

DATALAD

DECENTRALIZED MANAGEMENT OF DIGITAL OBJECTS FOR OPEN SCIENCE

Dr. Adina Wagner

 mas.to/@adswa

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



Slides: DOI [10.5281/zenodo.15193934](https://doi.org/10.5281/zenodo.15193934) (Scan the QR code)
files.inm7.de/adina/talks/html/nhr

ACKNOWLEDGEMENTS

DataLad software & ecosystem

- Psychoinformatics Lab,
Research Centre Jülich
- Center for Open
Neuroscience,
Dartmouth College
- Joey Hess (git-annex)
- *>100 additional contributors*

Funders



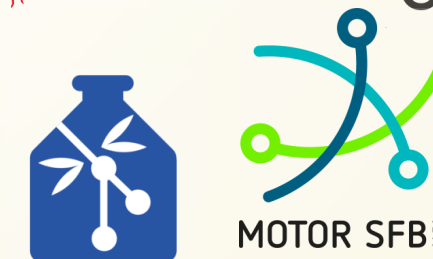
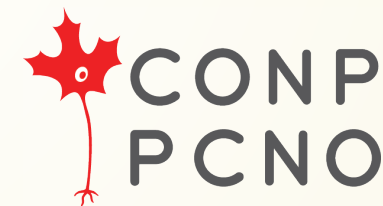
NSF 1429999

