DATA MANAGEMENT FOR OPEN SCIENCE

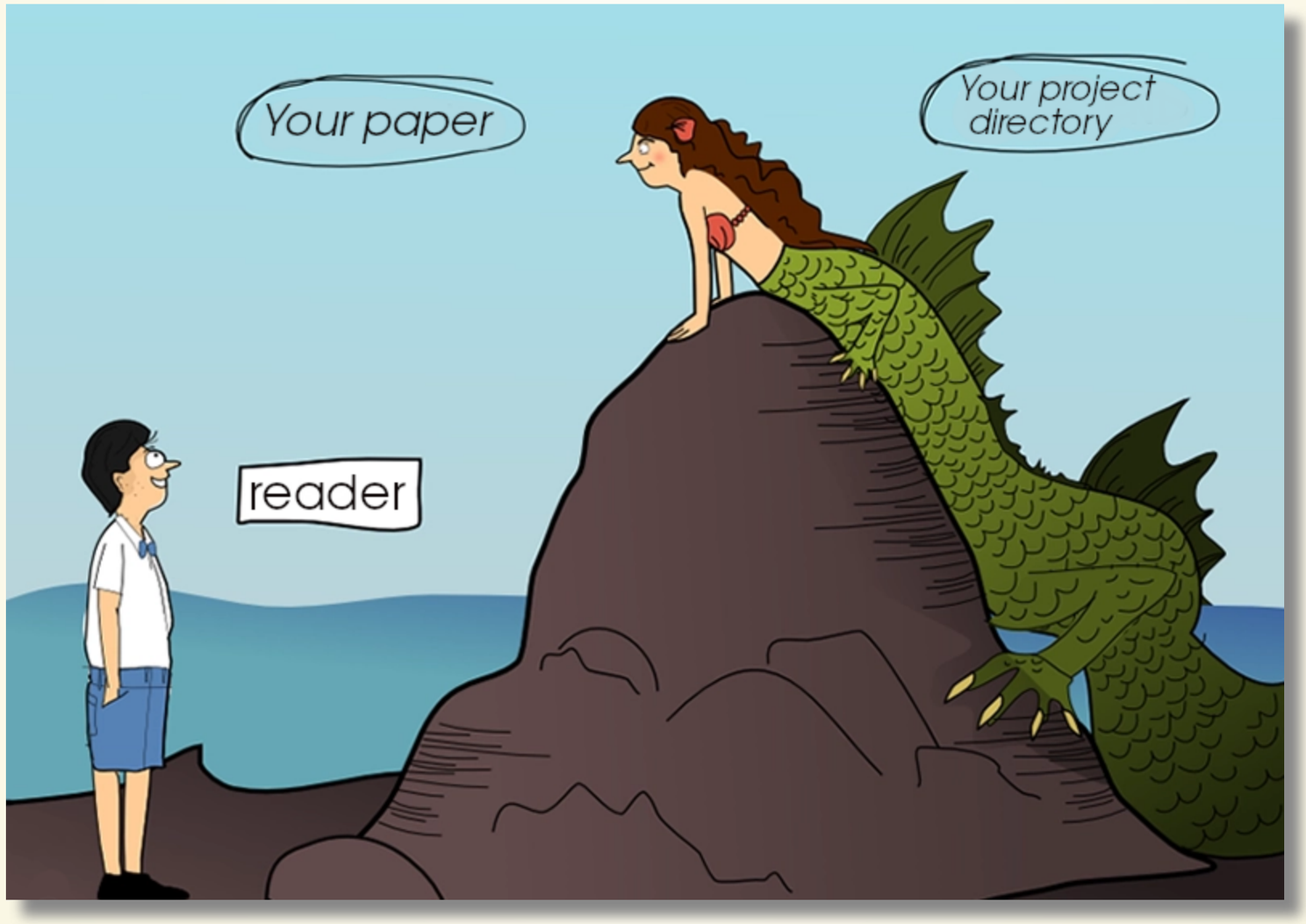
Dr. Adina Wagner
mas.to/@adswa

JÜLICH Forschungszentrum

Institute of Neuroscience and Medicine, Brain & Behavior (INM-7)
Research Center Jülich

DOI: doi.org/10.5281/zenodo.10869053
Slides: files.inm7.de/adina/talks/html/neuraltraces.html

RESEARCH DATA MANAGEMENT?



DataLad

- Domain-agnostic **command-line tool** (+ **graphical user interface**), built on top of Git & Git-annex
- 10+ year open source project (100+ contributors), available for all major OS
- Major features:

  Version-controlling arbitrarily large content

  Version control data & software alongside to code!

  Transport mechanisms for sharing, updating & obtaining data

  Consume & collaborate on data (analyses) like software

  (Computationally) reproducible data analysis

  Track and share provenance of all digital objects

  (... and much more)

The building blocks of a scientific result are rarely static

Analysis code, manuscripts, ...

evolve

(Rewrite, fix bugs, add functions, refactor, extend, ...)

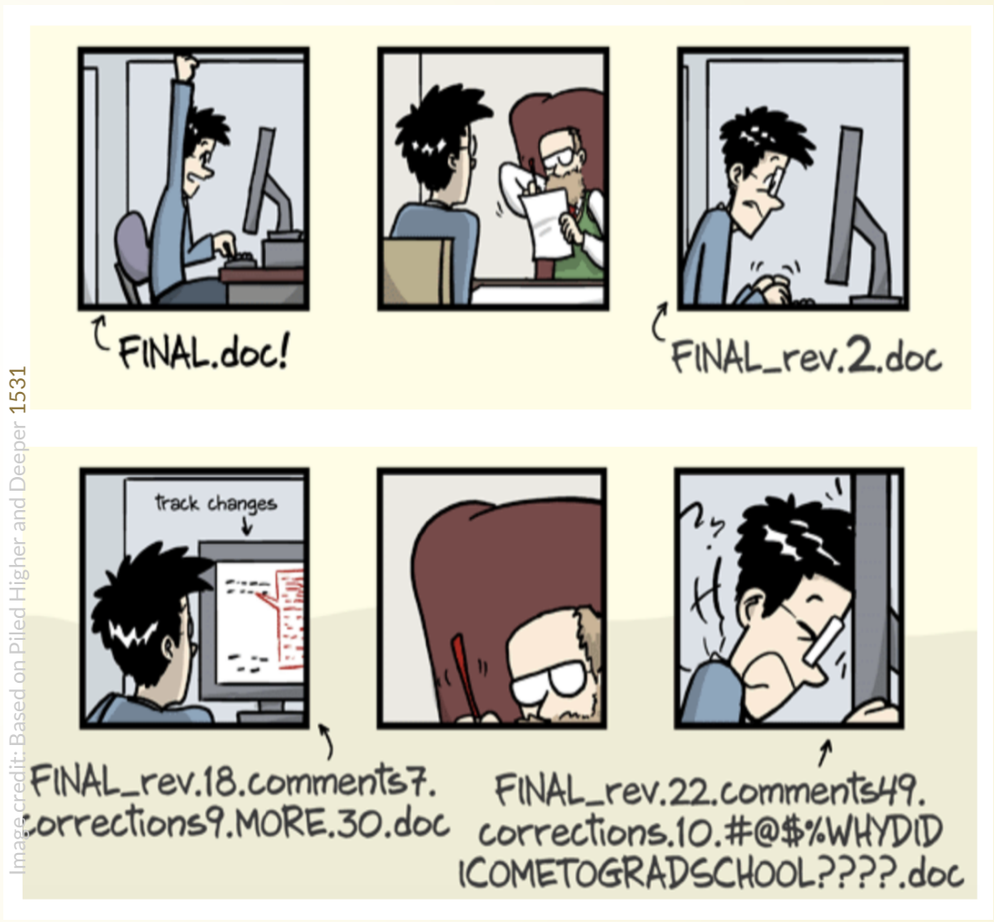


Image credit: Based on Piled Higher and Deeper 1531

The building blocks of a scientific result are rarely static

Data changes, too

(errors are fixed, data is extended, naming standards change, an analysis requires only a subset of your data...)

