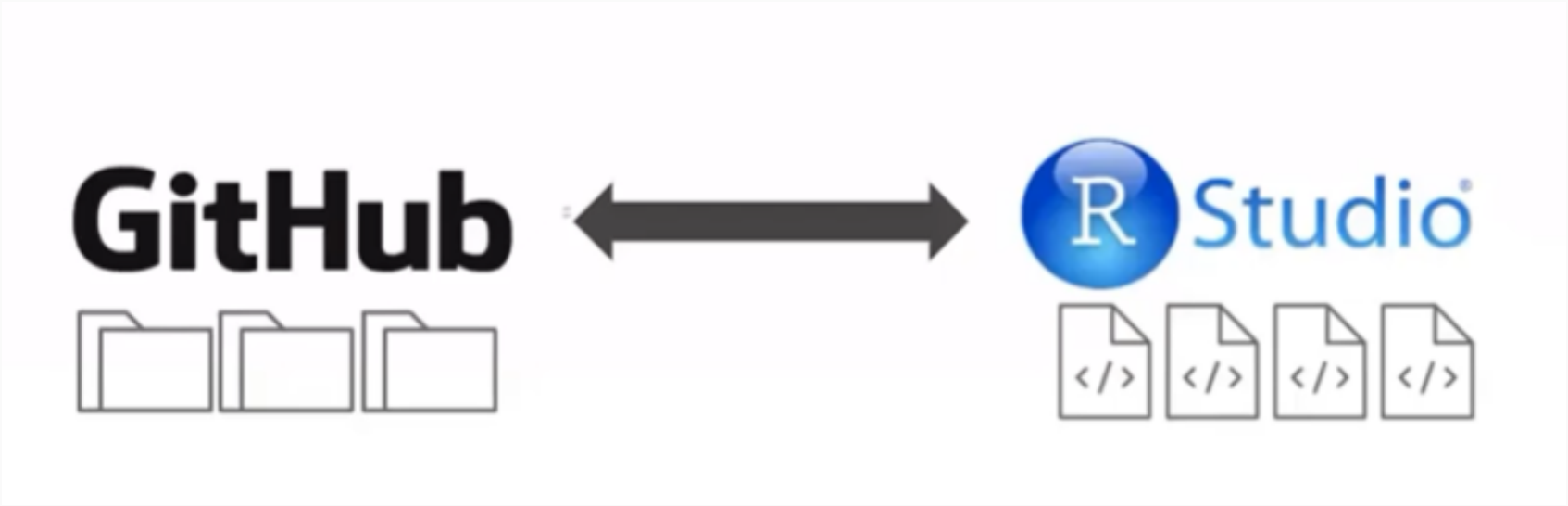
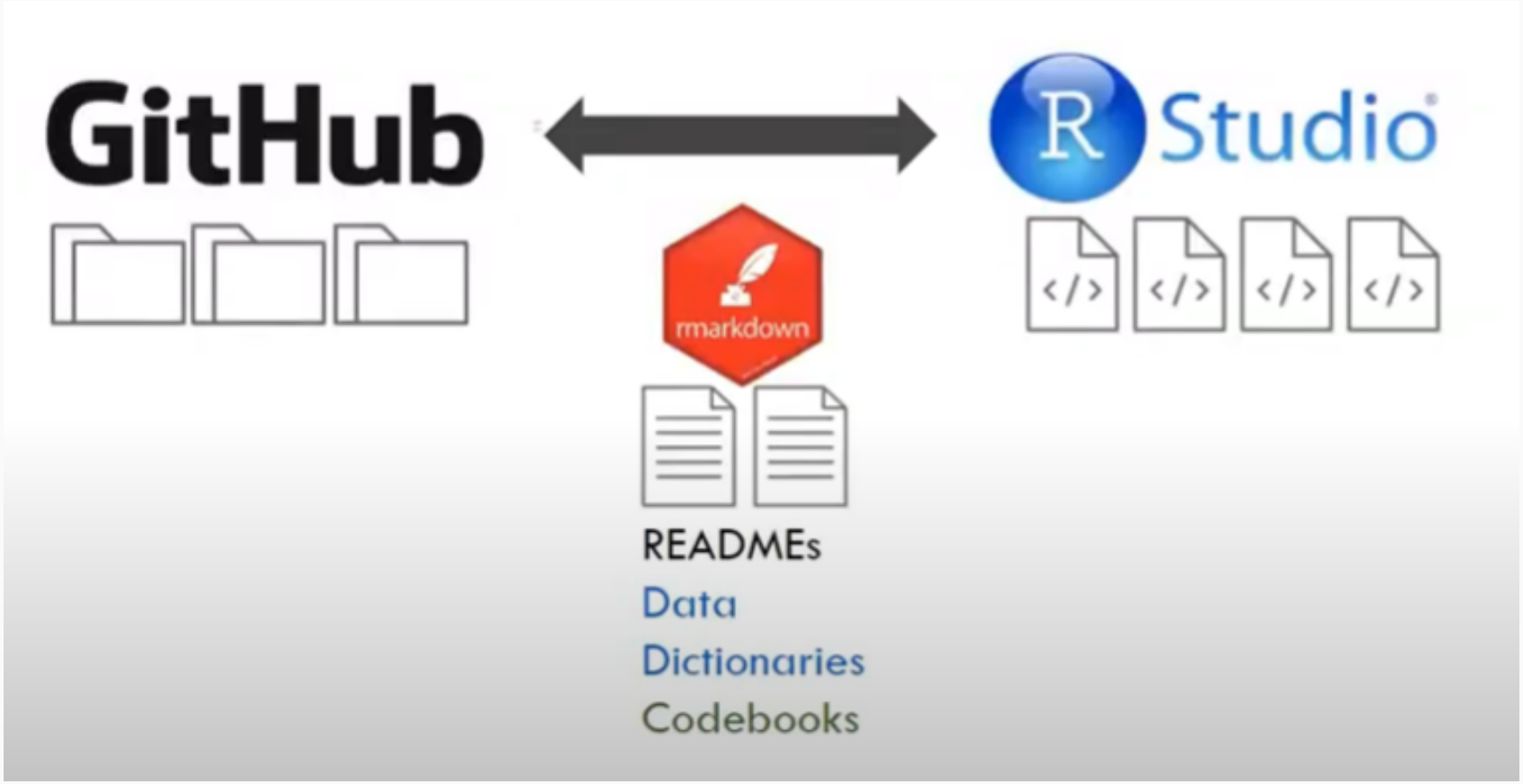


