Opinionated practices for teaching reproducibility:

motivation, guided instruction and practice

Tiffany Timbers (UBC)
Joel Ostblom (UBC)

2023-07-27

JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2022, VOL. 30, NO. 3, 241–250
https://doi.org/10.1080/26939169.2022.2074922

Taylor & Francis
Taylor & Francis Group

Opinionated Practices for Teaching Reproducibility: Motivation, Guided Instruction and Practice

Joel Ostblom[a] and Tiffany Timbers[b]

[a]Department of Computer Science, University of British Columbia, Vancouver, Canada; [b]Department of Statistics, University of British Columbia, Vancouver, Canada

ABSTRACT

In the data science courses at the University of British Columbia, we define data science as the study, development and practice of reproducible and auditable processes to obtain insight from data. While reproducibility is core to our definition, most data science learners enter the field with other aspects of data science in mind, for example predictive modeling, which is often one of the most interesting topics to novices. This fact, along with the highly technical nature of the industry standard reproducibility tools currently employed in data science, present out-of-the gate challenges in teaching reproducibility in the data science classroom. Put simply, students are not as intrinsically motivated to learn this topic, and it is not an easy one for them to learn. What can a data science educator do? Over several iterations of teaching courses focused on reproducible data science tools and workflows, we have found that providing extra motivation, guided instruction and lots of practice are key to effectively teaching this challenging, yet important subject. Here we present examples of how we motivate, guide, and provide ample practice opportunities to data science students to effectively engage them in learning about this topic.

ARTICLE HISTORY
Received September 2021
Accepted April 2022

KEYWORDS
Curriculum; Data science; Education; Reproducibility

Reproducibility in data science courses at UBC

Courses where we *explicitly* teach reproducibility for data science

Course	Title	Level	Reproducibility topic(s)
DSCI 100	Introduction to data science	undergraduate	code for analysis, version control
DSCI 310	Reproducible and trustworthy workflows for data science	undergraduate	file naming & organization, version control, reproducible reports, environments, data analysis pipelines, software packages
DSCI 521	Computing platforms for data science	graduate	file naming & organization, version control, reproducible reports, environments
DSCI 522	Data science workflows	graduate	version control, reproducible reports, environments, data analysis pipelines
DSCI 524	Collaborative software development	graduate	version control, software packages

Data Science:

the study, development and practice

of reproducible and auditable processes

to obtain insight from data

Reproducible analysis:

reaching the same result given the same input, computational methods and conditions[1].

Auditable analysis,

a readable record of the steps used to carry out the analysis as well as a record of how the analysis methods evolved.[2]

[1] National Academies of Sciences, 2019

[2] Parker, 2017 and Ram, 2013

Why do we adopt this definition?

- It is important that insights from data science are trustworthy
- Reproducible and auditable methods are one of the expectations for trustworthy data science

Source: Image by jcomp on Freepik

We cannot trust non-reproducible and non-auditable analyses because they:

1. lack evidence that the results could be regenerated
2. we don't know enough details of how they were created
3. there is an insufficient record of how and why analysis decisions were made

Examples of things that can go wrong without reproducible practices

- An interesting result that you cannot recreate 😞
- Your email inbox is full of information related to the project that only you have access too 😫
- A small change to the analysis code requires re-running the entire thing, *and takes hours*... 😯
- Activation time to becoming productive after taking a break from the project is hours to days 😴
- Code that can only be run on one machine, *and you don't know why*... 😵

