


DATA MANAGEMENT FOR OPEN & REPRODUCIBLE SCIENCE

Adina Wagner

 @AdinaKrik

Psychoinformatics lab,
Institute of Neuroscience and Medicine, Brain & Behavior (INM-7)
Research Center Jülich
ReproNim/INCF fellow

Slides: [DOI 10.5281/zenodo.5702023](https://doi.org/10.5281/zenodo.5702023) (Scan the QR code)

Sources: github.com/datalad-handbook/course

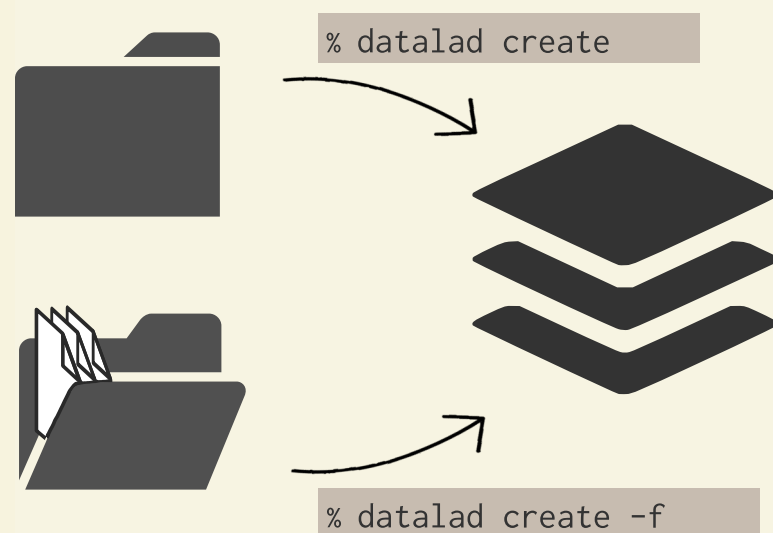


- A command-line tool, available for all major operating systems (Linux, macOS/OSX, Windows), free & open source
- Build on top of **Git** and **Git-annex**
- **Main features:**
 - Version control for arbitrarily large content**
version control data and software alongside to code!
 - Transport logistics 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
- Completely domain-agnostic

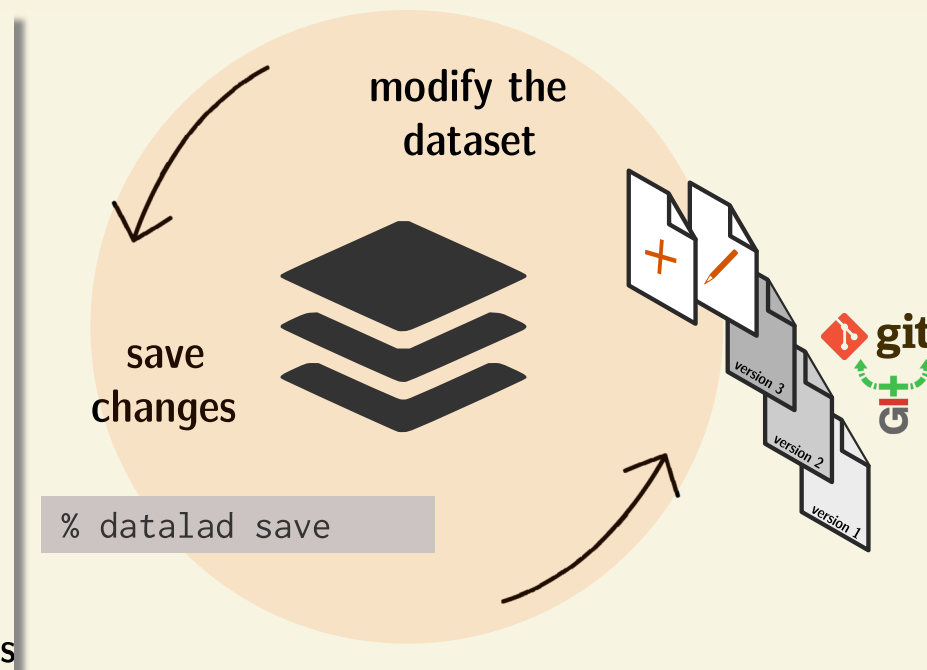
VERSION CONTROL

- DataLad knows two things: Datasets and files

create new, empty datasets to populate...



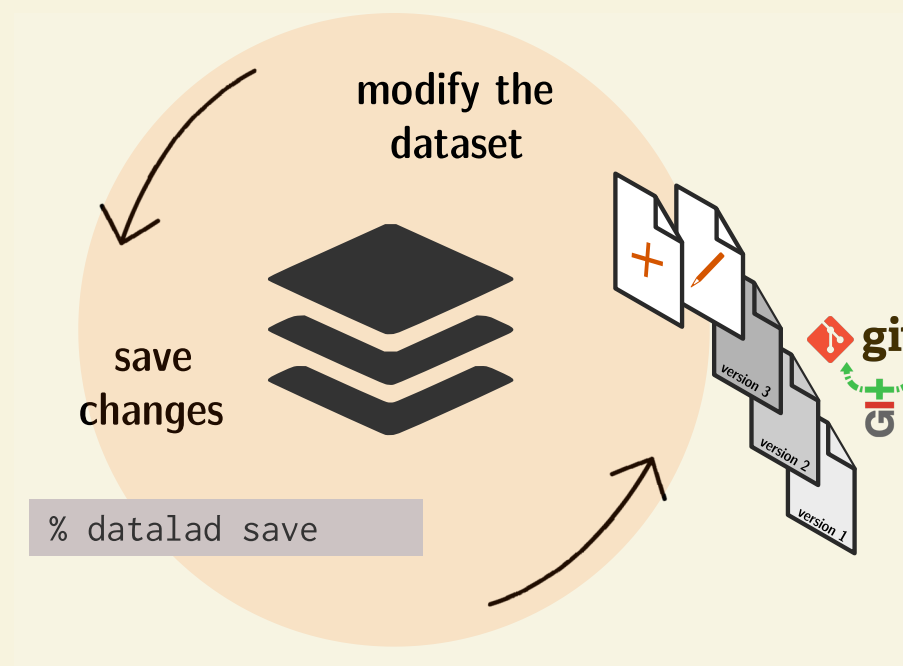
.... or transform existing directories into datasets



