Penn Medicine

BGS Orientation 2024

Keeping an electronic lab notebook when your lab is your laptop

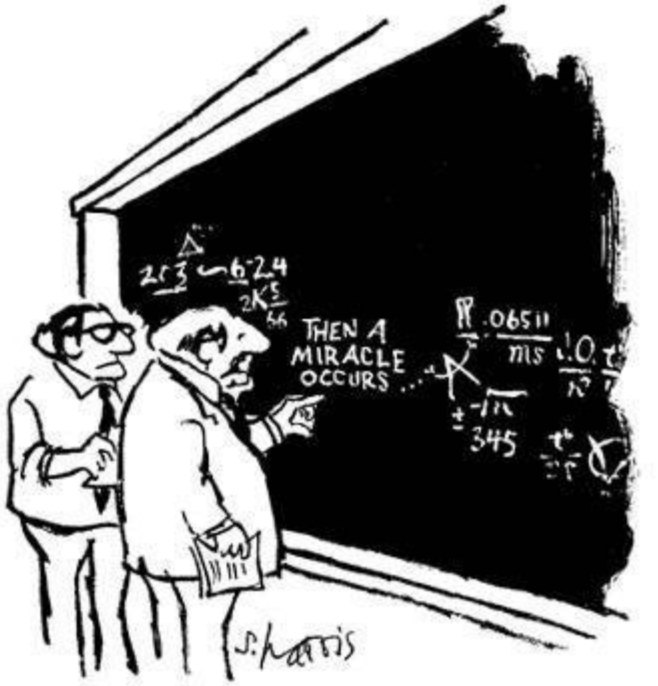
PhD Candidate in Biostatistics: Jeremy Rubin

Adviser: Jarcy Zee, PhD

08/20/2024

I’m just going to leave this proof out to dry…

- In a dry lab setting, our “methods” consist of some combination of **mathematical/on-paper calculations as well code that either analyzes real data or simulations some real-world application of interest**
- The focus of this presentation will be on considerations for documenting your code, but here are a few considerations for **mathematical derivations**
  - Check the conventions in your field for publishing technical reports that may be collections of proofs rather than complete journal articles (Example paper: 0912.4045.pdf (arxiv.org))
    - Ensuring your work is on arXIv or other field-specific repository ensures that your theorems and corresponding proofs can be **credited to you!**
    - Perhaps a “begin with the end in mind mentality” – if you know your intended journal’s proof inclusion and formatting criteria, then as you work on your derivations you can format them in the way they’ll ultimately need to be presented
  - **Overleaf**: Overleaf, Online LaTeX Editor
    - A cloud-based google doc-like text editor for LaTeX – a mathematical typesetting language
    - Allows for easy sharing of mathematical derivation development and even draft manuscripts with a of mathematical language between student/adviser/collaborators

