Knit, Commit, and Push

Teaching version control in undergraduate statistics courses

Maria Tackett
Duke University

Toronto Workshop on Reproducibility
February 24, 2022

🔗 bit.ly/teach-version-control

Introduction
+
Motivation

Teaching
Version Control

Lessons Learned

Introduction
+
Motivation

Teaching
Version Control

Lessons Learned

Teaching a reproducible workflow

- Part of replicating a scientific study is the ability to **reproduce the statistical analysis**
- Students need facility with **professional statistical analysis software** - 2014 ASA Curriculum Guidelines for Undergraduate Programs
- **Workflow and reproducibility** are important components of "data acumen" - 2018 National Academies report *Data Science for Undergraduates: Opportunities and Options*
- **Version control** as key topic for statistical analysis, particularly for coordinating work across a team - Nolan and Temple Lang (2010)

Key terms

Version control: recording changes to a file or set of files over time

Git: Version control software system

GitHub: Commercial hosting service for Git repositories (folders)

Why teach version control in statistics courses?

✓Important component of reproducible workflow

- More easily track analysis history and provide analysis provenance
- Keep track of files being modified by multiple people

✓Equip students with important computing skills

- Students can do a lot using basic functionality
- Develop skills throughout curriculum if introduced in early courses

Beckman et al. (2021)

- Instructors from multiple institutions share experience implementing version control using Git and GitHub in their courses
  - Represents courses throughout statistics curriculum from introductory to graduate-level
- Focus on implementation: computing toolkit, first exposure in class, assessment
- Discussion on pedagogical approach

This talk focuses on implementation in a second semester course

Introduction
+
Motivation

Teaching
Version Control

Lessons Learned

Course description

~ 90 students who have taken introductory statistics, data science, or probability course

Topics: Linear regression, logistic regression, and ANOVA with focus on application

Activities: In-class exercises, computing labs, homework, quizzes, group project

Computing toolkit

- R Markdown for write up
- Run Git commands using point-and-click interface
- Server-based RStudio*
  - Git already configured
  - Same set up for all students

- Assign and submit assignments
- Facilitates collaboration on group assignments
- Course management using **`ghclass`** R package (or GitHub Classroom**)

*Çetinkaya-Rundel, M., and Rundel, C. (2018), "Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum," The American Statistician, 72, 58–65,

**Fiksel, J., Jager, L. R., Hardin, J. S., and Taub, M. A. (2019), "Using GitHub Classroom to Teach Statistics," Journal of Statistics Education, 27, 100–119.

First exposure in class

- Lecture introducing reproducible workflow and computing toolkit
  - Help students understand value early on
- Start with individual assignments and using scaffolding to ease students into the new workflow

Individual assignment

This is another good place to knit, commit, and push changes to your remote lab-01 repo on GitHub. Write an informative commit message (e.g. “Completed exercises 5 - 8”), and push every file to GitHub by clicking the checkbox next to each file in the Git pane. After you push the changes, the Git pane in RStudio should be empty.

Group assignment

✅ ⬆️ **Team Member 1**: Knit, commit and push your changes to GitHub with an appropriate commit message again. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.

All other team members: **Pull** to get the updated documents from GitHub. Click on the .Rmd file, and you should see the responses to exercises 1- 4.

Team Member 2: It’s your turn! Type the team’s response to exercises 5 - 7.

