A full-semester course on “Version control of code and data using Git and DataLad”

Symposium on “Reproducible Research: Education and Teaching Formats Reports
From the Reproducibility Networks” at QUEST Center Berlin, Germany

Talk by Dr. Lennart Wittkuhn

2023-05-11

About

Me

👨‍🔬 **Position:** PostDoc & Lab Manager at University of Hamburg & MPI for Human Development Berlin

🎓 **Education:** BSc Psychology & MSc Cognitive Neuroscience (TU Dresden), PhD Psychology (FU Berlin)

🔬 **Research:** I study the role of fast neural memory reactivation ("replay") in the human brain using fMRI

🕸️ **GRN:** Member of the MPIB's working group on research data management & open science (GRN member)

🔗 **Contact:** You can connect with me via email, Twitter, Mastodon, GitHub or LinkedIn

ℹ️ **Info:** Find out more about my work on my website, Google Scholar and ORCiD

This presentation

🏗️ **WIP:** The presented teaching project is work in progress!

💻 **Slides:** Slides are publicly available at lennartwittkuhn.com/ddlitlab-presentation

📦 **Software:** Reproducible slides built with Quarto and deployed to GitHub Pages using GitHub Actions

Source: Source code is publicly available on GitHub at github.com/lnnrtwttkhn/ddlitlab-presentation

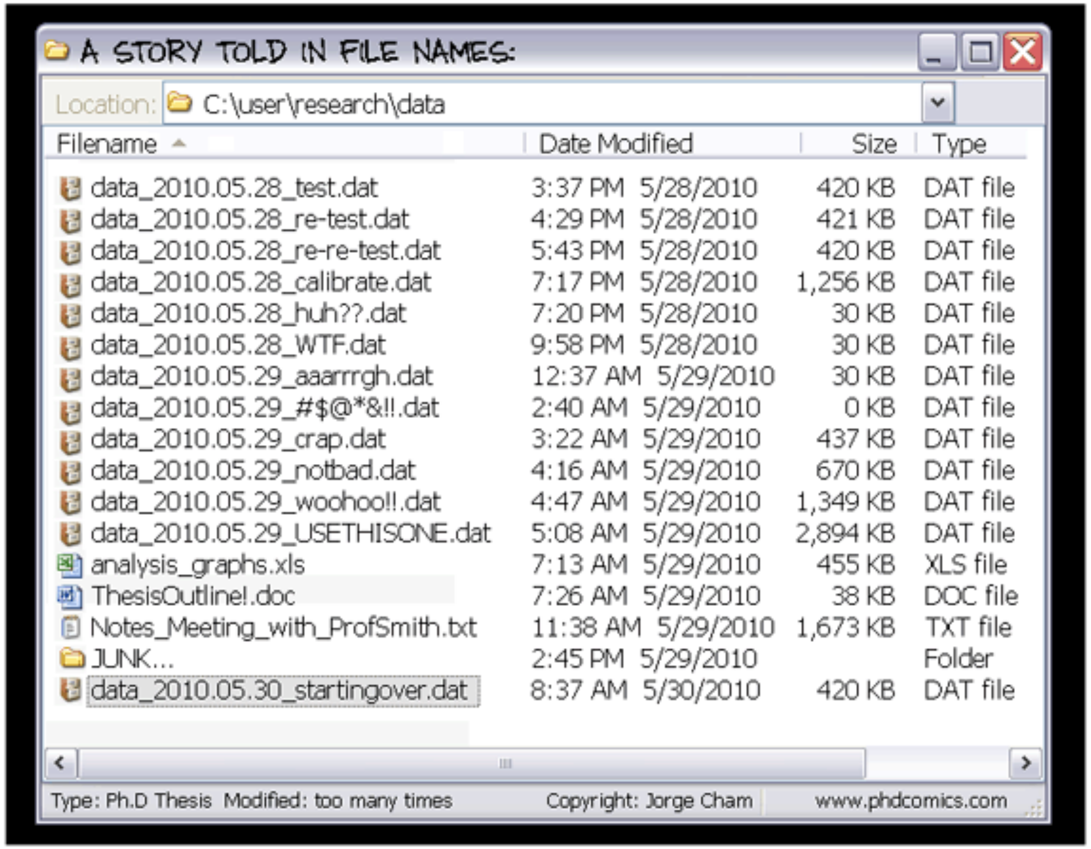
🙏 **Contact:** I am happy for any feedback or suggestions via email or GitHub issues. Thank you!