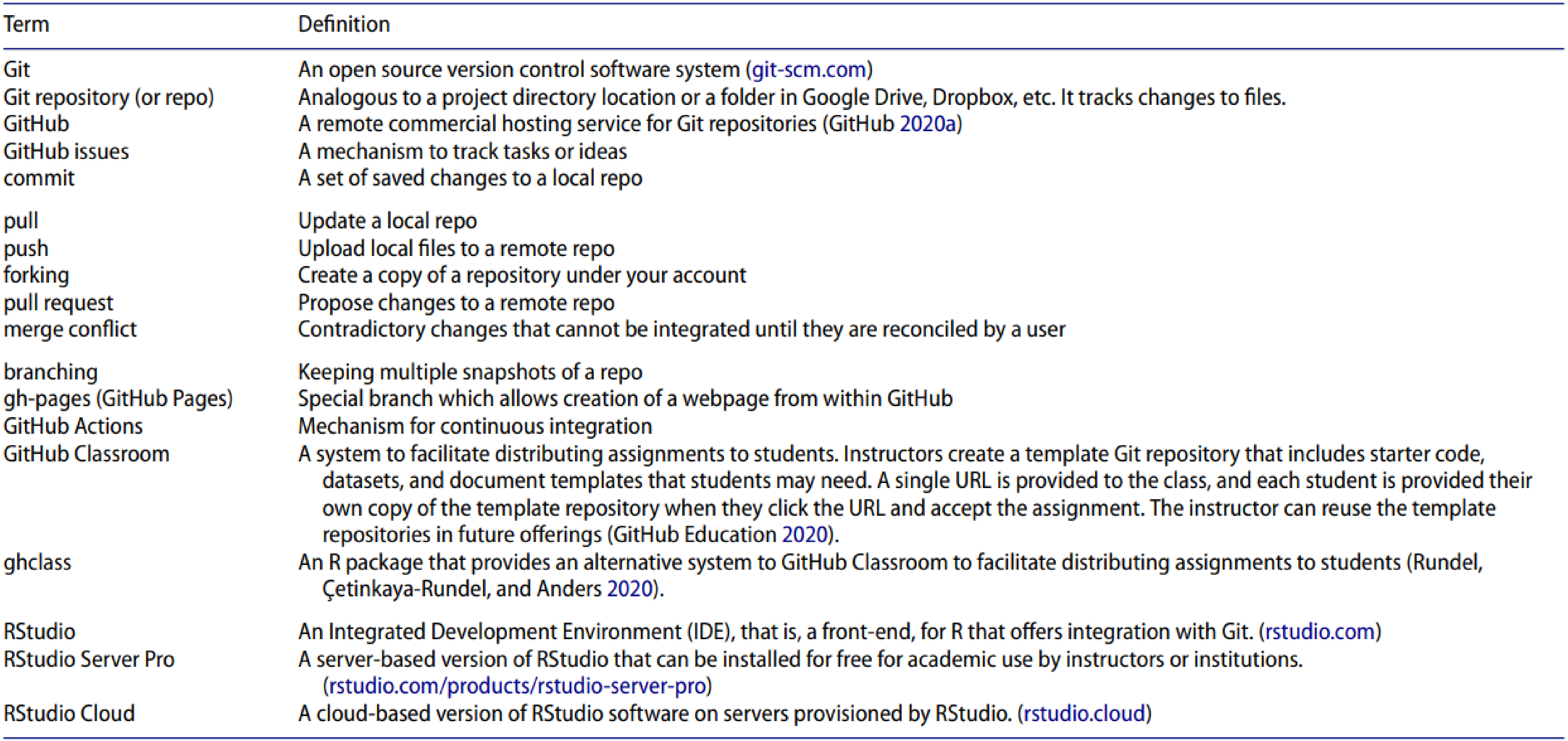


“I think you should be more explicit here in step two.”

Hopefully you like this version of my talk…

- **Three most popular web-hosting platforms for projects version control**: GitHub, GitLab, and Bitbucket
  - **GitHub** is recognized as the industry standard platform for hosting and collaborating on version controlled files via Git
    - **Records changes to a file or set of files over time so that changes can be tracked and specific versions of a file can be recalled later**
  - **Ability to use version control systems is a highly desired skill in industry when writing code is part of the job**
- **Challenges with GitHub**
  - Even when code is in private repositories, it is on GitHub servers, so data leaves the university – **must check data use agreements for projects to determine whether it can leave Penn networks**

Table 1. Definitions of common terms.



Term	Definition
Git	An open source version control software system (git-scm.com)
Git repository (or repo)	Analogous to a project directory location or a folder in Google Drive, Dropbox, etc. It tracks changes to files.
GitHub	A remote commercial hosting service for Git repositories (GitHub 2020a)
GitHub issues	A mechanism to track tasks or ideas
commit	A set of saved changes to a local repo
pull	Update a local repo
push	Upload local files to a remote repo
forking	Create a copy of a repository under your account
pull request	Propose changes to a remote repo
merge conflict	Contradictory changes that cannot be integrated until they are reconciled by a user
branching	Keeping multiple snapshots of a repo
gh-pages (GitHub Pages)	Special branch which allows creation of a webpage from within GitHub
GitHub Actions	Mechanism for continuous integration
GitHub Classroom	A system to facilitate distributing assignments to students. Instructors create a template Git repository that includes starter code, datasets, and document templates that students may need. A single URL is provided to the class, and each student is provided their own copy of the template repository when they click the URL and accept the assignment. The instructor can reuse the template repositories in future offerings (GitHub Education 2020).
ghclass	An R package that provides an alternative system to GitHub Classroom to facilitate distributing assignments to students (Rundel, Çetinkaya-Rundel, and Anders 2020).
RStudio	An Integrated Development Environment (IDE), that is, a front-end, for R that offers integration with Git. (rstudio.com)
RStudio Server Pro	A server-based version of RStudio that can be installed for free for academic use by instructors or institutions. (rstudio.com/products/rstudio-server-pro)
RStudio Cloud	A cloud-based version of RStudio software on servers provisioned by RStudio. (rstudio.cloud)

But like is there something I can do that requires less effort?

- **General tools/strategies**
  - **Box** – Free Box account through Penn!
    - A ton of storage
    - May be more comfortable to collaborators who haven't used GitHub
  - **Journaling** – email threads with collaborators, meeting minutes, etc.
- **R**
  - **Rmarkdown** - Can produce code, output, and written analysis using the rmarkdown package
  - **Renv package** - keeps track of versions of packages you've used so you consistently are using the same environment for analyses
  - **Targets package** – helps organize workflow
  - **Verified, public repositories** – CRAN, bioconductor
- **Python**
  - **Juptyer Notebook** – Open-source web application to make and share documents with live code, equations, visualizations, and text
  - **Google Colab** – Cloud-based platform for writing an executing Python code

Thanks for listening!