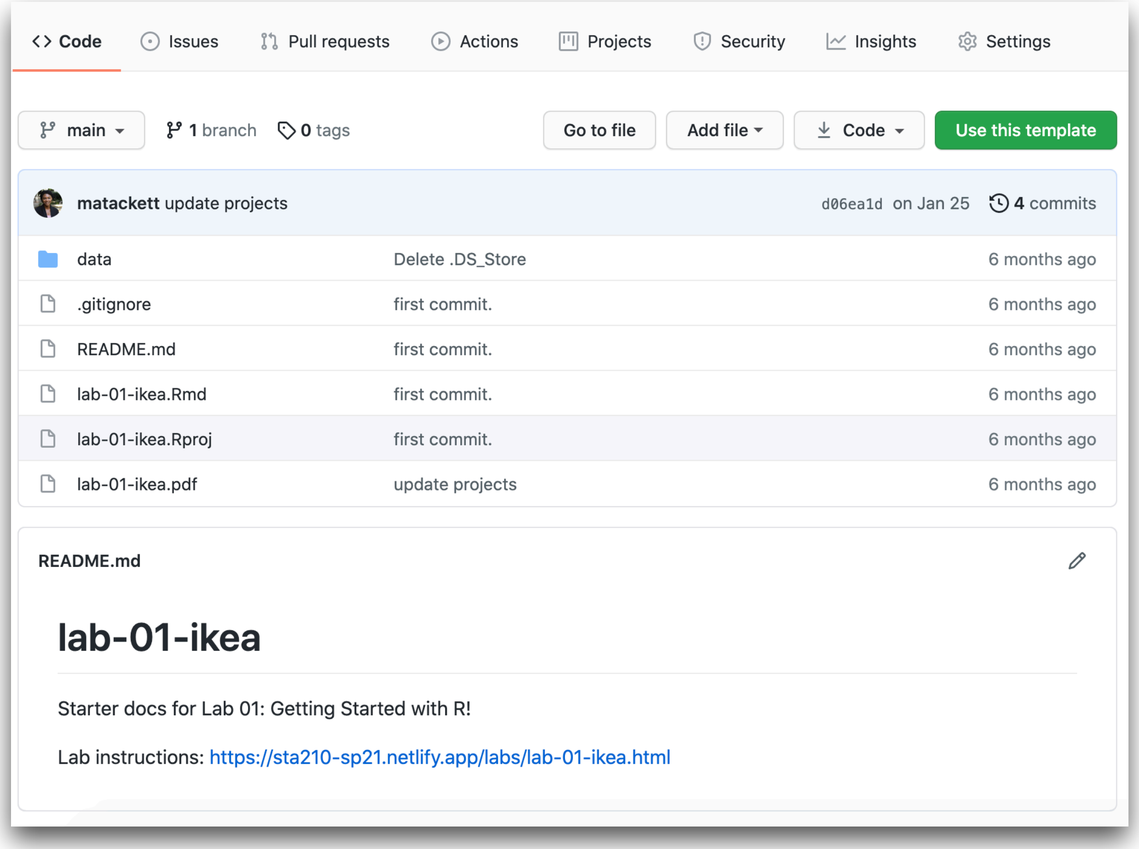
ghclass

- R package for managing courses on GitHub
- Use R functions to distribute assignments, manage teams, collect student work
- Reproducible administrative workflow

