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!)

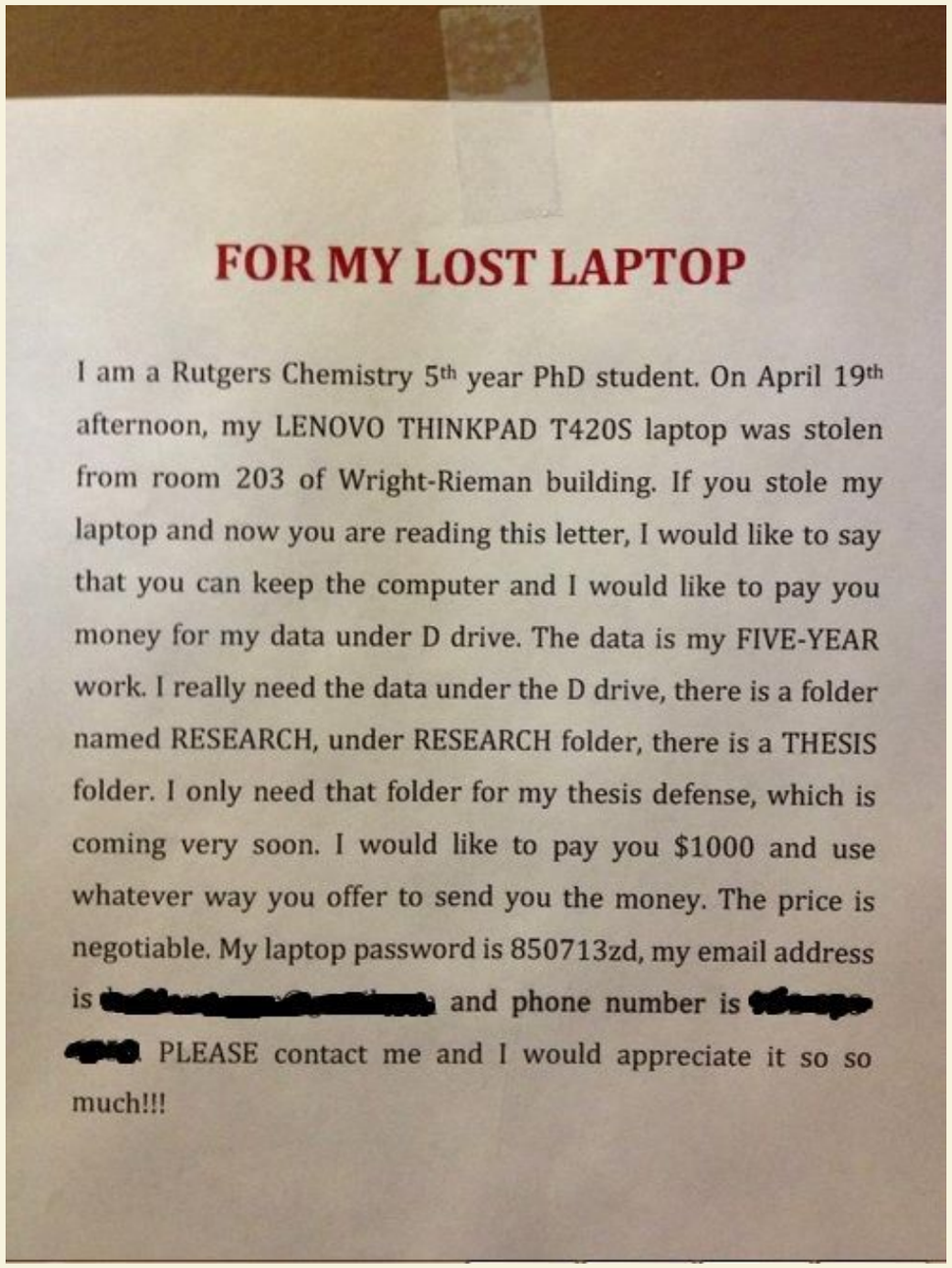
MOTIVATION

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.