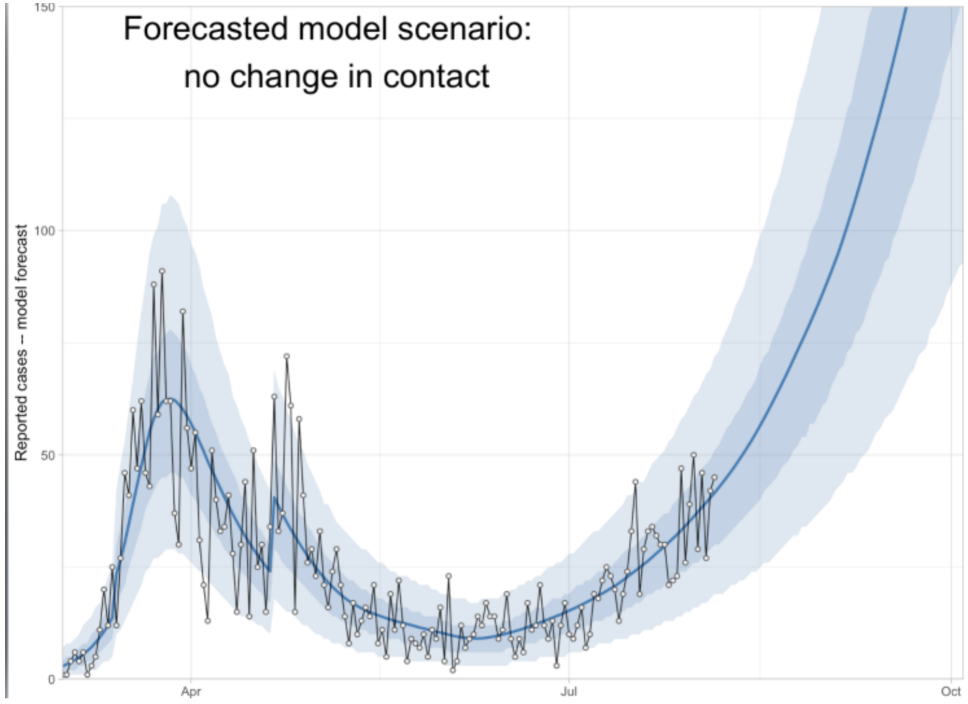
The lagniappe[1] of reproducibile data science methods:

more effective collaboration!

[1] lagniappe: a small gift given with a purchase to a customer, by way of compliment or for good measure; bonus.

So if reproducibility is so important for data science, why is it hard to teach it?

Students are more excited about generating insights from data than reproducibility?



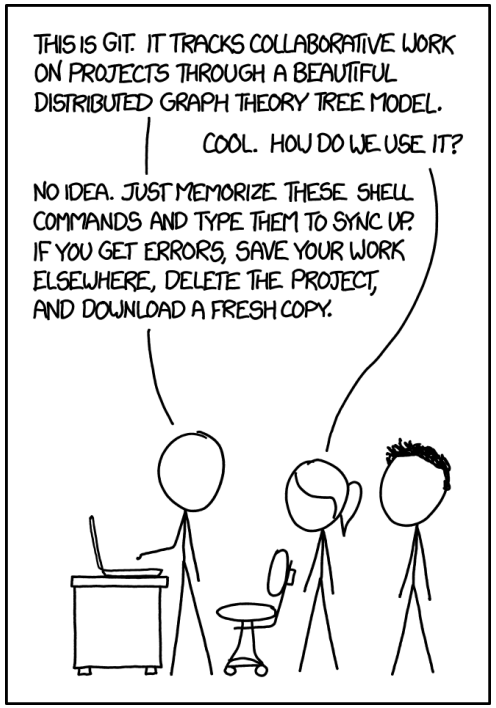
Source: CTV BC News article on COVID case predictions

Lack of awareness of the problems with non-reproducible analyses?

Source: https://cheezburger.com/6600205568/untitled

Reproducibility tools are not necessarily smooth and easy to learn?

For example, Git is hard!