Developed by Colin Rundel & Mine Çetinkaya-Rundel

rundel.github.io/ghclass

Assignment workflow



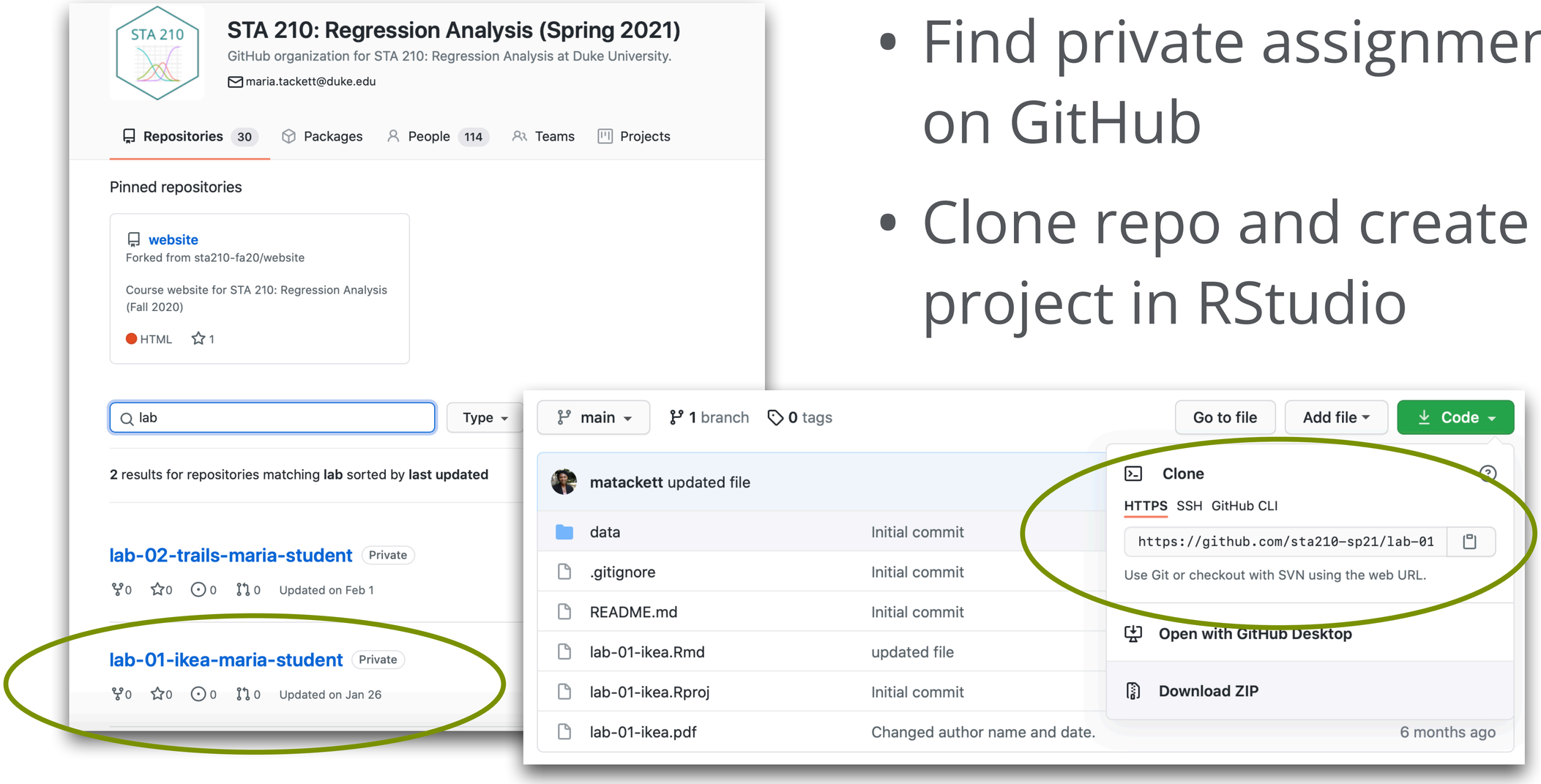
Instructor

- Create starter repo in GitHub.
  - Includes R Markdown documents, data set, etc.
- Make a copy of the starter repo for each student (or group) using **`ghclass`**

