
Cloud Storage
Services in
Reproducible Science

Cloud Storage: What and Why?

- Online platforms allowing individuals and organizations to store, manage, and access digital files and data remotely over the internet.

- Cloud storage services offer a convenient and secure way to store, synchronize, and share files, photos, videos, documents, and code across multiple devices.

- Importance in reproducible science due to open-source collaboration.

Cloud Storage: What and Why?

- Maintaining a well-organized research file structure and ensuring accessibility for fellow researchers is essential for scientific reproducibility.

- Adopting version control systems is imperative for managing data, recording changes, and reverting to previous versions.

- Leveraging cloud storage enhances collaboration, protects against data loss, and ensures accessibility of data.

Version Control in Research

- Safeguards against inadvertent file loss or difficulties in replicating analyses.

- Allows for collaborative development.

- Know who made what changes and when.

- Revert any changes and go back to a previous state.

Popular Cloud Storage Services: Large Format File-Sharing

- **Google Drive**: User-friendly (debatable!), integrates with Google Workspace. Not ideal for code repositories.

- **Dropbox**: Straightforward, easy to use, automated desktop syncing. Limited version control features, not designed for code repositories.

- **Microsoft OneDrive**: Integrates with Office applications, lacks specialized features for effective code collaboration.

Bits & Gits:
Specialized for Code Management

- **GitLab**: Open-source comprehensive DevSecOps platform. Code review with live preview, extended project management, security tools. Focus on roles beyond developers, emphasis on security.

Bitbucket: Integration with Atlassian product suite. Bitbucket Pipelines, apps on the Atlassian Marketplace, versatility in language support.

What's a Git?

- Software tool created in 2005 by Linus Torvalds to manage development of Linux; huge project that involved thousands of independent programmers.

- **Designed for fine-grained, line-by-line monitoring of changes in source code ("commits").**

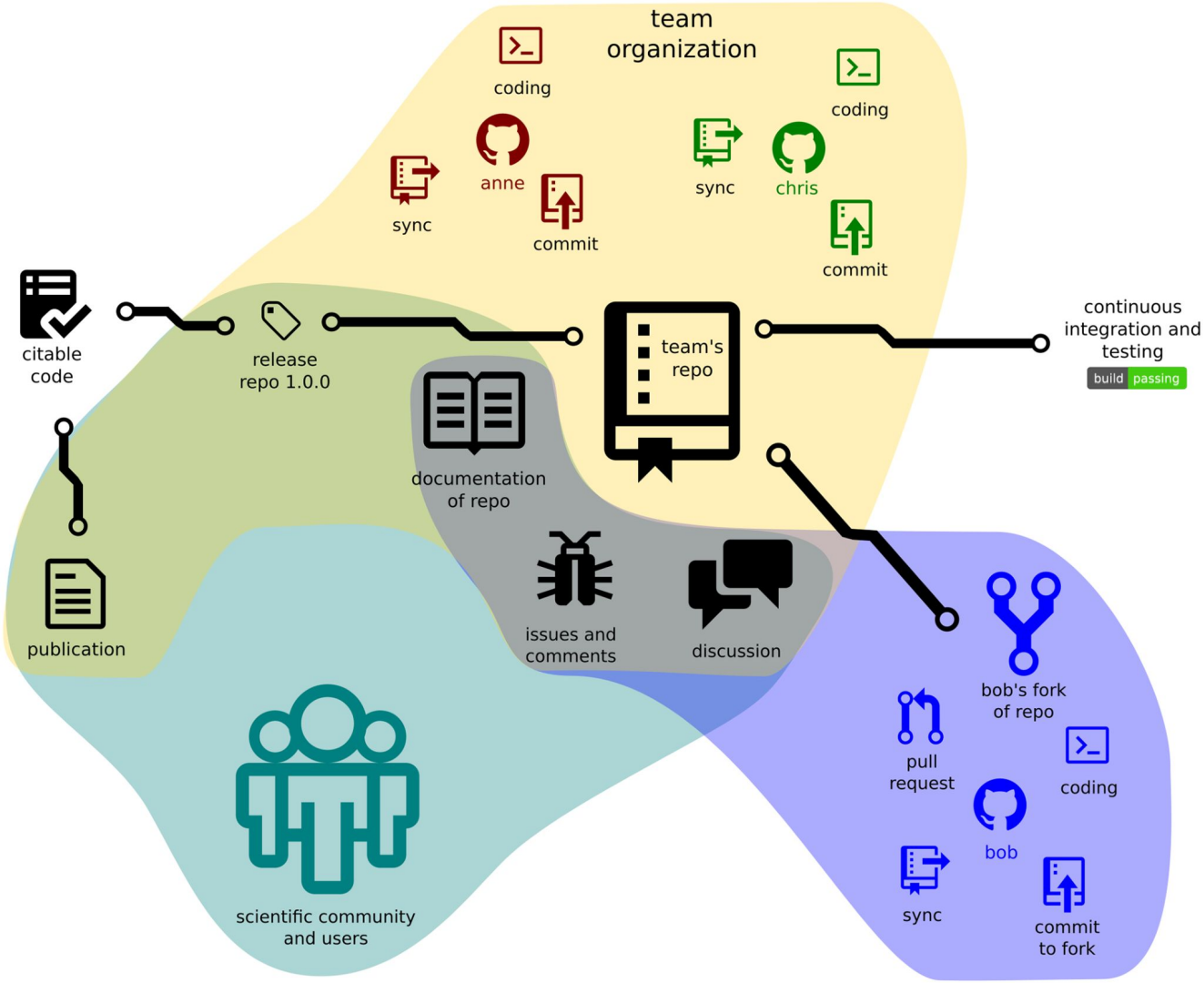
- Crucial component of version control and reproducibility.