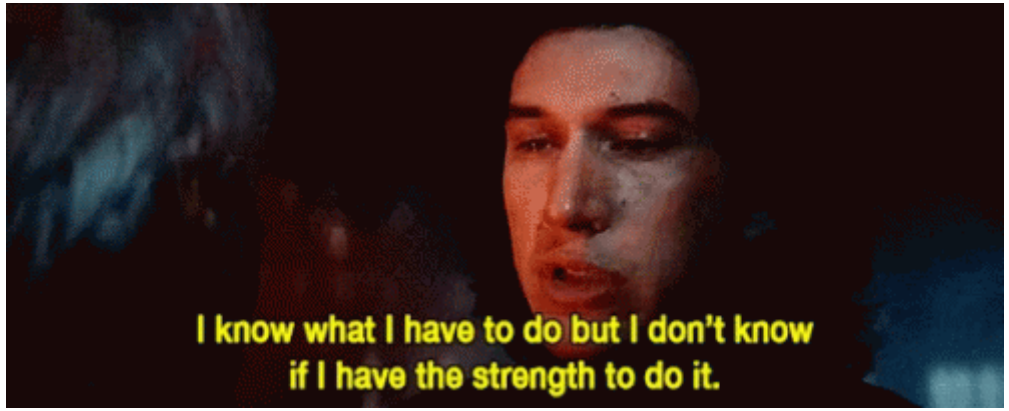


Source: https://xkcd.com/1597/

Our expectations of students' learning?

We want to actually change their habits or behaviours!

Habits protect individuals from motivational lapses, where a desired good behavior is not expressed due to a momentary lack of willpower[1].



Source: Star Wars: The Force Awakens

[1] Gardner, Benjamin and Amanda L. Rebar (Apr. 2019). “Habit Formation and Behavior Change”. en. In: Oxford Research Encyclopedia of Psychology. Oxford University Press. isbn: 978-0-19-023655-7. doi: 10.1093/acrefore/9780190236557.013.129.

Key things for teaching reproducibility

1. placing extra emphasis on motivation
2. guided instruction
3. lots of practice!!!

Key things for teaching reproducibility

1. placing extra emphasis on motivation
2. guided instruction
3. lots of practice!!!

Strategies for placing extra emphasis on motivation

1. Let them fail (in a controlled manner)
2. Study cases of failures with real world consequences
3. Tell stories from the trenches

Tell stories from the trenches

Exercise prompt:

1. Think and write down a non-reproducible, or non-auditable, workflow you have used before at work, on a personal project, or in course work, that negatively impacted your work somehow (make sure to include this in the story). Here's an example:

   As a Masters student, I started to use R to do my statistical analysis. I obtained the results I needed from running my code in the R console and copying the results into the word document that was my manuscript. Six months later we were working on revisions requested by the reviewers and I could not remember which version of the code I ran to get my results. I eventually figured it out through much trial and error, but the process was inefficient and very stressful.
   –- Tiffany Timbers

2. When prompted, paste your story in the Google doc (link to be shared in class)

Tell stories from the trenches

Sample student story:

"In my line of work, my colleagues and I heavily relied on SQL queries to withdraw data from database. She always forgot to save the queries while just sending me an excel file to read. Sometimes (Most of the time) when I cross checked the data, the result didn't match. It usually required hours of troubleshooting to figure out where the gap was between my query and her data. ¯_(ツ)_/¯"

-- Master of Data Science student, UBC

- Reproducibility failure > methods not recorded as code
- Impact of reproducibility failure > loss of time

Tell stories from the trenches

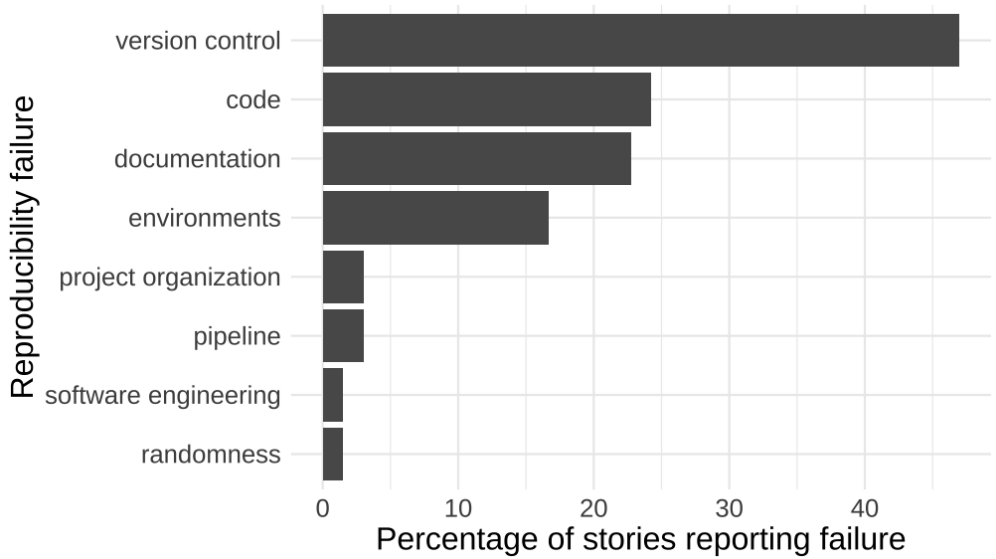
Sample student story:

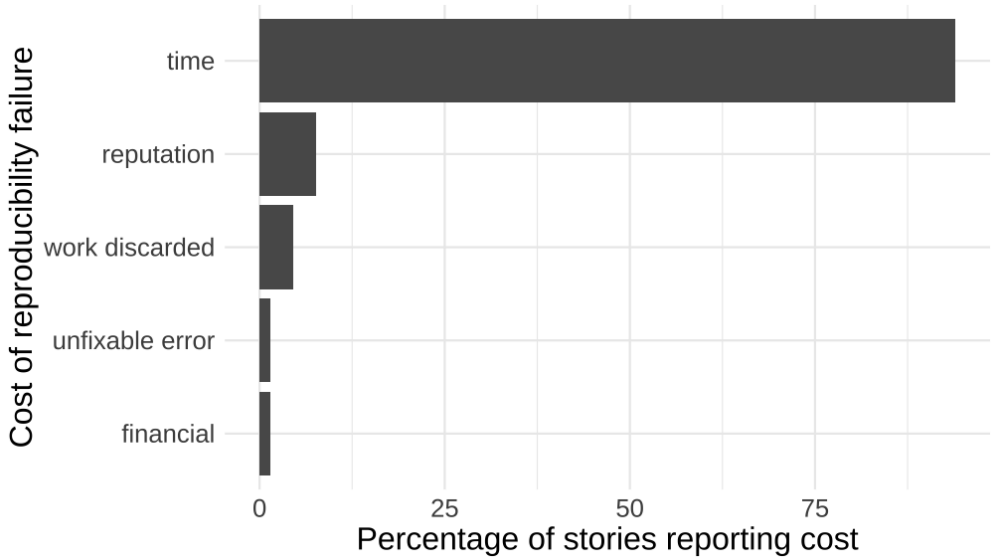
“When I was working in my last job creating a database, I was working between two different computer systems. I had written the code on my local computer (a windows) but it was being deployed on a Linux command line system. To debug my code it was easier to do on my computer but certain lines of code were specific to the linux machine vs. the window machine. While I was using version control I had a really hard time keeping track of which script was which and it caused me to waste lots of time and energy trying to fix the issue”

-- Master of Data Science student, UBC

- Reproducibility failure > lack of a shippable and shareable computational environment
- Impact of reproducibility failure > loss of time

Causes and impacts of reproducibility failures from students' stories from the trenches





Tell stories from the trenches

Exercise prompt for further engagement:

Follow the instructions below for each story share in the class stories from the trenches Google document.

- Read the story and reflect on which of the themes listed below was likely the biggest cause of the reproducibility or transparency failure described in the story.
- Label the story with the emoji corresponding to the chosen theme from the table below:

Reproducibility and transparency themes	Emoji label
Code	⌨️ (keyboard)
Computational environments	💻 (laptop)
Software design	💾 (floppy disk)
Data analysis pipeline	➡️ (right arrow)
Documentation	📄 (page facing up)
Project organization	🗃️ (card file box)
Randomness	🎲 (game die)
Version control	📜 (scroll)

Exercise prompt for further engagement (cont'd):

- Re-read the story and reflect on what the primary cost was from the reproducibility or transparency failure described in the story.
- Label the story with the emoji corresponding to the chosen cost from the table below:

Reproducibility and transparency failure cost	Emoji label
Financial	💸 (money with wings)
Reputational	🎖️ (military medal)
Time	🕰️ (mantelpiece clock)
Unfixable error	💥 (collision)
Work discarded.	🗑️ (wastebasket)

Tell stories from the trenches

Why I really like this exercise:

- Self reflection
- Community building ("you are not alone!")
- Engaging peer learning
- Relatable

UBC MDS student classroom

Strategies for lots of practice!!!