Figure 4: Peng, R. D. (2021). Annual Review of Public Health

Overview of workflow



Overview of workflow

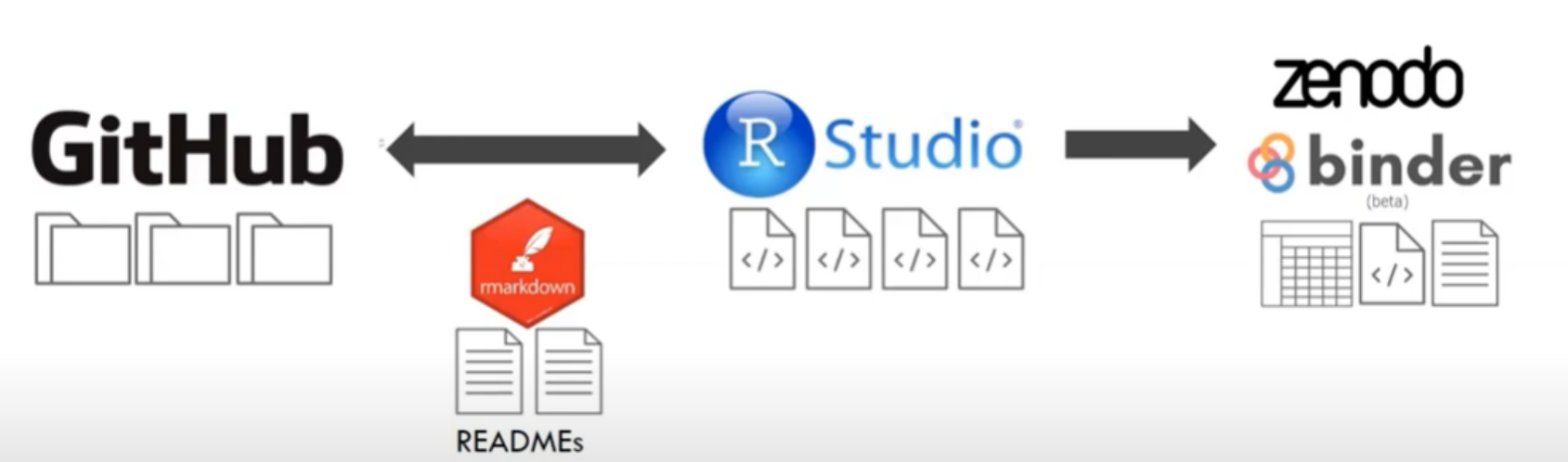


Overview of workflow

Content of data (i.e., variables, values, missing data, units of measurement, etc.)

