ITB Tech Talks

Dr Matthias König
Humboldt University Berlin,
Institute for Theoretical Biology

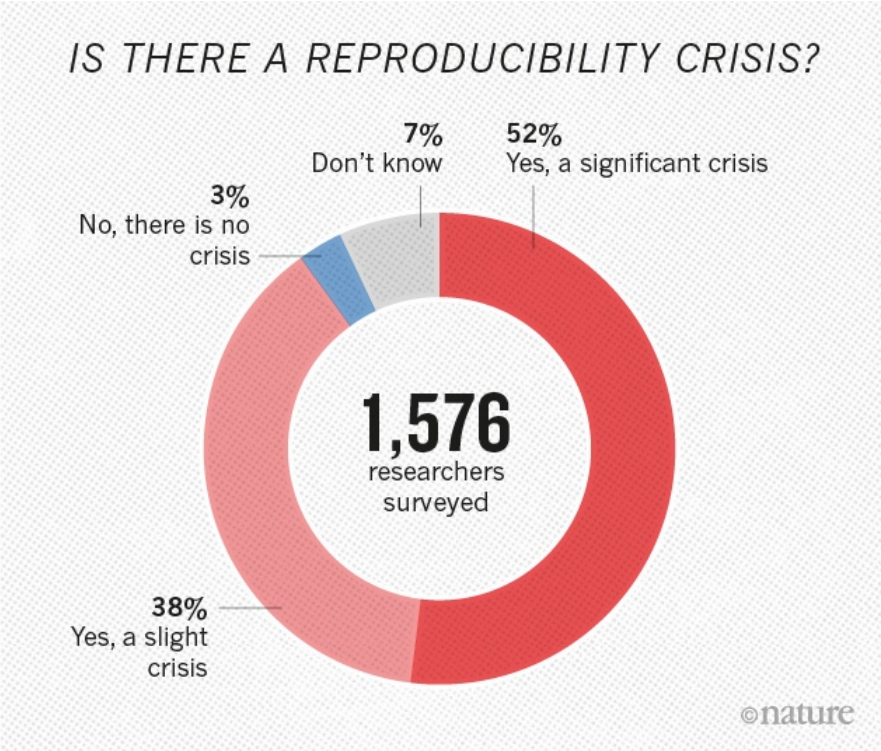
livermetabolism.com
@konigmatt

01 A Quick Introduction to Version Control with Git and GithHub

ITB Tech Talks

- **Motivation**: Improve Reproducibility, Reusability, Open Science & Quality of Computational Biology
- **Content**: Technologies for Computational & Theoretical Biology
- **Format**: Short talks 10 minutes, **life demo required**
- **Core technologies**
  - git
  - jupyter notebooks/lab
  - virtual environments & pip
  - docker
- **Core libraries (python/R)**
  - data science (pandas, numpy, tensorflow, scikit learn)
  - visualization (matplotlib, plotly)
  - reports (knitr)

Reproducibility/Reusability Crisis



"Really Reproducible Research" (1992) inspired by Stanford Professor Jon Claerbout:

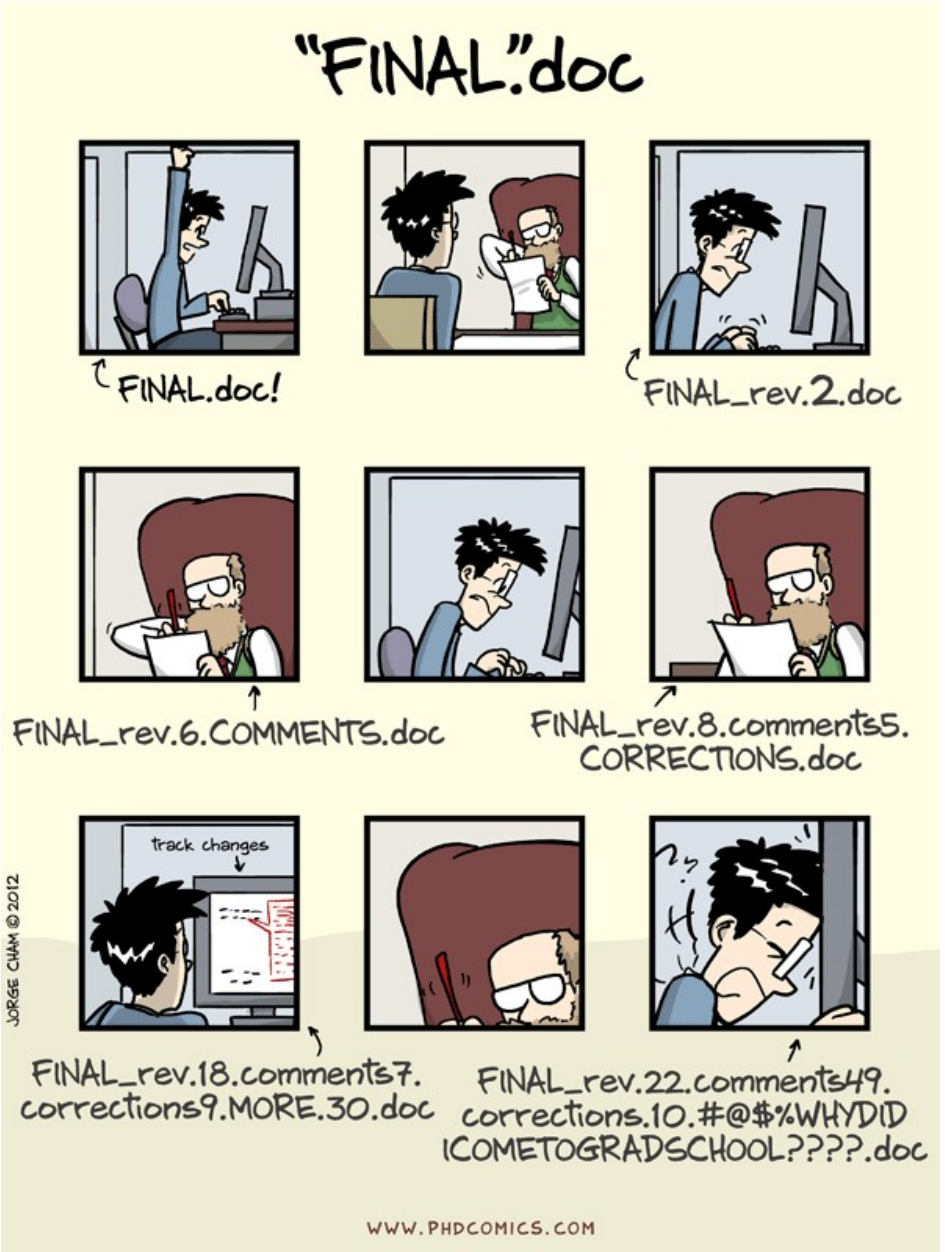
"The idea is: **An article** about computational science in a scientific publication is not the scholarship itself, it **is merely advertising of the scholarship**. The **actual scholarship is the complete ... set of instructions [and data]** which generated the figures."

David Donoho, 1998

How long would it take you to reproduce your results from your last (computational) paper?

https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
https://web.stanford.edu/~vcs/talks/MSKCC-Sept15-2015-STODDEN.pdf

Versioning Crisis



git to the rescue

- **Git** is a free and open source **distributed version control system**.
  - tiny footprint
  - lightning fast performance
  - works for everything from small to very large projects

- **GitHub** is a **code hosting platform for version control and collaboration**.
  - social coding
  - alternatives: GitLab, BitBucket, SourceForge

What can git do for you?

- Work anywhere & offline
- Decentralized backups
- Version control & track changes
  - diffs & branches
- Revert & experiment
  - revisions & branches
- Collaborative editing/work
  - Pull requests
- Reproducible research
- Releases & snapshots
  - citable code
- Issue tracker
- Continuous integration
  - unit tests
  - commit hooks