- A DataLad dataset is a **Git repository**:
 - keep track of changes
 - revert changes or go back to previous states
 - collect and share digital provenance

VERSION CONTROL: DATA

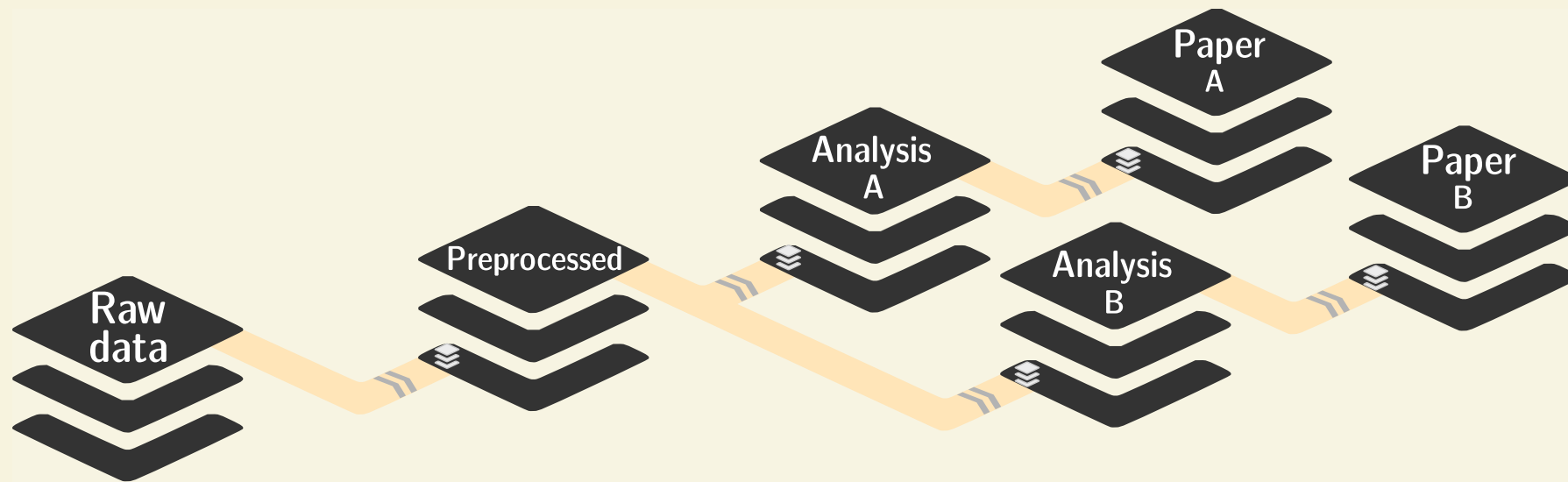
- Datasets have an optional annex for (large or sensitive) data (or text/code).
- Identity (hash) and location information is put into Git, rather than file content. The annex, and transport to and from it is managed with `git-annex` (git-annex.branchable.com)
→ decentralized version control for files of any size.
- DataLad works towards wrapping Git and git-annex into a non-complex core-API (helpful for data management novices).



- Flexibility and commands of Git and git-annex are preserved (useful for experienced Git/git-annex users).

VERSION CONTROL: NESTING

- Link datasets as "dependencies":