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

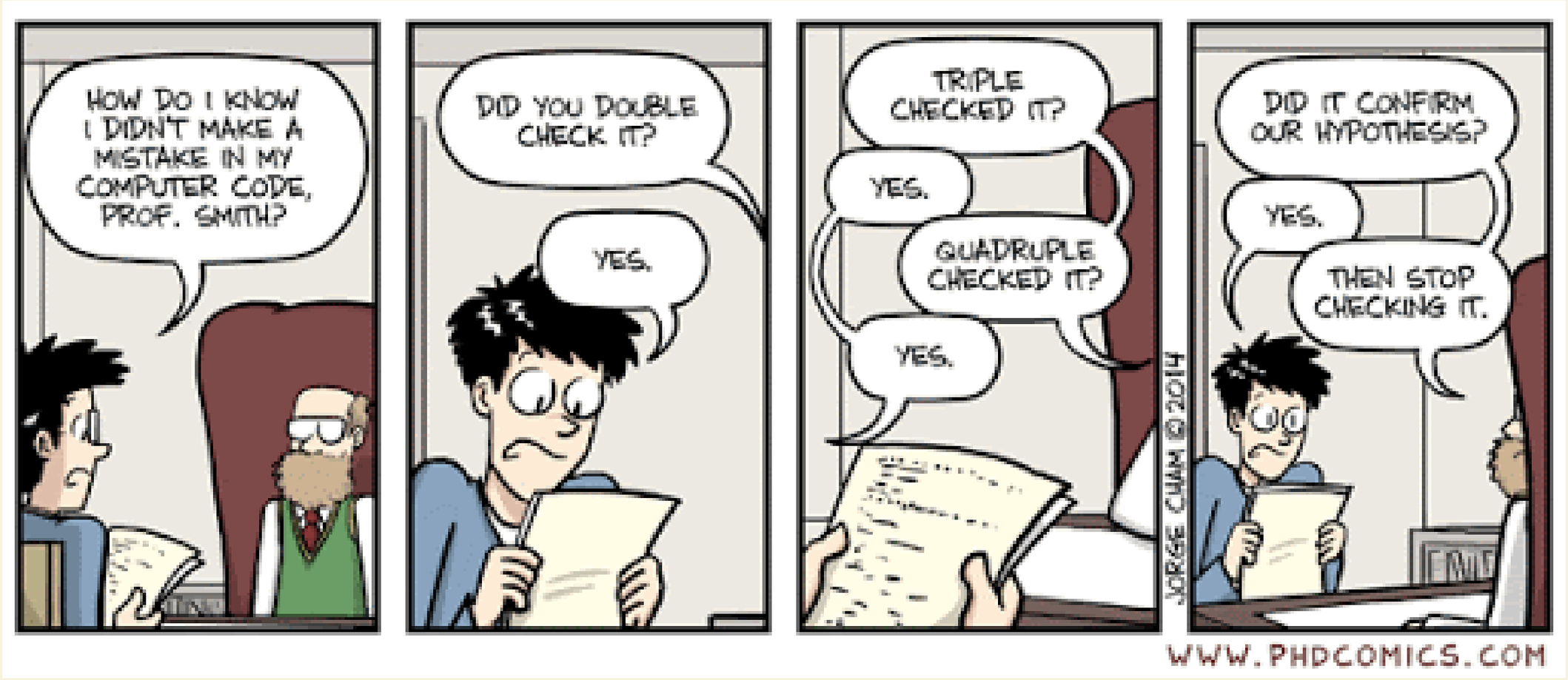


FOR MY LOST LAPTOP

I am a Rutgers Chemistry 5th year PhD student. On April 19th afternoon, my LENOVO THINKPAD T420S laptop was stolen from room 203 of Wright-Rieman building. If you stole my laptop and now you are reading this letter, I would like to say that you can keep the computer and I would like to pay you money for my data under D drive. The data is my FIVE-YEAR work. I really need the data under the D drive, there is a folder named RESEARCH, under RESEARCH folder, there is a THESIS folder. I only need that folder for my thesis defense, which is coming very soon. I would like to pay you $1000 and use whatever way you offer to send you the money. The price is negotiable. My laptop password is 850713zd, my email address is [redacted] and phone number is [redacted] PLEASE contact me and I would appreciate it so so much!!!

