

Have a !

RESEARCH DATA MANAGEMENT



WITH DATALAD

Adina Wagner Michał Szczepanik

 @AdinaKrik

Psychoinformatics lab,
Institute of Neuroscience and Medicine (INM-7)
Research Center Jülich

Slides: <https://github.com/datalad-handbook/course/>

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website.
- Feel free to take collaborative, public notes at etherpad.wikimedia.org/p/RDM_with_DataLad. You can also use this pad for anonymous questions.
- We are using a JupyterHub. You should have received credentials in advance via email - if not, please raise your hand!
- Find the workshop contents (and more) at psychoinformatics-de.github.io/rdm-course/
- Let us introduce the workshop organizers...
- Some guidelines for the virtual workshop venue...
 - Please mute yourself when you don't speak
 - Make use of the "Raise hand" feature
 - Drop out and re-join as you please
 - Adhere to the code of conduct

QUESTIONS/INTERACTION THROUGHOUT THE WORKSHOP

- If you have a question during a lecture, please first type your questions in the chat. There are no stupid questions :)
- It would be great to have lively discussions - unless its interrupting others, please feel encouraged to unmute/turn on your video to interact with us.
- We're happy to discuss specific use cases at the end. Please make a note about them in the "Shared notes"

QUESTIONS/INTERACTION AFTER THE WORKSHOP

If you have a question after the workshop, you can reach out for help:

Reach out to to the DataLad team via

- [Matrix](#) (free, decentralized communication app, no app needed). We run a weekly Zoom office hour (Thursday, 4pm Berlin time) from this room as well.
- [the development repository on GitHub](#)

Reach out to the user community with

- A question on [neurostars.org](#) with a `datalad` tag

Find more user tutorials or workshop recordings

- On [DataLad's YouTube channel](#)
- In the [DataLad Handbook](#)
- In the [DataLad RDM course](#)
- In the [Official API documentation](#)

RESOURCES AND FURTHER READING

Comprehensive user documentation in the
DataLad Handbook (handbook.datalad.org)

- High-level function/command overviews,
Installation, Configuration, Cheatsheet
- Narrative-based code-along course
- Independent on background/skill level,
suitable for data management novices
- Step-by-step solutions to common
data management problems, like
how to make a reproducible paper

Overview of most tutorials, talks, videos, ... at github.com/datalad/tutorials

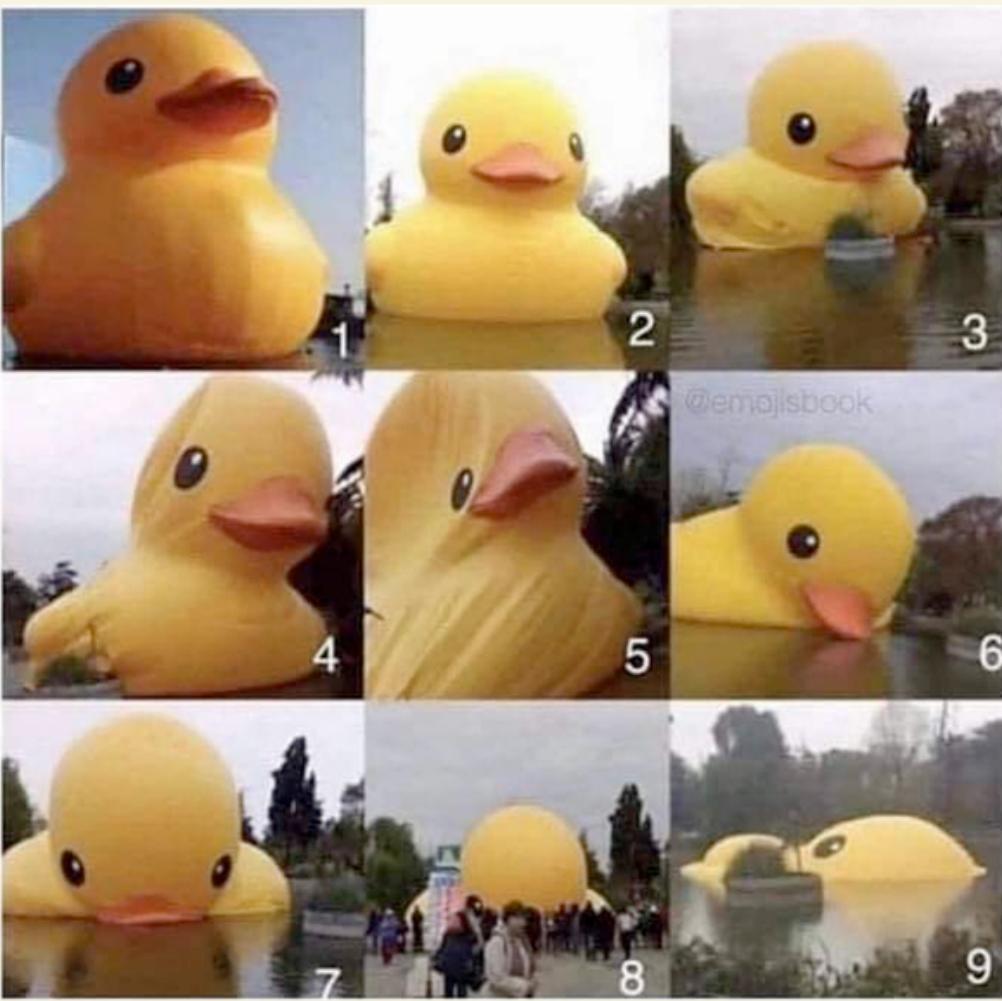
LIVE POLLING SYSTEM

Please use your phone to scan to QR code, or open the link in a new browser window

WHAT'S YOUR MOOD TODAY?



WHAT'S YOUR LEVEL OF EXCITEMENT?