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

- hierarchies of datasets in super-/sub-dataset relationships
- ✓ Scalability

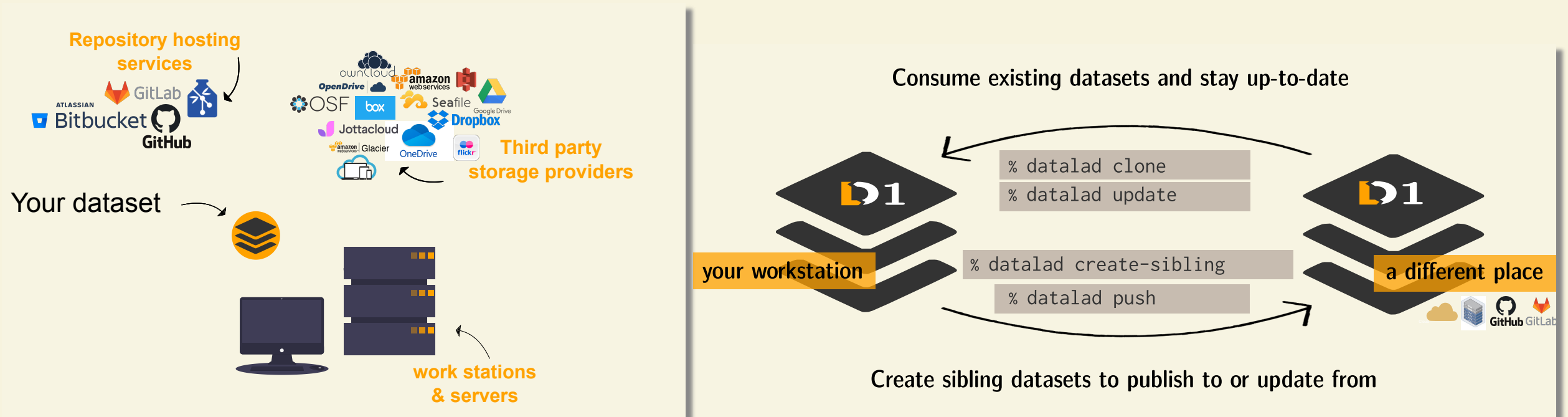
```
adina@bulk1 in /ds/hcp/super on git:master › datalad status --annex -r
15530572 annex'd files (77.9 TB recorded total size)
nothing to save, working tree clean
```

(github.com/datalad-datasets/human-connectome-project-openaccess)

- ✓ Modularizes research components for transparency, reuse, and access management

TRANSPORT LOGISTICS

- Share datasets easily
- Datasets can be "cloned", "pushed", and "updated" from and to local paths, remote hosting services, cloud services, ...



TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean:

```
$ datalad clone git@github.com:datalad-datasets/machinelearning-books.git
install(ok): /tmp/machinelearning-books (dataset)
$ cd machinelearning-books && du -sh
348K  .
```

```
$ ls
A.Shashua-Introduction_to_Machine_Learning.pdf
B.Efron_T.Hastie-Computer_Age_Statistical_Inference.pdf
C.E.Rasmussen_C.K.I.Williams-Gaussian_Processes_for_Machine_Learning.pdf
D.Barber-Bayesian_Reasoning_and_Machine_Learning.pdf
[...]
```

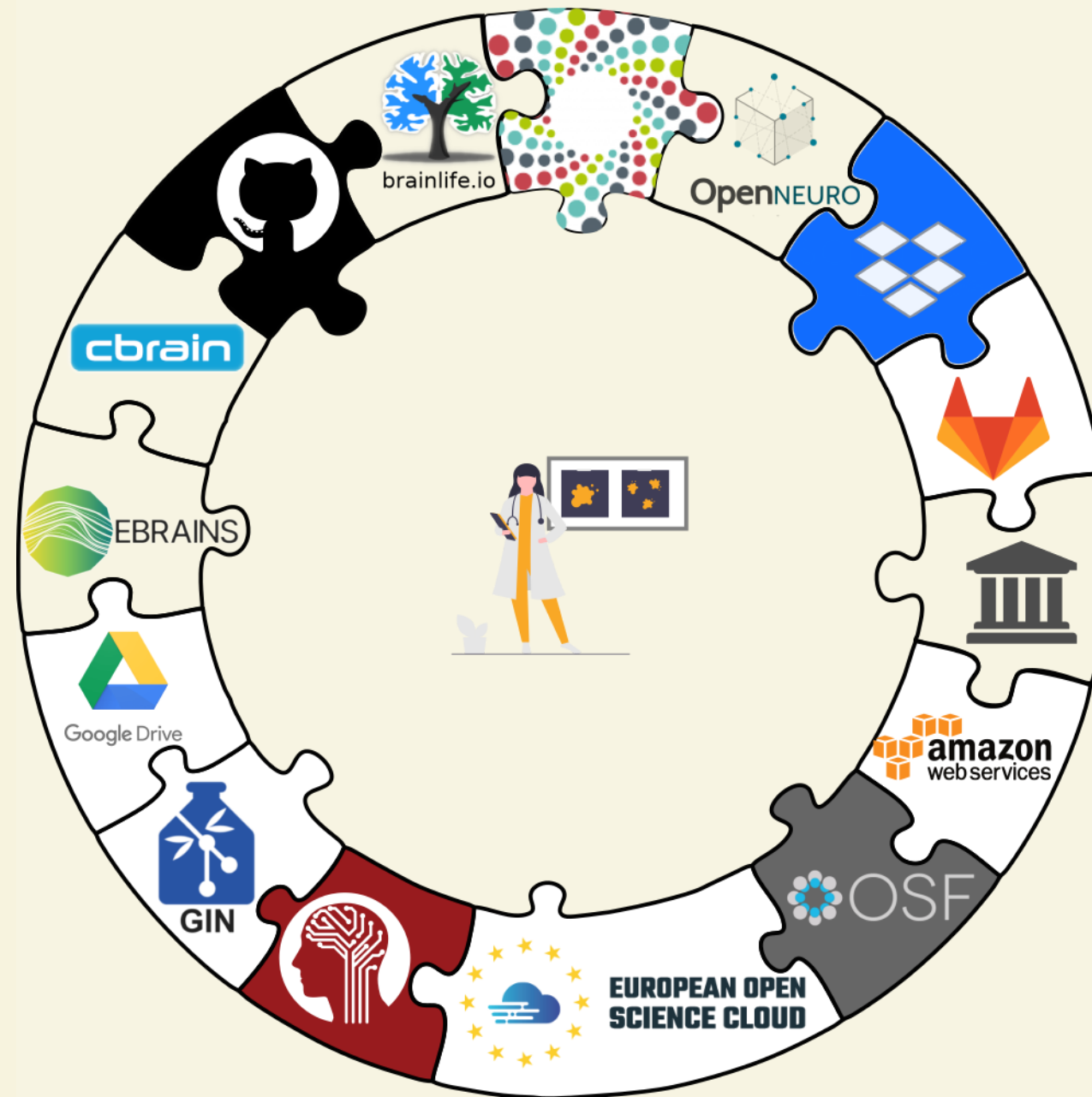
- file contents are retrieved & dropped on demand on up to per-file granularity:

```
$ datalad get A.Shashua-Introduction_to_Machine_Learning.pdf
get(ok): /tmp/machinelearning-books/A.Shashua-Introduction_to_Machine_Learning.pdf (file) [from web...]
```

```
$ datalad drop A.Shashua-Introduction_to_Machine_Learning.pdf
drop(ok): /tmp/machinelearning-books/A.Shashua-Introduction_to_Machine_Learning.pdf (file) [checking https
```