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

NSF 1912266

germany-usa

Federal Ministry of Education and Research

BMBF 01GQ1905

JÜLICH Forschungszentrum

EUROPEAN UNION
European Regional Development Fund

cbbs center for behavioral brain sciences

SACHSEN-ANHALT
Ministerium für Wissenschaft und Wirtschaft

Collaborators

Human Brain Project

CONP PCNO

VirtualBrainCloud

OpenNEURO

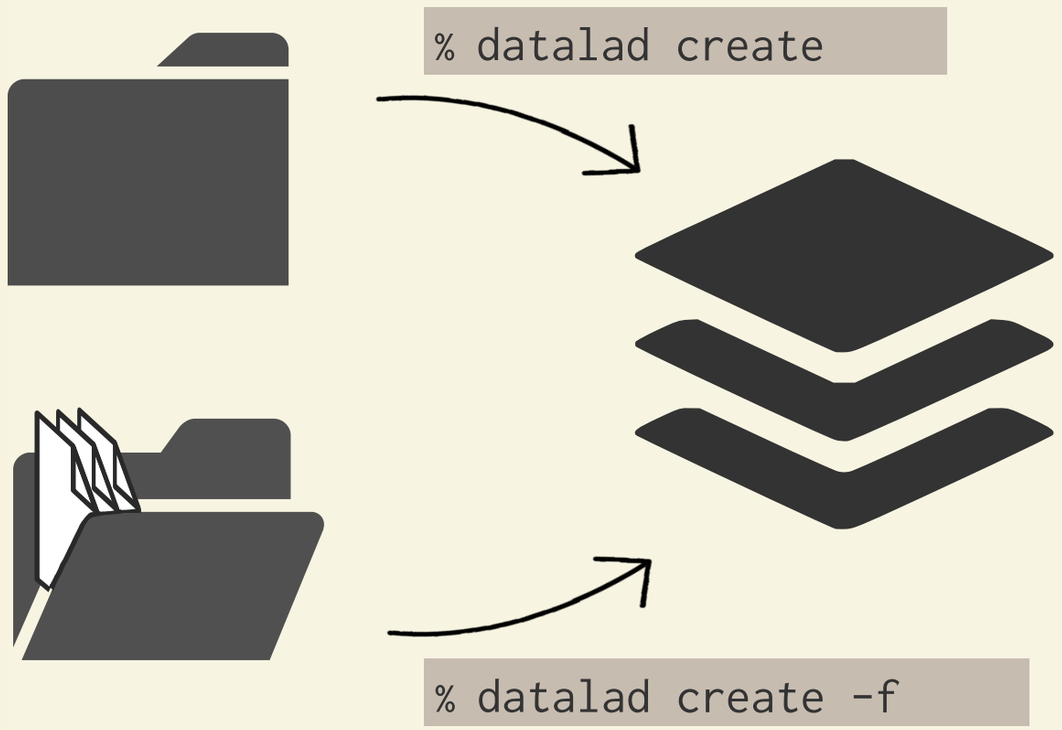
cbrain

brainlife.io

CORE CONCEPTS & FEATURES

EVERYTHING HAPPENS IN DATALAD DATASETS

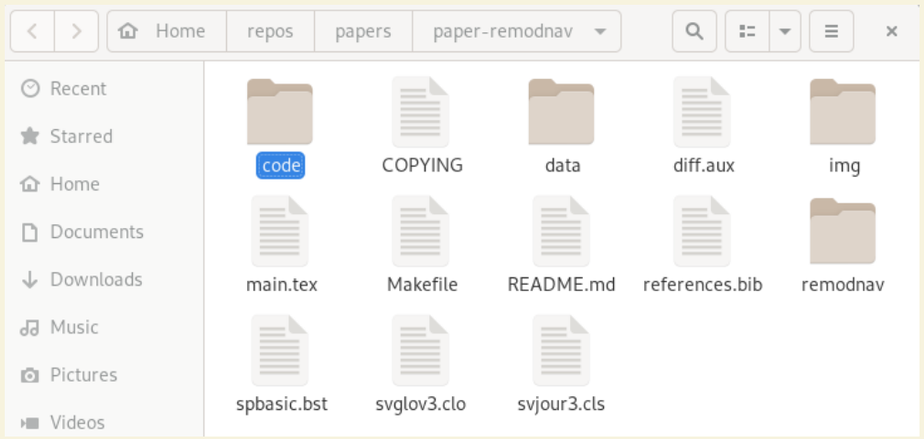
create new, empty datasets to populate...



.... or transform existing directories into datasets

DATASET = GIT/GIT-ANNEX REPOSITORY

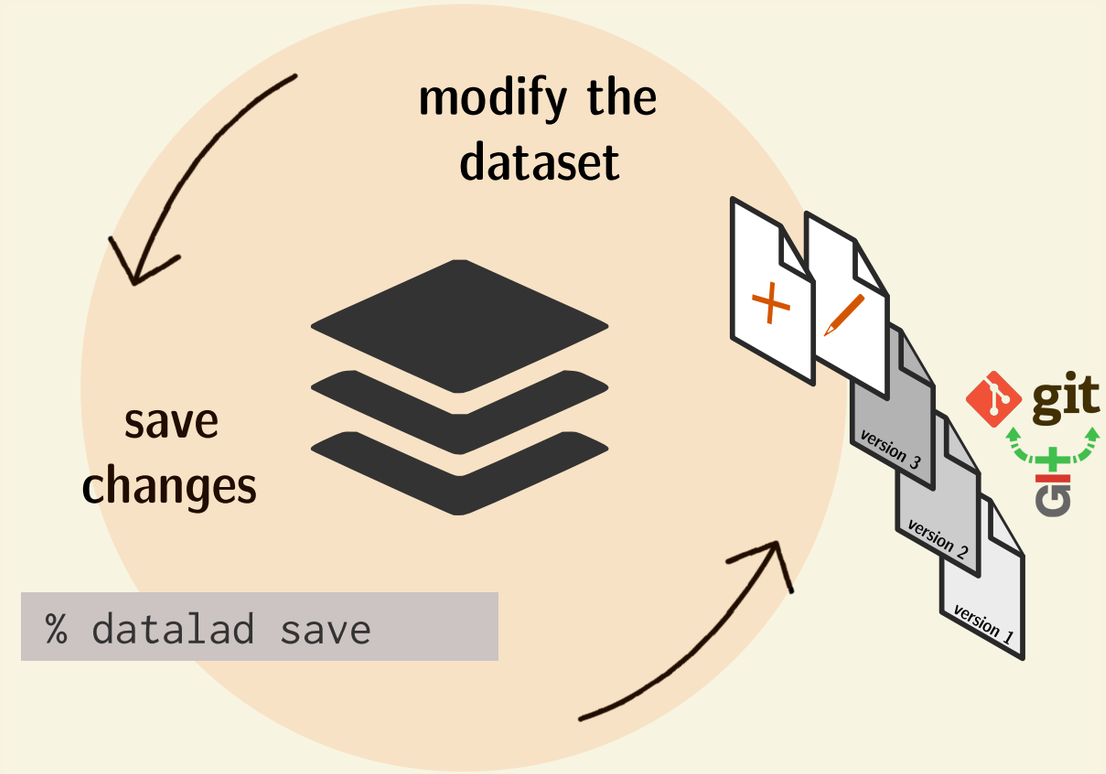
- content agnostic
- no custom data structures
- complete decentralization
- Looks and feels like a directory on your computer:




```
adina@muninn ~/repos/papers/paper-remodnav on master!
➤ ls
code  remodnav  main.tex   references.bib  svjour3.cls
data  COPYING   Makefile   spbasic.bst
img   diff.aux  README.md  svglov3.clo
```