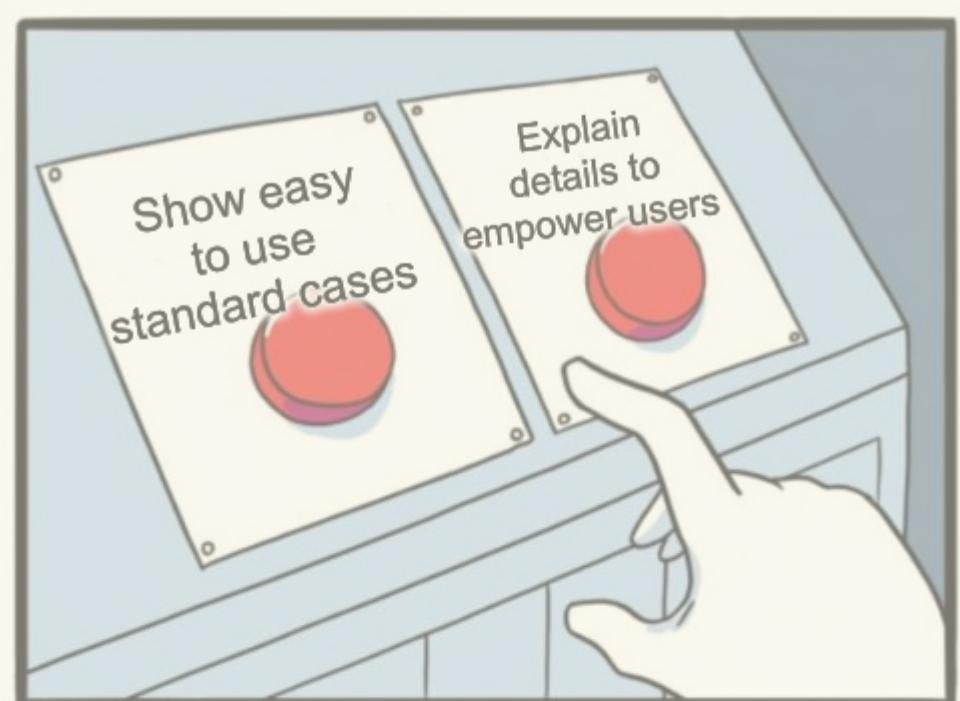


VIDEO RECORDINGS

The recording would be edited or stopped to exclude certain or all discussions.
This poll is unanimous - only if everyone votes "yes" the workshop will be recorded

WHAT WILL WE DO TODAY?

- The workshop centers around DataLad (version 0.16)
- We aim to do more than a standard introduction by providing in-depth explanations, hands-on exercises, and discussions throughout the workshop
- (Help us by asking any question that comes up!)



MOTIVATION

COMMON PROBLEMS IN SCIENCE

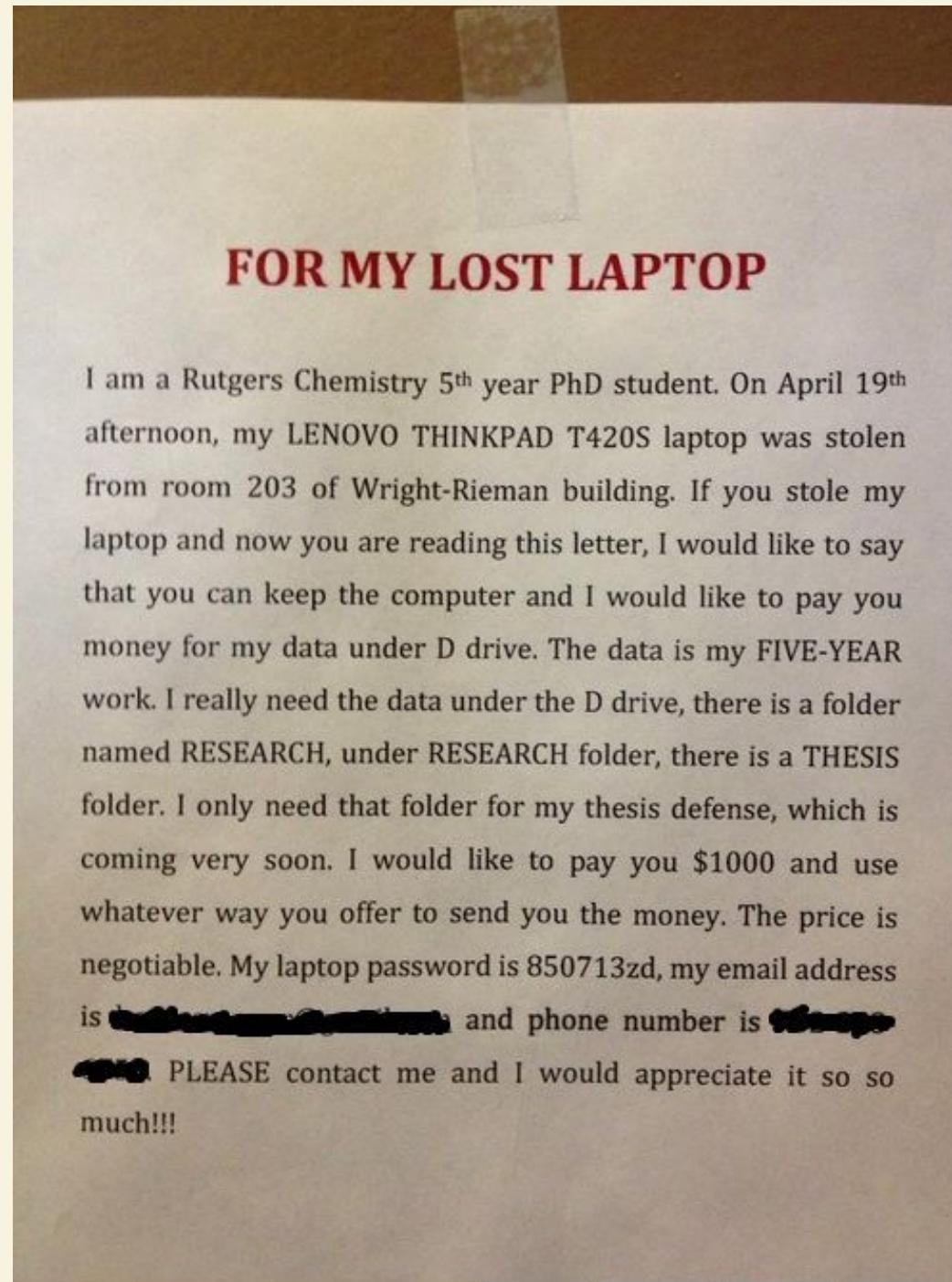
You write a paper about an algorithm, stay up late to generate good-looking figures, but you have to tweak parameters and display options to make it work AND look good. The next morning, you have no idea which parameters produced which figures, and which of the figures fits to what you report in the paper.