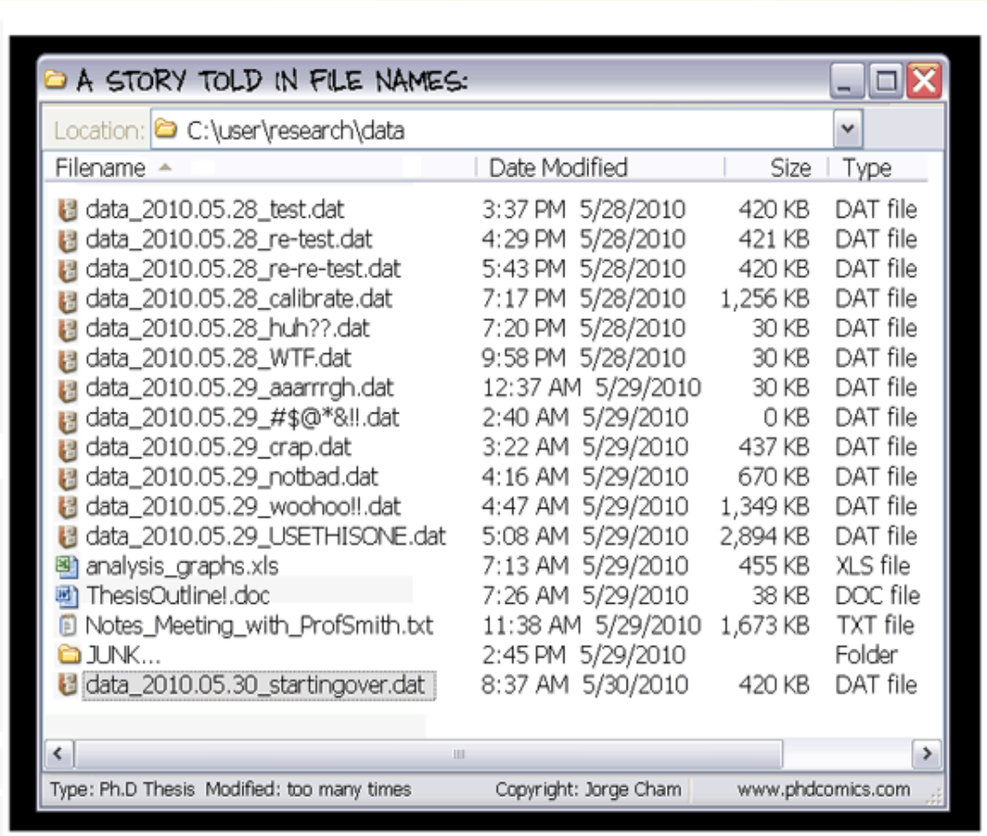


Image credit: Piled Higher and Deeper 1323

VERSION CONTROL - BEYOND TEXT FILES

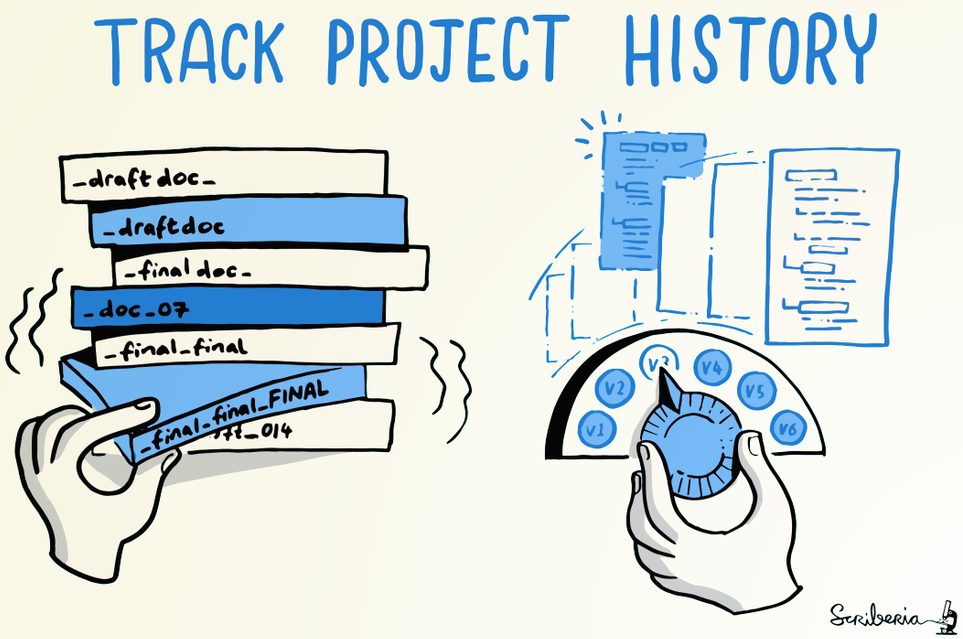


Image credit: CC-BY Scriberia & The Turing Way

- keep things organized
- keep track of changes
- revert changes or go back to previous states
- collect and share digital provenance
- industry standard: Git

Building up on Git and git-annex, DataLad version controls **any** data

```
2020-03-13 10:46 +0100 Adina Wagner     o [DATALAD RUNCMD] add non-defaced
2020-03-13 10:29 +0100 Adina Wagner     o [DATALAD RUNCMD] reconvert DICOM
2018-05-11 09:23 +0200 Michael Hanke    o [master] {origin/HEAD} {origin/m
2018-05-11 09:19 +0200 Michael Hanke    o Enable DataLad metadata extracto
2018-05-11 09:17 +0200 Michael Hanke    o [DATALAD] new dataset
2018-05-11 09:17 +0200 Michael Hanke    o [DATALAD] Set default backend fo
2018-01-19 14:19 +0100 Michael Hanke    o <v1.5> Update changelog for 1.5
2018-01-19 14:09 +0100 Michael Hanke    o BF: Re-import respiratory trace
2018-01-14 18:59 +0100 Michael Hanke    o Fix type in physio log converter
2017-01-10 10:10 +0100 Michael Hanke    o ENH: Report per-stimulus events
2016-12-10 20:18 +0100 Michael Hanke    o Add BIDS-compatible stimuli/ dir
2016-11-15 07:04 +0100 Michael Hanke    o Minor tweaks to gaze overlay scr
2016-10-30 11:03 +0100 Michael Hanke    o Add "TaskName" meta data field f
2016-09-21 08:33 +0200 Michael Hanke    o Add task-*_physio.json files
2016-09-21 08:23 +0200 Michael Hanke    o BF: Fix task label in file names
2016-08-04 13:14 +0200 Michael Hanke    o Update changelog
2016-08-03 22:22 +0200 Michael Hanke    o Add cut position information to
2016-05-27 17:35 +0200 Michael Hanke    o {origin/ } Mention openfmri as d
```