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

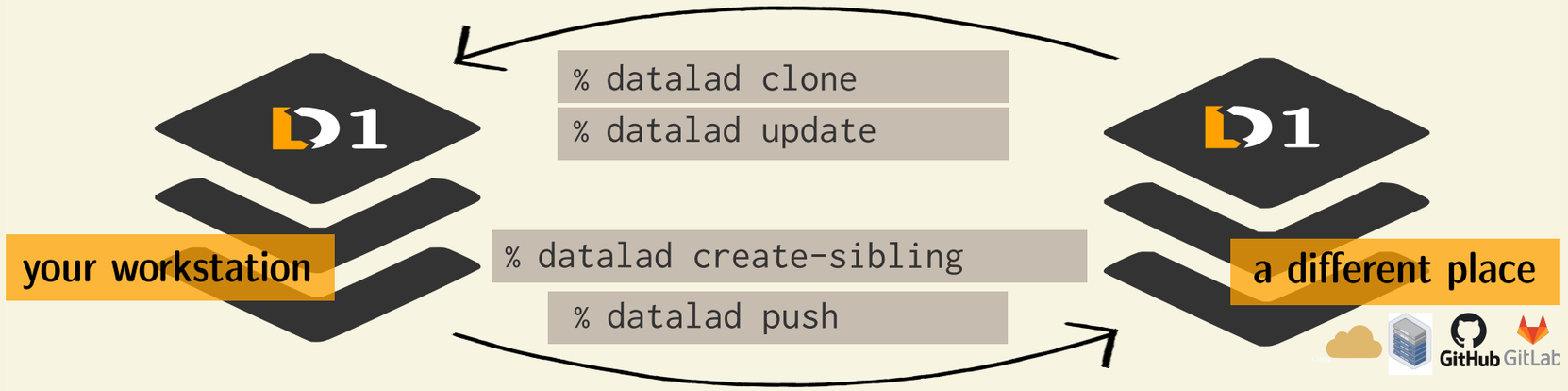
```
2020-06-05 10:58 +0200 Adina Wagner  M─┐ [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-d
2020-06-05 08:24 +0200 Adina Wagner  │ o [finalround] {upstream/finalround} add results from computing with mean instead of median
2020-06-05 09:09 +0200 Michael Hanke 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
```

Date Author Change summary (commit message)

- 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

capture the origin of files obtained from web sources

```
% datalad download-url
```

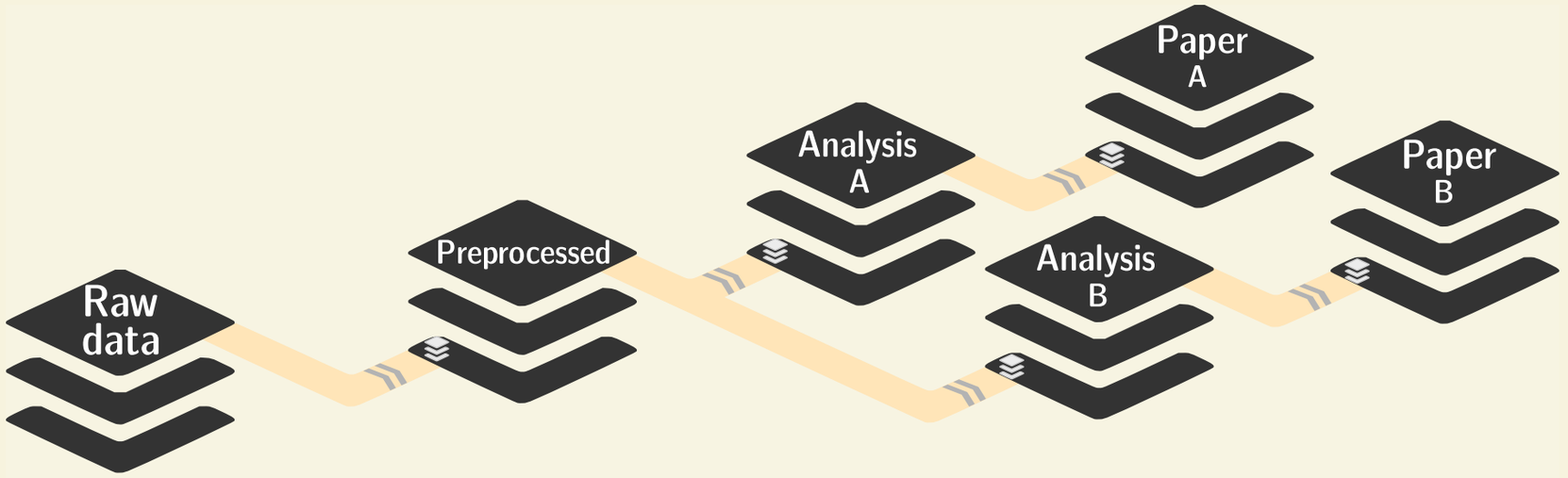
link input, code, containerized software environments, and output, or re-run previous executions


```
% datalad run
```

```
% datalad rerun
```