Image credit: Illustration adapted from Scriberia and The Turing Way



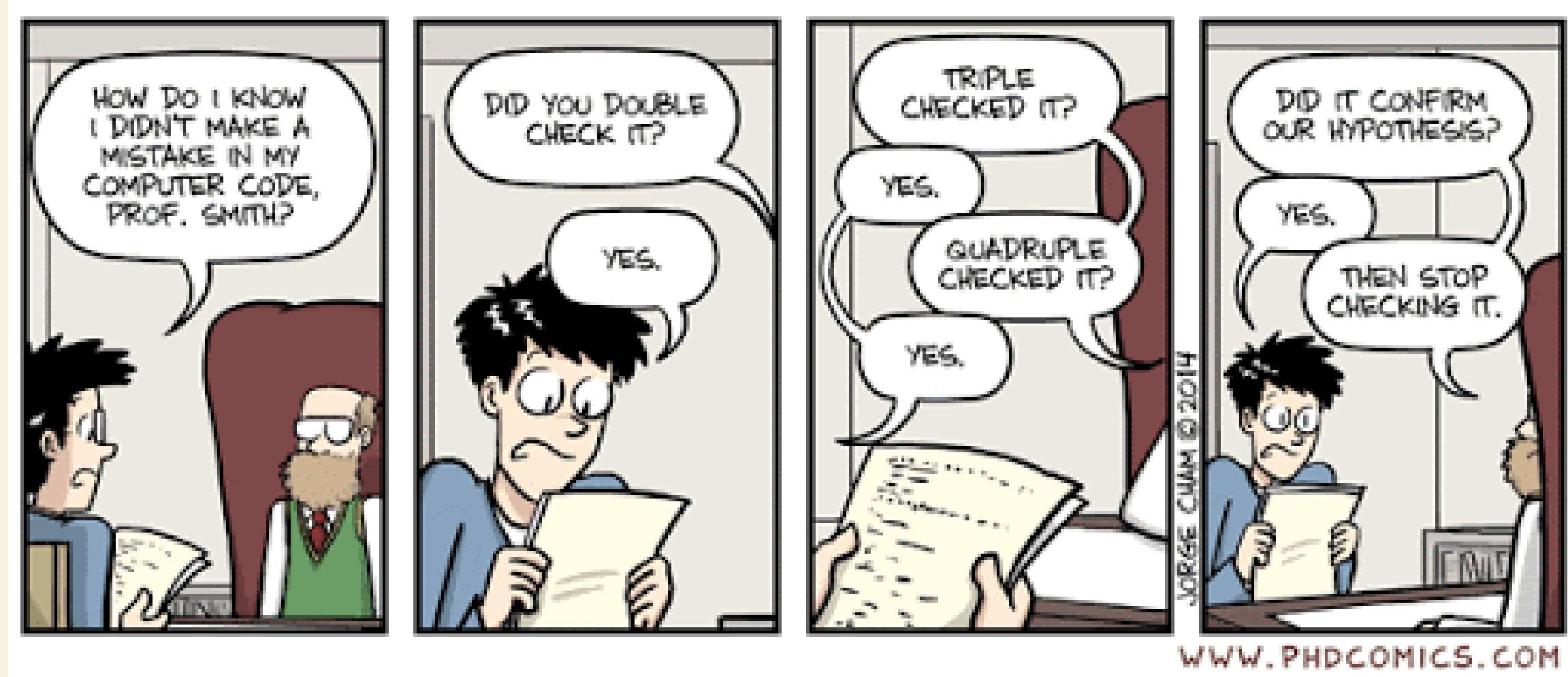
COMMON PROBLEMS IN SCIENCE

Your research project produces phenomenal results, but your laptop, the only place that stores the source code for the results, is stolen/breaks



COMMON PROBLEMS IN SCIENCE

A graduate student approaches their supervisor, complaining that the supervisors research idea does not work. After weeks of discussion, it becomes apparent that oral communication doesn't suffice - the student can't sufficiently explain the environment (data, algorithms, ...) they constructed, and if the supervisor can't enter and use the students project there's no way to find a fix.



COMMON PROBLEMS IN SCIENCE

A Post-doc wrote a script during the PhD that applied a specific method to a dataset. Now, with new data and a new project, they try to reuse the script, but forgot how it worked.



COMMON PROBLEMS IN SCIENCE

You try to recreate results from another lab's published paper. You base your re-implementation on everything reported in their paper, but the results you obtain look nowhere like the original.



COMMON OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, 1995
Let's do better!



**DataLad can help
with small or large-scale
data management**

Free,
open source,
command line tool & Python API



- A command-line tool, available for all major operating systems (Linux, macOS/OSX, Windows), MIT-licensed
- Build on top of Git and Git-annex
- Allows...
 - ... **version-controlling arbitrarily large content**
version control data and software alongside to code!
 - ... **transport mechanisms for sharing and obtaining data**
consume and collaborate on data (analyses) like software
 - ... **(computationally) reproducible data analysis**
Track and share provenance of all digital objects
 - ... **and much more**
- Completely domain-agnostic