Context of research (i.e., source of data, methods, processing, relationship of files, software version, etc.)

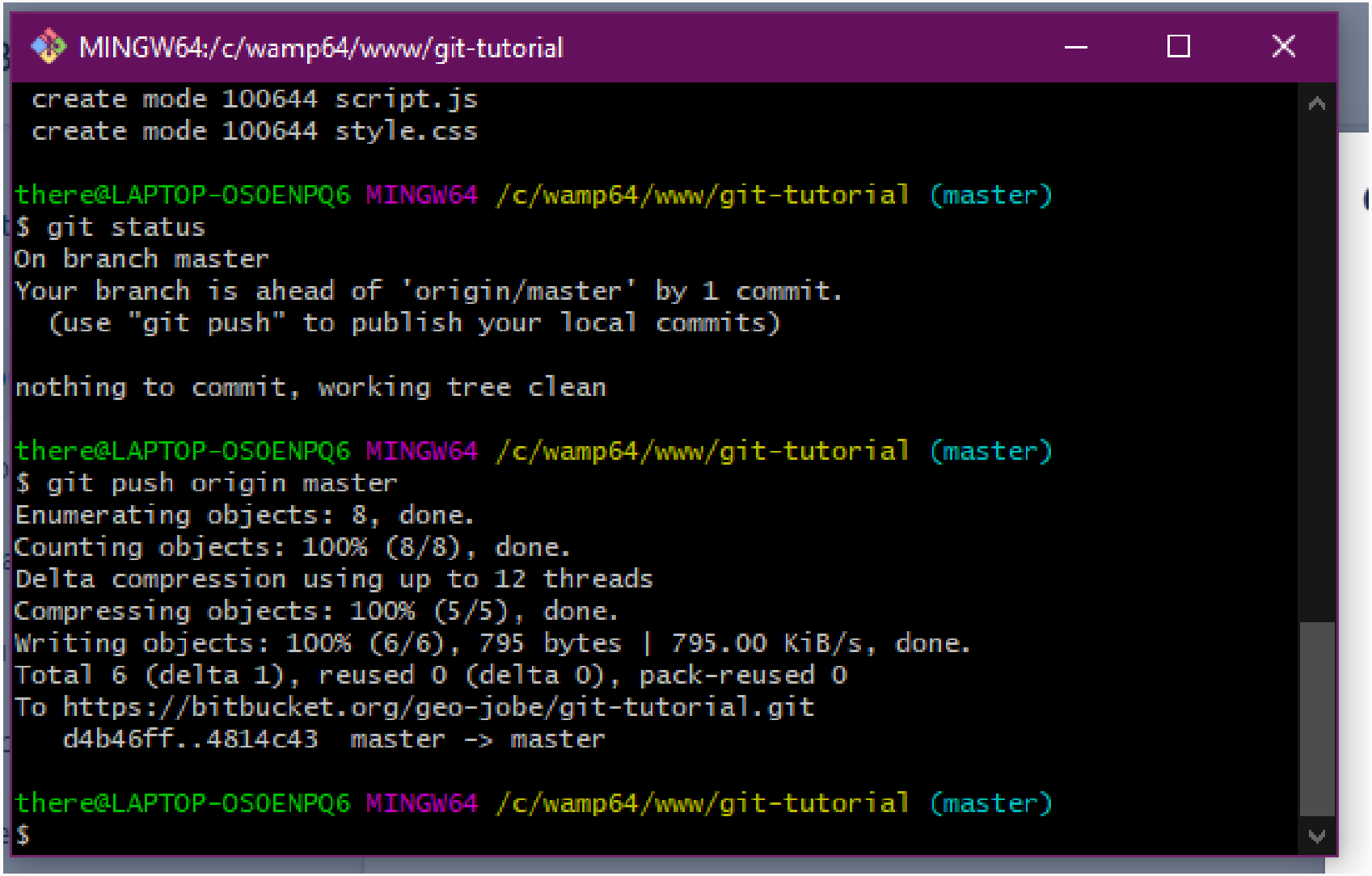
Overview of workflow



What is git? And, why git?