Github!

- Code repository built for large, distributed teams of developers.

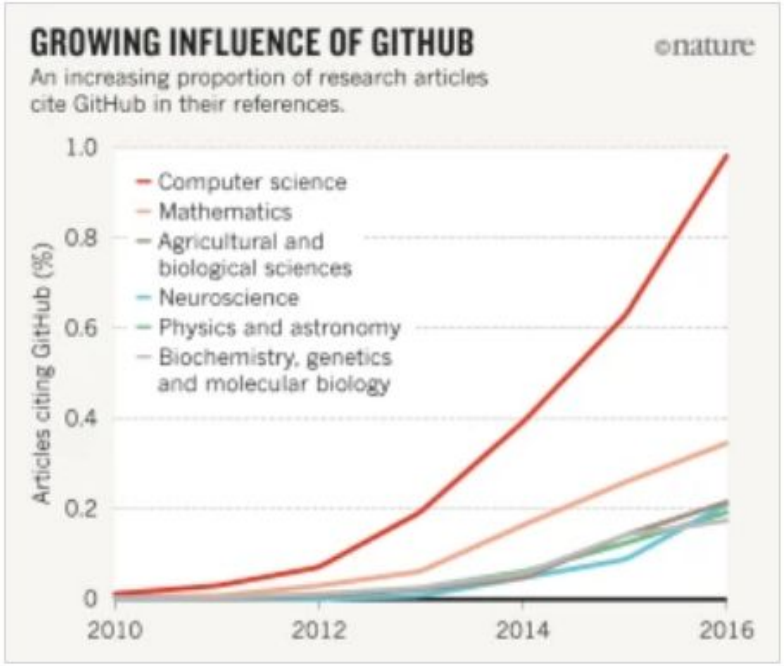
- Owned by Microsoft (not open-source).