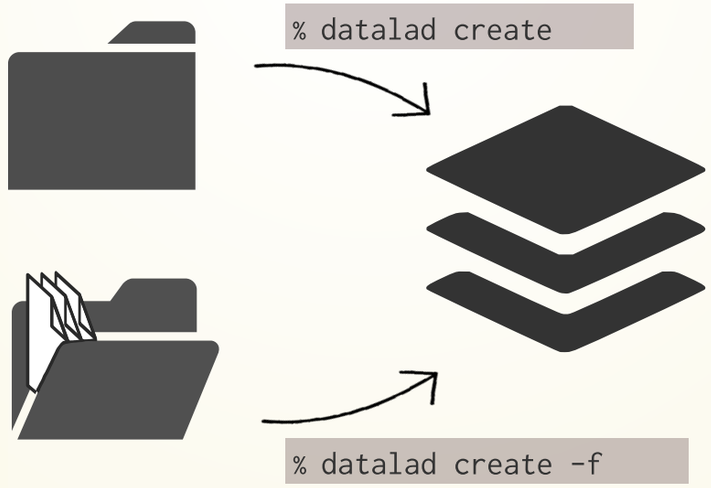
```
commit 6da25fb6fee2c698d35f52066698b6f94850f4d2
Refs: v1.0-19-g6da25fb6
Author:     Michael Hanke <michael.hanke@gmail.com>
AuthorDate: Fri Jan 19 14:09:53 2018 +0100
Commit:     Michael Hanke <michael.hanke@gmail.com>
CommitDate: Fri Jan 19 14:11:23 2018 +0100

    BF: Re-import respiratory trace after bug fix in converter (fixes gh-
---
 ...er_task-movielocalizer_run-1_recording-cardresp_physio.tsv.gz | 2 +-
 ..._task-objectcategories_run-1_recording-cardresp_physio.tsv.gz | 2 +-
 ..._task-objectcategories_run-2_recording-cardresp_physio.tsv.gz | 2 +-
 ..._task-objectcategories_run-3_recording-cardresp_physio.tsv.gz | 2 +-
 ..._task-objectcategories_run-4_recording-cardresp_physio.tsv.gz | 2 +-
 ...calizer_task-retmapccw_run-1_recording-cardresp_physio.tsv.gz | 2 +-
 ...calizer_task-retmapclw_run-1_recording-cardresp_physio.tsv.gz | 2 +-
 ...calizer_task-retmapcon_run-1_recording-cardresp_physio.tsv.gz | 2 +-
 ...calizer_task-retmapexp_run-1_recording-cardresp_physio.tsv.gz | 2 +-
```

VERSION CONTROL

- DataLad knows two things: Datasets and files
- A DataLad dataset is an Git repository:
  - Content and domain agnostic
  - Minimization of custom procedures or data structures (**user must not lose data or data access if DataLad vanishes)**
  - **Uncomprimised decentralization**

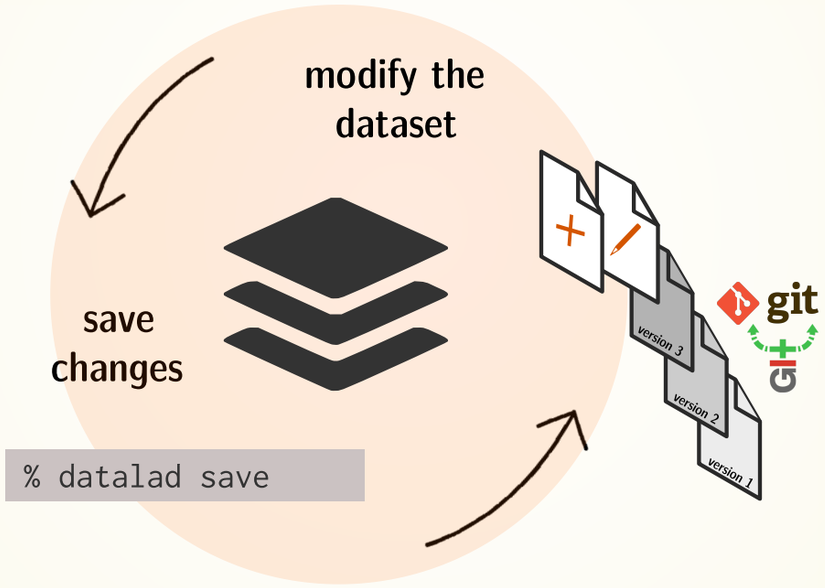
create new, empty datasets to populate...



.... or transform existing directories into datasets

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 → **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).

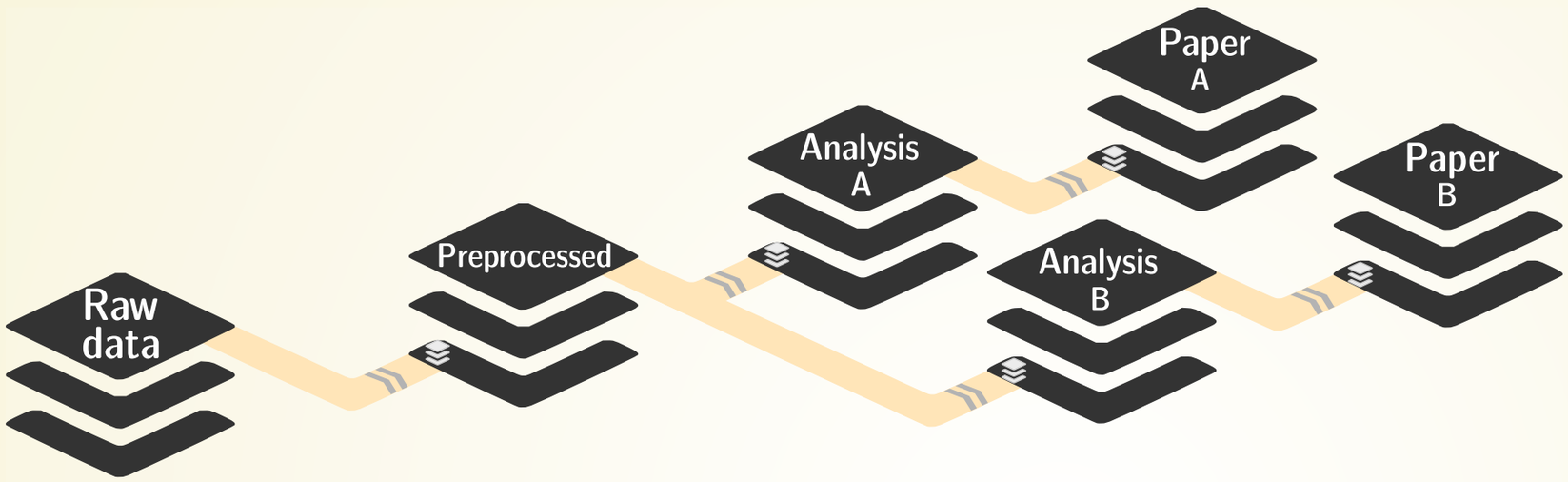


- Flexibility and commands of Git and git-annex are preserved (useful for experienced Git/git-annex users).