Why we need version control …

… for **code** (text files)

"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
track changes
FINAL_rev.18.comments7.corrections9.MORE.30.doc
FINAL_rev.22.comments49.corrections.10.#@$%WHYDIDICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

… for **data** (binary files)



Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

© Jorge Cham (phdcomics.com)

What is version control?

“Version control is a systematic approach to record changes made in a […] set of files, over time. This allows you and your collaborators to track the history, see what changed, and recall specific versions later […]” (Turing Way)

- keep track of changes in a directory (a “repository”)
- take snapshots (“commits”) of your repo at any time
- know the history: what was changed when by whom
- compare commits and go back to any previous state
- work on parallel “branches” & flexibly “merge” them

- “push” your repo to a “remote” location & share it
- share repos on platforms like GitHub or GitLab
- work together on the same files at the same time
- others can read, copy, edit and suggest changes
- make your repo public and openly share your work

by Scriberia for The Turing Way community (CC-BY 4.0)

by Scriberia for The Turing Way community (CC-BY 4.0)

What are git and DataLad?

git-scm.com

- most popular version control system
- free, open-source command-line tool
- graphical user interfaces exist, e.g., GitKraken
- standard tool for most (all?) software developers
- 100 million GitHub users [1]

datalad.org

- “git for (large) data”
- free, open-source command-line tool
- builds on top of git and git-annex
- allows to version control arbitrarily large datasets [2]
- graphical user interface exists: DataLad Gooey

Course details

Overview

- Full-semester seminar (~ 12 sessions of 90 mins)
- Winter semester 2023/24 (October to January)
- University of Hamburg (virtual option TBD)
- MSc and PhD students (research focus)
- Psychology and Cognitive Neuroscience