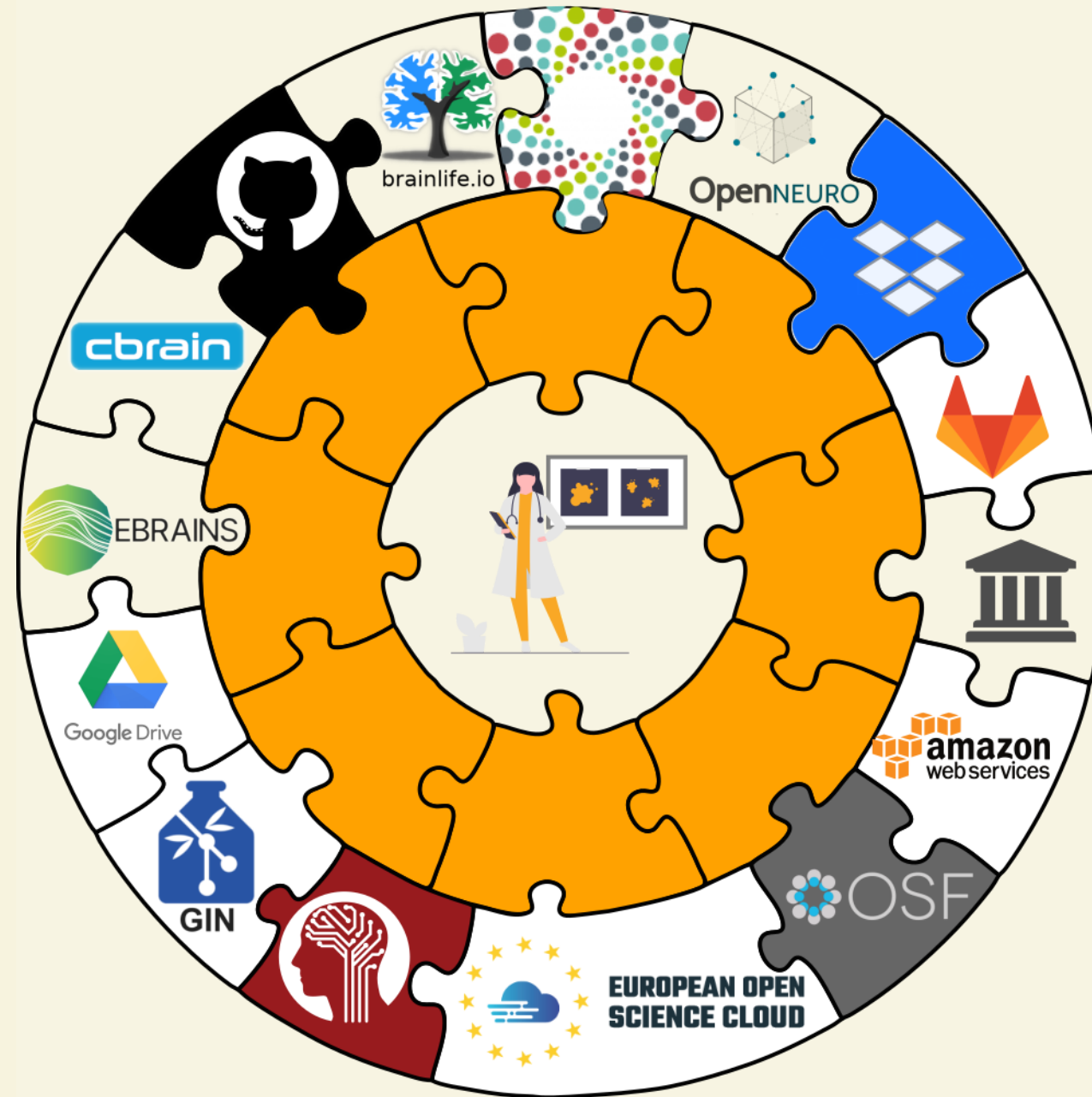
INTEROPERABILITY

- DataLad is built to maximize interoperability and use with hosting and storage technology