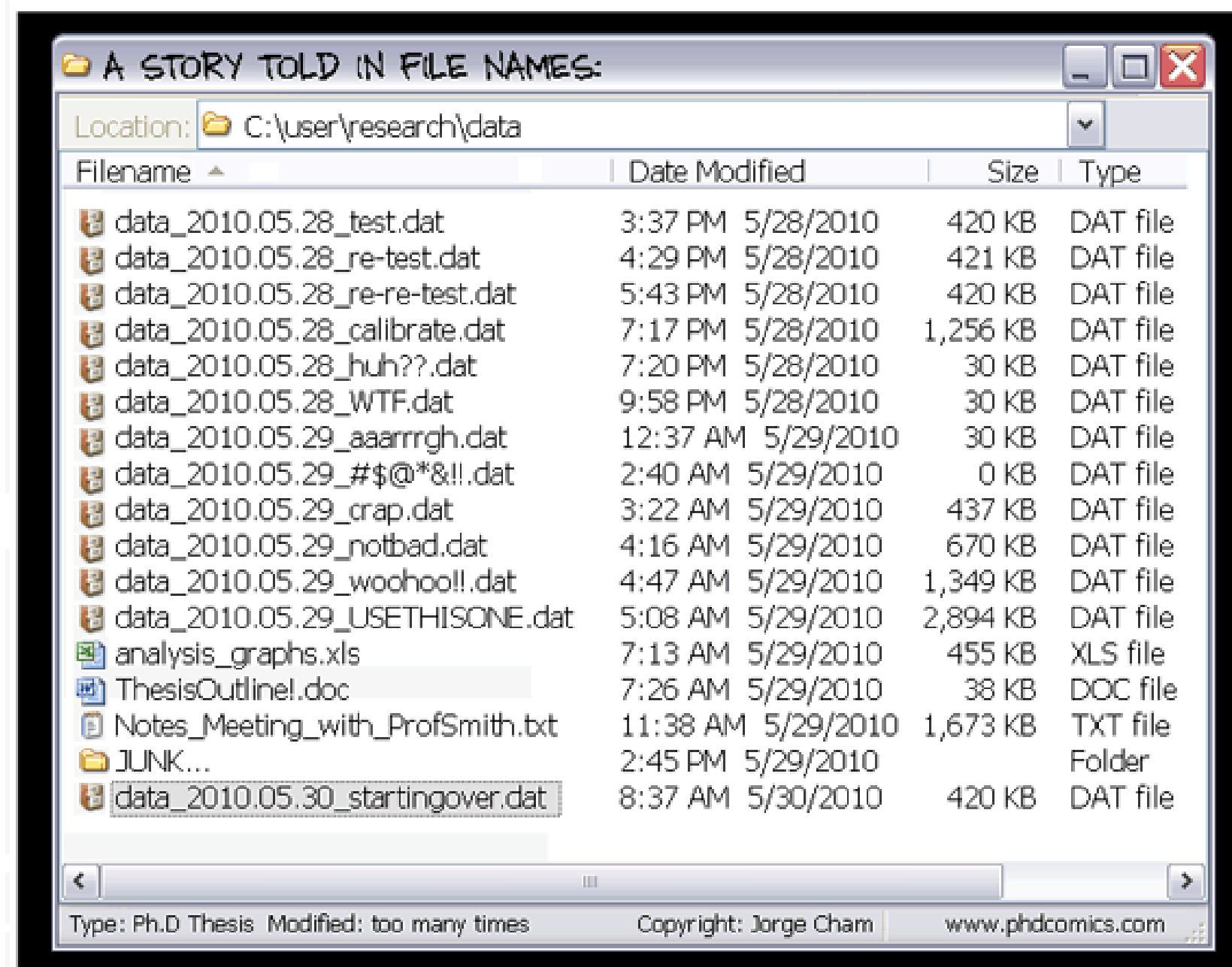


BMBF 01GQ1411

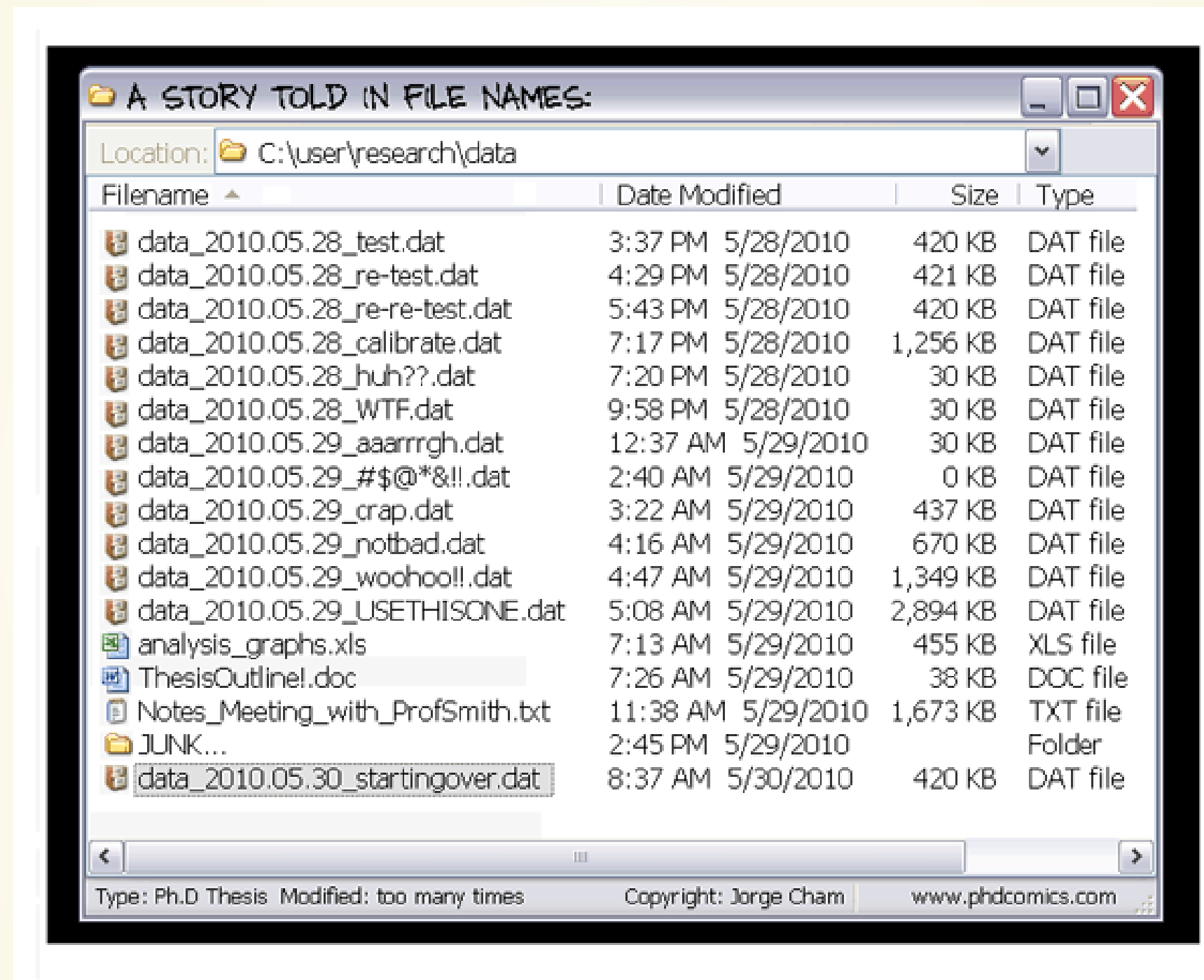


Collaborators





THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE RARELY STATIC



Data changes

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

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE RARELY STATIC

Mar 2019

ABCD Data Release 2.0 available now on the NIMH Data Archive!

The second annual curated ABCD data release 2.0 is available now on the [NIMH Data Archive](#). Data Release 2.0 includes baseline data on the full participant cohort, ages 9-10 years. This release contains much of the same type of data as in Data Release 1.1, and also includes genotypic data for the first time. Smokescreen genotyping array data are available on almost 11,000 participants. These include common variations, as well as variations associated with addiction, smoking behavior and nicotine metabolism.

ABCD Study data will be released annually. The next data release will be in early summer 2020 and will include the first longitudinal data from the 6-month and 1-year follow-up assessments.

source: abcdstudy.org/scientists/data-sharing-archive

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE RARELY STATIC

Spring 2019

Issues Identified with Data Release 2.0

Problems were identified with the data in 11 imaging-related files (5 diffusion tensor imaging (DTI) files and 6 fMRI Monetary Incentive Delay (MID) task files). See below for more detail. Corrected files will be re-uploaded to the NIMH Data Archive and available soon to users.

DTI subcortical and cortical ROIs and white fiber tracts (Part 1)

DTI subcortical and cortical ROIs and white matter fiber tracts (Part 4)

dMRI DTI Destrieux Parcellations Part 1

dMRI DTI Destrieux Parcellations Part 2

dMRI DTI Full Part 2

Task fMRI MID Average SEM Destrieux Parcellations Part 2

Task fMRI MID Run 1 Beta Weights Destrieux Parcellations Part 2

Task fMRI MID Run 2 Beta Weights Destrieux Parcellations Part 2

Task fMRI MID Average Beta Weights Destrieux Parcellations Part 2

Task fMRI MID Run 1 SEM Destrieux Parcellations Part 2

Task fMRI MID Run 2 SEM Destrieux Parcellations Part 2

An error was also discovered in imaging data collected from Siemens scanners between September 2017 and December 2017 where structural images are flipped left-right. These data will be updated in a patch release later this year.

source: abcdstudy.org/scientists/data-sharing-archive

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE RARELY STATIC

July 2019

ABCD Data Release 2.0.1 available now on the NIMH Data Archive

Due to reporting compilation and processing errors in 2.0 Data Release, a **2.0.1 Fix Release** has been issued. Please ensure curated data (datasheets, minimally processed data) from the original Data Release 2.0 are replaced with data from 2.0.1 Fix Release. The following release notes were updated to reflect these changes:

- NDA 2.0.1 Release Notes ABCD README FIRST
- NDA 2.0.1 Release Notes Imaging Instruments
- NDA 2.0.1 Changes and Known Issues Fix Release 2.0.1
- NDA 2.0.1 Diffusion Magnetic Resonance Imaging
- NDA 2.0.1 Task-Based Functional Magnetic Resonance Imaging
- NDA 2.0.1 Mental Health
- NDA 2.0.1 Genetics

source: abcdstudy.org/scientists/data-sharing-archive

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE RARELY STATIC

Dec 2019

Issues Identified with Data Release 2.0.1

Issues have been identified with ABCD Data Release 2.0.1 that impact some of the neuroimaging and genetics data. See below for details.

fMRI:

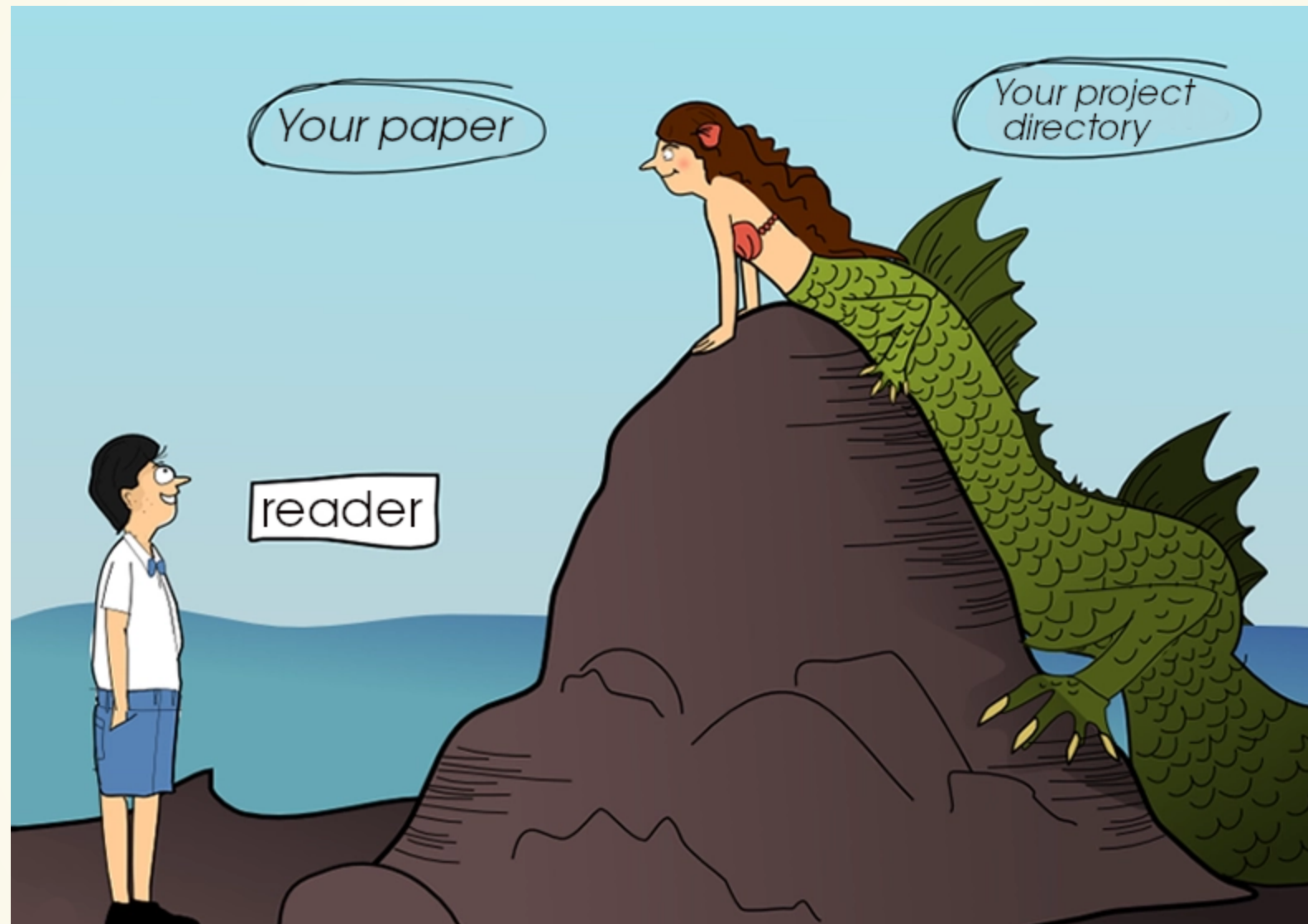
Due to incorrect post-processing, **all task and resting-state fMRI data obtained on Phillips scanners should be excluded from all analyses.** This includes tabulated and minimally processed data. This issue ***does not affect*** other modalities (sMRI, dMRI) or raw DICOM data (Fast Track). **Phillips fMRI data have been removed** from DEAP and RDS. They will be corrected and made available with **Data Release 3.0.**