Milestones

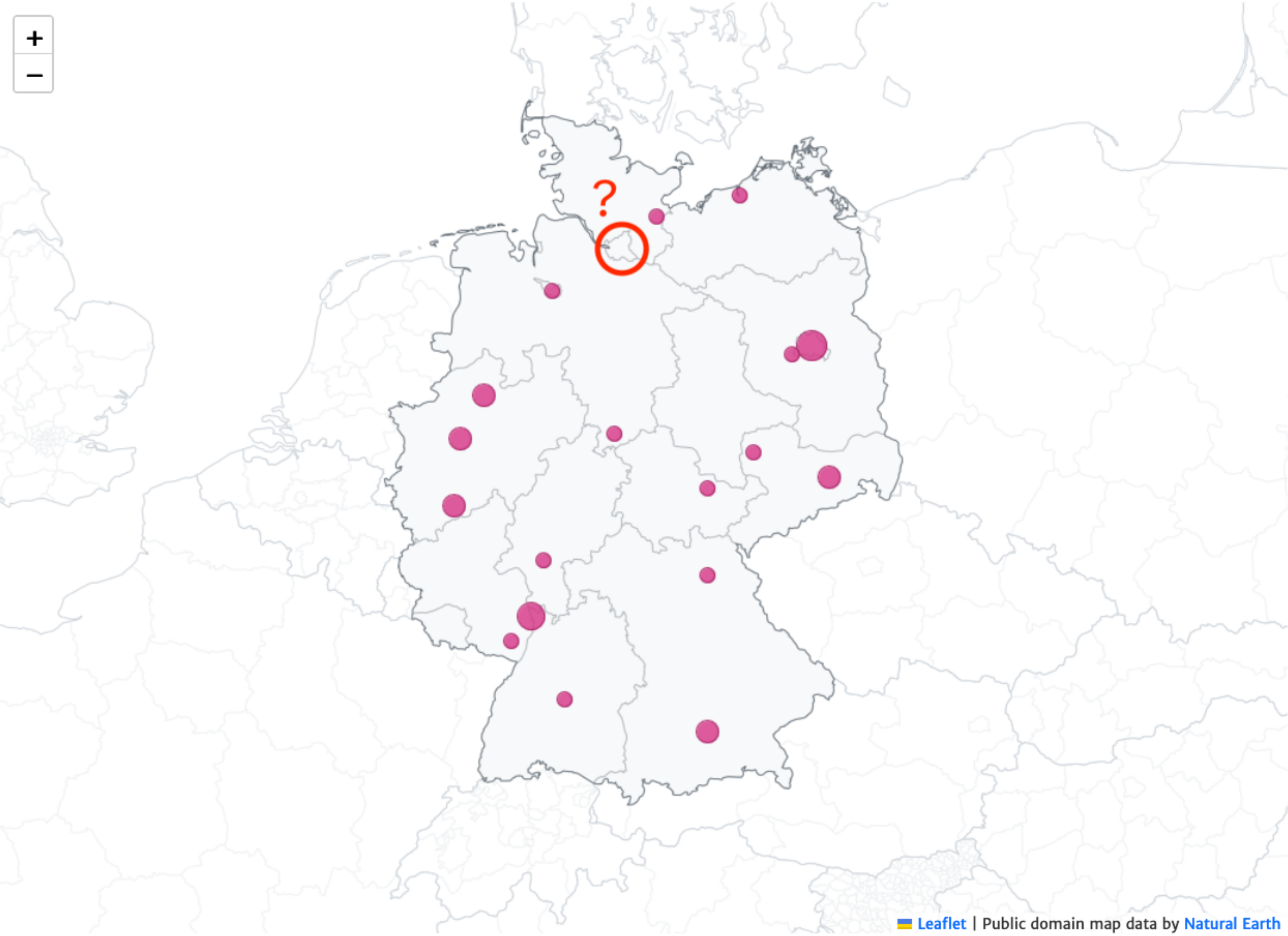
- ★ Project received funding!
- ★ Project website online!
- ★ Course website online!

→ Next up: Prepare course!

Implementation and Tools

- **Impulse lectures** & **live demonstrations**
- </> **Code-along** & **exercises** (individual and group)
- **Discussions** on reproducibility, open code & data
- Fixed computational environments on JupyterHub
- >_ Focus on **command-line** interaction
- Alternative use via **Graphical User Interfaces**
- **Quizzes** & continuous **evaluation** (in R Shiny)
- Reuse quiz & evaluation data as **example datasets**
- **Follow-up** research projects in summer semester
- Support by **research** and **teaching assistants**
- Materials shared as **Open Educational Resources**
- Integration with GRN, Carpentries Incubator, etc.?

Local GRN node in Hamburg?



from reproducibilitynetwork.de/members

Thank you!

Funding & Support

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Digital and Data Literacy in Teaching Lab (DDLitLab), an initiative by the ISA-Zentrum at University of Hamburg

Stiftung Innovation in der Hochschullehre

Stiftung Innovation in der Hochschullehre

People

Prof. Dr. Nicolas Schuck
(UHH & MPIB)

Carolin Scharfenberg
(UHH DDLitLab)

Contact

✉ lennart.wittkuhn@uni-hamburg.de

⌂ lennartwittkuhn.com

Twitter Mastodon GitHub LinkedIn

Footnotes

1. (Source: Wikipedia)

2. see DataLad dataset of 80TB / 15 million files from the Human Connectome Project (see details)