- Social networking functionality.
  - **Public repositories allow other researchers to see how the work was done; clone it to their own computer to replicate the original work or apply the methods to their own data; open “Issues” to ask questions or give feedback on the project; send a “Pull request” with suggested changes to the text or code.**

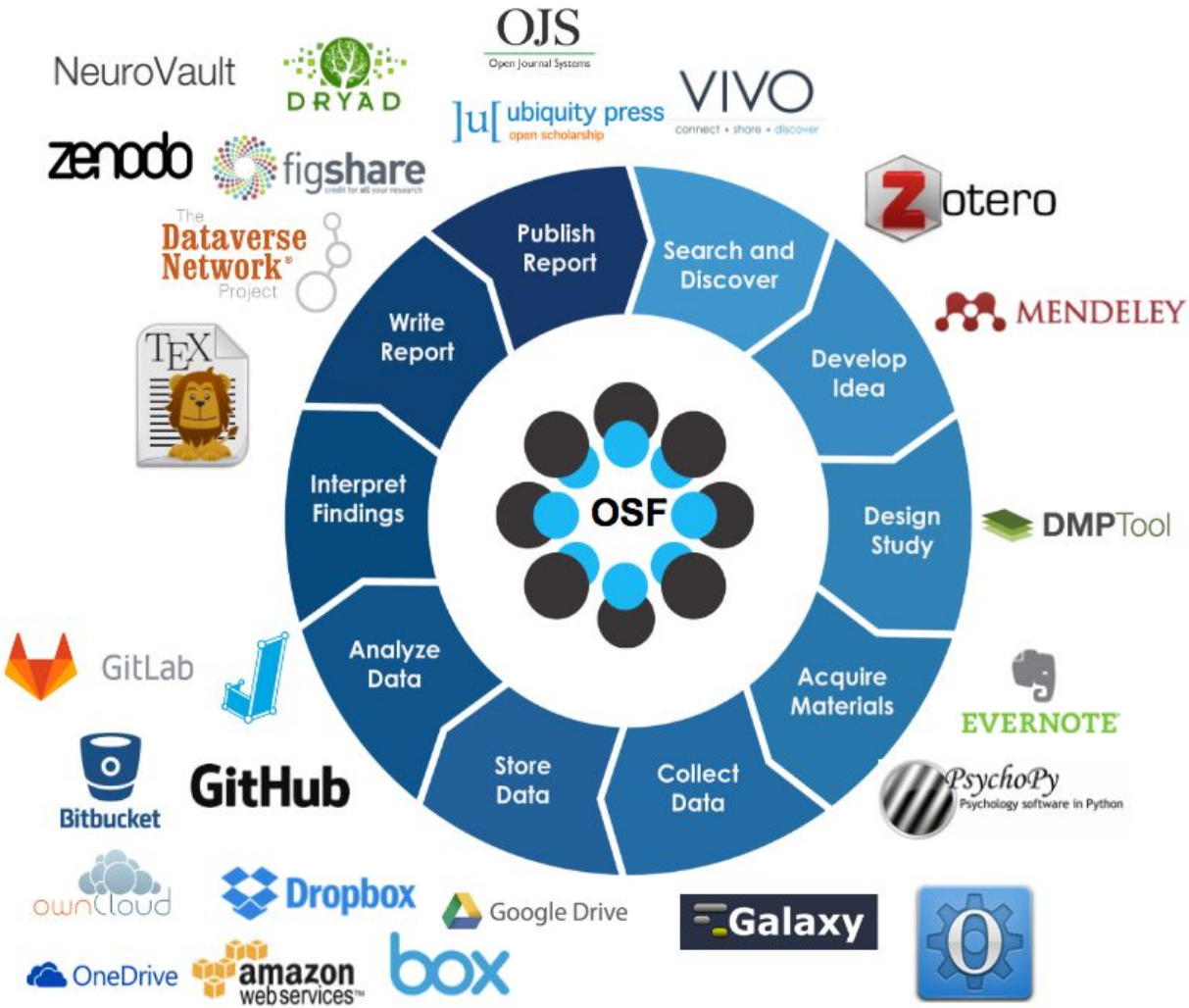

team
organization
coding
anne
sync
commit
coding
sync
chris
commit
citable
code
release
repo 1.0.0
team's
repo
continuous
integration and
testing
build passing
documentation
of repo
publication
issues and
comments
discussion
bob's fork
of repo
pull
request
coding
bob
sync
commit
to fork
scientific community
and users

Github!

