Motivation: Version Control with Git as a Learning Objective in Statistics Courses

Matthew Beckman
Penn State University

August 4, 2020
JSM Virtual Conference

Who cares?

- 2014 ASA Curriculum Guidelines for [. . .] Statistical Science
- 2016 Curriculum Guidelines for [. . .] Data Science (i.e., “Park City Report”)
- CS education calls for version control in the curriculum (e.g., Haaranen & Lehtinen, 2015; Zagalsky et al., 2015)
- 2017 Kaggle Study

Reproducibility:

- completely self-contained including. . .
  - source data
  - code book
  - all data wrangling/prep steps
  - recreate all analysis, models, visuals
  - final reporting
- easy to verify results or refresh if source data updates
- e.g., all code “just works” with no changes needed

Version control

- maintains the evolution of the project
- safely explore alternative solutions/ideas in parallel

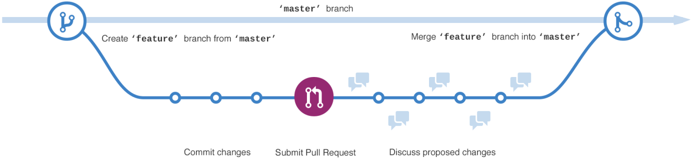
'master' branch
Create 'feature' branch from 'master'
Merge 'feature' branch into 'master'
Commit changes
Submit Pull Request
Discuss proposed changes

Figure 1: exploring parallel solutions (https://guides.github.com/activities/hello-world/)

Version control

- collaboration among users
- self-collaboration–e.g., RStudio Desktop and RStudio Server

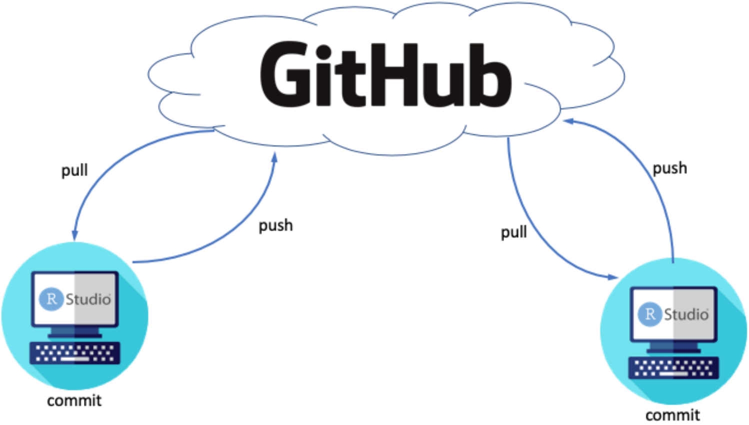

Figure 2: Collaboration schematic

Reproducibility $\neq$ Version Control

- Sometimes lumped together as if they're one in the same, and it's tempting to speak of Git(Hub) as a panacea. . .
- They aren't and it isn't. . .

Our motivation: invest in good habits with a professional workflow designed to streamline **both** virtues.

Ethical practice

- Any analysis may require hundreds of tiny decisions
- These decisions may necessarily be handled by a single person
- Work products are often intended for audience without technical expertise to scrutinize those decisions

With reproducibility & version control

- all decisions are documented
- all results can be checked
- proper scrutiny is possible (now or in future)

Industry & Academic Preparedness

- programming is a collaborative sport
- effective entry point for research participation

- Industry Preparedness
    - programming is a collaborative sport
    - quite common to refresh standard reports

Acknowledgments

References

1. American Statistical Association Undergraduate Guidelines Workgroup (2014). 2014 Curriculum guidelines for undergraduate programs in statistical science. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/curriculumguidelines.cfm
2. Beckman, M. D., Cetinkaya-Rundel, M., Horton, N., Rundel C., Sullivan A. J., & Tackett, M. (in review). Implementing version control with Git as a learning objective in statistics courses. Preprint URL: https://arxiv.org/pdf/2001.01988.pdf
3. De Veaux, R. D., et al. (2016). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and Its Application, 4:*2.1-2.16. URL: https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-060116-053930
4. Haaranen, L. & Lehtinen, T. (2015). Teaching git on the side: Version control system as a course platform, in *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*, ITiCSE '15, ACM, New York, NY, USA, pp. 87–92. URL: http://doi.acm.org/10.1145/2729094.2742608
5. Kaggle (2017). Kaggle machine learning & data science survey 2017. URL: https://www.kaggle.com/kaggle/kaggle-survey-2017
6. K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, & J. L. Olds (2015).

Q & A

Motivation: Version Control with Git as a Learning Objective in Statistics Courses

Matthew Beckman
Penn State University

August 4, 2020
JSM Virtual Conference

https://mdbeckman.github.io/JSM2020-Virtual/

Backup slide