Version control as a learning objective in statistics and data science courses

Maria Tackett
Duke University

63rd World Statistics Congress
July 2021

[Add link to slides]

Teaching a reproducible workflow

- Part of replicating a scientific study is the ability to **reproduce the statistical analysis**

- **Workflow and reproducibility** as important components of “data acumen” - 2018 National Academies report *Data Science for Undergraduates*

- *“Students need **facility with professional statistical analysis software**”* - 2014 ASA Curriculum Guidelines for Undergraduate Programs

- Nolan and Temple Lang (2010) promoted **version control** as key topic for statistical analysis

Why teach version control in statistics courses?

✓ **Important component of reproducible workflow**

- Version control makes it more feasible to track analysis history and provide analysis provenance
- Makes it more feasible to keep track of versions of analysis and data files being modified by multiple people

✓ **Equip students with computing skill widely in industry and academia**

- Students can do a lot using basic functionality
- Can learn more advanced functionality in later courses if version control introduced in early courses

Beckman et al. (2021)

- Instructors from multiple institutions share experience implementing version control in their courses
  - Represents courses throughout curriculum from intro to graduate-level
- Focus on implementation: computing toolkit, first exposure in class, assignments & assessments, additional remarks
- Discussion on pedagogical approach

This talk focuses on implementation in a second semester course

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

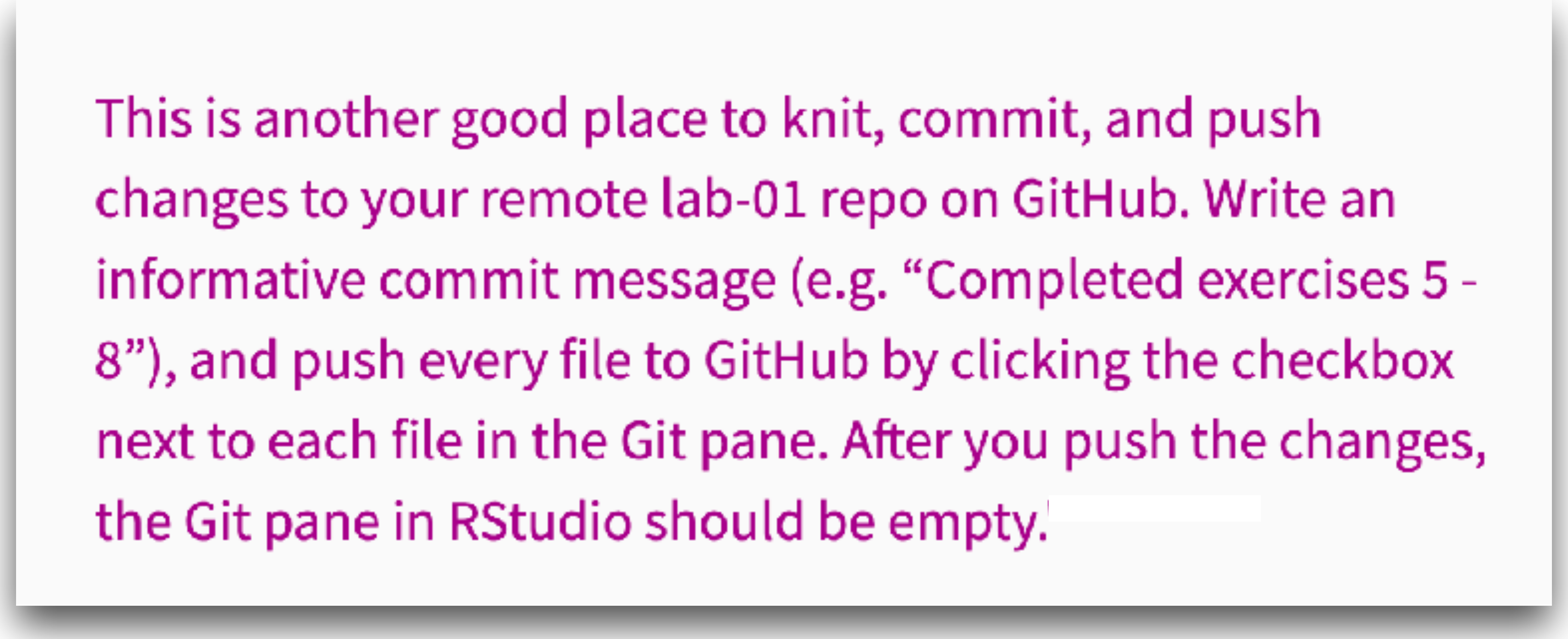
- R Markdown for analysis and write up
- Run git commands through git pane
- Server-based RStudio
  - Git already configured
  - Equitable computing capabilities