Delineation and advantages of decentral versus central RDM: Hanke et al., (2021). In defense of decentralized research data management

VERSION CONTROL: NESTING

- Seamless nesting mechanisms:



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
  - based on Git submodules, but more seamless
- Overcomes scaling issues with large amounts of files

```
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

USE DATALAD FOR ...

- ... self-descriptive, reusable projects

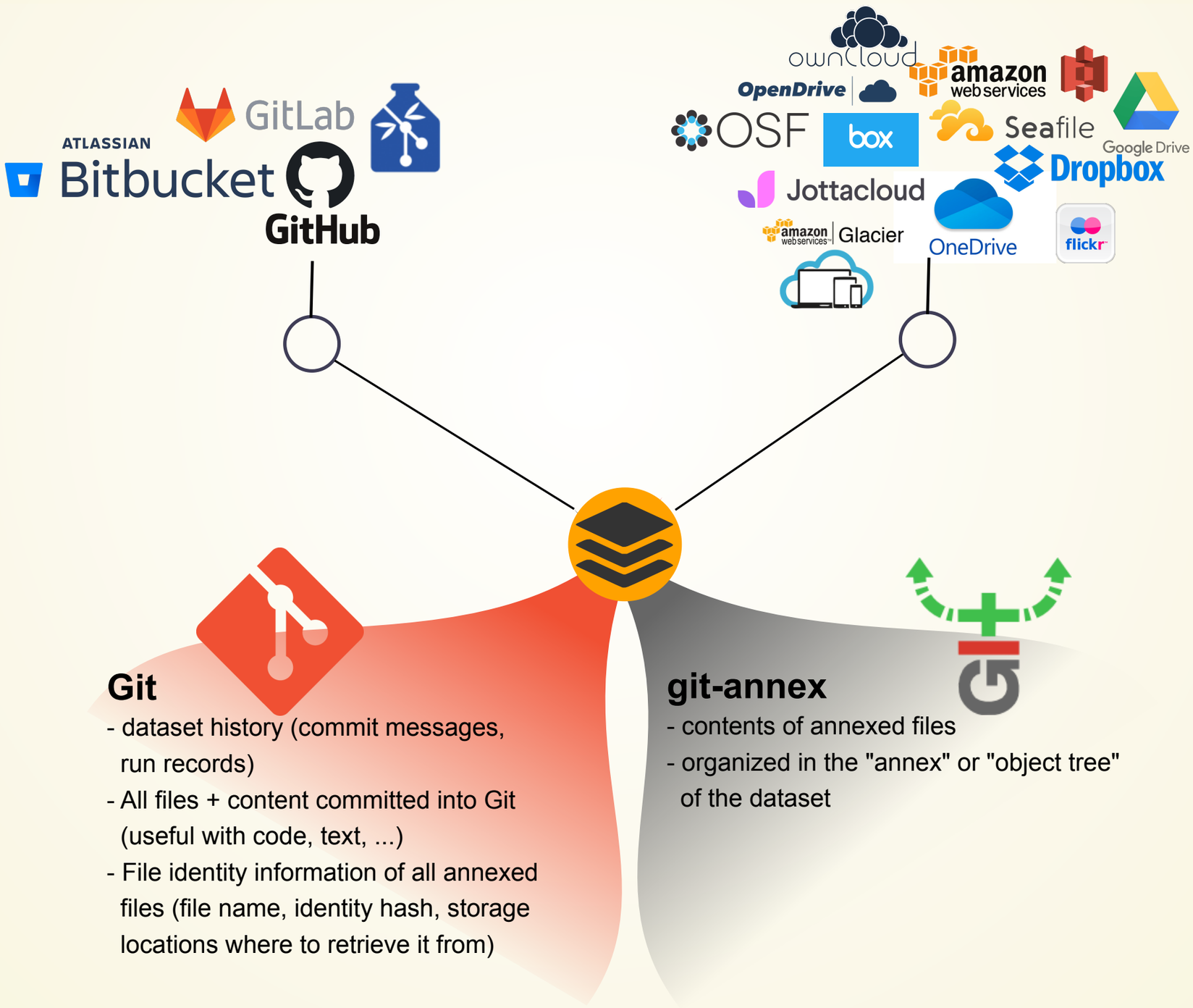
github.com/psychoinformatics-de/paper-remodnav

USE DATALAD TO ...

- ... share and consume data like source code

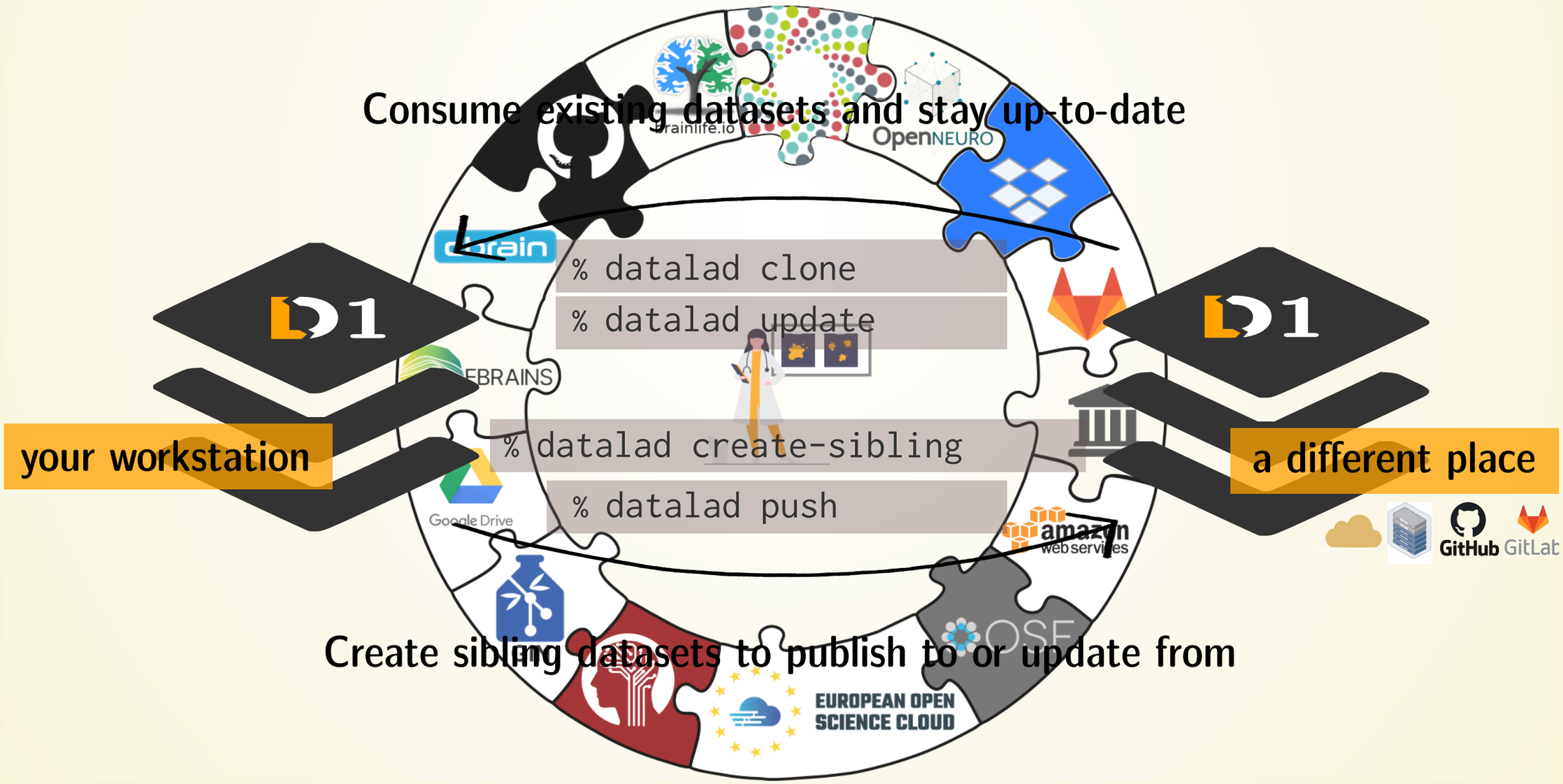
PUBLISHING DATASETS

- The best of both worlds: Publish to Git hosting services, storage providers, or both



TRANSPORT LOGISTICS

- Scientific workflows can be idiosyncratic across institutions / departments / labs / any two scientists
- DataLad is built to maximize interoperability and streamline routines across hosting and storage technology



LOTS OF DATA, LITTLE DISK-USAGE

- Cloned datasets are lean. "Meta data" (file names, availability) are present, but **no file content**:

```
$ datalad clone git@github.com:psychoinformatics-de/studyforrest-data-phase2.git
install(ok): /tmp/studyforrest-data-phase2 (dataset)
$ cd studyforrest-data-phase2 && du -sh
18M .
```

- files' contents can be retrieved on demand - and also dropped:

```
$ datalad get sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz          copy
get(ok): /tmp/studyforrest-data-phase2/sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_b
```

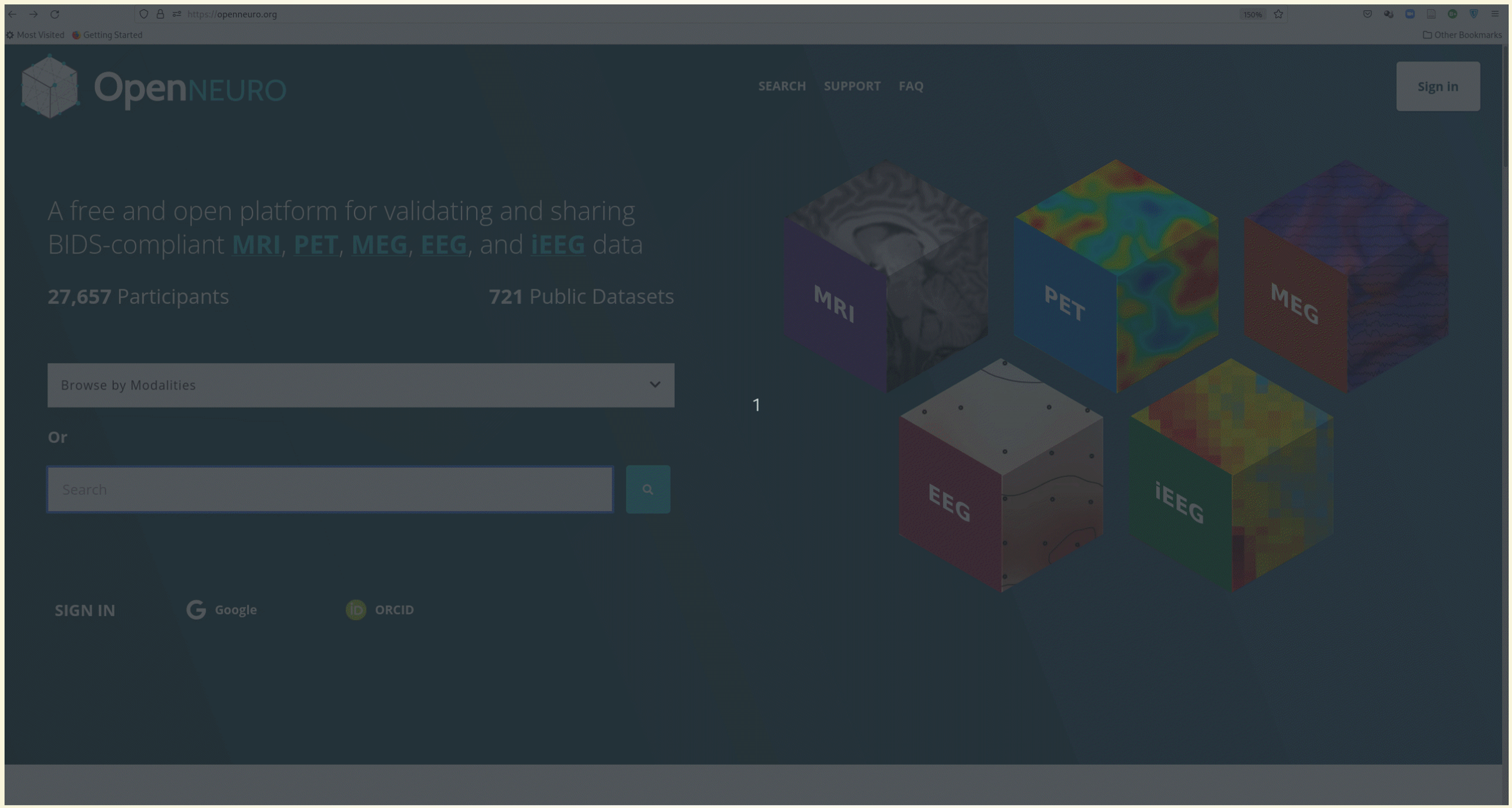
```
$ datalad drop sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz          copy
drop(ok): /tmp/studyforrest-data-phase2/sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_
```

- Have access to more data on your computer than you have disk-space:

```
# eNKI dataset (1.5TB, 34k files):                                                  copy
$ du -sh
1.5G  .
# HCP dataset (~200TB, >15 million files)
$ du -sh
48G .
```