1. "I do, you do"
2. Interleave reproduciblity practice with other data science tasks
3. Pre-bake exercises

Pre-bake exercises

Some reproducibility workflows require some time consuming, or complex prerequisite work to have been done before it makes sense to do them.

In such cases. "pre-baking" exercises to a particular point is helpful for providing practice opportunities.

Source: https://peopletv.com/video/martha-bakes-angel-food-cake/

Examples include:

- automating the running of scripts via a data analysis pipeline (e.g., Makefile)
- using `bookdown` or Jupyter book to automate figure and table numbering
- organizing GitHub issues into milestones and a project board (example exercise)
- performing a code review on a pull requests (example exercise)

Pre-baked review my pull request exercise

Reviewing pull requests is a hard task to get a lot of practice on!

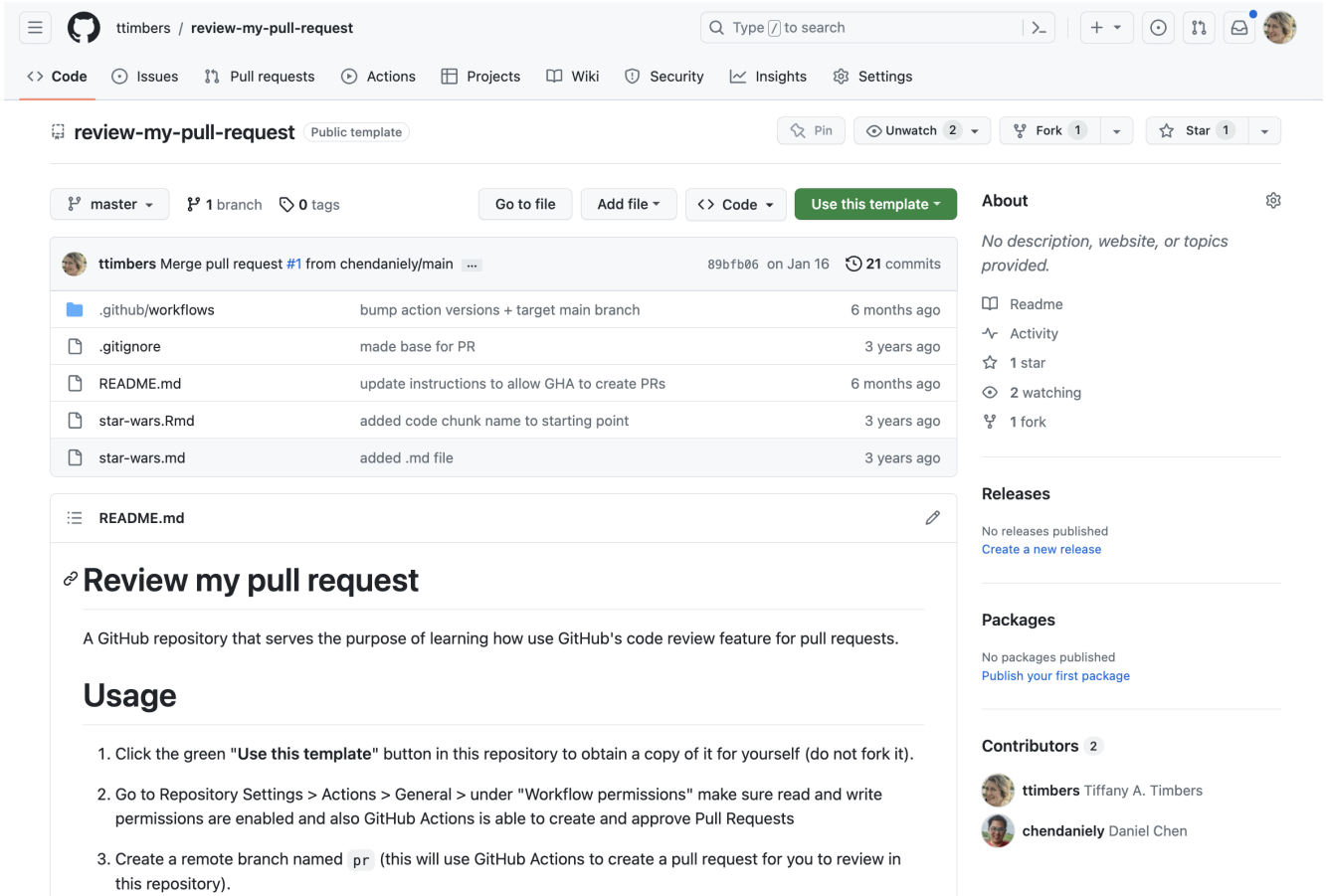
Why?