ACKNOWLEDGEMENTS

Software

- Joey Hess (git-annex)
- The DataLad team & contributors

Illustrations

- The Turing Way project & Scriberia



Funders

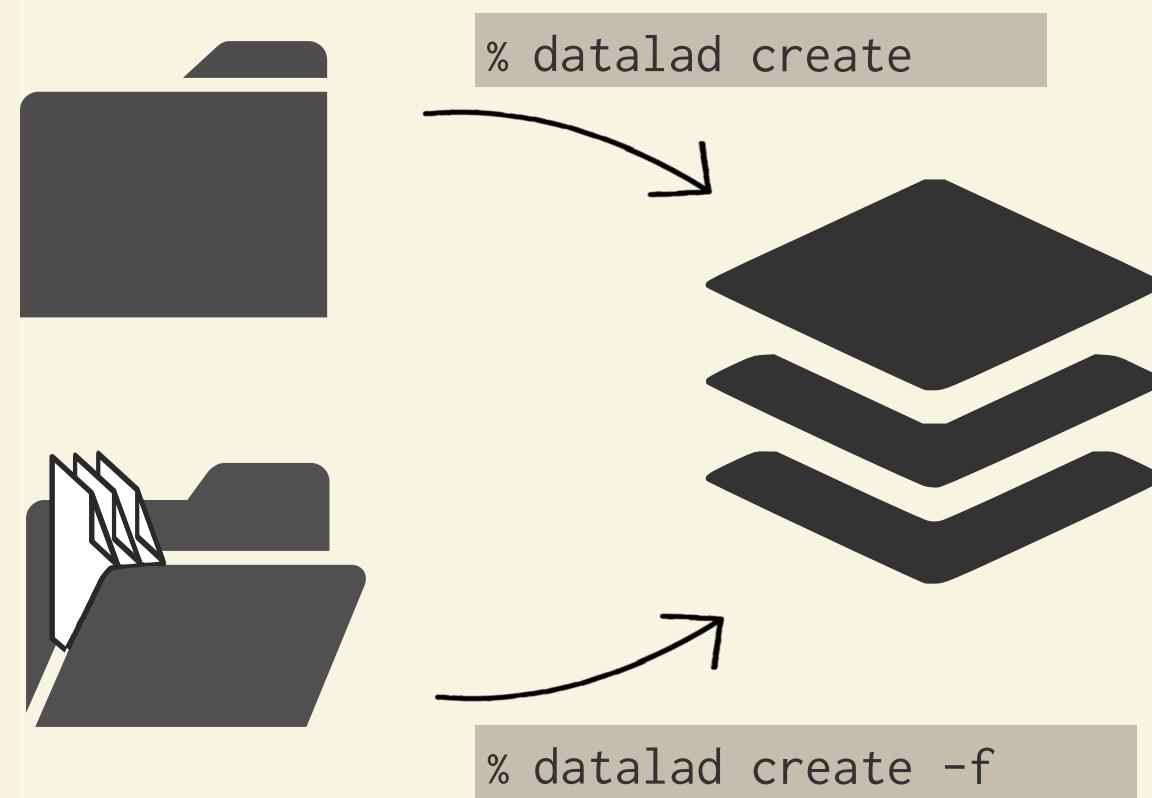
Collaborators



CORE CONCEPTS & FEATURES

EVERYTHING HAPPENS IN DATALAD DATASETS

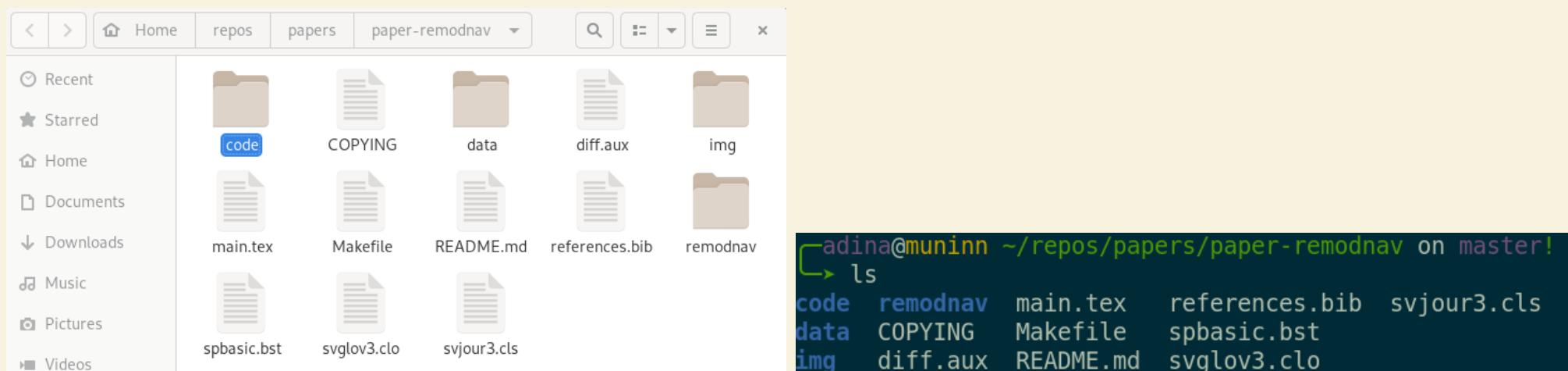
create new, empty datasets to populate...



.... or transform existing directories into datasets

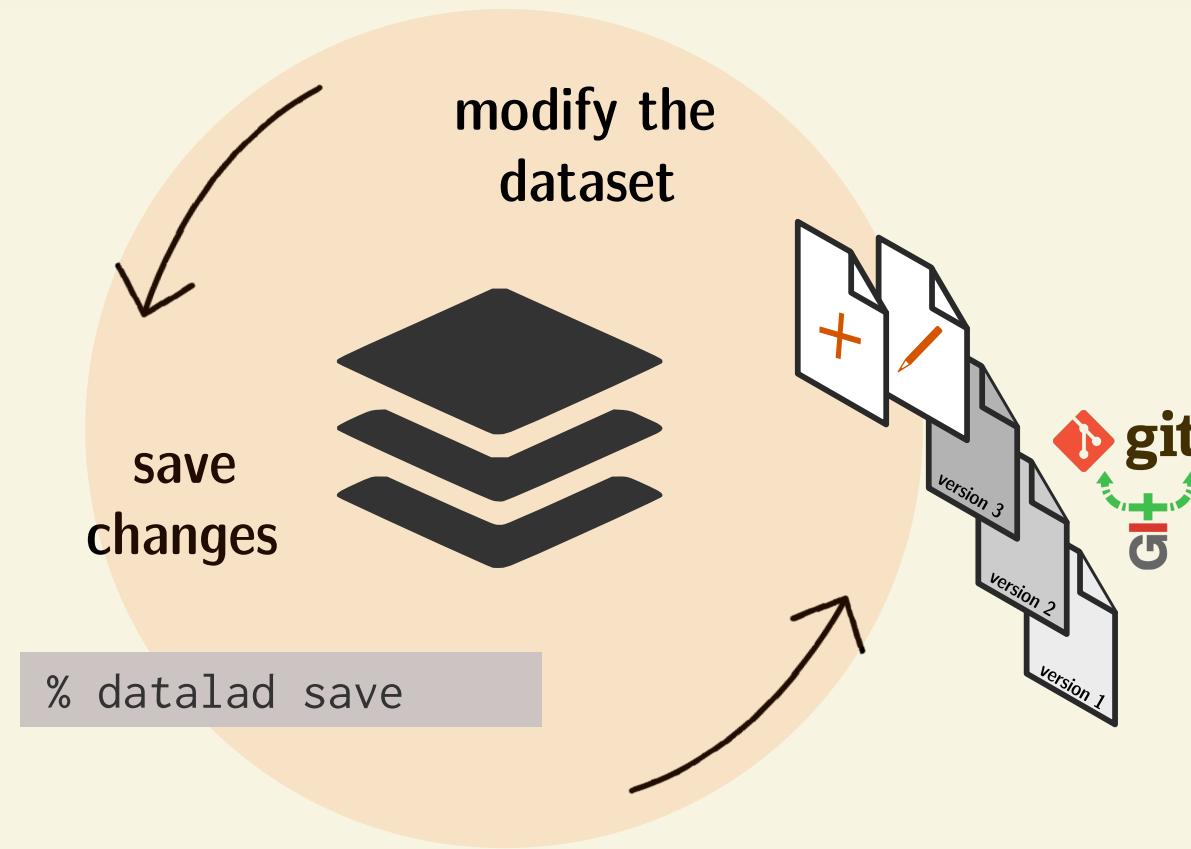
DATASET = GIT/GIT-ANNEX REPOSITORY

- content agnostic
- no custom data structures
- complete decentralization
- Looks and feels like a directory on your computer:



File viewer and terminal view of a DataLad dataset

VERSION CONTROL ARBITRARILY LARGE FILES



Stay flexible:

- Non-complex DataLad core API (easy for data management novices)
- Pure Git or git-annex commands (for regular Git or git-annex users, or to use specific functionality)

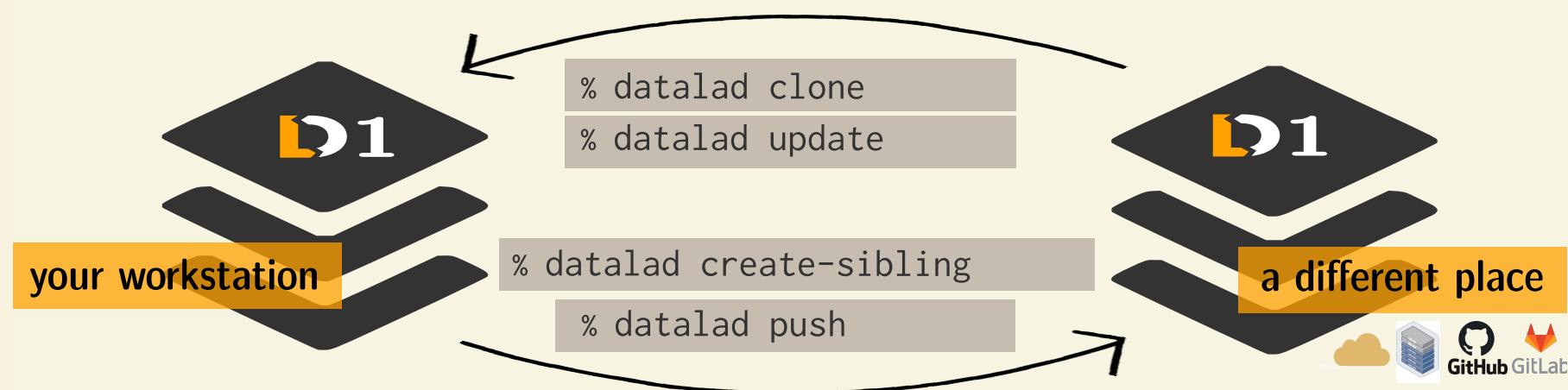
USE A DATASETS' HISTORY

Date	Author	Change summary (commit message)
2020-06-05 10:58 +0200	Adina Wagner	M [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-de
2020-06-05 08:24 +0200	Adina Wagner	o [finalround] {upstream/finalround} add results from computing with mean instead of median
2020-06-05 09:09 +0200	Michael Hanke	o Change wording, clarify comment
2020-06-05 07:26 +0200	Michael Hanke	M Merge remote-tracking branch 'gh-mine/finalround'
2020-05-28 16:39 +0200	Asim H Dar	o Added datalad.get() so S2SRMS() pulls data and can run standalone
2020-05-18 08:25 +0200	Adina Wagner	o {gh-asim/finalround} S2SRMS: example implementation of the S2SRMS method suggested by R2
2020-05-01 17:38 +0200	Adina Wagner	o Minor edits as suggested by reviewer 2
2020-05-29 09:04 +0200	Adina Wagner	M Merge pull request #13 from psychoinformatics-de/adswa-patch-1
2020-05-24 09:15 +0200	Adina Wagner	o {upstream/adswa-patch-1} Fix installation instructions
2020-05-24 09:53 +0200	Adina Wagner	M Merge pull request #14 from psychoinformatics-de/bf-data
2020-05-24 09:33 +0200	Adina Wagner	o [bf-data] One-time datalad import
2020-05-24 09:32 +0200	Adina Wagner	o install and get relevant subdataset data
2020-03-18 10:19 +0100	Michael Hanke	M Merge pull request #8 from psychoinformatics-de/adswa-patch-1
2019-12-19 10:22 +0100	Adina Wagner	o {gh-asim/adswa-patch-1} add sklearn to requirements
2020-03-18 10:13 +0100	Michael Hanke	o Tune new figure caption
2020-03-18 10:03 +0100	Michael Hanke	M Merge pull request #11 from ElectronicTeaCup/revision_2
2020-03-18 09:59 +0100	Adina Wagner	M [revision_2] {gh-asim/revision_2} Merge branch 'revision_2' of github.com:ElectronicTe
2020-03-18 09:58 +0100	Michael Hanke	o Last detections
2020-03-18 09:59 +0100	Adina Wagner	o name parameter in caption

- reset your dataset (or subset of it) to a previous state,
- revert changes or bring them back,
- find out what was done when, how, why, and by whom
- Identify precise versions: Use data in the most recent version, or the one from 2018, or...