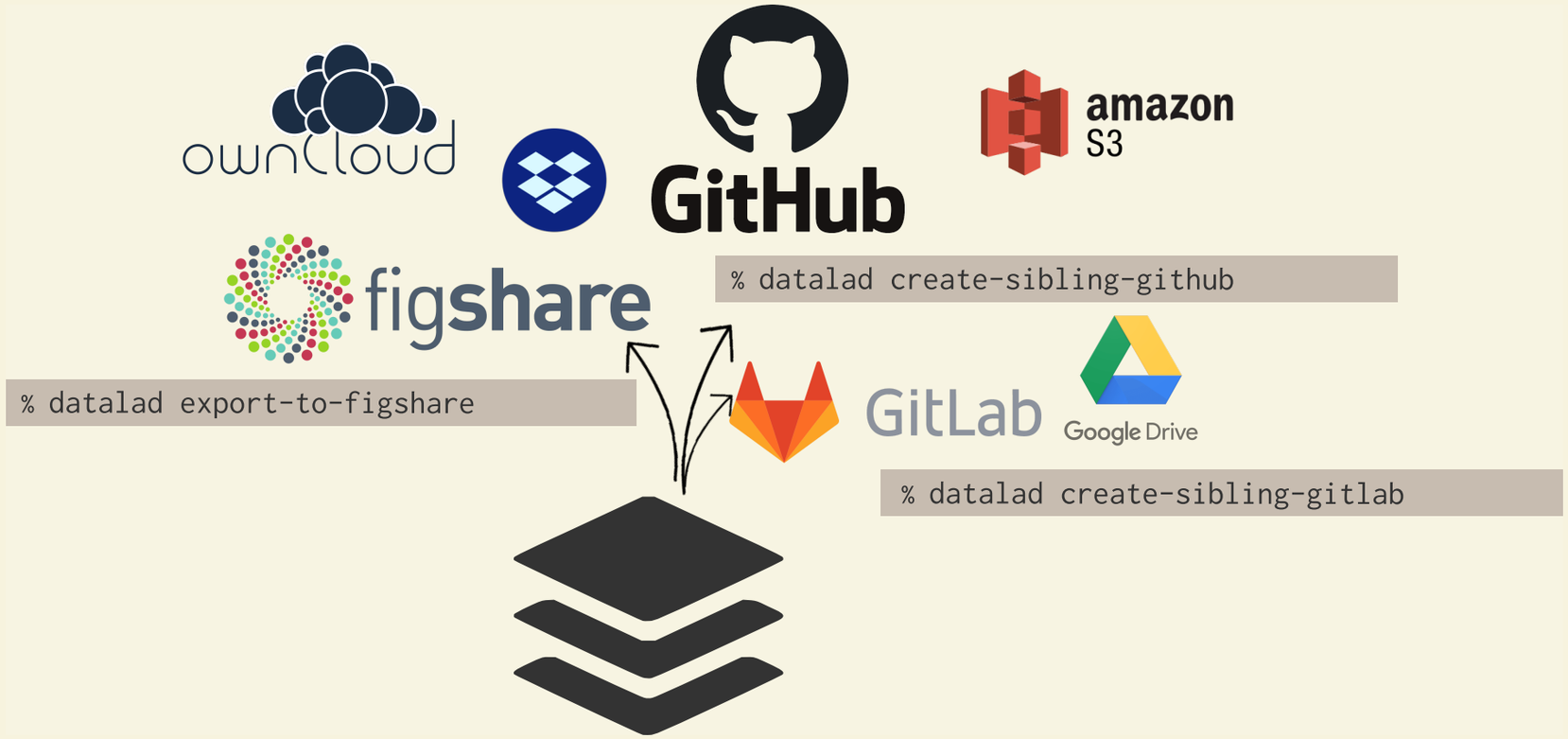
```
% datalad run-procedure
```

SEAMLESS NESTING AND DATASET LINKAGE



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

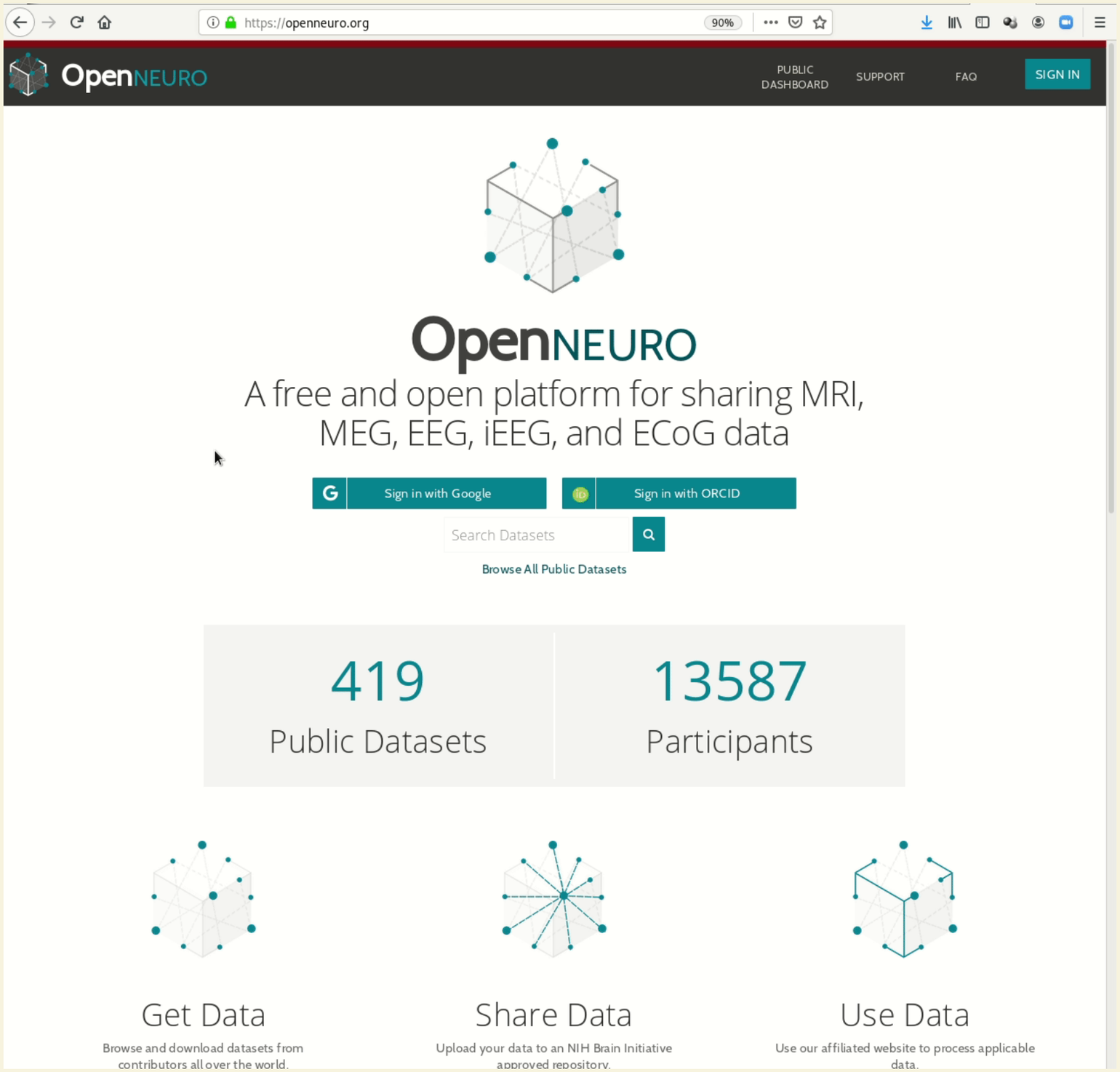