Before a pull request can be reviewed, a lot needs to happen to set the stage:

- Setup a GitHub repository
- Commit some files
- Make a change on another branch (or in a forked repo)
- Open a pull request

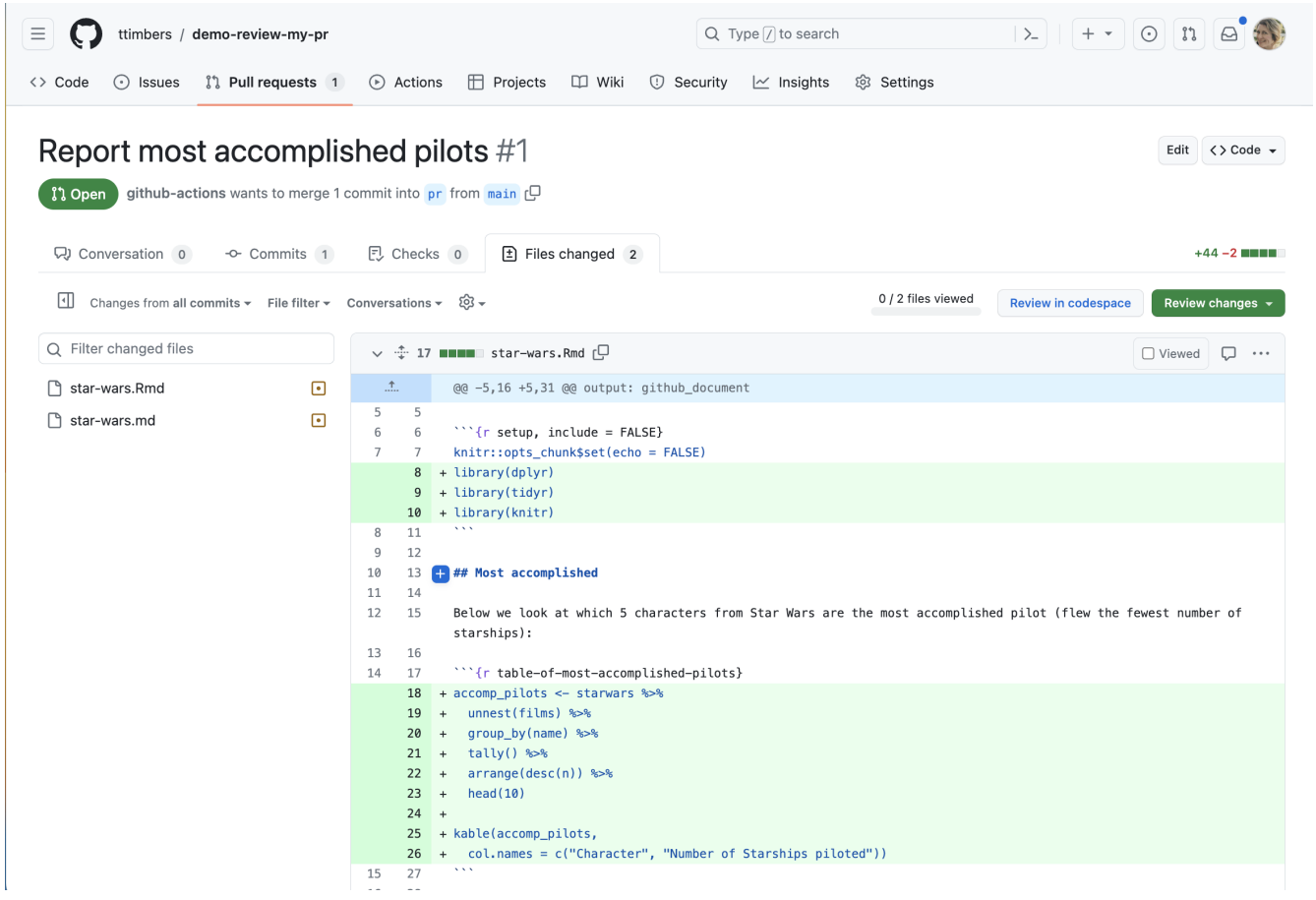
Source: Image by jcomp on Freepik

Exercise: Pre-baked review my pull request exercise



review-my-pull-request GitHub repository

Exercise: Pre-baked review my pull request exercise



review-my-pull-request GitHub repository

Pre-baked review my pull request exercise

Why I really like this exercise:

- Allows focused practice on what students are supposed to be learning
- Mistakes on earlier steps do not prevent practice
- Very scalable

UBC MDS student classroom

Wrap up

Key things for teaching reproducibility

1. Extra emphasis on motivation

2. Guided instruction

3. Lots of practice!!!

Source: https://giphy.com/gifs/opinion-the-big-lebowski-x1yhRwlqxBiQU

Anectodal evidence

- reduced technical debt in UBC Master of Data Science Capstone projects as we have gained more experience teaching reproducibility

Feedback from Alumni

- *"I had my [COMPANY] Git training today for the data engineers/scientists and I wanted to tell you that everything I learned in your class about git was extremely helpful. You did such an amazing job teaching everything we need to know about git and working in technology, (I personally think I learned much more in your class than in my training here in SF)."*
- *"...the use of Git and Docker and the strong focus on reproducibility are of the best technical skills our program offer."*