- Assign and submit assignments
- Collaboration on group assignments
- Course management using **`ghclass`** R package or GitHub classroom

First exposure in class

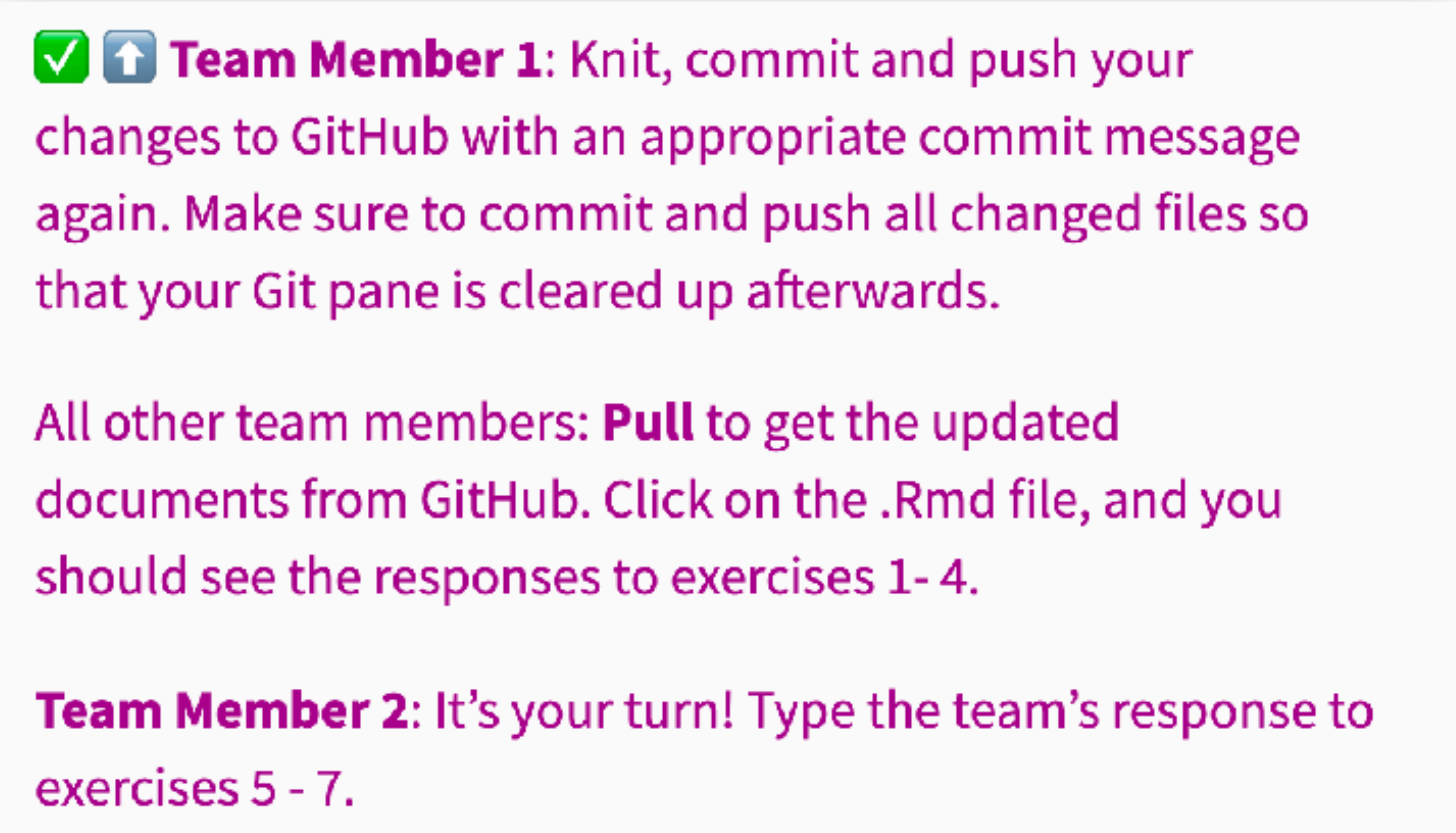
- Lecture introducing reproducible workflow and computing toolkit
  - Help students understand value early on
- Start with individual assignments and using scaffolding to ease students into the new workflows