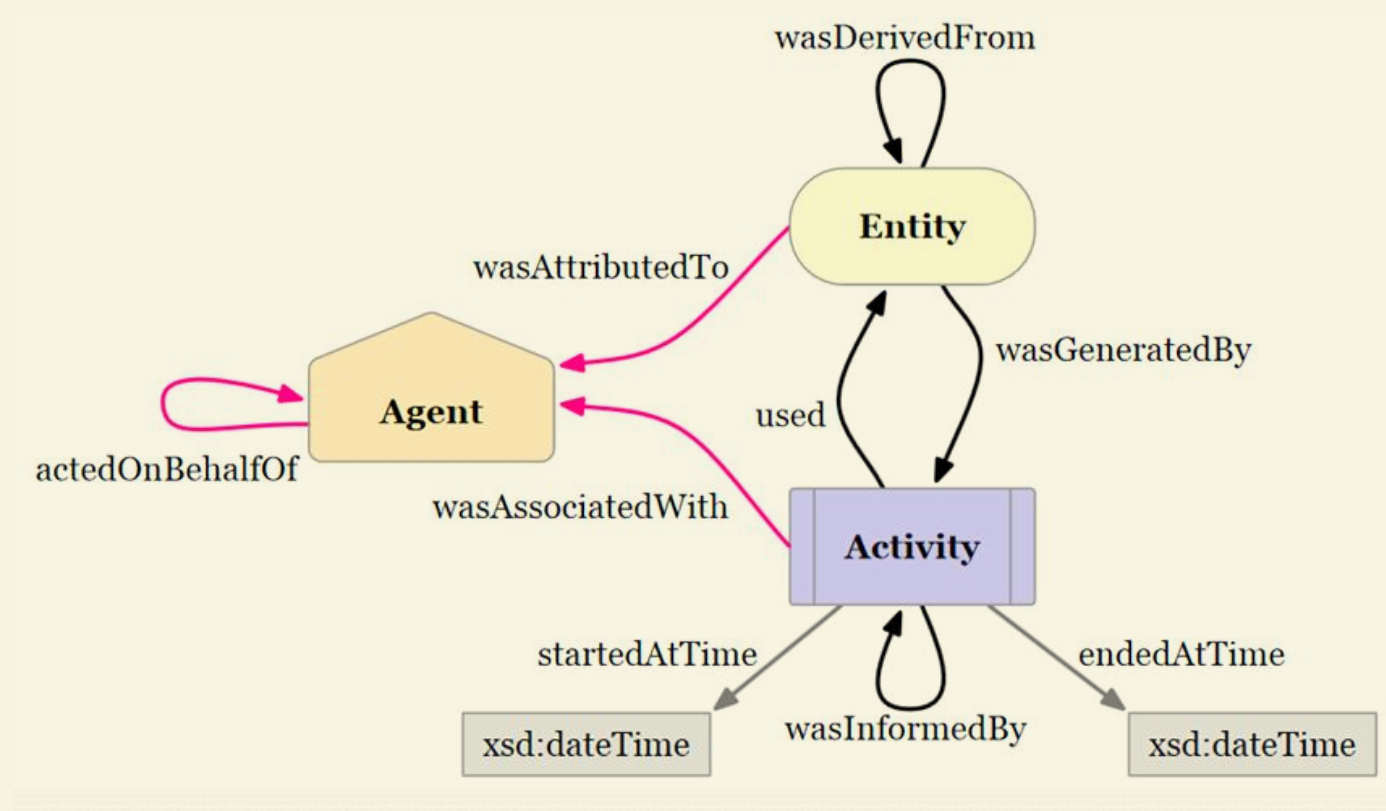
INTEROPERABILITY

- DataLad is built to maximize interoperability and use with hosting and storage technology



PROVENANCE CAPTURE

- Datasets can capture dataset **transformations** and their **cause** in order to track the entire evolution and lineage of files in datasets



- "How did this file come to be?", "What steps were undertaken to transform the raw data into the published result?", "Can you recompute this for me?"

PROVENANCE CAPTURE

- **Basic provenance:** DataLad can capture arbitrary dataset transformations (e.g., from computing analysis results) and record the cause of such a change

```
$ datalad run -m "Perform eye movement event detection"\  
  --input 'raw_data/*.tsv.gz' --output 'sub-*' \  
  bash code/compute_all.sh  
  
-- Git commit -- Michael Hanke <... @gmail.com>; Fri Sep 21 22:00:47 2019  
  [DATAHAD RUNCMD] Perform eye movement event detection  
  === Do not change lines below ===  
  {  
    "cmd": "bash code/compute_all.sh",  
    "dsid": "d2b4b72a-7c13-11e7-9f1f-a0369f7c647e",  
    "exit": 0,  
    "inputs": ["raw_data/*.tsv.gz"],  
    "outputs": ["sub-*"],  
    "pwd": "."  
  }  
  ^^^ Do not change lines above ^^^  
---  
sub-01/sub-01_task-movie_run-1_events.png | 2 +-  
sub-01/sub-01_task-movie_run-1_events.tsv | 2 +-  
...
```

PROVENANCE CAPTURE

- **Computational provenance:** Datasets can track software containers, and perform and record computations inside it:

```
$ datalad containers-run -n neuroimaging-container \  
  --input 'mri/*_bold.nii --output 'sub-*/LC_timeseries_run-*.csv' \  
  "bash -c 'for sub in sub-*; do for run in run-1 ... run-8;  
    do python3 code/extract_lc_timeseries.py \${sub} \${run}; done; done'"  
  
-- Git commit -- Michael Hanke < ... @gmail.com>; Fri Jul 6 11:02:28 2019  
  [DATALAD RUNCMD] singularity exec --bind {pwd} .datalad/e...  
  === Do not change lines below ===  
  {  
    "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..",  
    "dsid": "92ealfaa-632a-11e8-af29-a0369f7c647e",  
    "inputs": [  
      "mri/*.bold.nii.gz",  
      ".datalad/environments/nilearn.simg"  
    ],  
    "outputs": ["sub-*/LC_timeseries_run-*.csv"],  
    ...  
  }  
  ^^^ Do not change lines above ^^^  
  ---  
  sub-01/LC_timeseries_run-1.csv | 1 +  
  ...
```