References

Gardner, Benjamin and Rebar, Amanda L. (Apr. 2019). “Habit Formation and Behavior Change”. en. In: Oxford Research Encyclopedia of Psychology. Oxford University Press. isbn: 978-0-19-023655-7. doi: 10.1093/acrefore/9780190236557.013.129.

National Academies of Sciences, Engineering, and Medicine and others (2019). Reproducibility and replicability in science. National Academies Press.

Ostblom, Joel and Timbers, T.A. (2022). Opinionated Practices for Teaching Reproducibility: Motivation, Guided Instruction and Practice. Journal of Statistics and Data Science Education 30(3) pp. 241–250. https://doi.org/10.1080/26939169.2022.2074922

Ram, Karthik (2013). “Git can facilitate greater reproducibility and increased transparency in science”. In: Source code for biology and medicine 8.1, pp. 1–8.

Credits

Title slide illustration was created by pch.vector - www.freepik.com.

Acknowledgements

UBC Master of Data Science teaching team (past and present)

- Tomas Beuzen
- Vincenzo Coia
- Giulio Valentino Dalla Riva
- Florencia D'Andrea
- Mike Gelbart
- Gittu George
- Varada Kolhatkar
- Rodolfo Lourenzutti
- Firas Moosvi
- Quan Nguyen
- Joel Ostblom (co-author on this work)
- Alexi Rodríguez-Arelis
- Arman Seyed-Ahmadi

DSCI 100 teaching team

- Trevor Campbell
- Melissa Lee

2021-22 Master of Data Science teaching team

Questions?

@TiffanyTimbers

talk slides: *url-TBD*

paper: https://www.tandfonline.com/doi/pdf/10.1080/26939169.2022.2074922