THIRD PARTY INTEGRATIONS



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.

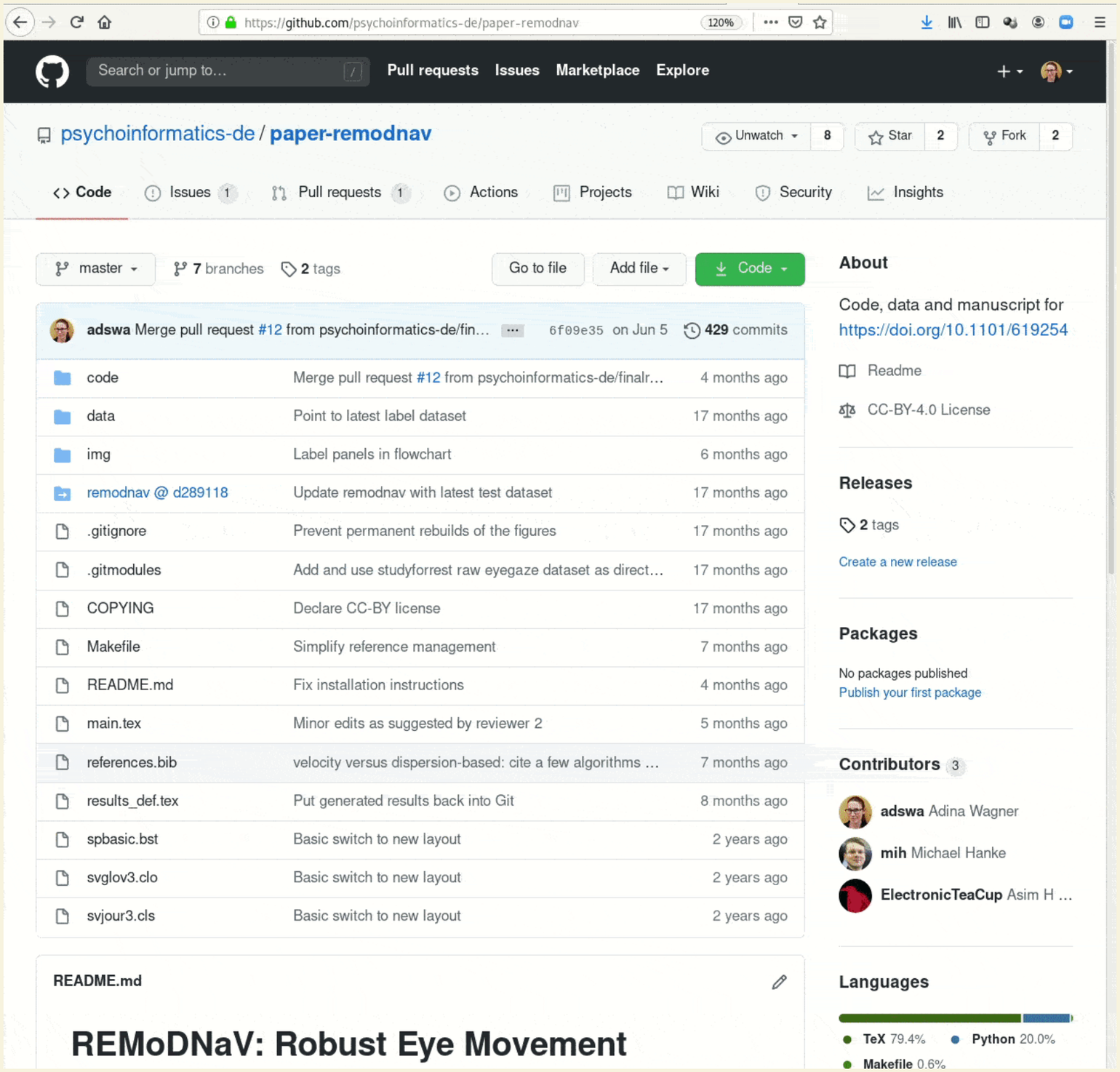
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