PROVENANCE CAPTURE

- All recorded transformations can be re-computed automatically

```
$ datalad rerun eee1356bb7e8f921174e404c6df6aadcc1f158f0
[INFO] == Command start (output follows) =====
[INFO] == Command exit (modification check follows) =====
add(ok): sub-01/LC_timeseries_run-1.csv (file)
...
save(ok): . (dataset)
action summary:
  add (ok: 45)
  save (notneeded: 45, ok: 1)
  unlock (notneeded: 45)
...
```

- Aid with the reproducibility of a result and verify it (via content hash)
- Use complete capture and automatic re-computation as alternative to storage and transport

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Publish or consume datasets via GitHub, GitLab, OSF, or similar services

The screenshot shows a GitHub repository page for 'psychoinformatics-de / studyforrest-data-phase2'. The page includes a navigation bar with 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below the repository name, there are buttons for 'Unwatch', 'Unstar', and 'Fork'. The main content area shows a list of files and folders, including '.datalad', 'code', 'src', 'stimuli', and a series of 'sub-01' through 'sub-20' folders. Each folder has a description and a commit date. On the right side, there is an 'About' section with a description of the data, a 'Releases' section with a 'First public release' on Mar 26, 2016, and a 'Contributors' section listing 'mih Michael Hanke', 'dakot Daniel Kottke', and 'adswa Adina Wagner'.

https://github.com/psychoinformatics-de/studyforrest-data-phase2

Search or jump to... Pull requests Issues Marketplace Explore

psychoinformatics-de / studyforrest-data-phase2

Unwatch 1 Unstar 6 Fork 8

Code Issues 4 Pull requests 1 Actions Projects Security Insights

master 2 branches 1 tag

Go to file Add file Code

mih Merge pull request #15 from adswa/ENH/README b5306e2 on May 7 77 commits

.datalad	[DATALAD] dataset aggregate metadata update	2 years ago
code	Fix type in physio log converter (fixes gh-11)	3 years ago
src	Recover lost segment from eyetracker (closes gh-3)	5 years ago
stimuli	Add BIDS-compatible stimuli/ directory (with symlinks)	4 years ago
sub-01	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-02	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-03	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-04	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-05	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-06	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-09	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-10	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-14	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-15	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-16	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-17	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-18	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-19	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-20	BF: Re-import respiratory trace after bug fix in converte...	3 years ago

About

studyforrest.org: Phase2 data (movie, eyetracking, retmapping, visual localizers) [BIDS]