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

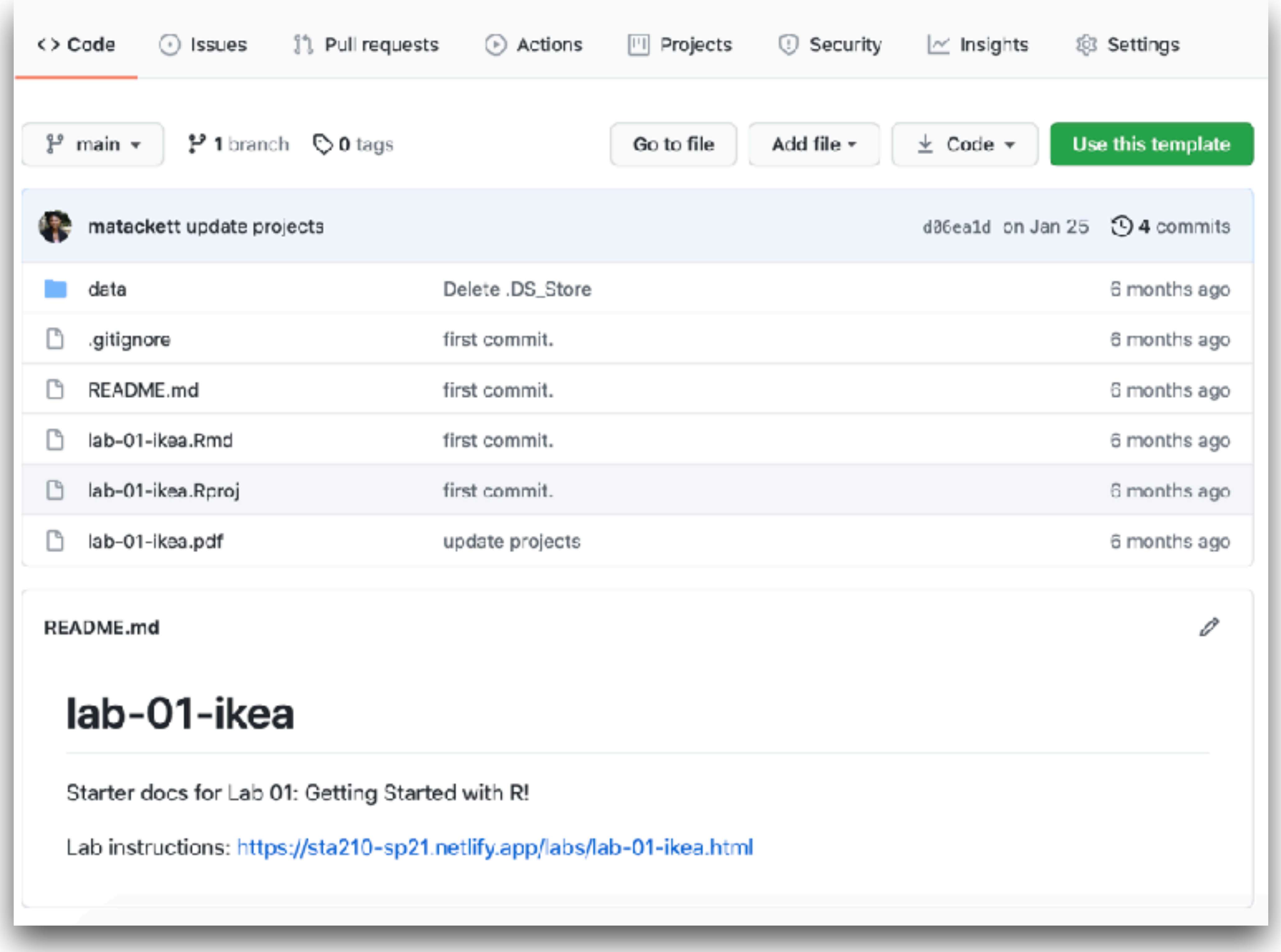


✅ ⬆️ **Team Member 1**: Knit, commit and push your changes to GitHub with an appropriate commit message again. Make sure to commit and push all changed files so that your Git pane is cleared up afterwards.

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.

Assignment workflow



matackett update projects d86ea1d on Jan 25 4 commits

data	Delete .DS_Store	6 months ago