Philips fMRI ***minimally processed*** and ***tabulated data*** can be excluded by using the ABCD MRI Info instrument (abcd_mri01) and exclude by: mri_info_manufacturer = "Philips Medical Systems". An R script is available at <https://github.com/ABCD-STUDY/fMRI-cleanup> to remove Philips fMRI data from tabulated data.

We recommend that all **users re-run fMRI related analyses** without the Philips fMRI data. **For published studies, we encourage users to confirm findings without the Philips fMRI data and provide a corrigendum with the updated results.**

source: abcdstudy.org/scientists/data-sharing-archive

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE RARELY STATIC



source: abcdstudy.org/scientists/data-sharing-archive



- Domain-agnostic **command-line tool** (+ graphical user interface), built on top of **Git & Git-annex**
- Open source (MIT) research software developed since 2013
- Available for all major operating systems
- Major features:
 - Version-controlling arbitrarily large content**
Version control data & software alongside to code!
 - Transport mechanisms for sharing & obtaining data**
Consume & collaborate on data (analyses) like software
 - (Computationally) reproducible data analysis**
Track and share provenance of all digital objects
 - (... and much more)**



A DataLad dataset is a joint Git/git-annex repository that can version control any file

```
# turn any directory into a dataset  
# with version control  
  
% datalad create <directory>
```

```
# save a new state of a dataset with  
# file content of any size  
  
% datalad save
```



- text document
- source code
- binary file
- regular file
- file reference:
identity and
availability

Which data (at which version), with which code, running with what parameterization in which computational environment, to generate what?

```
# execute any command and capture its output
# while recording all input versions too

% datalad run --input ... --output ... <command>
```



Decentral data transport to Git hosting, local or remote infrastructure, or external hosting services

```
# transfer data and metadata to other sites and services  
# with fine-grained access control for dataset components  
  
% datalad push --to <site-or-service>
```



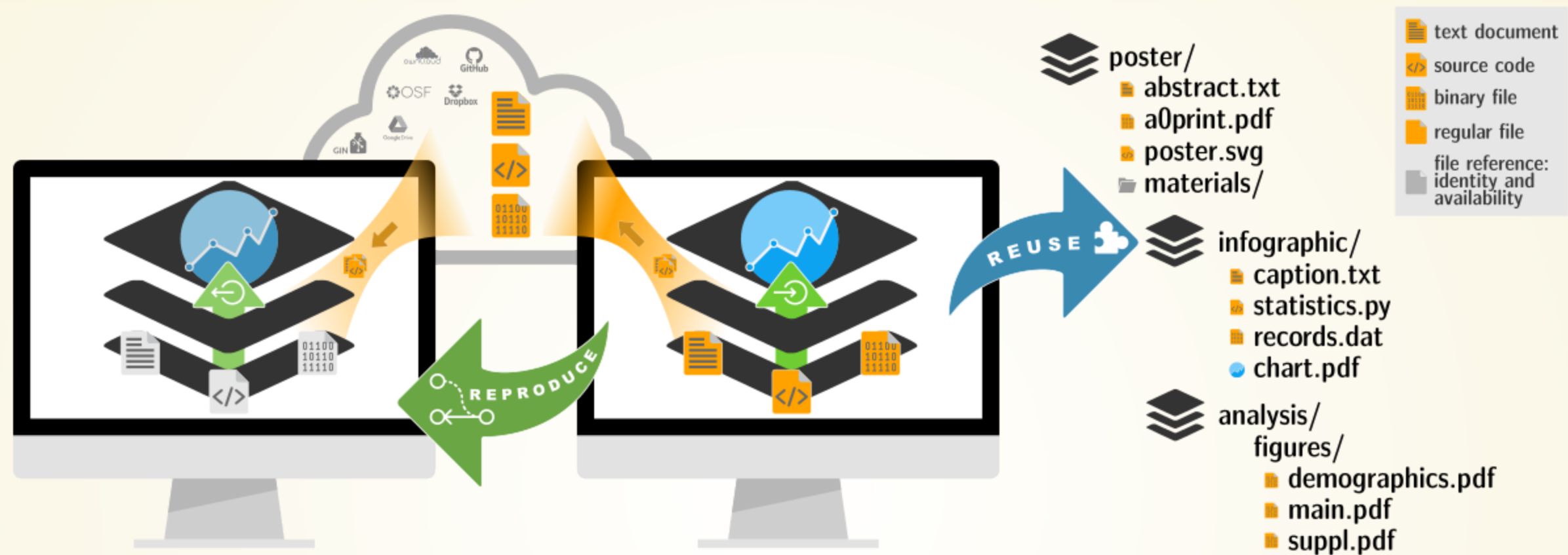
Outcomes can be validated. This enables audits, promotes accountability, and streamlines automated "upgrades" of outputs

```
# obtain dataset (initially only identity,
# availability, and provenance metadata)

% datalad clone <url>
```

```
# immediately actionable provenance records
# full abstraction of input data retrieval

% datalad rerun <commit|tag|range>
```



Datasets can be (re-)used as modular components in larger contexts — propagating their traits. They are verifiable, portable, self-contained data structures

```
# declare a dependency on another dataset and
# re-use it a particular state in a new context

% datalad clone -d <superdataset> <url> <path-in-dataset>
```

VERSION CONTROL BEYOND TEXT FILES

VERSION CONTROL BEYOND TEXT FILES



Using `git-annex`, DataLad version controls large data



VERSION CONTROL BEYOND TEXT FILES



Using git-annex, DataLad version controls large data

```
2020-03-13 10:46 +0100 Adina Wagner    o [DATALAD RUNCMD] add non-defaced anatomical images
2020-03-13 10:29 +0100 Adina Wagner    o [DATALAD RUNCMD] reconvert DICOMs without defacing
2018-05-11 09:23 +0200 Michael Hanke   o [master] {origin/HEAD} {origin/master} {origin/synced/master} [DATALAD] dataset aggregate metadata update
2018-05-11 09:19 +0200 Michael Hanke   o Enable DataLad metadata extractors
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 for all files to be MD5E
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 after bug fix in converter (fixes gh-11)
2018-01-14 18:59 +0100 Michael Hanke   o Fix type in physio log converter (fixes gh-11)
2017-01-10 10:10 +0100 Michael Hanke   o BF: Report per-stimulus events (fixes gh-6)
2016-12-10 20:18 +0100 Michael Hanke   o Add BIDS-compatible stimuli/ directory (with symlinks)
2016-11-15 07:04 +0100 Michael Hanke   o Minor tweaks to gaze overlay script
2016-10-30 11:03 +0100 Michael Hanke   o Add "TaskName" meta data field for compliance with BIDS 1.0.0
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 of contracting retmap run.
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 allow for timing verification of generated stimulus files
2016-05-27 17:35 +0200 Michael Hanke   o {origin/_} Mention openfmri as download source
2016-04-04 09:31 +0200 Michael Hanke   o Update publication links
2016-03-31 11:26 +0200 Michael Hanke   o Disable invalid test
[main] d93ed6e3b1a00f1dba18b603e668b7a060cdcf77 - commit 48 of 79
```