```
dl.get('input/sub-01')                                                              copy
[really complex analysis]
dl.drop('input/sub-01')
```

HAVE YOURSELF SOME DATA



> 500TB of open data available at datasets.datalad.org

REUSING PAST WORK

... isn't necessarily simple

Your past self is the worst collaborator:

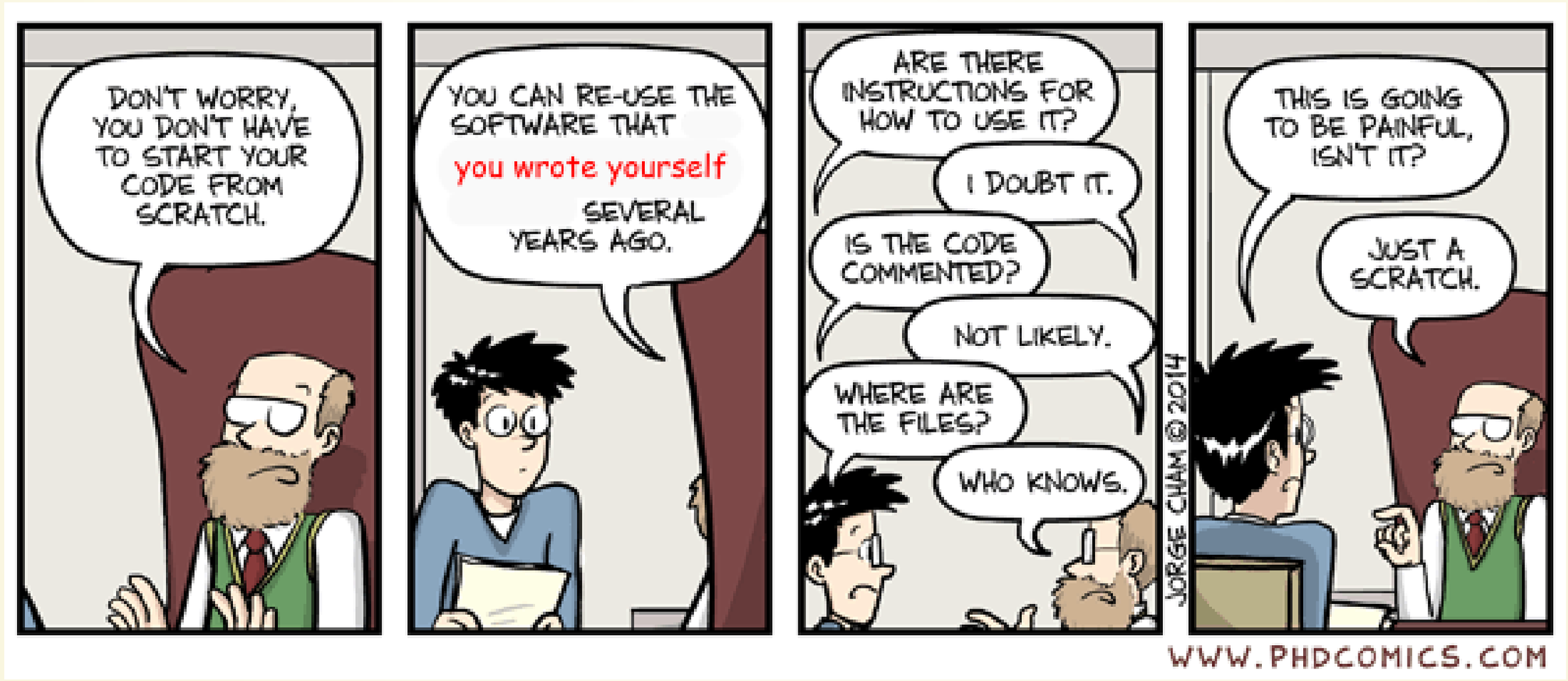
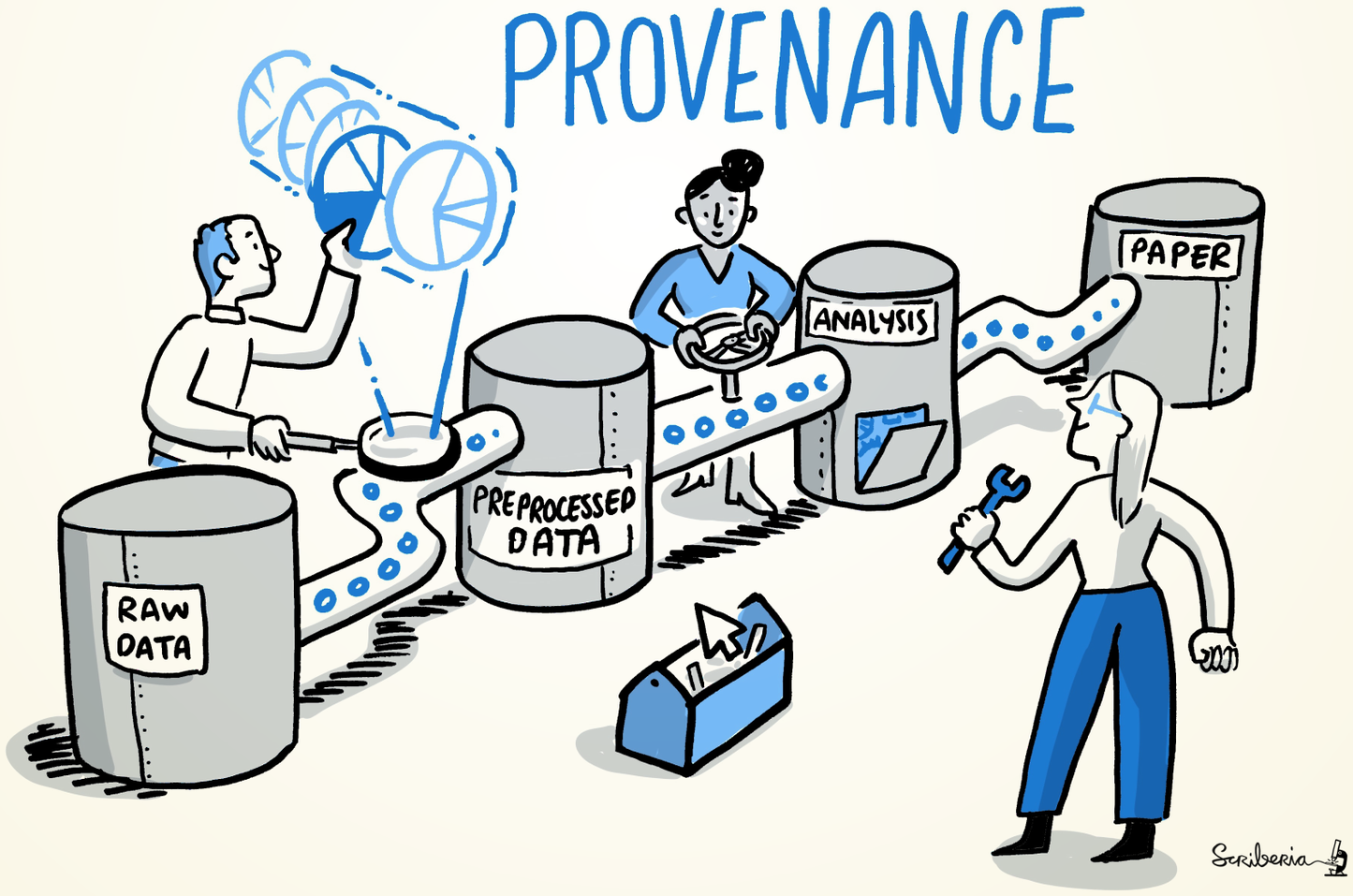


Image credit: Full comic at http://phdcomics.com/comics.php?f=1979

LACK OF PROVENANCE CAN BE DEVASTATING

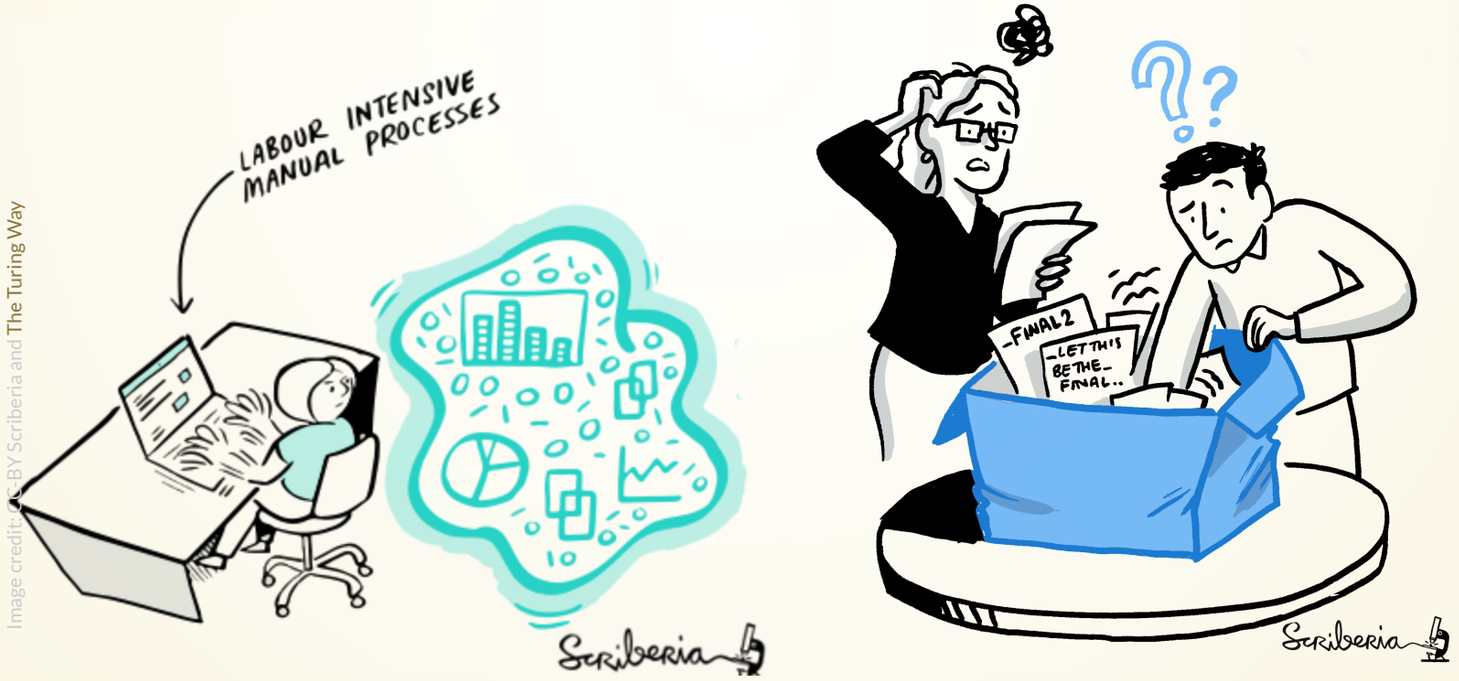
- Data analyses or data wrangling is complex
  - Move/Copy/Rename/Reorganize/Transform/Compute/... data
- Mistakes propagate through the complete analysis pipeline - especially those early ones are hard to find!



LEAVING A TRACE

"Shit, which version of which script produced these outputs from which version of what data?"

"Shit, why buttons did I click and in which order did I use all those tools?"

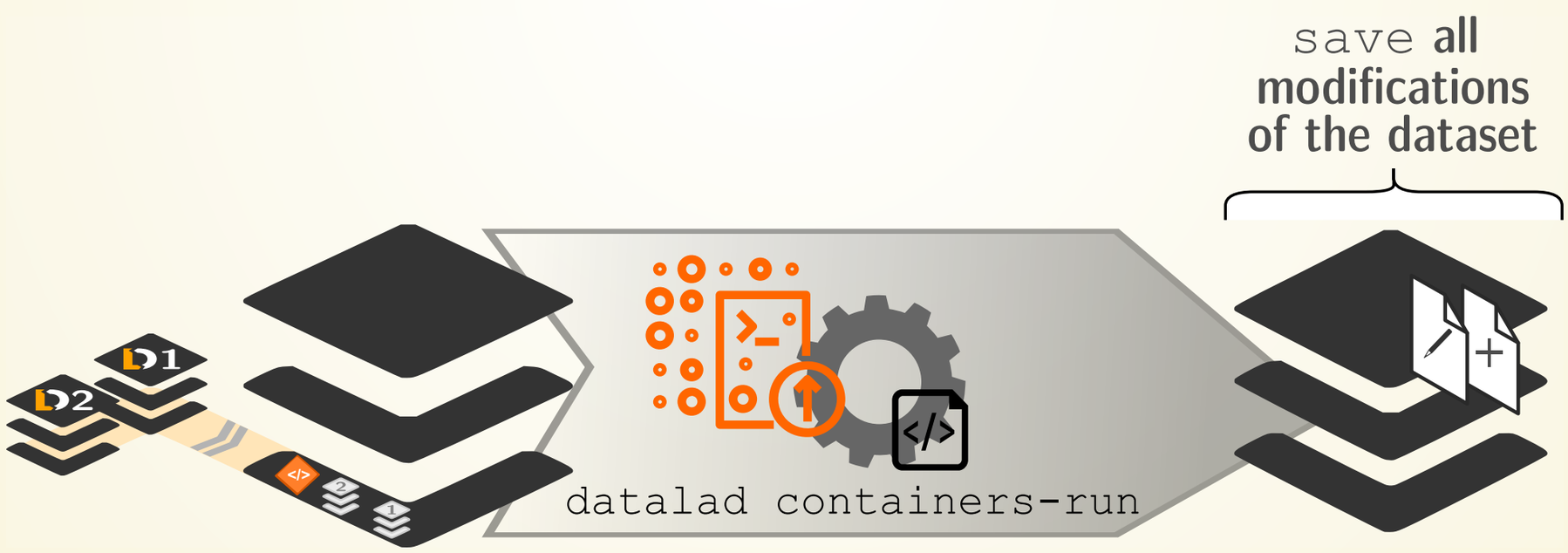


LEAVING A TRACE

datalad run wraps around anything expressed in a command line call and saves the dataset modifications resulting from the execution.

datalad rerun repeats captured executions. If the outcomes differ, it saves a new state of them.

datalad containers-run executes command line calls inside a tracked software container and saves the dataset modifications resulting from the execution.



DATA ANALYSIS PROVENANCE