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



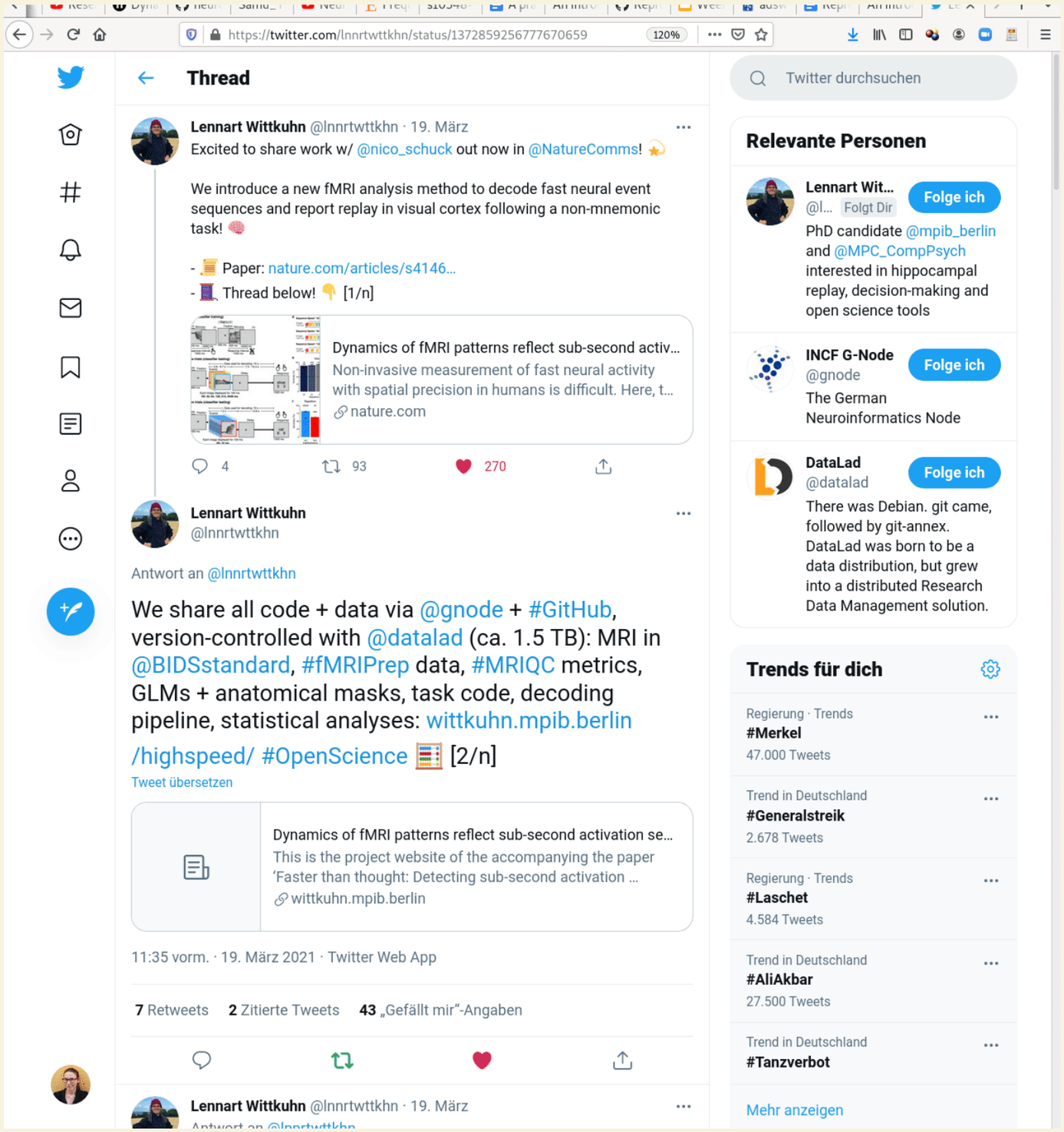
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

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



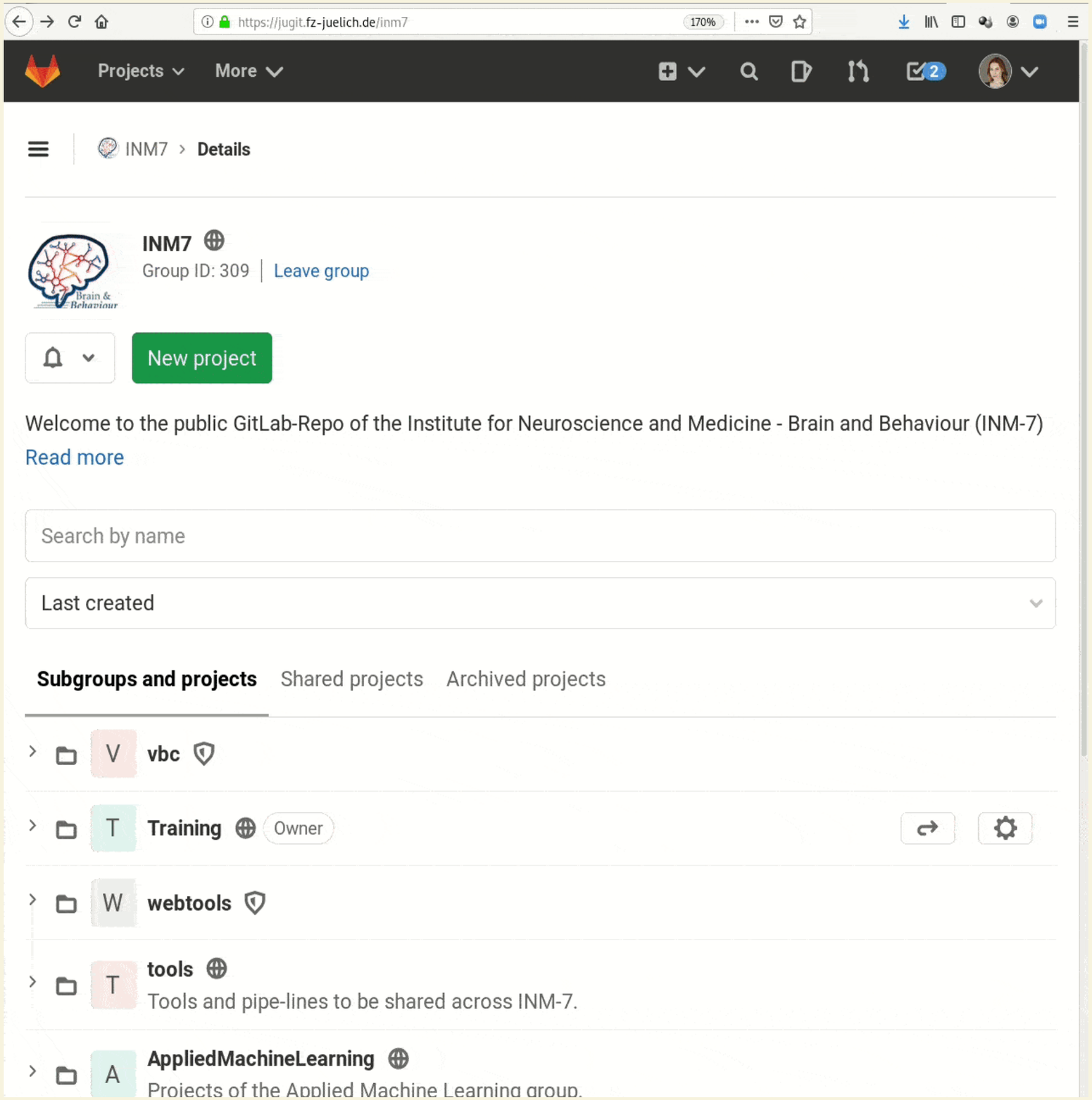
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