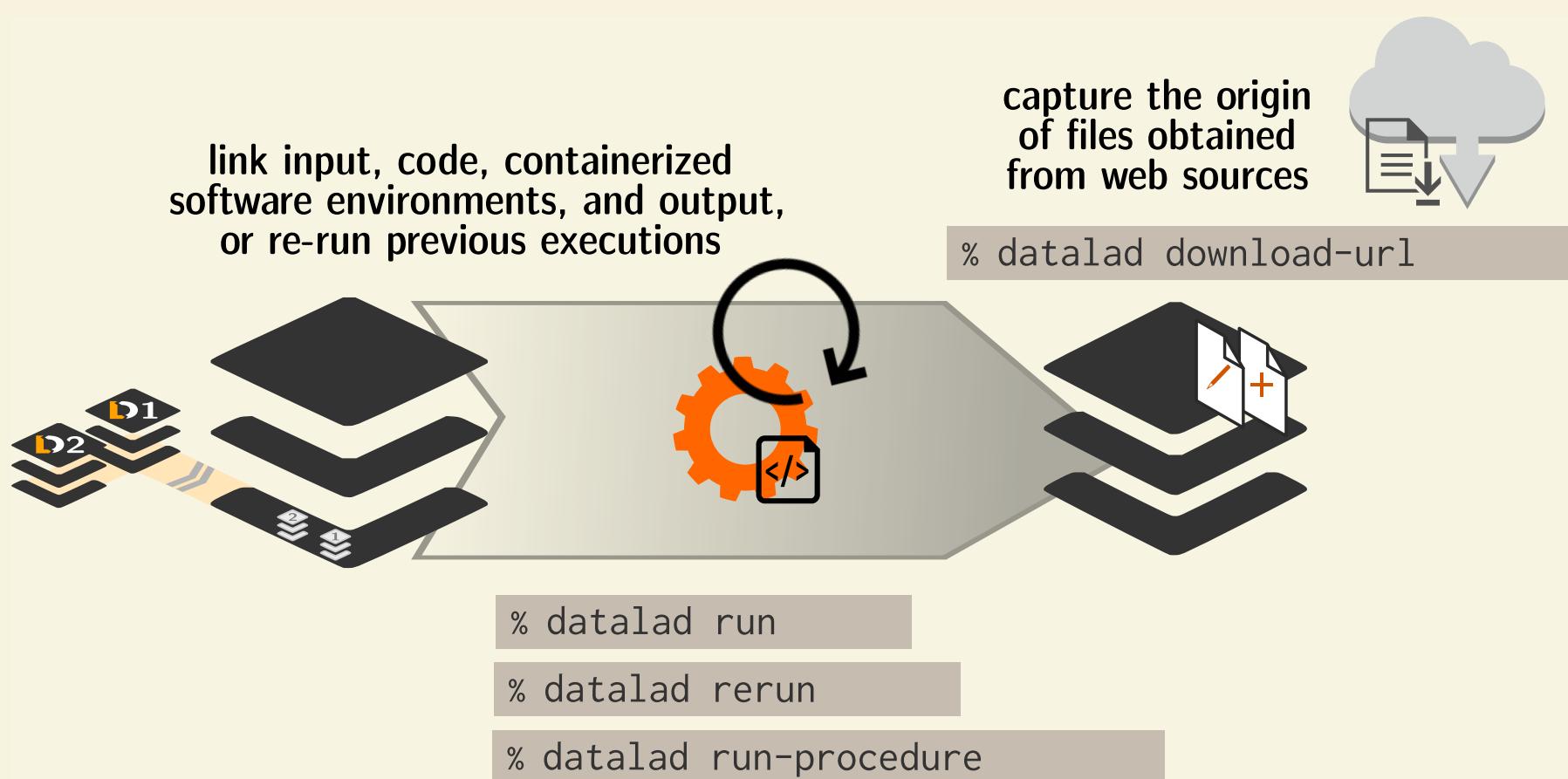
CONSUME AND COLLABORATE

Consume existing datasets and stay up-to-date

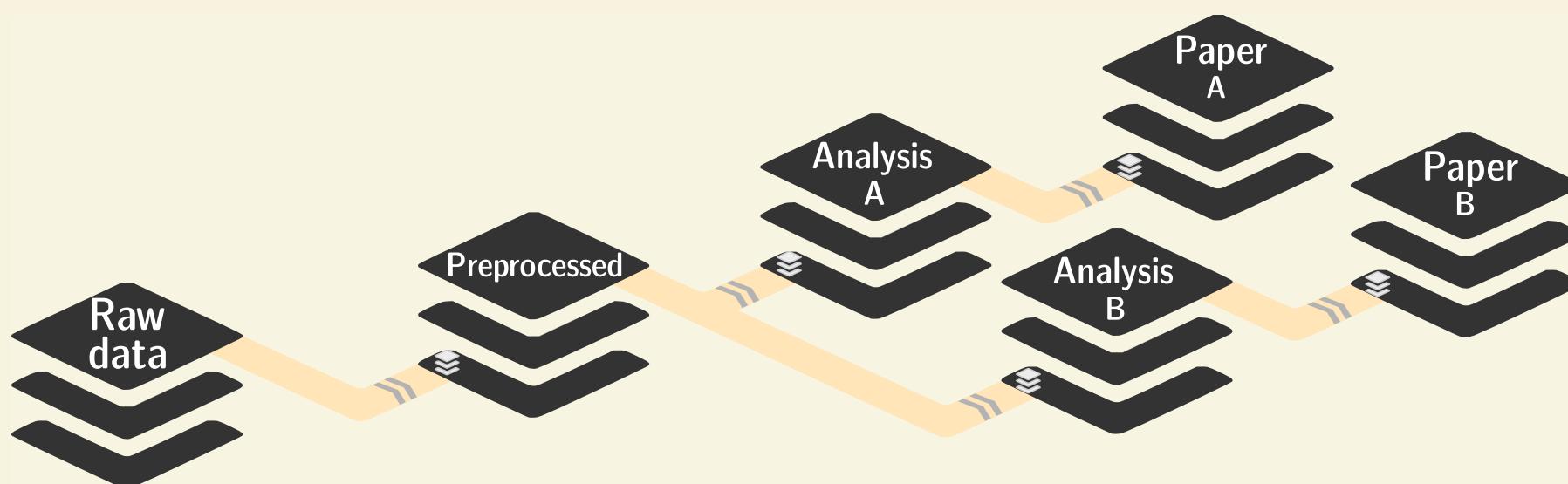


Create sibling datasets to publish to or update from

MACHINE-READABLE, RE-EXECUTABLE PROVENANCE

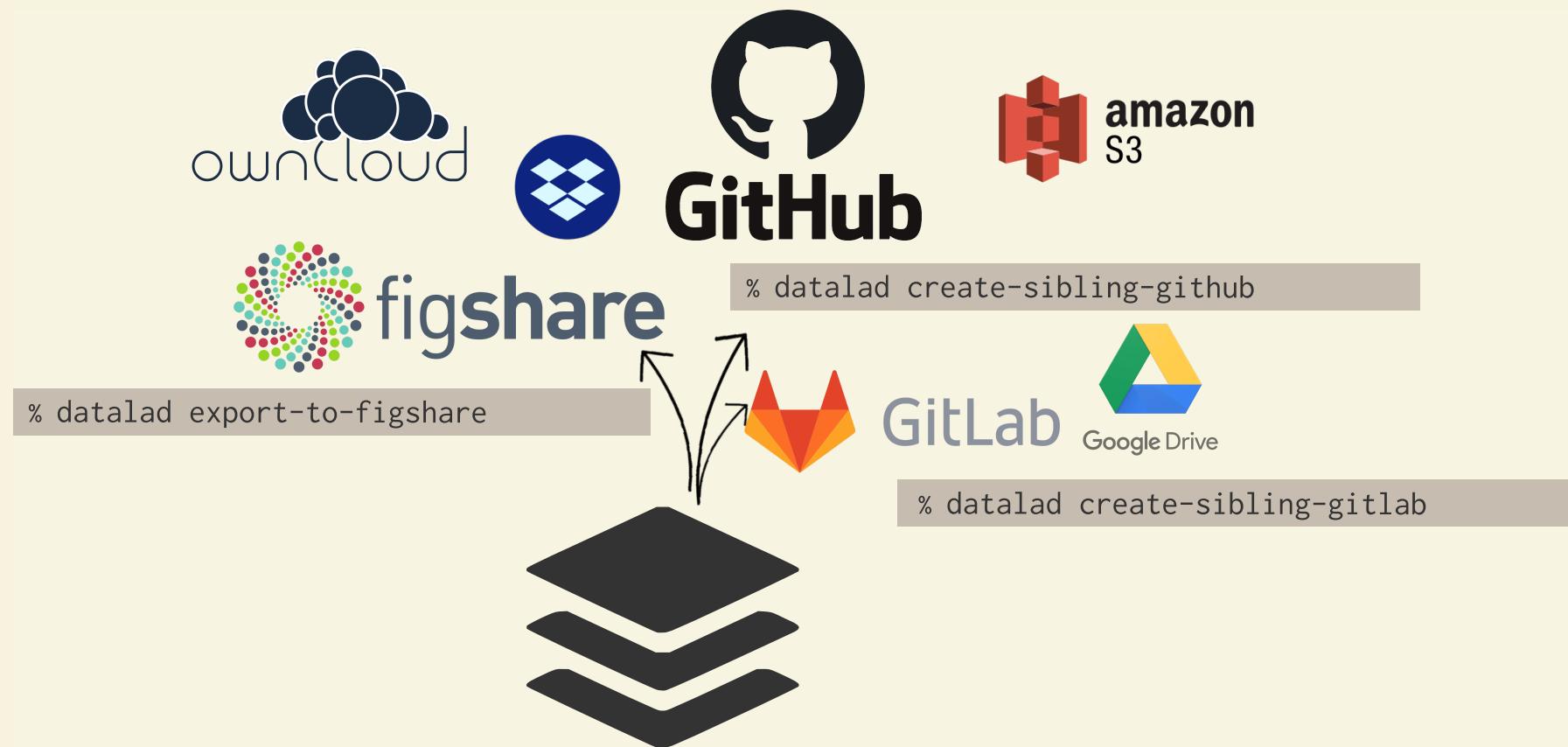


SEAMLESS NESTING AND DATASET LINKAGE



Nest modular datasets to create a linked hierarchy of datasets,
and enable recursive operations throughout the hierarchy

THIRD PARTY INTEGRATIONS



Apart from **local computing infrastructure** (from private laptops to computational clusters), datasets can be hosted in major **third party repository hosting and cloud storage** services. More info: Chapter on **Third party infrastructure**.

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Behind-the-scenes infrastructure component for data transport and versioning (e.g., used by OpenNeuro, brainlife.io , the Canadian Open Neuroscience Platform (CONP), CBRAIN)

The screenshot shows the homepage of the OpenNeuro platform. At the top, there is a navigation bar with links for 'PUBLIC DASHBOARD', 'SUPPORT', 'FAQ', and 'SIGN IN'. Below the navigation bar, there is a large graphic of a brain network. The main title 'OpenNEURO' is displayed prominently, followed by the subtitle 'A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data'. Below the subtitle, there are two sign-in buttons: 'Sign in with Google' and 'Sign in with ORCID'. There is also a search bar labeled 'Search Datasets' and a link to 'Browse All Public Datasets'. Two large statistics are displayed: '419 Public Datasets' and '13587 Participants'. At the bottom, there are three sections: 'Get Data', 'Share Data', and 'Use Data', each accompanied by a brain network icon and a brief description.

https://openneuro.org

PUBLIC DASHBOARD SUPPORT FAQ SIGN IN

OpenNEURO

A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data

G Sign in with Google

id Sign in with ORCID

Search Datasets Q

Browse All Public Datasets

419 Public Datasets

13587 Participants

Get Data

Share Data

Use Data

Browse and download datasets from contributors all over the world.

Upload your data to an NIH Brain Initiative approved repository.

Use our affiliated website to process applicable data.

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Creating and sharing reproducible, open science: Sharing data, software, code, and provenance