studyforrest.org

Readme

View license

Releases 1

First public release Latest on Mar 26, 2016

Packages

No packages published

Contributors 3

mih Michael Hanke

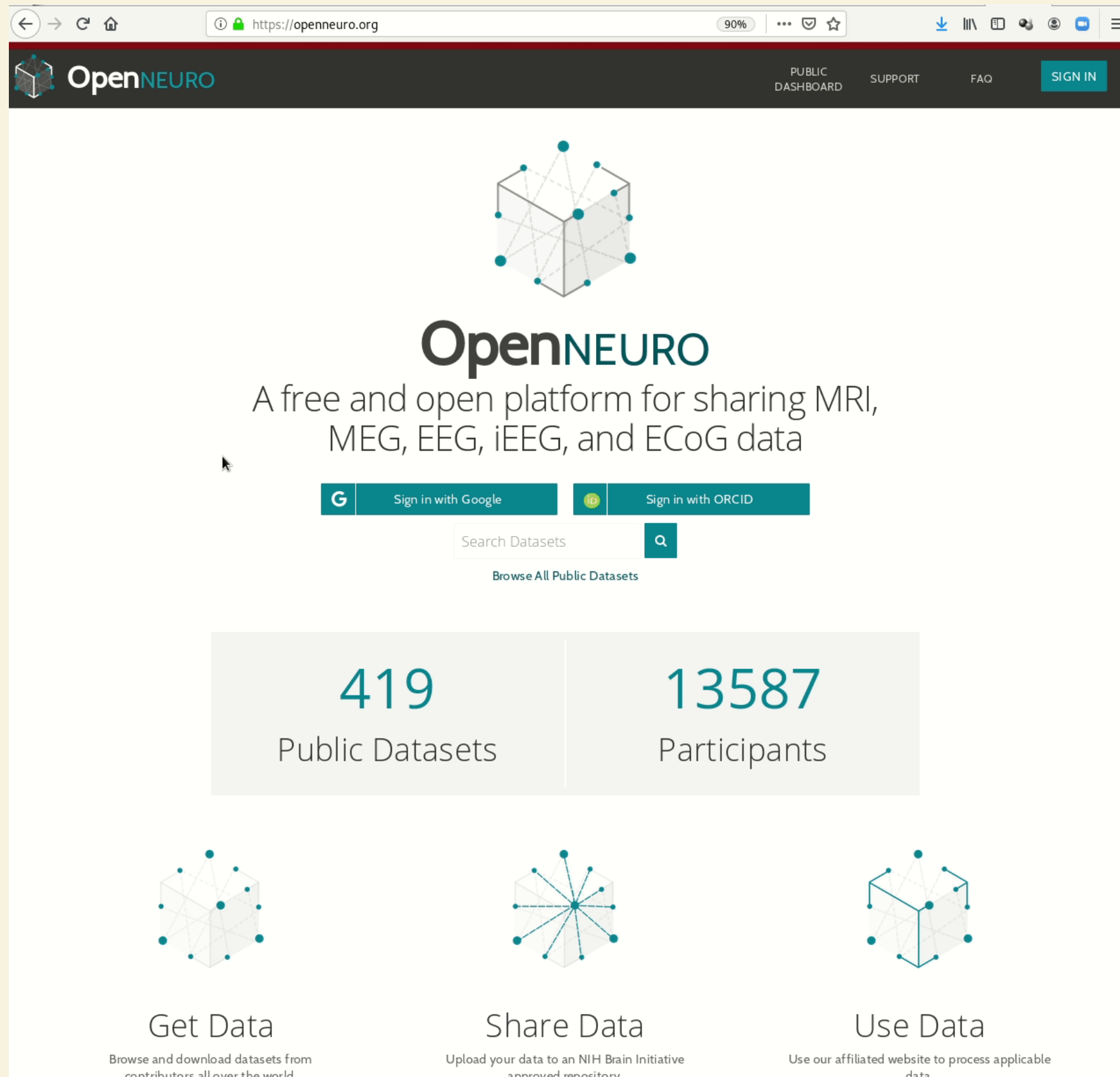
dakot Daniel Kottke

adswa Adina Wagner

Languages

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Behind-the-scenes infrastructure component for data transport and versioning



The screenshot shows the OpenNEURO website interface. At the top, there is a navigation bar with the OpenNEURO logo, a 'PUBLIC DASHBOARD' link, 'SUPPORT' and 'FAQ' links, and a 'SIGN IN' button. The main content area features a large network diagram icon, the OpenNEURO logo, and a tagline: 'A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data'. Below this are two sign-in buttons: 'Sign in with Google' and 'Sign in with ORCID'. A search bar labeled 'Search Datasets' is present, along with a 'Browse All Public Datasets' link. Two large statistics are displayed: '419 Public Datasets' and '13587 Participants'. At the bottom, three main actions are highlighted with icons and text: 'Get Data' (Browse and download datasets from contributors all over the world.), 'Share Data' (Upload your data to an NIH Brain Initiative approved repository.), and 'Use Data' (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 Twitter thread on a mobile web interface. The main tweet is from Lennart Wittkuhn (@lnnrtwttkhn) dated March 19, 2021. It discusses a new fMRI analysis method and includes a link to a paper on Nature. A reply from the same user explains that all code and data are shared via @gnode and #GitHub, version-controlled with @datalad (ca. 1.5 TB). The reply lists various data types: MRI in @BIDSstandard, #fMRIPrep data, #MRIQC metrics, GLMs + anatomical masks, task code, decoding pipeline, and statistical analyses. A link to wittkuhn.mpib.berlin is provided. The thread also includes a link to a project website for the paper 'Faster than thought: Detecting sub-second activation ...'.

Thread

Lennart Wittkuhn @lnnrtwttkhn · 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...](https://www.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](https://www.nature.com)

4 replies · 93 retweets · 270 likes

Lennart Wittkuhn @lnnrtwttkhn
Antwort an @lnnrtwttkhn

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]

Dynamics of fMRI patterns reflect sub-second activation se...
This is the project website of the accompanying the 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 @lnnrtwttkhn · 19. März
Antwort an @lnnrtwttkhn

Relevante Personen

- Lennart Wit...** @l... Folgt Dir Folge ich
PhD candidate @mpib_berlin and @MPC_CompPsych interested in hippocampal replay, decision-making and open science tools
- INCF G-Node** @gnode Folge ich
The German Neuroinformatics Node
- DataLad** @datalad Folge ich
There was Debian. git came, followed by git-annex. DataLad was born to be a data distribution, but grew into a distributed Research Data Management solution.