- Git helps you manage work done on projects.
- Git lets you tell the story of your project.
- Git uses a **repository** or **repo** to store all your files.
- When you make changes and update your work, you create a **commit**.
- Git let's you travel in time!

GIT IS VERY UNFRIENDLY!



Underneath all jargon, git is really simple!

Underneath all jargon, git is really simple!

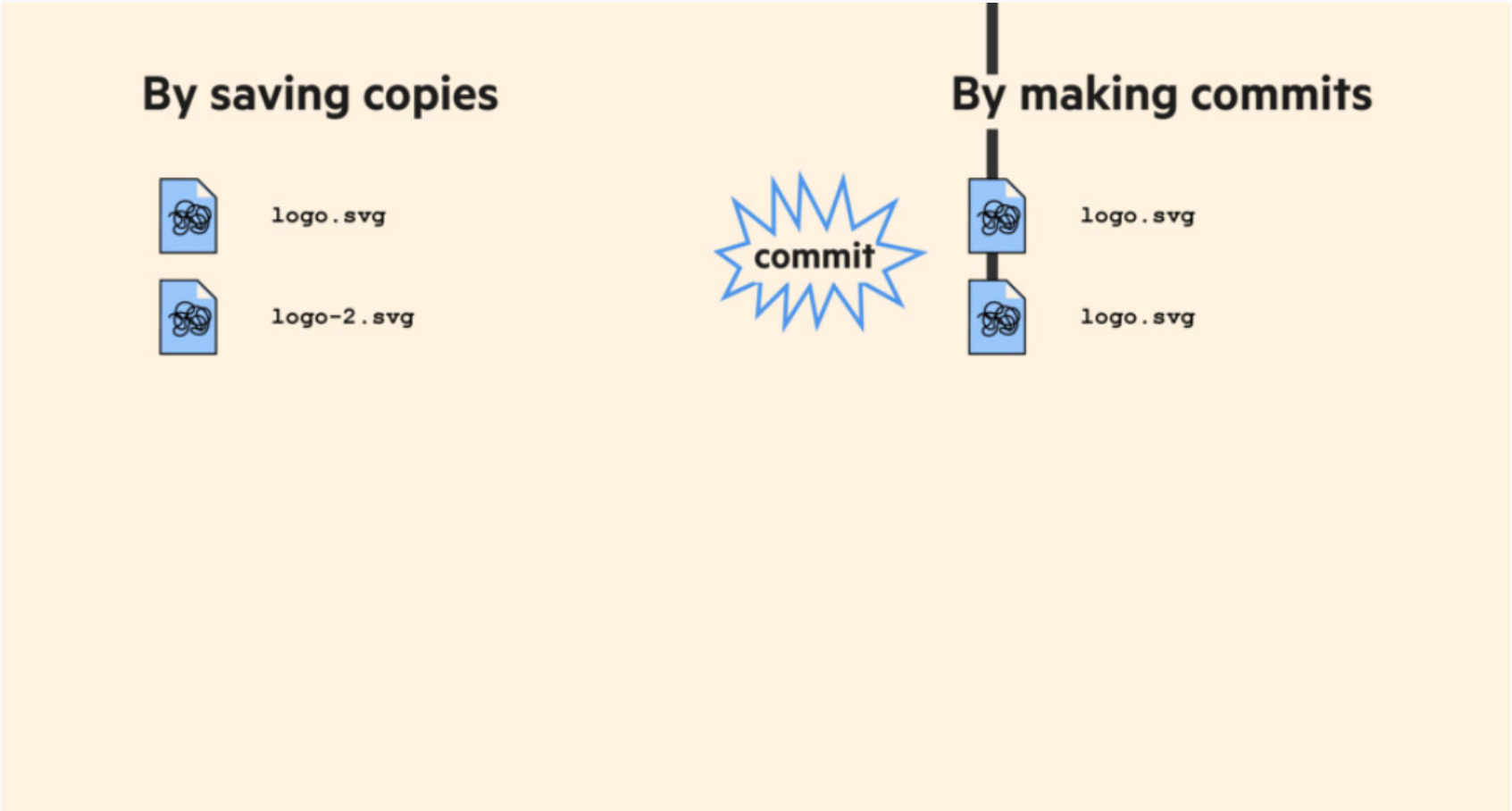
By saving copies

`logo.svg`