.gitignore	first commit.	6 months ago
README.md	first commit.	6 months ago
lab-01-ikea.Rmd	first commit.	6 months ago
lab-01-ikea.Rproj	first commit.	6 months ago
lab-01-ikea.pdf	update projects	6 months ago

README.md

lab-01-ikea

Starter docs for Lab 01: Getting Started with R!

Lab instructions: https://sta210-sp21.netlify.app/labs/lab-01-ikea.html

Instructor

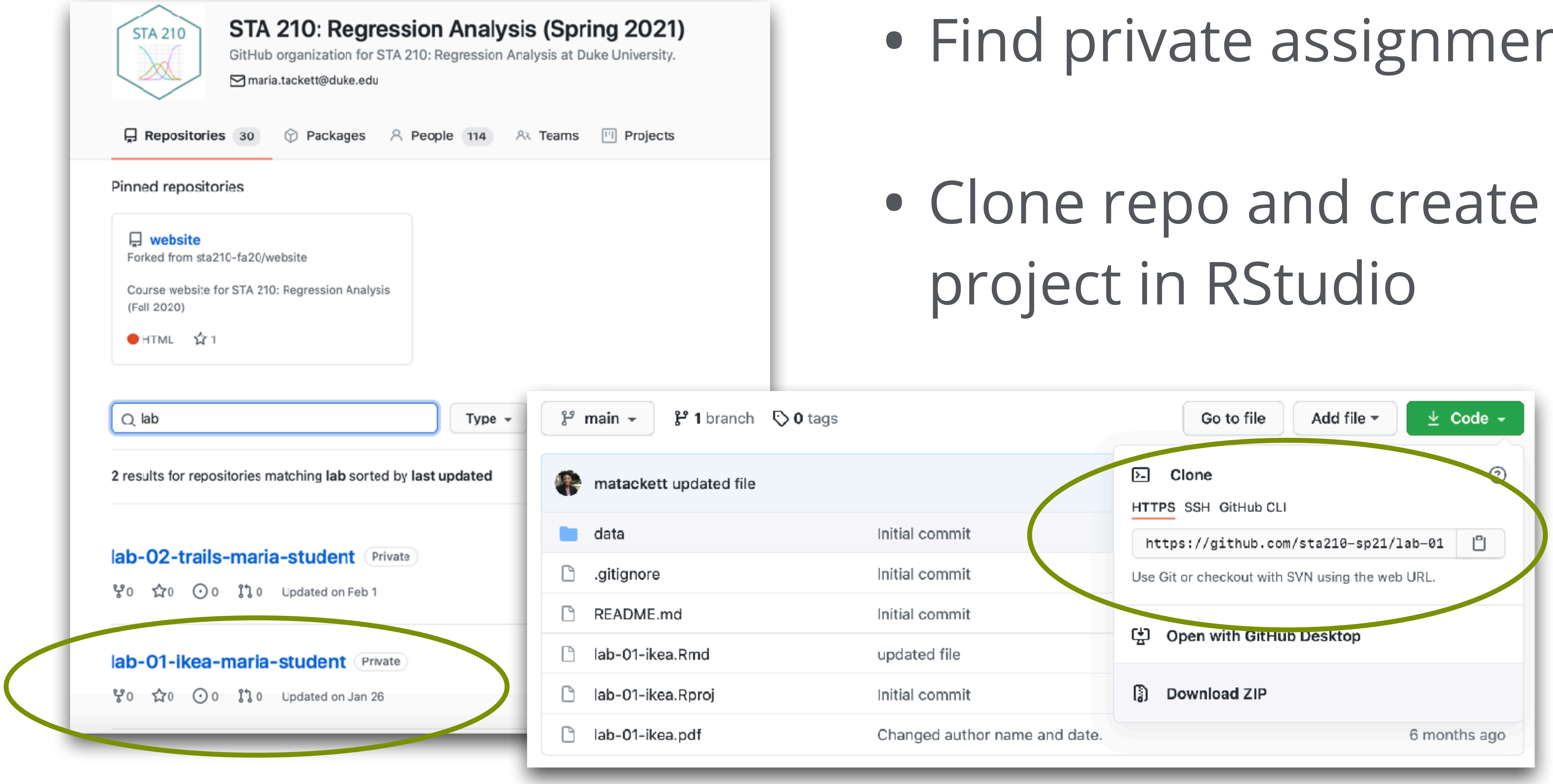
- Create starter repo in GitHub.
  - Includes R Markdown documents and data sets