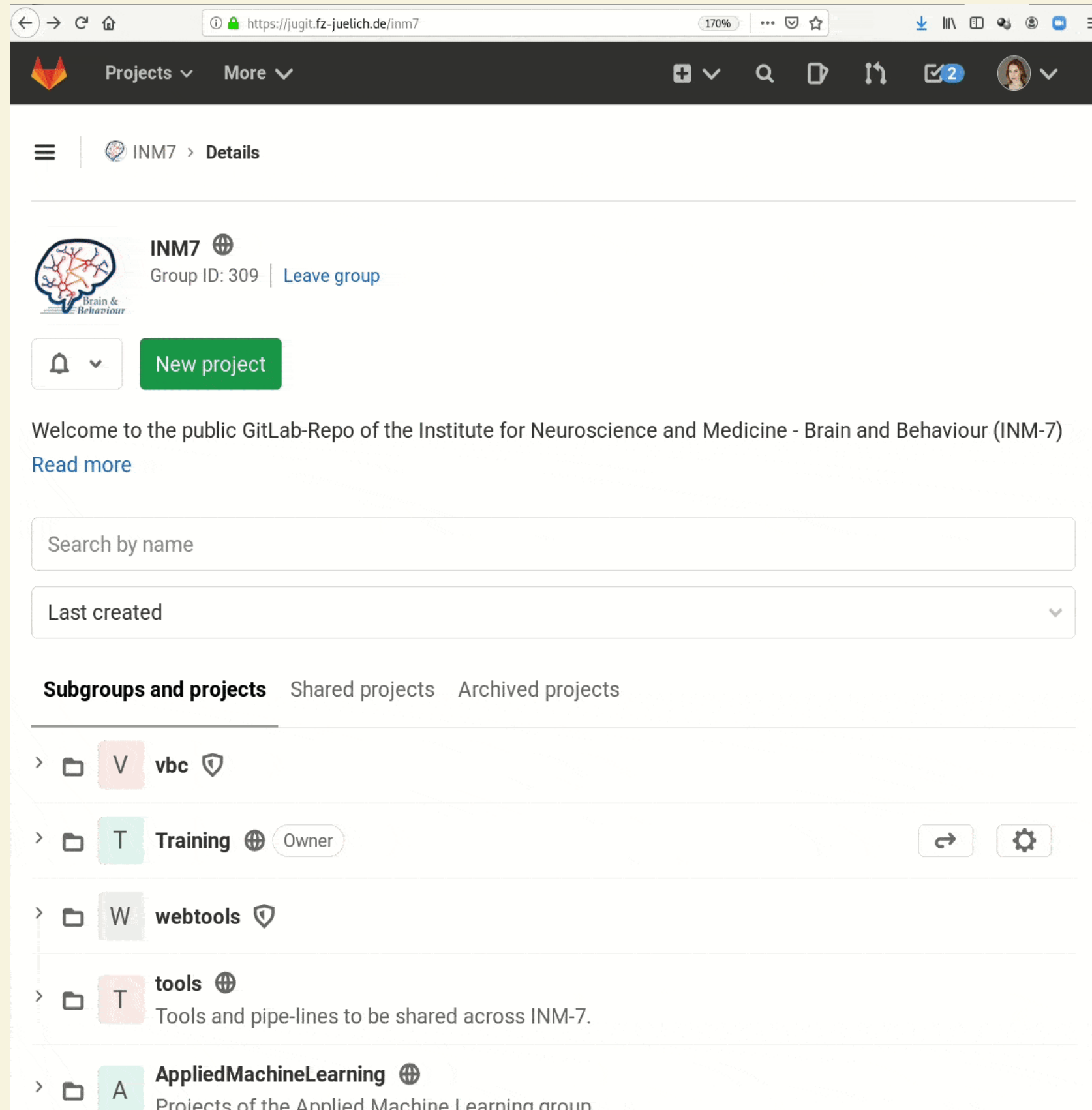
Trends für dich

- Regierung · Trends **#Merkel** 47.000 Tweets
- Trend in Deutschland **#Generalstreik** 2.678 Tweets
- Regierung · Trends **#Laschet** 4.584 Tweets
- Trend in Deutschland **#AliAkbar** 27.500 Tweets
- Trend in Deutschland **#Tanzverbot**

[Mehr anzeigen](#)

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Central data management and archival system



The screenshot shows a web browser window displaying the GitLab repository page for the INM7 group. The browser's address bar shows the URL <https://jugit.fz-juelich.de/inm7>. The page header includes navigation links for 'Projects' and 'More', along with search and user profile icons. The main content area features the INM7 group logo, name, and ID (309), with a 'Leave group' link. A 'New project' button is prominently displayed. Below this, a welcome message states: 'Welcome to the public GitLab-Repo of the Institute for Neuroscience and Medicine - Brain and Behaviour (INM-7)'. A search bar labeled 'Search by name' and a dropdown menu for 'Last created' are provided for filtering projects. The 'Subgroups and projects' section is active, showing a list of subgroups and projects:

- vbc** (Private)
- Training** (Public, Owner)
- webtools** (Private)
- tools** (Public) - Tools and pipe-lines to be shared across INM-7.
- AppliedMachineLearning** (Public) - Projects of the Applied Machine Learning group.

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Reproducible computation at the largest scale
FAIRly big: A framework for computationally reproducible
processing of large-scale data
(doi.org/10.1101/2021.10.12.464122)

FURTHER INFORMATION

- User documentation & tutorials: handbook.datalad.org
- Source code, issue tracker: github.com/datalad/datalad
- Technical docs: docs.datalad.org
- Video tutorials: www.youtube.com/datalad
- User support: [DataLad Matrix Channel](#)
- "DataLad Office Hour" (weekly): [DataLad Office Hour Matrix Channel](#)
- DataLad Paper: doi.org/10.21105/joss.03262

Use it on Hilbert:

```
module load datalad
```

Install it on your own hardware: handbook.datalad.org/r.html?install

ACKNOWLEDGEMENTS

Funders



EUROPEAN UNION
European Regional Development Fund



DataLad software

- Yaroslav Halchenko
- Joey Hess (git-annex)
- Kyle Meyer
- Benjamin Poldrack
- *32 additional contributors*

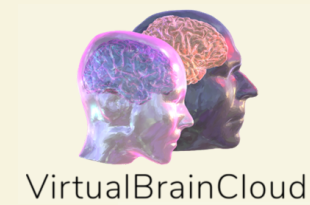
DataLad handbook

- Adina Wagner
- *41 additional contributors*

Collaborators



Human Brain Project



THANKS!