27%

VERSION CONTROL BEYOND TEXT FILES



Using git-annex, DataLad version controls large data

```
2020-03-13 10:46 +0100 Adina Wagner    o [DATALAD RUNCMD] add non-defaced anatomical images
2020-03-13 10:29 +0100 Adina Wagner    o [DATALAD RUNCMD] reconvert DICOMs without defacing
2018-05-11 09:23 +0200 Michael Hanke   o [master] {origin/HEAD} {origin/master} {origin/synced/master} [DATALAD] dataset aggregate metadata update
2018-05-11 09:19 +0200 Michael Hanke   o Enable DataLad metadata extractors
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 for all files to be MD5E
2018-01-19 14:19 +0100 Michael Hanke   o <v1.0> Update changelog for 1.5
2018-01-19 14:09 +0100 Michael Hanke   o BF: Re-import respiratory trace after bug fix in converter (fixes gh-11)
2018-01-14 18:59 +0100 Michael Hanke   o Fix type in physio log converter (fixes gh-11)
2017-01-10 10:10 +0100 Michael Hanke   o BF: Report per-stimulus events (fixes gh-6)
2016-12-10 20:18 +0100 Michael Hanke   o Add BIDS-compatible stimuli/ directory (with symlinks)
2016-11-15 07:04 +0100 Michael Hanke   o Minor tweaks to gaze overlay script
2016-10-30 11:03 +0100 Michael Hanke   o Add "TaskName" meta data field for compliance with BIDS 1.0.0
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 of contracting retmap run.
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 allow for timing verification of generated stimulus files
2016-05-27 17:35 +0200 Michael Hanke   o {origin/_} Mention openfmri as download source
2016-04-04 09:31 +0200 Michael Hanke   o Update publication links
2016-03-31 11:26 +0200 Michael Hanke   o Disable invalid test
[main] d93ed6e3b1a00f1dba18b603e668b7a060cdcf77 - commit 48 of 79
```

27%

VERSION CONTROL BEYOND TEXT FILES



Using git-annex, DataLad version controls large 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
2016-04-04 09:31 +0200 Michael Hanke   o Update publication links
2016-03-31 11:26 +0200 Michael Hanke   o Disable invalid test
[main] 6da25fb6fee2c698d35f52066698b6f94850f4d2 - commit 10 of 79    27%
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 +-
...2_ses-movie_task-movie_run-1_recording-cardresp_physio.tsv.gz | 2 +-
...2_ses-movie_task-movie_run-2_recording-cardresp_physio.tsv.gz | 2 +-
[diff] 6da25fb6fee2c698d35f52066698b6f94850f4d2 - line 1 of 2391    0%
```

VERSION CONTROL BEYOND TEXT FILES

- Datasets have an **annex** to track files without placing their content into Git
- Rather than content, **identity** (hash) and **location** information is put into Git:

VERSION CONTROL BEYOND TEXT FILES

- Datasets have an **annex** to track files without placing their content into Git
- Rather than content, **identity** (hash) and **location** information is put into Git:
 - Where the filesystem allows it, annexed files are symlinks:

```
$ ls -l sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
lrwxrwxrwx 1 adina adina 142 Jul 22 19:45 sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz ->
../../../../.git/annex/objects/kZ/K5/MD5E-s24180157--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz/MD5E-s241
--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz
```

copy

(PS: especially useful in datasets with many identical files)

VERSION CONTROL BEYOND TEXT FILES

- Datasets have an **annex** to track files without placing their content into Git
- Rather than content, **identity** (hash) and **location** information is put into Git:
 - Where the filesystem allows it, annexed files are symlinks:

```
$ ls -l sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
lrwxrwxrwx 1 adina adina 142 Jul 22 19:45 sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz ->
../../.git/annex/objects/kZ/K5/MD5E-s24180157--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz/MD5E-s241
--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz
```

(PS: especially useful in datasets with many identical files)

- The symlink reveals: This internal data organization based on identity hash

```
$ md5sum sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
aeb0e5f2e2d5fe4ade97117a8cc5232f  sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
```

VERSION CONTROL BEYOND TEXT FILES

- Datasets have an **annex** to track files without placing their content into Git
- Rather than content, **identity** (hash) and **location** information is put into Git:
 - Where the filesystem allows it, annexed files are symlinks:

```
$ ls -l sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
lrwxrwxrwx 1 adina adina 142 Jul 22 19:45 sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz ->
../../.git/annex/objects/kZ/K5/MD5E-s24180157--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz/MD5E-s241
--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz
```

(PS: especially useful in datasets with many identical files)

- The symlink reveals: This internal data organization based on identity hash

```
$ md5sum sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
aeb0e5f2e2d5fe4ade97117a8cc5232f  sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
```

- The (tiny) symlink instead of the (potentially large) file content is committed - version controlling precise file identity without checking contents into Git

```
diff --git a/sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz b/sub-02/func/sub-02_task-oneback_run-01_bold.nii.
new file mode 120000
index 0000000..398e7f1
--- /dev/null
+++ b/sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
@@ -0,0 +1 @@
+../../.git/annex/objects/kZ/K5/MD5E-s24180157--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz/MD5E-s24180157--aeb0e5f2e2
```

VERSION CONTROL BEYOND TEXT FILES

- Datasets have an **annex** to track files without placing their content into Git
- Rather than content, **identity** (hash) and **location** information is put into Git:
 - Where the filesystem allows it, annexed files are symlinks:

```
$ ls -l sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
lrwxrwxrwx 1 adina adina 142 Jul 22 19:45 sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz ->
../../.git/annex/objects/kZ/K5/MD5E-s24180157--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz/MD5E-s241
--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz
```

(PS: especially useful in datasets with many identical files)

- The symlink reveals: This internal data organization based on identity hash

```
$ md5sum sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
aeb0e5f2e2d5fe4ade97117a8cc5232f  sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
```

- The (tiny) symlink instead of the (potentially large) file content is committed - version controlling precise file identity without checking contents into Git

