

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 about reproducibility & VC?

- American Statistical Association ^{1 2 3}
- National Science Foundation ^{4 5}
- CS education (e.g., ACM SIGCSE) ^{6 7}
- Employers, Practitioners, & Students ⁸

¹ASA Undergraduate Guidelines Workgroup (2014)

²DeVeaux et al. (2016); a.k.a., The Park City Report

³Broman et al. (2017); ASA reproducibility recommendations

⁴Broman et al. (2017); ASA reproducibility recommendations

⁵Bollen et al. (2015); Cmte recommendations to NSF

⁶Haaranen & Lehtinen (2015); SIGCSE teaching with VC

⁷Zagalsky et al. (2015); SIGSCE teaching VC

⁸Kaggle (2017); user survey

The Kubler-Ross model ⁹

... better known as the 5 stages of grief:

- **Denial:** We never taught Git before, what's the big deal?

The Kubler-Ross model ⁹

... better known as the 5 stages of grief:

- **Denial:** We never taught Git before, what's the big deal?
- **Anger:** ANOTHER learning objective?! *SERIOUSLY??*

The Kubler-Ross model ⁹

... better known as the 5 stages of grief:

- **Denial:** We never taught Git before, what's the big deal?
 - **Anger:** ANOTHER learning objective?! *SERIOUSLY??*
 - **Bargaining:** Can't students just pick it up on their own?
-

The Kubler-Ross model ⁹

... better known as the 5 stages of grief:

- **Denial:** We never taught Git before, what's the big deal?
 - **Anger:** ANOTHER learning objective?! *SERIOUSLY??*
 - **Bargaining:** Can't students just pick it up on their own?
 - **Depression:** I don't know *anything* about Git...
-

The Kubler-Ross model ⁹

... better known as the 5 stages of grief:

- **Denial:** We never taught Git before, what's the big deal?
- **Anger:** ANOTHER learning objective?! *SERIOUSLY??*
- **Bargaining:** Can't students just pick it up on their own?
- **Depression:** I don't know *anything* about Git...
- **Acceptance:** Wait, there are buttons in RStudio? Maybe it won't be so bad after all...

⁹Kübler-Ross E (2005). *On Grief and Grieving: Finding the Meaning of Grief Through the Five Stages of Loss*. Simon & Schuster.

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 (1/2)

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

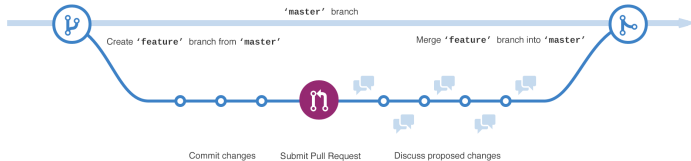


Figure 1: exploring parallel solutions
(<https://guides.github.com/activities/hello-world/>)

Version control (2/2)

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

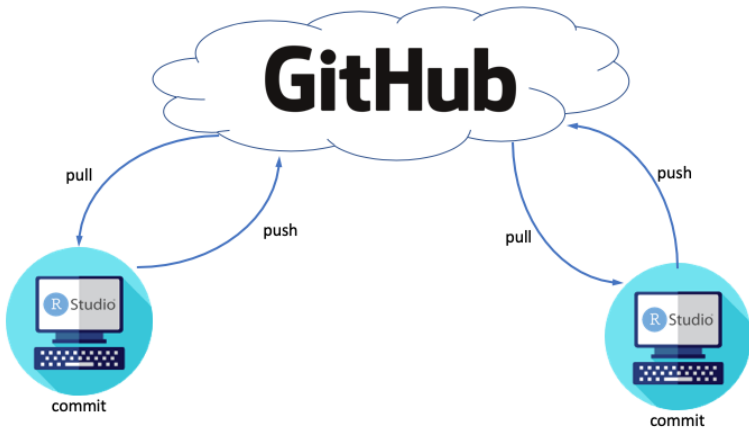


Figure 2: Collaboration schematic

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)

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 K. Bollen, J. T. Cacioppo, R. Kaplan, J. Krosnick, & J. L. Olds (2015). Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science. National Science Foundation. Arlington, VA.
- 4 Broman, K., Cetinkaya-Rundel, M., Nussbaum, A., Paciorek, C., Peng, R., Turek, D., & Wickham, H. (2017). Recommendations to Funding Agencies for Supporting Reproducible Research. Alexandria, VA: American Statistical Association. URL: <https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf>
- 5 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>
- 6 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>
- 7 Kaggle (2017). Kaggle machine learning & data science survey 2017. URL: <https://www.kaggle.com/kaggle/kaggle-survey-2017>
- 8 Zagalsky, A., Feliciano, J., Storey, M.-A., Zhao, Y. & Wang, W. (2015). The emergence of GitHub as a collaborative platform for education, in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, ACM, New York, NY, USA, pp. 1906–1917.

Thank You

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/>