- Make a copy of the starter repo for each student (or team) using **`ghclass`** R package.

Assignment workflow

Student

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



STA 210

STA 210: Regression Analysis (Spring 2021)

GitHub organization for STA 210: Regression Analysis at Duke University.

maria.tackett@duke.edu

Repositories 30 · Packages · People 114 · Teams · Projects

Pinned repositories

website
Forked from sta210-fa20/website

Course website for STA 210: Regression Analysis (Fall 2020)

● HTML ☆ 1

lab · Type

2 results for repositories matching **lab** sorted by **last updated**

lab-02-trails-maria-student Private
0 ☆0 0 0 Updated on Feb 1

lab-01-ikea-maria-student Private
0 ☆0 0 0 Updated on Jan 26

main · 1 branch · 0 tags · Go to file · Add file · Code

matackett updated file

data	Initial commit
.gitignore	Initial commit
README.md	Initial commit
lab-01-ikea.Rmd	updated file
lab-01-ikea.Rproj	Initial commit
lab-01-ikea.pdf	Changed author name and date.

6 months ago

Clone

HTTPS SSH GitHub CLI

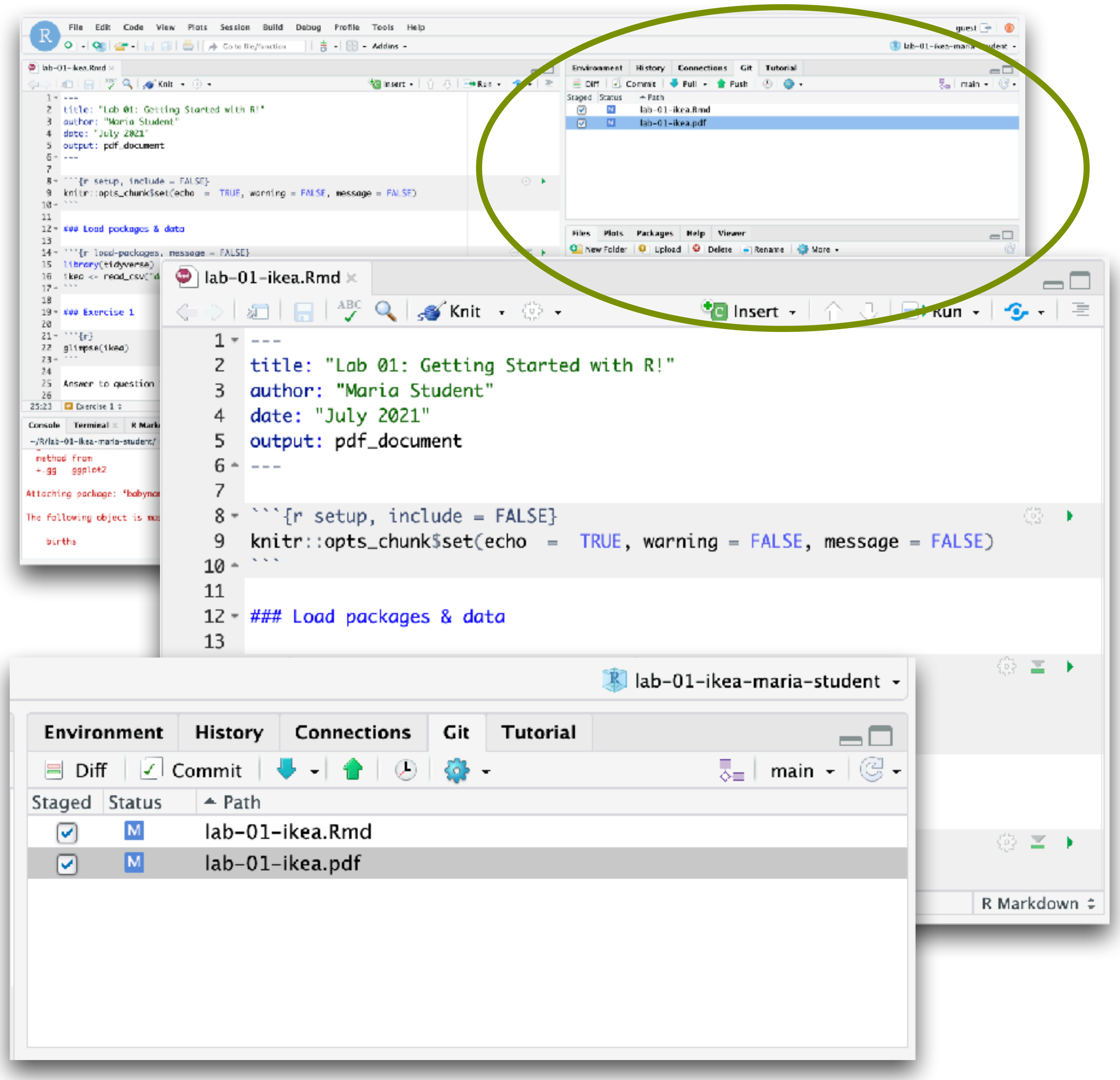
```
https://github.com/sta210-sp21/lab-01
```

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

Assignment workflow



~~~~
---
title: "Lab 01: Getting Started with R!"
author: "Maria Student"
date: "July 2021"
output: pdf_document
---

```{r setup, include = FALSE}
knitr::opts_chunk$set(echo  = TRUE, warning = FALSE, message = FALSE)
```


### Load packages & data

~~~~

Staged	Status	Path
☑	M	lab-01-ikea.Rmd
☑	M	lab-01-ikea.pdf

Student

- Repeat the following while completing assignment:
  - Write code and narrative in R Markdown file.
  - *"Knit, commit, and push"* work to GitHub repo
- Final submission on Gradescope, online grading platform

Assessment

Individual assignments

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

Group assignments

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

Getting students over the learning curve

✓ **Help students understand the value of a reproducible workflow**

✓ **Keep it simple** and focus only on functionality needed for the course (commit, push, pull, dealing with merge conflicts)

- Use functionality through Git pane in RStudio

✓ **Start with a few individual assignments** before introducing group work

- Use scaffolding to ease students into the new workflow

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

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

Thank You!

✉ maria.tackett@duke.edu

@MT_statistics

[Add link to slides]