The screenshot shows a GitHub repository page for 'psychoinformatics-de / paper-remodnav'. The repository has 429 commits, 7 branches, and 2 tags. The 'About' section includes a link to a manuscript at <https://doi.org/10.1101/619254>. The 'Releases' section shows 2 tags. The 'Languages' section indicates 79.4% TeX, 20.0% Python, and 0.6% Makefile.

About
Code, data and manuscript for
<https://doi.org/10.1101/619254>

Releases
2 tags

Packages
No packages published
[Publish your first package](#)

Contributors (3)

- adswa Adina Wagner
- mih Michael Hanke
- ElectronicTeaCup Asim H ...

Languages

- TeX 79.4%
- Python 20.0%
- Makefile 0.6%

Code

master · 6f09e35 · Jun 5 · 429 commits

Issues 1 · **Pull requests** 1 · **Actions** · **Projects** · **Wiki** · **Security** · **Insights**

Code · **Add file** · **Code**

adswa Merge pull request #12 from psychoinformatics-de/fin... · 6f09e35 · on Jun 5 · 429 commits

File	Description	Time Ago
code	Merge pull request #12 from psychoinformatics-de/finalr...	4 months ago
data	Point to latest label dataset	17 months ago
img	Label panels in flowchart	6 months ago
remodnav @ d289118	Update remodnav with latest test dataset	17 months ago
.gitignore	Prevent permanent rebuilds of the figures	17 months ago
.gitmodules	Add and use studyforrest raw eyegaze dataset as direct...	17 months ago
COPYING	Declare CC-BY license	17 months ago
Makefile	Simplify reference management	7 months ago
README.md	Fix installation instructions	4 months ago
main.tex	Minor edits as suggested by reviewer 2	5 months ago
references.bib	velocity versus dispersion-based: cite a few algorithms ...	7 months ago
results_def.tex	Put generated results back into Git	8 months ago
spbasic bst	Basic switch to new layout	2 years ago
svglov3 clo	Basic switch to new layout	2 years ago
svjour3 cls	Basic switch to new layout	2 years ago

README.md

REMoDNaV: Robust Eye Movement

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Creating and sharing reproducible, open science: Sharing data, software, code, and provenance

The screenshot shows a Twitter thread from Lennart Wittkuhn (@lnnrtwttkh). The first tweet discusses a new fMRI analysis method for decoding fast neural event sequences. It includes a link to a paper on nature.com and a link to a follow-up thread. The second tweet provides more details about the fMRI analysis and links to the project website (wittkuhn.mpib.berlin). The third tweet links to the project's DataLad repository (<https://github.com/wittkuhn/fMRI>). The sidebar shows relevant users, trends, and a search bar.

