

Utilizing Open Source Resources to Teach Introductory Data Science

Tyler George

tgeorge@cornellcollege.edu

 @stats-tgeorge

 stats_tgeorge

Slides: bit.ly/3zvy8L6

About Cornell College

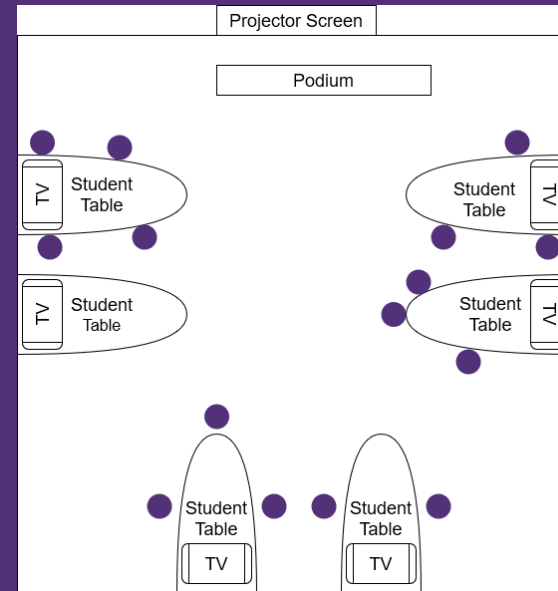
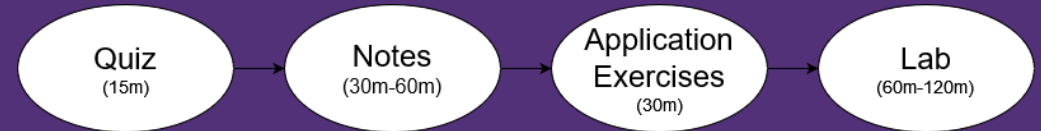
- Small liberal arts college of about 1000 students
- One-Course-at-A-Time block schedule
 - Each class occurs over 3.5 weeks including 18 days of 4 hours of instructional time and three weekends
- **New majors in Applied Statistics and Data Science**

A Day of Class

- This course has a pre-requisite of *either* introductory statistics or computer science (python)
- Each day consists of 2, 2 hour sessions

Student homeworks:

- R homework similar to labs
- Readings related to ethics
- Group project

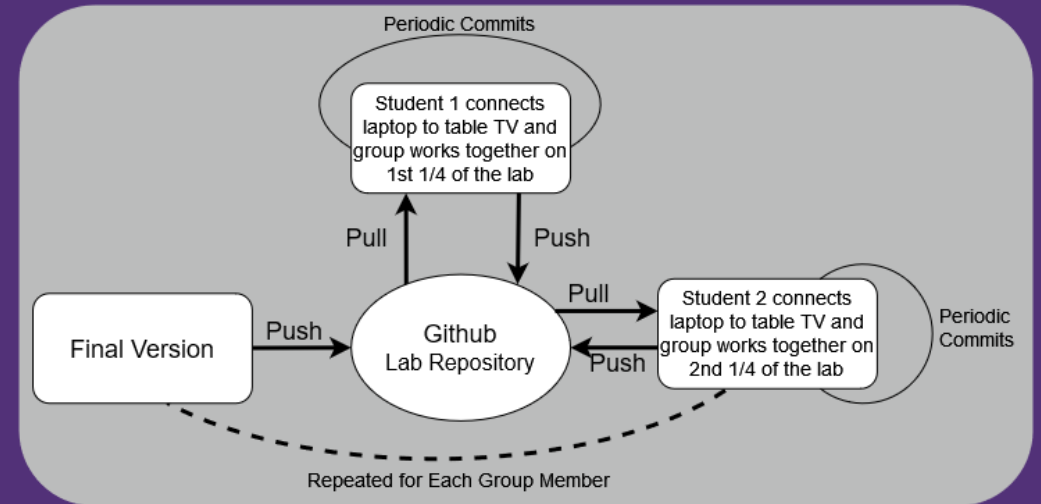


Utilized Resources

- [Data Science in a Box](#) by Mine Çetinkaya-Rundel (2021)
- Data science ethics assignments from the [Quantitative Analysis Institute](#) at Wellesley College (Pattanayak, Gan, Li, Liang, and Wong, 2022)
- Workshops by Julia Silge available at <https://juliasilge.github.io/tidytext-tutorial/site/>
- Many R packages were used throughout the class but [ghclass](#) was particularly useful for teaching/running the course

Teaching Pedagogies

- Collaborative learning (Roseth, Garfield, and Ben-Zvi, 2008)
- Using tidyverse (Çetinkaya-Rundel, Hardin, Baumer, McNamara, Horton, and Rundel, 2022)
- Version control to lead to reproducible work (Beckman, Çetinkaya-Rundel, Horton, Rundel, Sullivan, and Tackett, 2021)



Learning the Bare Necessities

- I was new to Git, Github, RStudio Server (Linux), and a handful of the tidyverse packages and functions
- What did I *need* to learn to teach this class in this way?
 - Git and Github: Basic Git functions such as push, pulling, and handling git merge conflicts in RStudio
 - *ghclass* R package made using Git/Github for distributing content much smoother. See the [R script I created](#)

All course materials can be found at bit.ly/3Qq99Qb



Learning more than needed

- RStudio server implementation and management required learning some basic Linux commands. See the [short list of commands I ended up needing](#)
- Learning the Xaringan R package to modify slides (now I'm using it for this!)
 - I converted the notes to pdf's using *xaringanBuilder* package. See my [short script](#). *I will host these using Github pages next time.*

All course materials can be found at bit.ly/3Qq99Qb

Student Comments

- "I enjoy working with my group on the project, although it did take countless hours outside of the classroom I enjoyed it a lot and felt accomplished after presenting."
- "Working on labs and application exercises were beneficial for me. After completing those I could refer back to them which helped a lot. When working alone I often would get stuck on a exercise but when working in groups we were able to put our heads together and solve the problems. I really enjoyed that."
- "There were far too many notes in this class."

Future Updates

- Expanding the ethics integration in the course
 - Collaboration with Dr. Megan Altman, to integrate CARVE(R) principals into the data science discussions
 - **integrated ethics LABS**
 - Data science ethics modules from (Baumer, Garcia, Kim, Kinnaird, and Ott, 2022)
- Trying out the new **gradetools** R package to make grading quicker while giving useful feedback to students (Ricci, Medina, and Dogucu, 2022)
- Less statistical inference and more data wrangling
- Higher project and assessment expectations. The estimated time required was way off.
- Github actions to help students know if they submitted something that others can compile (reproducible)
- Quarto

JMS 2022

References

See the bibliography at bit.ly/3bCuh6L

Acknowledgment

Dr. Ajit Chavan, Assistant Professor of Computer Science at Cornell College, for setting up the cluster used to run the RStudio Server and answering many of my questions about its use.

Questions?

Thank You!

JMS 2022



Cornell College