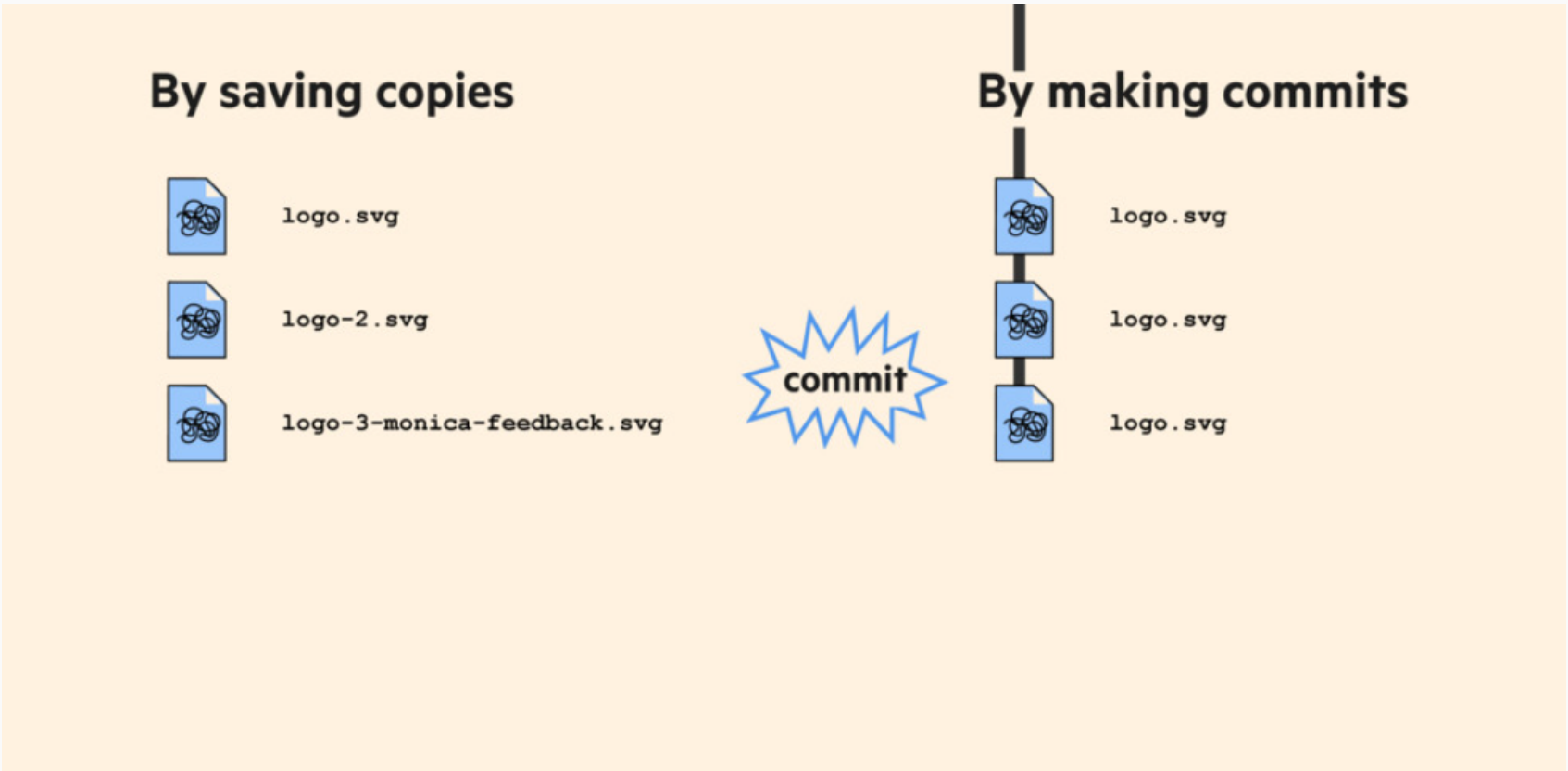
By making commits

`logo.svg`

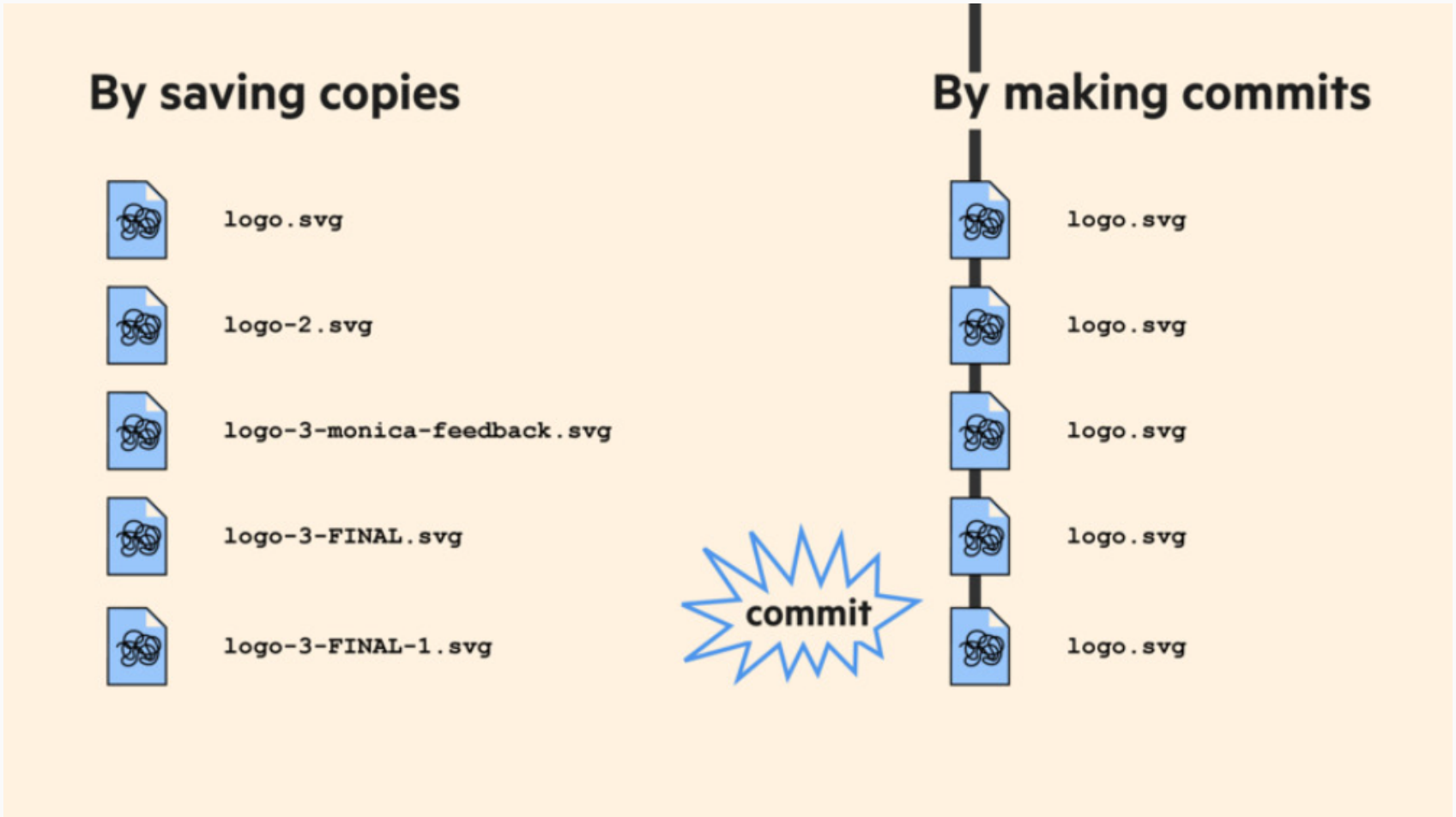
Underneath all jargon, git is really simple!



Underneath all jargon, git is really simple!



Underneath all jargon, git is really simple!



Git's advantages

- When you commit, you can also save information about your actions
- Who, when, why . . .

Underneath all jargon, git is really simple!

Let's look at some examples

- https://www.prio.org/journals/jpr/replicationdata
- https://www.emilyhenckenritter.com/research

Now, let's create a system for reproducible research!

Some resources

- Download Git from https://git-scm.com/
- Open a GitHub account https://github.com/
- https://happygitwithr.com/
- https://www.carlosivanrodriguez.com/guides/workflow/connect-an-existing-rstudio-project-to-github/
- Connect existing project to GitHub
- Git and R Studio Connection