Lennart Wittkuhn @lnnrtwttkh · 19. März
Excited to share work w/ @nico_schuck out now in @NatureComms! 🎉

We introduce a new fMRI analysis method to decode fast neural event sequences and report replay in visual cortex following a non-mnemonic task! 🧠

- 📄 Paper: nature.com/articles/s4146...
- Thread below! 🎉 [1/n]

Dynamics of fMRI patterns reflect sub-second activ... Non-invasive measurement of fast neural activity with spatial precision in humans is difficult. Here, t... nature.com

4 93 270

Lennart Wittkuhn @lnnrtwttkh
Antwort an @lnnrtwttkh

We share all code + data via [@gnode](#) + [GitHub](#), version-controlled with [@datalad](#) (ca. 1.5 TB): MRI in [@BIDSstandard](#), [#fMRIPrep](#) data, [#MRIQC](#) metrics, GLMs + anatomical masks, task code, decoding pipeline, statistical analyses: wittkuhn.mpib.berlin /highspeed/ [#OpenScience](#) [2/n]

Tweet übersetzen

Dynamics of fMRI patterns reflect sub-second activation se... This is the project website of the accompanying paper 'Faster than thought: Detecting sub-second activation ... wittkuhn.mpib.berlin

11:35 vorm. · 19. März 2021 · Twitter Web App

7 Retweets 2 Zitierte Tweets 43 „Gefällt mir“-Angaben

Lennart Wittkuhn @lnnrtwttkh · 19. März
Antwort an @lnnrtwttkh

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Central data management and archival system

The screenshot shows a web browser displaying a GitLab group page for 'INM7'. The URL in the address bar is <https://jugit.fz-juelich.de/inm7>. The page title is 'INM7' with a globe icon. Below it, it says 'Group ID: 309' and a 'Leave group' link. A 'New project' button is visible. The main content area includes a search bar labeled 'Search by name' and a dropdown menu labeled 'Last created'. Below these are tabs for 'Subgroups and projects', 'Shared projects', and 'Archived projects'. The 'Subgroups and projects' tab is selected, showing a list of projects: 'vbc' (highlighted with a pink background), 'Training' (highlighted with a green background and marked as 'Owner'), 'webtools' (highlighted with a grey background), 'tools' (highlighted with a pink background), and 'AppliedMachineLearning' (highlighted with a light blue background). Each project entry includes a brief description and small icons for more options.

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Scalable computing framework for reproducible science

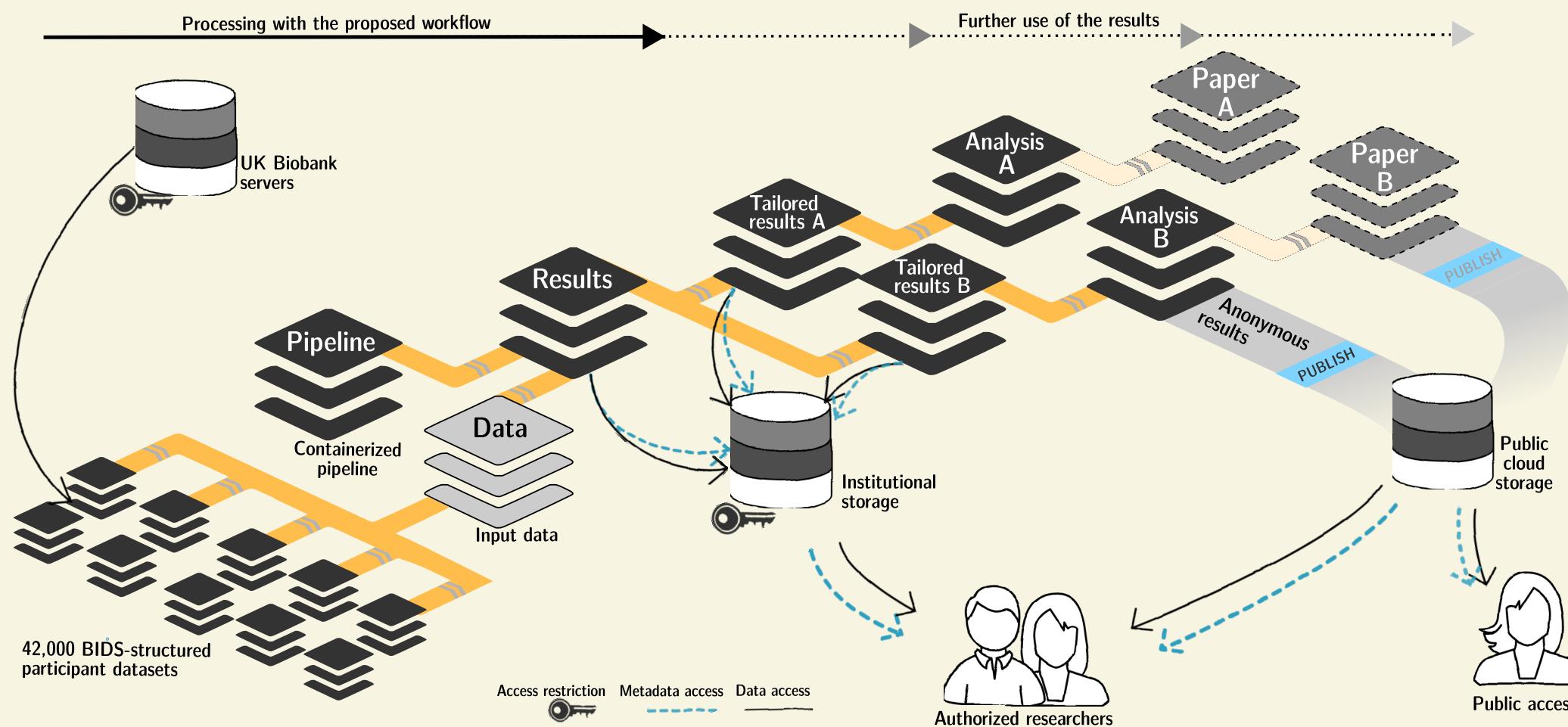
Article | [Open Access](#) | [Published: 11 March 2022](#)

FAIRly big: A framework for computationally reproducible processing of large-scale data

[Adina S. Wagner](#) , [Laura K. Waite](#), [Małgorzata Wierzba](#), [Felix Hoffstaedter](#), [Alexander Q. Waite](#), [Benjamin Poldrack](#), [Simon B. Eickhoff](#) & [Michael Hanke](#)

[Scientific Data](#) 9, Article number: 80 (2022) | [Cite this article](#)

813 Accesses | 23 Altmetric | [Metrics](#)



... AND MANY MORE!

Let's fire up a terminal and get started with the Basics