```
diff --git a/sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz b/sub-02/func/sub-02_task-oneback_run-01_bold.nii.
new file mode 120000
index 0000000..398e7f1
--- /dev/null
+++ b/sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
@@ -0,0 +1 @@
+../../.git/annex/objects/kZ/K5/MD5E-s24180157--aeb0e5f2e2d5fe4ade97117a8cc5232f.nii.gz/MD5E-s24180157--aeb0e5f2e2
```

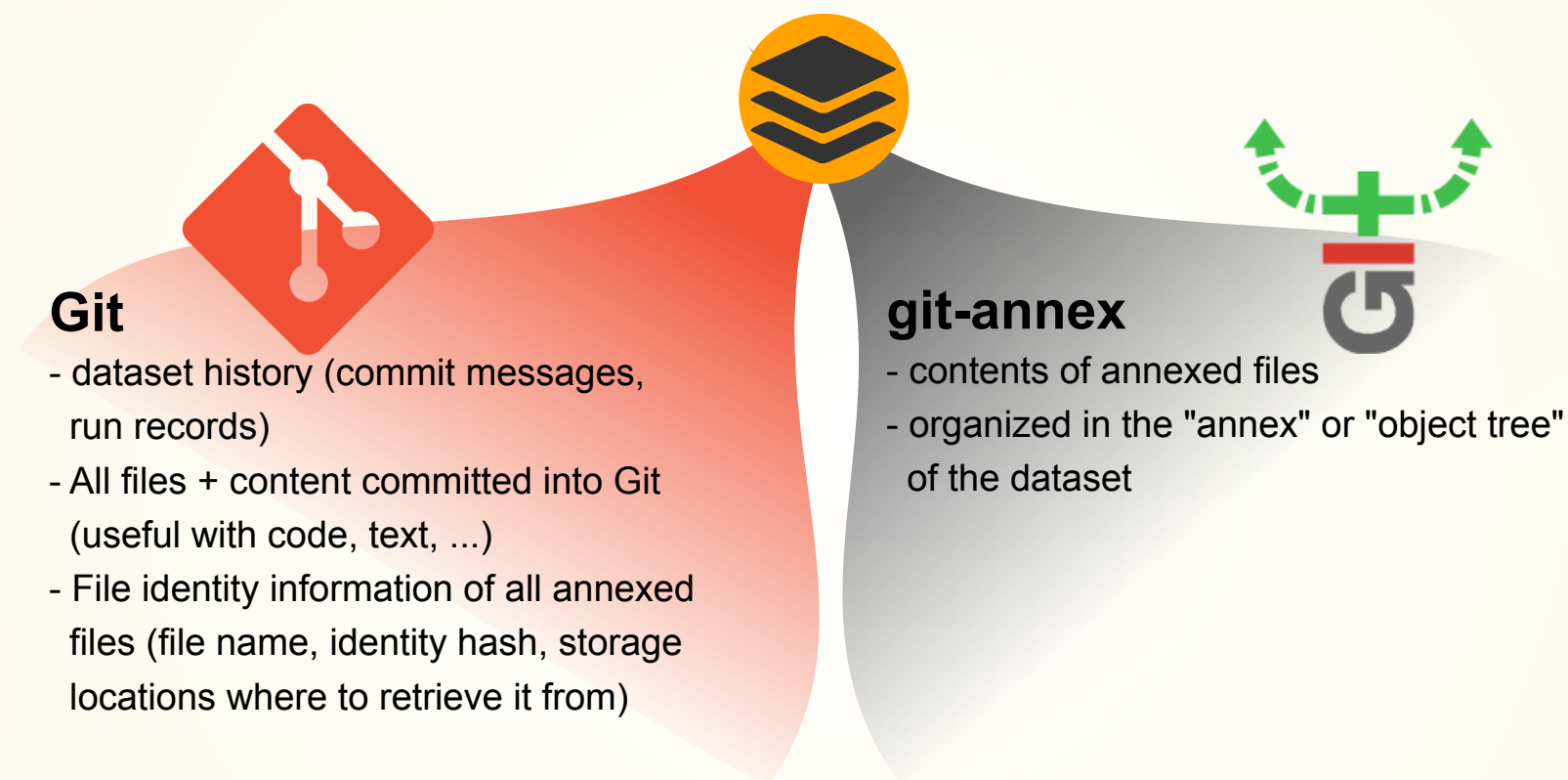
- File availability information is stored to record a decentral network of file content. A file can exist in multiple different locations.

```
$ git annex whereis sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz
whereis sub-02/func/sub-02_task-oneback_run-01_bold.nii.gz (2 copies)
  8c3680dd-6165-4749-adaa-c742232bc317 -- git@8242caf9acd8:/data/repos/adswa/bidsdata.git [git]
  fff8fdbbc-3185-4b78-bd12-718717588442 -- adina@muninn:~/bids-data [here]
ok
```

GIT VERSUS GIT-ANNEX

Data in datasets is either stored in Git or git-annex

By default, everything is annexed, i.e., stored in a dataset annex



Git	git-annex
handles small files well (text, code)	handles all types and sizes of files well
file contents are in the Git history and will be shared upon git/datalad push	file contents are in the annex. Not necessarily shared
Shared with every dataset clone	Can be kept private on a per-file level when sharing the dataset
Useful: Small, non-binary, frequently modified, need-to-be-accessible (DUA, README) files	Useful: Large files, private files

(RAW) DATA MISMANAGEMENT

- Multiple large datasets are available on a compute cluster 🌲
- Each researcher creates their own copies of data ▲
- Multiple different derivatives and results are computed from it 🏔️
- Data, copies of data, half-baked data transformations, results, and old versions of results are kept - undocumented 🌋

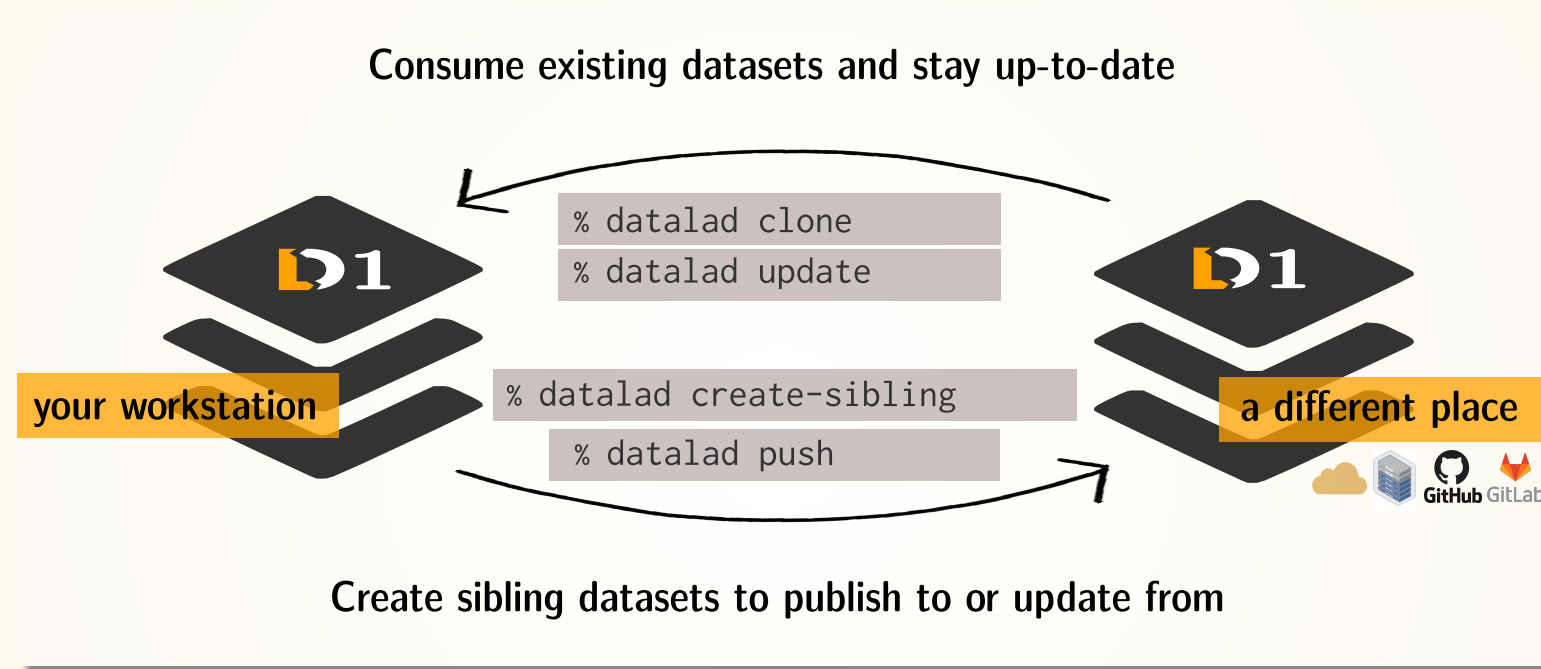
SHARE DATA LIKE SOURCE CODE

SHARE DATA LIKE SOURCE CODE

The screenshot shows the GitLab interface for the INM7 group. The left sidebar contains navigation links: Group information, Issues (366), Merge requests (10), and Packages & Registries. The main content area displays the INM7 group profile with a welcome message and a list of subgroups and projects. The subgroups and projects are listed in a table with columns for name, description, and statistics (issues, merge requests, and members).

Subgroups and projects	Shared projects	Archived projects	Search by name	Updated date
> public		1		0 4 1
> vbc				0 1 1
> Training Owner				0 2 3
> webtools				0 6 1
> tools Tools and pipe-lines to be shared across INM-7.				0 6 1
> AppliedMachineLearning Projects of the Applied Machine Learning group.				0 15 4
> JuTrack				0 1 1

SHARE DATA LIKE SOURCE CODE



TRANSPORT LOGISTICS: LOTS OF DATA, LITTLE DISK-USAGE

TRANSPORT LOGISTICS: LOTS OF DATA, LITTLE DISK-USAGE

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

TRANSPORT LOGISTICS: 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 .
```