```
$ datalad containers-run \
  --message "Time series extraction from Locus Coeruleus"
  --container-name nilearn \
  --input 'mri/*_bold.nii' \
  --output 'sub-*/LC_timeseries_run-*.csv' \
  "python3 code/extract_lc_timeseries.py"

-- Git commit --
    commit 5a7565a640ff6de67e07292a26bf272f1ee4b00e
    Author:     Adina Wagner adina.wagner@t-online.de
    AuthorDate: Mon Nov 11 16:15:08 2019 +0100
    Commit:     Adina Wagner adina.wagner@t-online.de
    CommitDate: Mon Nov 11 16:15:08 2019 +0100

    [DATALAD RUNCMD] Time series extraction from Locus Coeruleus
    === Do not change lines below ===
    {
     "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..",
     "dsid": "92ea1faa-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 +
...
```

DATA ANALYSIS PROVENANCE

```
$ datalad containers-run \
  --message "Time series extraction from Locus Coeruleus"
  --container-name nilearn \
  --input 'mri/*_bold.nii' \
  --output 'sub-*/LC_timeseries_run-*.csv' \
  "python3 code/extract_lc_timeseries.py"

-- Git commit --
    commit 5a7565a640ff6de67e07292a26bf272f1ee4b00e
    Author:     Adina Wagner adina.wagner@t-online.de
    AuthorDate: Mon Nov 11 16:15:08 2019 +0100
    Commit:     Adina Wagner adina.wagner@t-online.de
    CommitDate: Mon Nov 11 16:15:08 2019 +0100

    [DATALAD RUNCMD] Time series extraction from Locus Coeruleus
    === Do not change lines below ===
    {
     "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..",
     "dsid": "92ea1faa-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 +
...
```

DATA ANALYSIS PROVENANCE