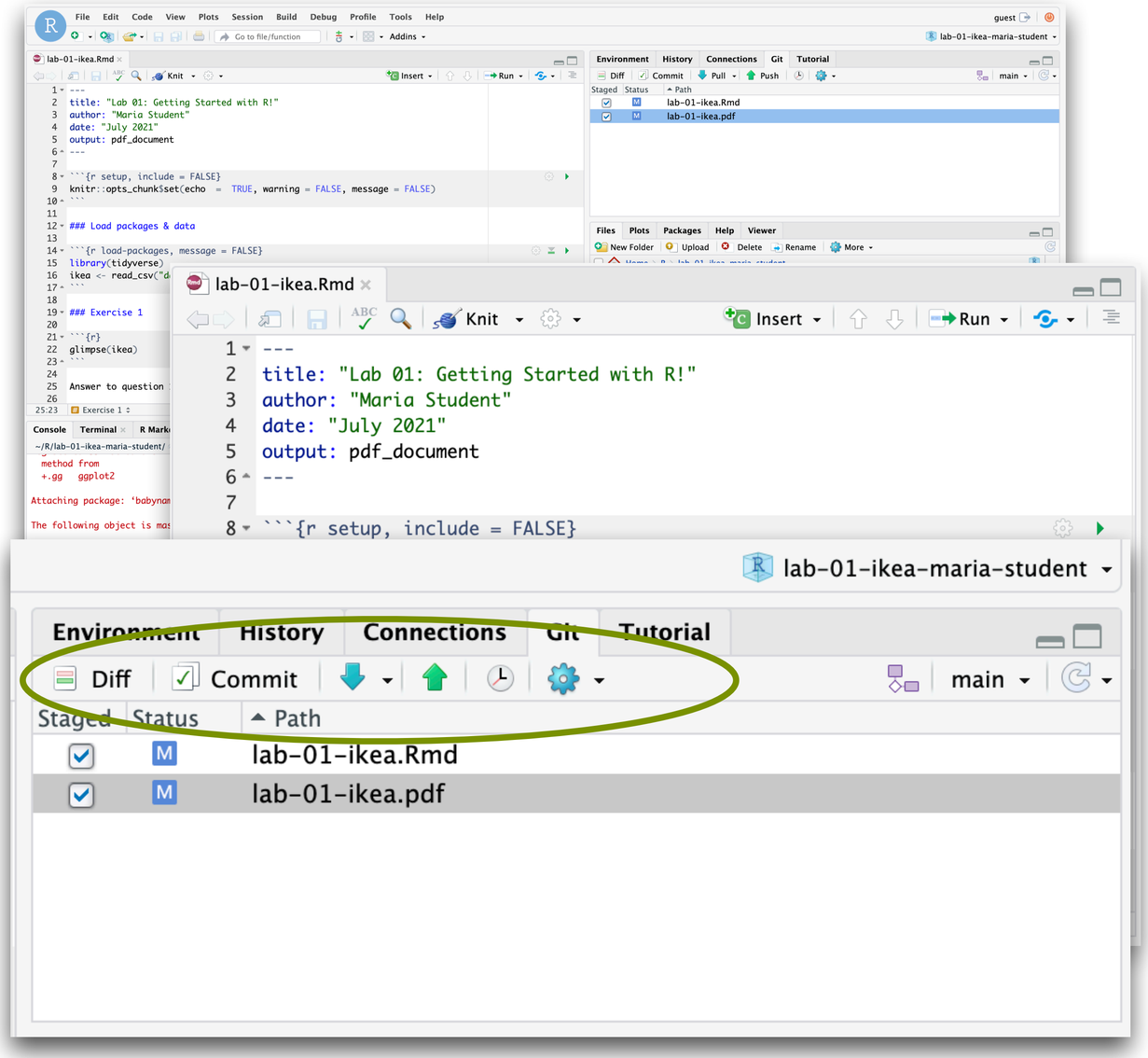
Assignment workflow

Student

- Find private assignment repo on GitHub
- Clone repo and create a new project in RStudio



Assignment workflow



Student

- As they work on assignment:
  - Write code and narrative in R Markdown document
  - ***"Knit, commit, and push"*** work to GitHub repo

- Submit PDF for grading on Gradescope

Assessment

Individual assignments

- ~ 10% of grade for formatting, reproducibility, and version control
- Assessed based on regular commits (3+) and informative commit messages

Group assignments

- ~ 10% of grade for formatting, reproducibility, version control, and collaboration
- Each group member’s contribution partially assessed based on commit history

Introduction
+
Motivation

Teaching
Version Control

Lessons Learned

Challenges + lessons learned

New workflow + computing tools

- Students may have an established workflow
- Emphasize the value of a reproducible workflow, particularly for collaboration

Getting over learning curve

- Git can have a steep learning curve
- Keep it simple (basic functionality, start with individual assignments)

Remember the “why”

- Some disconnect when students submit work to grading system or LMS
- Reiterate the role of version control by sharing experiences from academia and industry

Resources for implementation

Pedagogy

- Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing version control with Git and GitHub as a learning objective in statistics and data science courses. Journal of Statistics and Data Science Education, 29, 132-144. DOI: 10.1080/10691898.2020.1848485
- Çetinkaya-Rundel, M. (2020), "Data Science in a Box," available at https://www.datasciencebox.org

Computing

- Bryan, J. (2018), "Happy Git and GitHub for the useR," GitHub, available at https://happygitwithr.com.
- Çetinkaya-Rundel, M., and Rundel, C. (2018), "Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum," The American Statistician, 72, 58–65, DOI: 10.1080/00031305.2017.1397549.
- Fiksel, J., Jager, L. R., Hardin, J. S., and Taub, M. A. (2019), "Using GitHub Classroom to Teach Statistics," Journal of Statistics Education, 27, 100–119. DOI: 10.1080/10691898.2019.1617089
- Rundel, C., Çetinkaya-Rundel, M., and Anders, T. (2020), "ghclass: Tools for Managing Classes With GitHub," available at http://github.com/rundel/ghclass

Additional resources

Computing in statistics and data science curriculum

- Nolan, D., and Temple Lang, D. (2010), “Computing in the Statistics Curriculum,” *The American Statistician*, 64, 97–107. DOI: 10.1198/tast.2010.09132

- National Academies of Science, Engineering, and Medicine (2018), “Data Science for Undergraduates: Opportunities and Options,” available at https://nas.edu/envisioningds

- American Statistical Association (2014), “Curriculum Guidelines for Undergraduate Programs in Statistical Science,” available at http://www.amstat.org/education/curriculumguidelines.cfm

- (2021). Computing in the Statistics and Data Science Curriculum [Special issue]. *Journal of Statistics and Data Science Education*, 29(sup1), available at https://www.tandfonline.com/toc/ujse21/29/sup1

- **Minimal GitHub website**: https://nicholasjhorton.github.io/Minimal-GitHub/

Thank You!

maria.tackett@duke.edu

@MT_statistics

bit.ly/teach-version-control