TRANSPORT LOGISTICS: 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:

TRANSPORT LOGISTICS: 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:

```
$ datalad get sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz
get(ok): /tmp/studyforrest-data-phase2/sub-01/ses-movie/func/
sub-01_ses-movie_task-movie_run-1_bold.nii.gz (file) [from mddatasrc...]
```

copy

TRANSPORT LOGISTICS: 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:

```
$ datalad get sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz
get(ok): /tmp/studyforrest-data-phase2/sub-01/ses-movie/func/
sub-01_ses-movie_task-movie_run-1_bold.nii.gz (file) [from mddatasrc...]
```

copy

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

TRANSPORT LOGISTICS: 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:

```
$ datalad get sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz
get(ok): /tmp/studyforrest-data-phase2/sub-01/ses-movie/func/
sub-01_ses-movie_task-movie_run-1_bold.nii.gz (file) [from mddatasrc...]
```

copy

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

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

PLENTY OF DATA, BUT LITTLE DISK-USAGE

PLENTY OF DATA, BUT LITTLE DISK-USAGE

Drop file content that is not needed:

PLENTY OF DATA, BUT LITTLE DISK-USAGE

Drop file content that is not needed:

```
$ datalad drop sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz  
drop(ok): /[...]/sub-01_ses-movie_task-movie_run-1_bold.nii.gz (file)
```

copy

PLENTY OF DATA, BUT LITTLE DISK-USAGE

Drop file content that is not needed:

```
$ datalad drop sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz  
drop(ok): /[...]/sub-01_ses-movie_task-movie_run-1_bold.nii.gz (file)
```

copy

Only "meta data" stays behind, and files can be re-obtained on demand. This allows for disk-space-aware computing workflows:

PLENTY OF DATA, BUT LITTLE DISK-USAGE

Drop file content that is not needed:

```
$ datalad drop sub-01/ses-movie/func/sub-01_ses-movie_task-movie_run-1_bold.nii.gz  
drop(ok): /[...]/sub-01_ses-movie_task-movie_run-1_bold.nii.gz (file)
```

copy

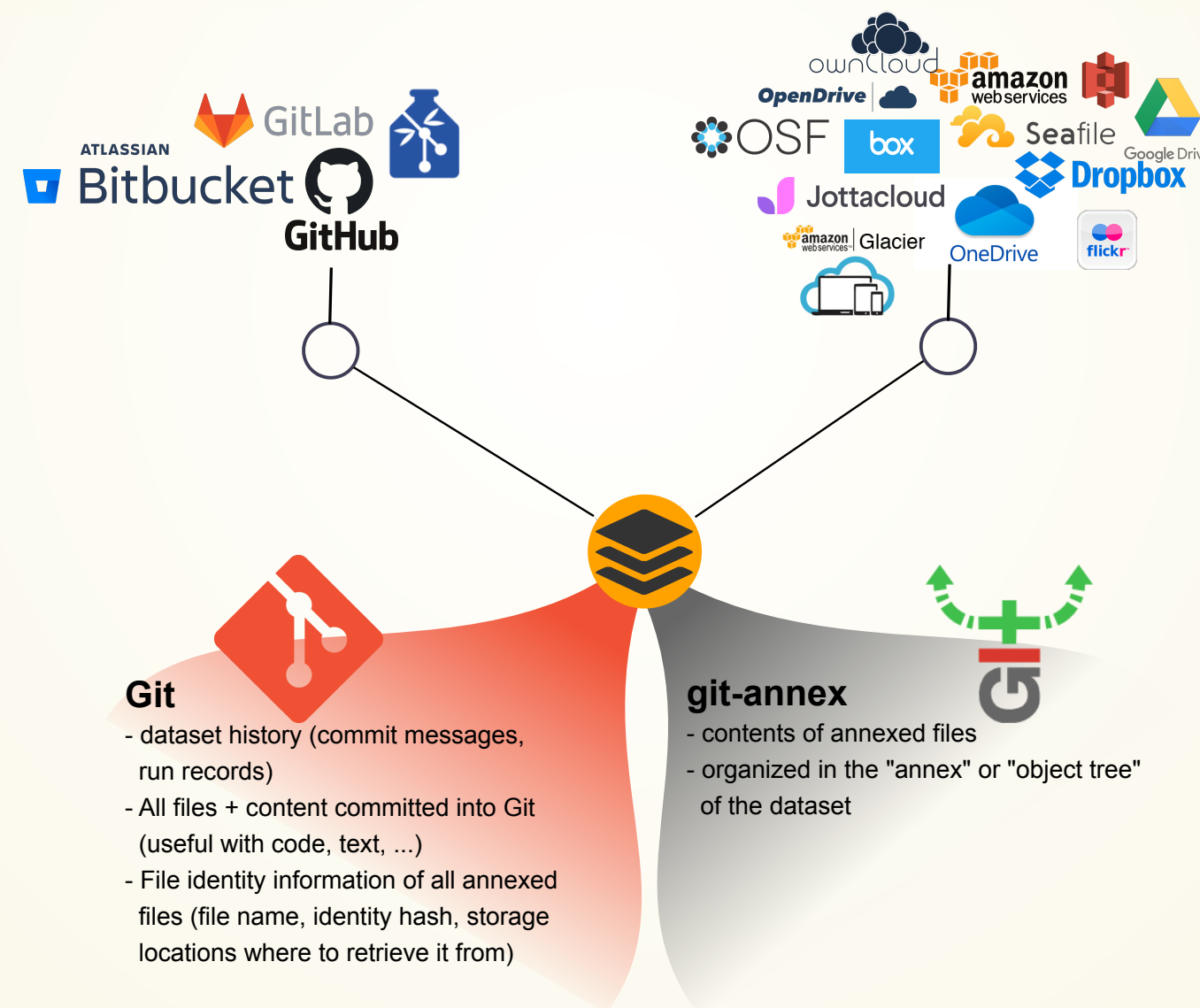
Only "meta data" stays behind, and files can be re-obtained on demand. This allows for disk-space-aware computing workflows:

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

copy

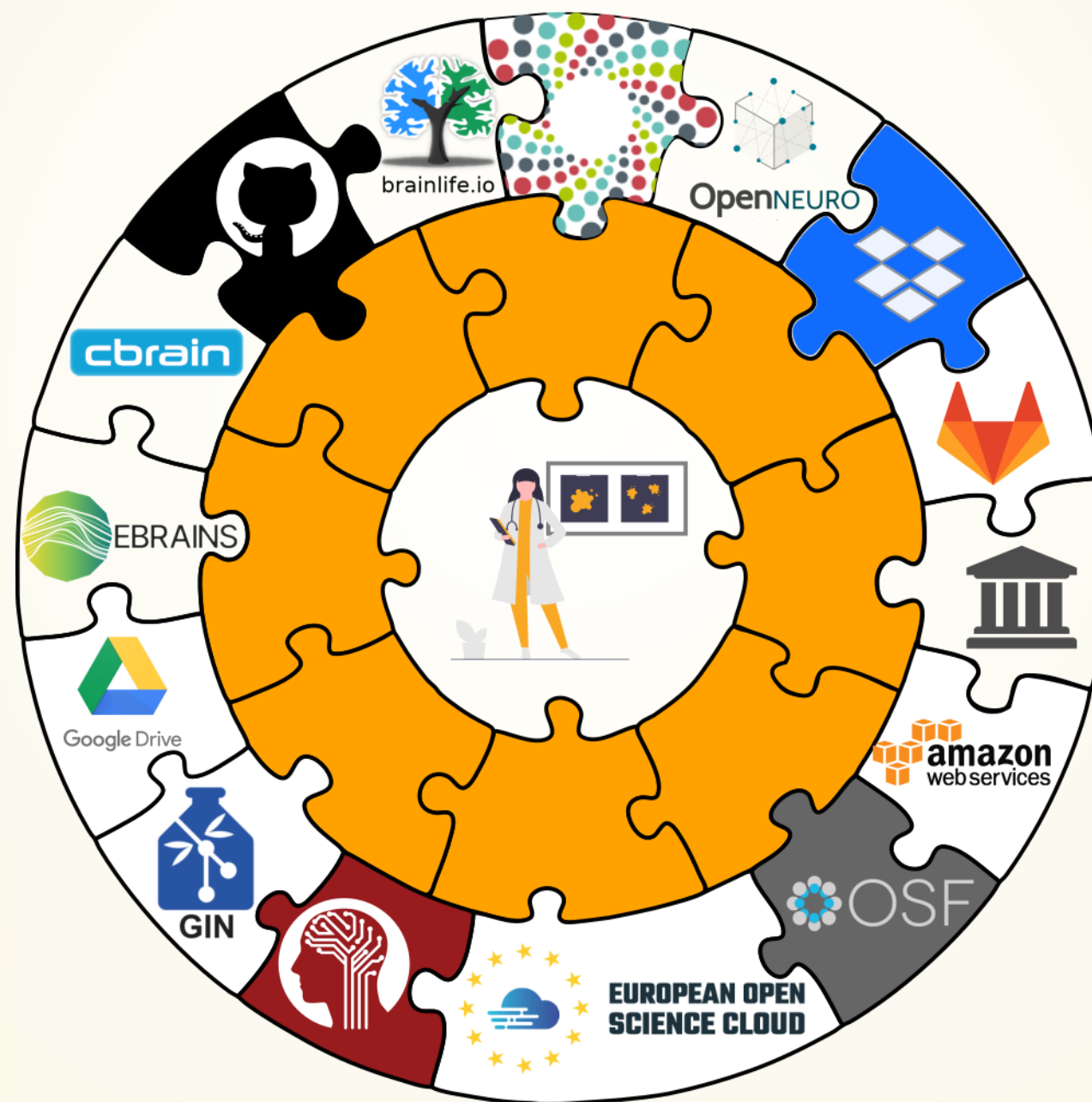
PUBLISHING DATASETS

Publish datasets, their annexed contents, or both to infrastructure of your choice

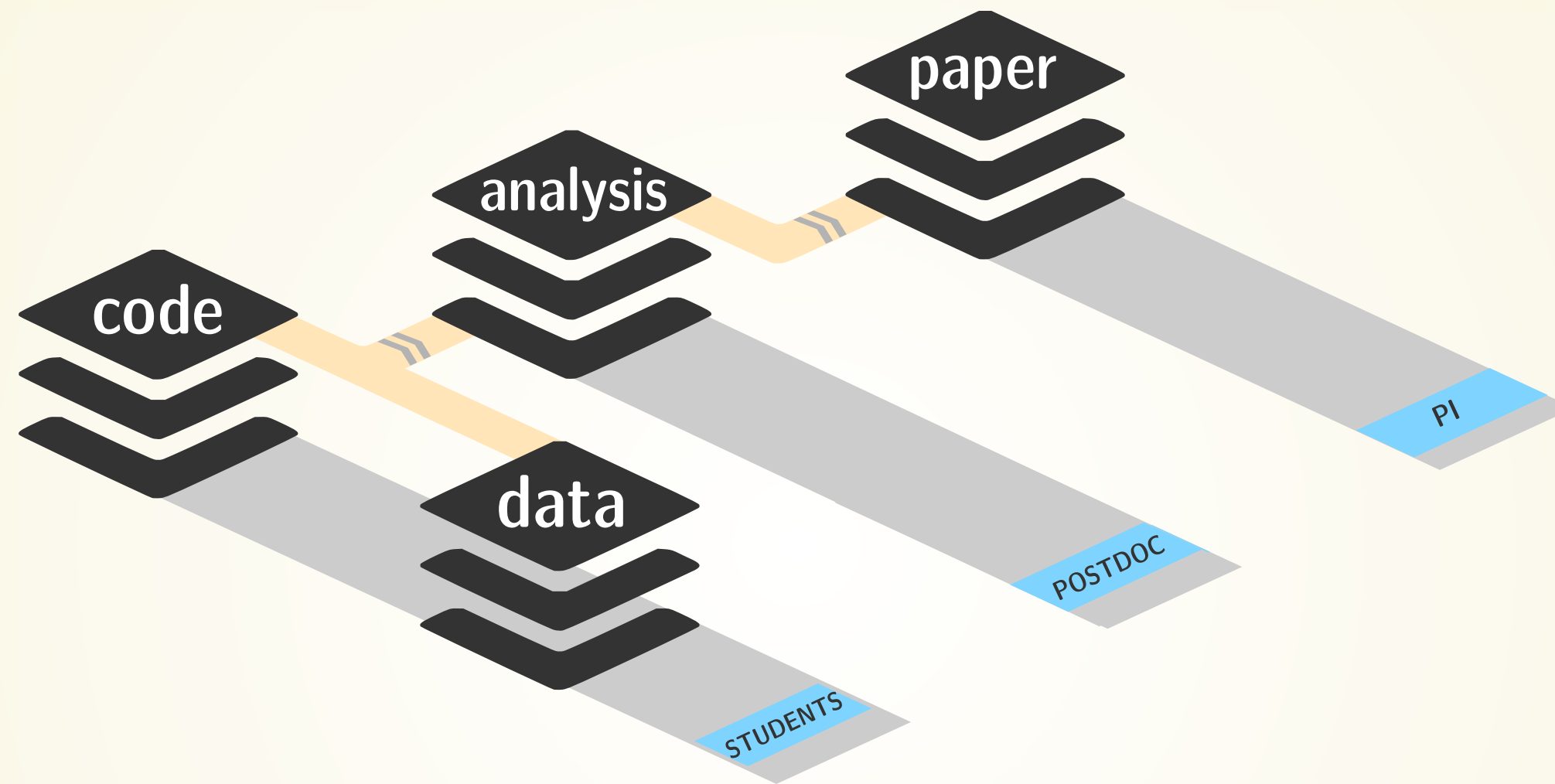


INTEROPERABILITY

- DataLad is built to maximize interoperability and streamline routines across hosting and storage technology