```
$ datalad containers-run \
  --message "Time series extraction from Locus Coeruleus"
  --container-name nilearn \
  --input 'mri/*_bold.nii' \
  --output 'sub-*/LC_timeseries_run-*.csv' \
  "python3 code/extract_lc_timeseries.py"

-- Git commit --
    commit 5a7565a640ff6de67e07292a26bf272f1ee4b00e
    Author:     Adina Wagner adina.wagner@t-online.de
    AuthorDate: Mon Nov 11 16:15:08 2019 +0100
    Commit:     Adina Wagner adina.wagner@t-online.de
    CommitDate: Mon Nov 11 16:15:08 2019 +0100

    [DATALAD RUNCMD] Time series extraction from Locus Coeruleus
    === Do not change lines below ===
    {
     "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..",
     "dsid": "92ea1faa-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 +
...
```

DATA ANALYSIS PROVENANCE

```
$ datalad rerun 5a7565a640ff6de67
[INFO   ] run commit 5a7565a640ff6de67; (Time series extraction from Locus Coeruleus)
[INFO   ] Making sure inputs are available (this may take some time)
get(ok): mri/sub-01_bold.nii (file)
get(ok): mri/sub-02_bold.nii (file)
        [...]
[INFO   ] == Command start (output follows) =====
[INFO   ] == Command exit (modification check follows) =====
add(ok): sub-01/LC_timeseries_run-*.csv(file)
add(ok): sub-02/LC_timeseries_run-*.csv (file)
        [...]
action summary:
  add (ok: 30)
  get (ok: 30)
  save (ok: 2)
  unlock (ok: 30)
```

DATA ANALYSIS PROVENANCE

```
$ datalad rerun 5a7565a640ff6de67
[INFO   ] run commit 5a7565a640ff6de67; (Time series extraction from Locus Coeruleus)
[INFO   ] Making sure inputs are available (this may take some time)
get(ok): mri/sub-01_bold.nii (file)
get(ok): mri/sub-02_bold.nii (file)
        [...]
[INFO   ] == Command start (output follows) =====
[INFO   ] == Command exit (modification check follows) =====
add(ok): sub-01/LC_timeseries_run-*.csv(file)
add(ok): sub-02/LC_timeseries_run-*.csv (file)
        [...]
action summary:
  add (ok: 30)
  get (ok: 30)
  save (ok: 2)
  unlock (ok: 30)
```

PROVENANCE AT THE LARGEST SCALE:

RESEARCH DATA MANAGEMENT IS TIED TO REPRODUCIBILITY

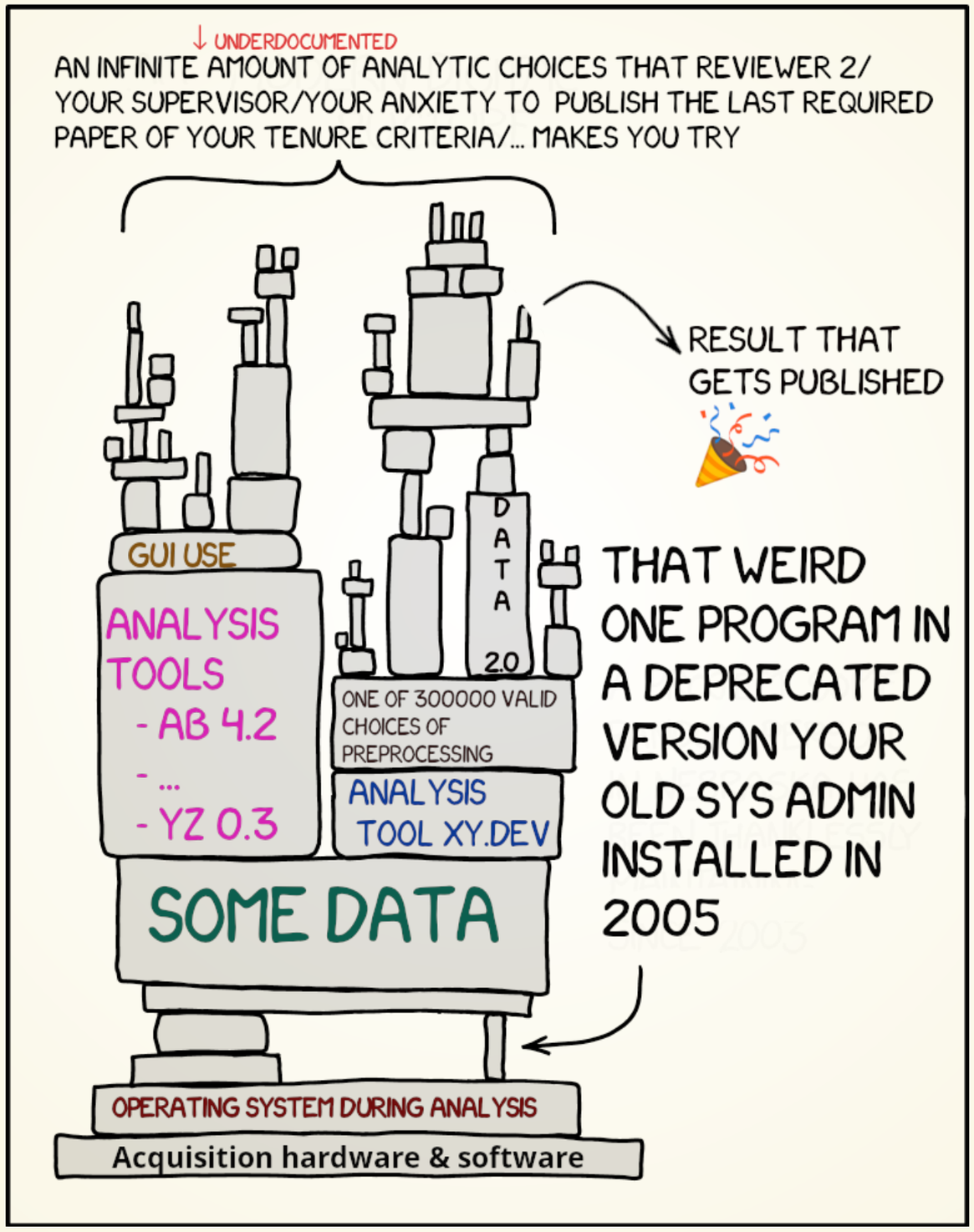


Image credit: Based on xkcd.com/2347/ (CC-BY)

Reproducibility Management in Neuroscience - Specific Issues and Solutions (DOI 10.5281/zenodo.4285927)

FURTHER INFORMATION

Reach out to to the DataLad team via

- Matrix (free, decentralized communication app, no app needed). We run a weekly Zoom office hour (Tuesday, 4pm Berlin time) from this room as well.
- the development repository on GitHub (github.com/datalad/datalad)

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 (www.youtube.com/channel/datalad)
- In the DataLad Handbook (handbook.datalad.org)
- In the DataLad RDM course (psychoinformatics-de.github.io/rdm-course)
- In the Official API documentation (docs.datalad.org)

- On the advantages of decentralized research data management: doi.org/10.1515/nf-2020-0037

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

ACKNOWLEDGEMENTS

Software

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

Funders

NSF 1912266

germany-usa

SPONSORED BY THE
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

THANKS!

(scan the QR code for slides)

Collaborators

Human Brain Project

CONP PCNO

VirtualBrainCloud

OpenNEURO

cbrain

brainlife.io