- Huge community: over 21 million Repositories.
  - “Explore” feature emphasizes open community, learning environment.
- Simultaneous, distributed workflow.
- Version control
  - Allows for experimental branches that can then be merged into main project, or revert to earlier versions.



Credit: Richard Van Noorden/Source: Elsevier Scopus database


NeuroVault
DRYAD
OJS
Open Journal Systems
ubiquity press
open scholarship
VIVO
connect • share • discover
zenodo
figshare
Zotero
The Dataverse Network Project
MENDELEY
TEX
Publish Report
Search and Discover
Develop Idea
Design Study
Acquire Materials
Collect Data
Store Data
Analyze Data
Interpret Findings
Write Report
OSF
DMPTool
GitLab
EVERNOTE
Bitbucket
GitHub
PsychoPy
Psychology software in Python
ownCloud
Dropbox
Google Drive
Galaxy
OneDrive
amazon web services
box

Remember Chapter 5?

6.3 Reproducible code

6.3.1 Introduction

The following steps have to be fully integrated in order to produce a reproducible code:

- **Step 1:** Establish a reproducible project workflow.
- **Step 2:** Organize project for reproducibility.
- **Step 3:** Ensure basic programming standards.
- **Step 4:** Document and manage dependencies.
- **Step 5:** Produce a reproducible report (with R Markdown).
- **Step 6: Implement a version control protocol (with Git).**
- **Step 7:** Ensure archiving and citation of code.

...step 6 will be covered in the bioinformatic tutorial associated to chapter 12

GitHub Tutorial

- Exercise to familiarize yourself with GitHub integration & workflow w/ RStudio
  - Got Git?
  - GitHub account?
  - Visual Studio?

Let’s begin!

References

Atlassian. (n.d.). *Bitbucket | Git solution for teams using Jira*. Bitbucket. Retrieved November 13, 2023, from https://bitbucket.org/product

Build software better, together. (n.d.). GitHub. Retrieved November 13, 2023, from https://github.com

Gandrud, C. (2018). *Reproducible Research with R and R Studio*. CRC Press. https://books.google.com/books?id=e6x-DwAAQBAJ

GitLab.com · GitLab. (n.d.). GitLab. Retrieved November 13, 2023, from https://gitlab.com/gitlab-com

Heller, M. (2018, April 2). *What is GitHub? More than Git version control in the cloud*. InfoWorld. https://www.infoworld.com/article/3267565/what-is-github-more-than-git-version-control-in-the-cloud.html

Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N., & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, *1*(6), Article 6. https://doi.org/10.1038/s41559-017-0160

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., Fufezan, C., Ternent, T., Eglen, S. J., Katz, D. S., Pollard, T. J., Konovalov, A., Flight, R. M., Blin, K., & Vizcaíno, J. A. (2016). Ten Simple Rules for Taking Advantage of Git and GitHub. *PLOS Computational Biology*, *12*(7), e1004947. https://doi.org/10.1371/journal.pcbi.1004947

Perkel, J. (2016). Democratic databases: Science on GitHub. *Nature*, *538*(7623), Article 7623. https://doi.org/10.1038/538127a

Van Lissa, C. J., Brandmaier, A. M., Brinkman, L., Lamprecht, A.-L., Peikert, A., Struiksma, M. E., & Vreede, B. M. I. (2021). WORCS: A workflow for open reproducible code in science. *Data Science*, *4*(1), 29–49. https://doi.org/10.3233/DS-210031