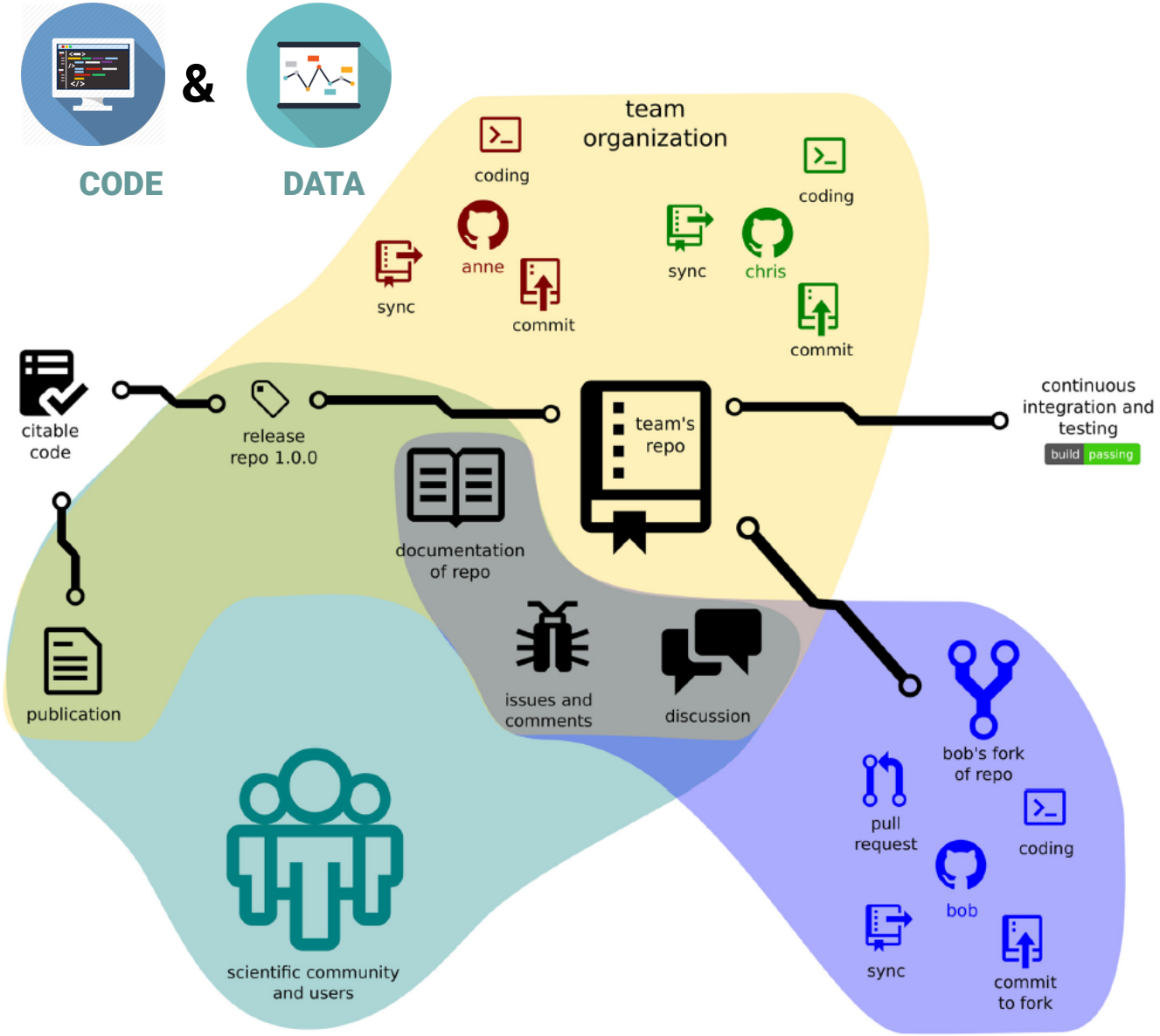
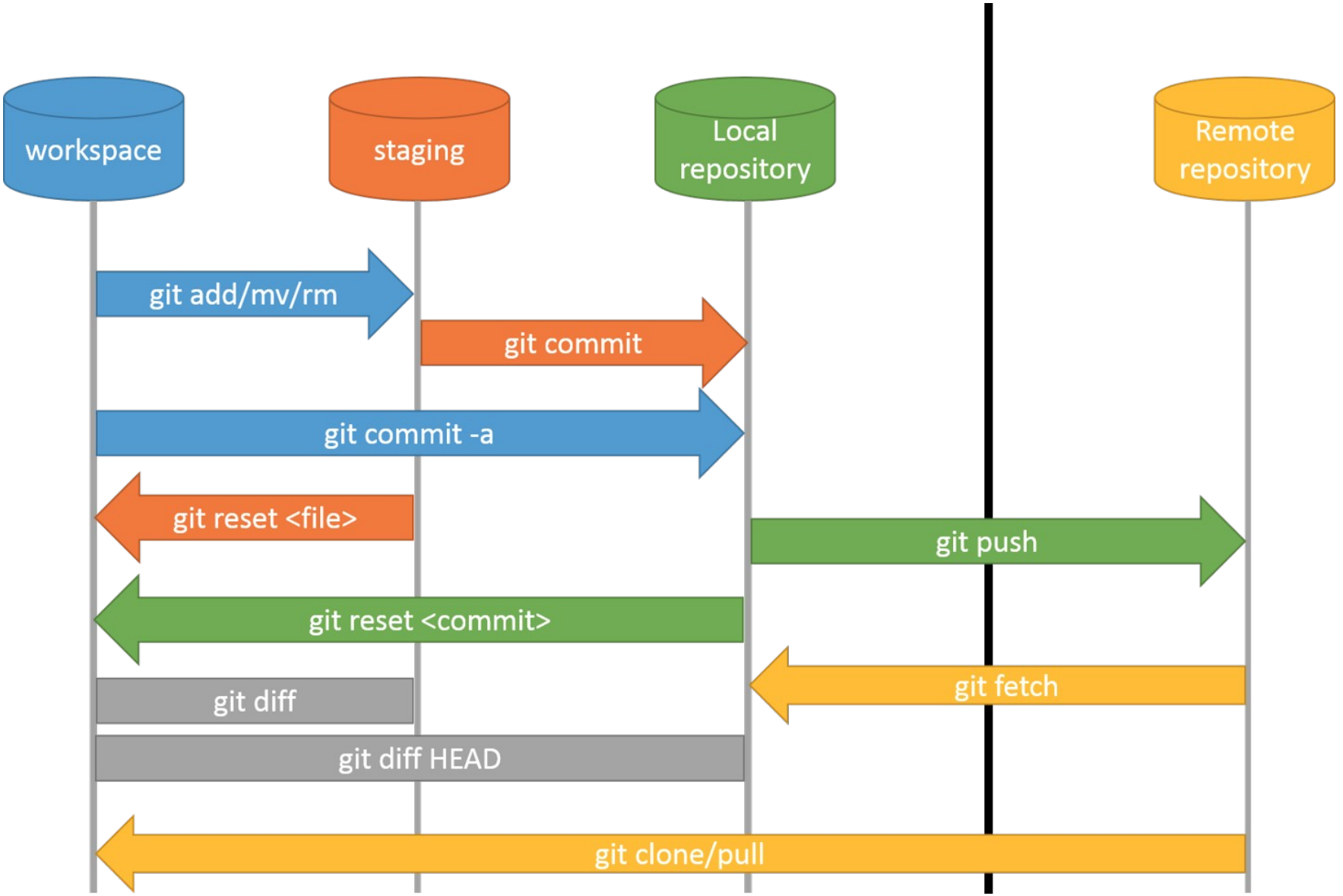


Fig 1. The structure of a GitHub-based project illustrating project structure and interactions with the community.

doi:10.1371/journal.pcbi.1004947.g001

Perez.Riverol2016

How does git work?



http://blog.podrezo.com/git-introduction-for-cvssvntfs-users/

References

- Perez-Riverol, Y.; Gatto, L.; Wang, R.; Sachsenberg, T.; Uszkoreit, J.; Leprevost, F. d. V.; Fufezan, C.; Ternent, T.; Eglen, S. J.; Katz, D. S.; Pollard, T. J.; Konovalov, A.; Flight, R. M.; Blin, K. & Vizcaíno, J. A.
  Ten Simple Rules for Taking Advantage of Git and GitHub.
  PLoS computational biology, 2016, 12, e1004947

- Blischak, J. D.; Davenport, E. R. & Wilson, G.
  A Quick Introduction to Version Control with Git and GitHub.
  PLoS computational biology, 2016, 12, e1004668

- Ram, K.
  Git can facilitate greater reproducibility and increased transparency in science.
  Source code for biology and medicine, 2013, 8, 7

- Tutorials
  https://try.github.io/levels/1/challenges/1

- Information/Books
  https://git-scm.com/
  https://git-scm.com/book/en/v2