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



EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Central data management** and archival system



EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Scalable computing framework** for reproducible science

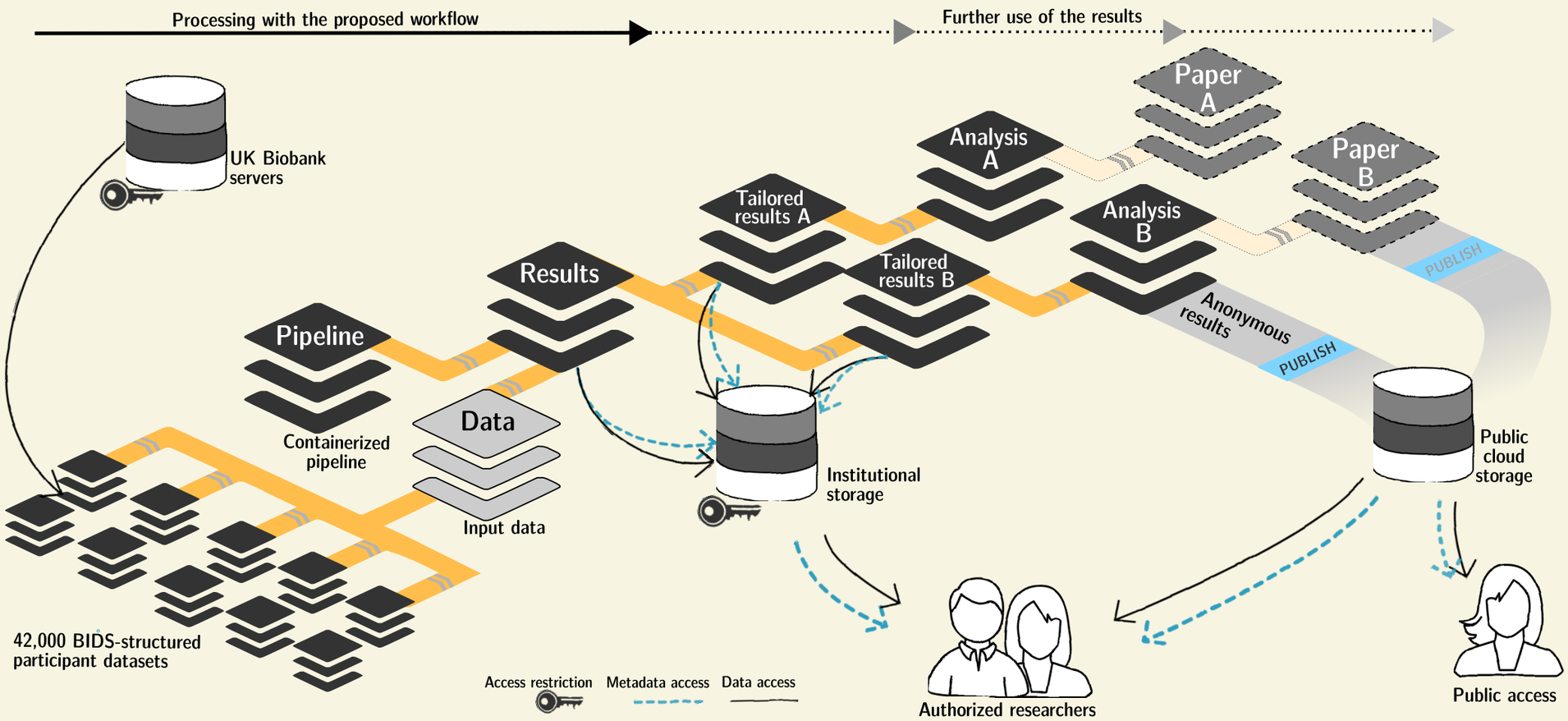
Article | Open Access | Published: 11 March 2022

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