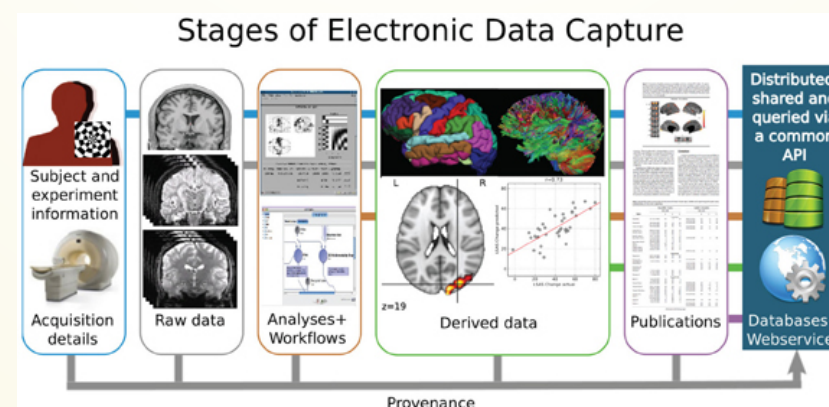
MODULARITY



- Typical workflow in science
 - Prior works (algorithm development, empirical data, etc.) are combined to produce novel results with to goal of a publication
 - **Aggregation across time and contributors**
 - Aiming for (but often failing) to be reproducible

VERSION CONTROL BEYOND SINGLE REPOSITORIES

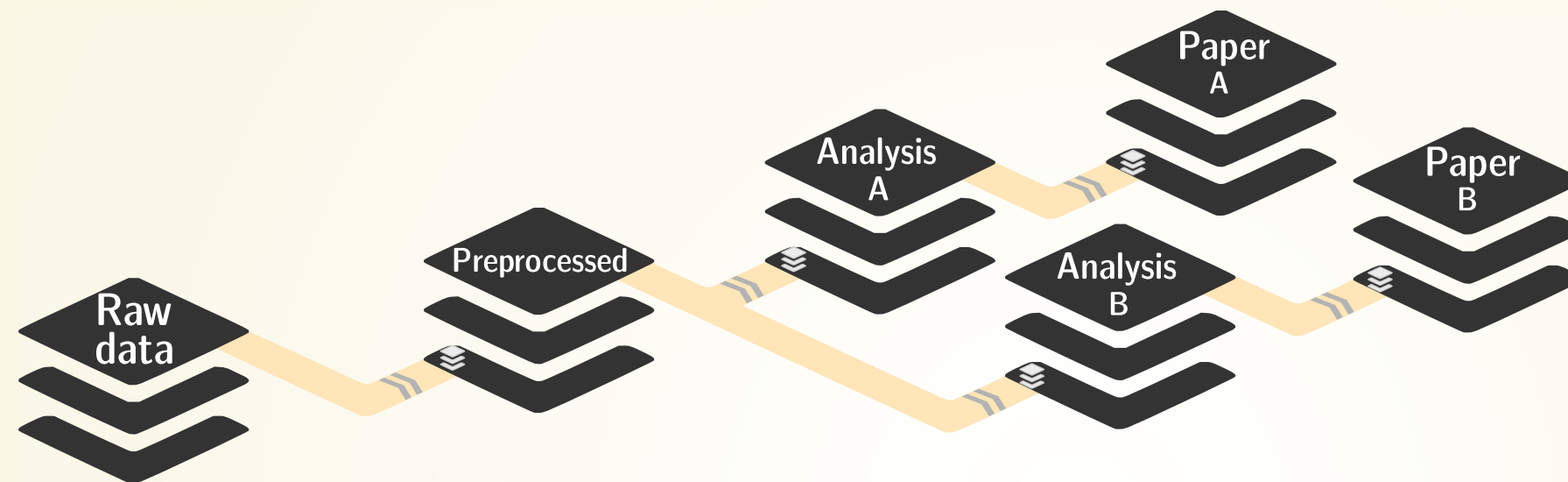
- **Why** are multiple repositories needed (in science)?
 - Size impacts I/O and logistics
 - Git can struggle with 1M+ files or 100k+ commits
 - Filesystems (licensing) can struggle with large numbers of inodes
 - Target audience is different
 - Public vs. private or personal vs. anonymized data
 - Pace of evolution or access patterns are different
 - "Factual" raw data vs. choices of (pre-)processing
 - Completed acquisition vs. ongoing study



- A single repository is not enough, but Git/Git-annex are not optimized for such use cases

DATASET NESTING

- Seamless nesting mechanisms:

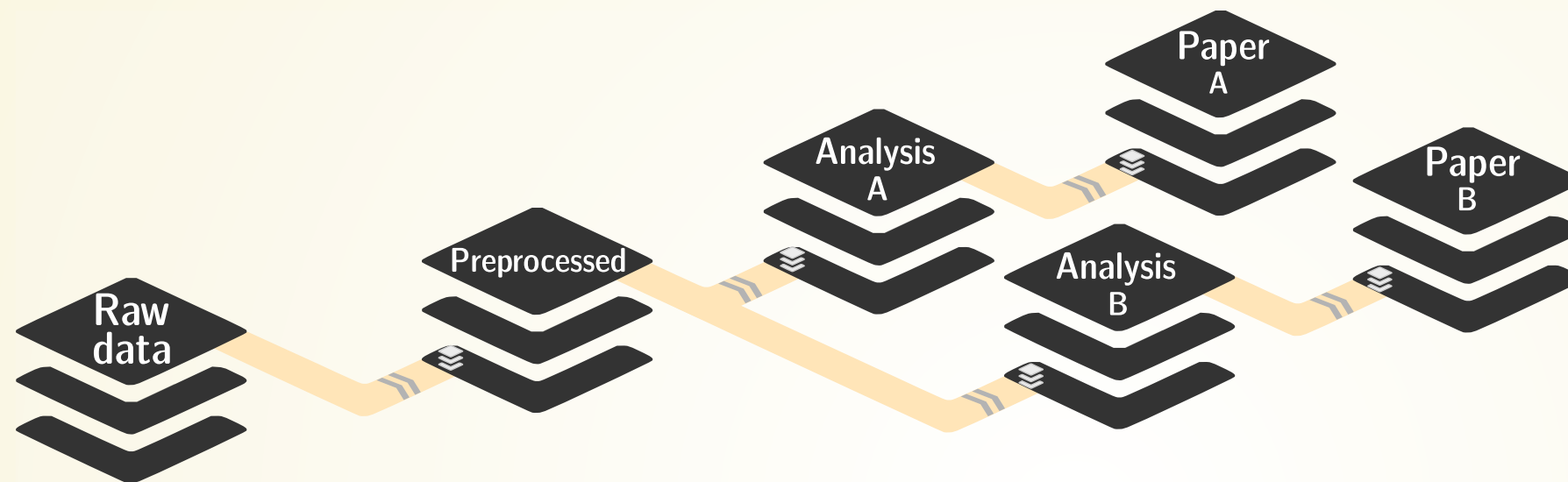


Nest modular datasets to create a linked hierarchy of datasets,
and enable recursive operations throughout the hierarchy

- hierarchies of datasets in super-/sub-dataset relationships
- based on Git submodules, but more seamless: Mono-repo feel thanks to recursive operations

DATASET NESTING

- Seamless nesting mechanisms:



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

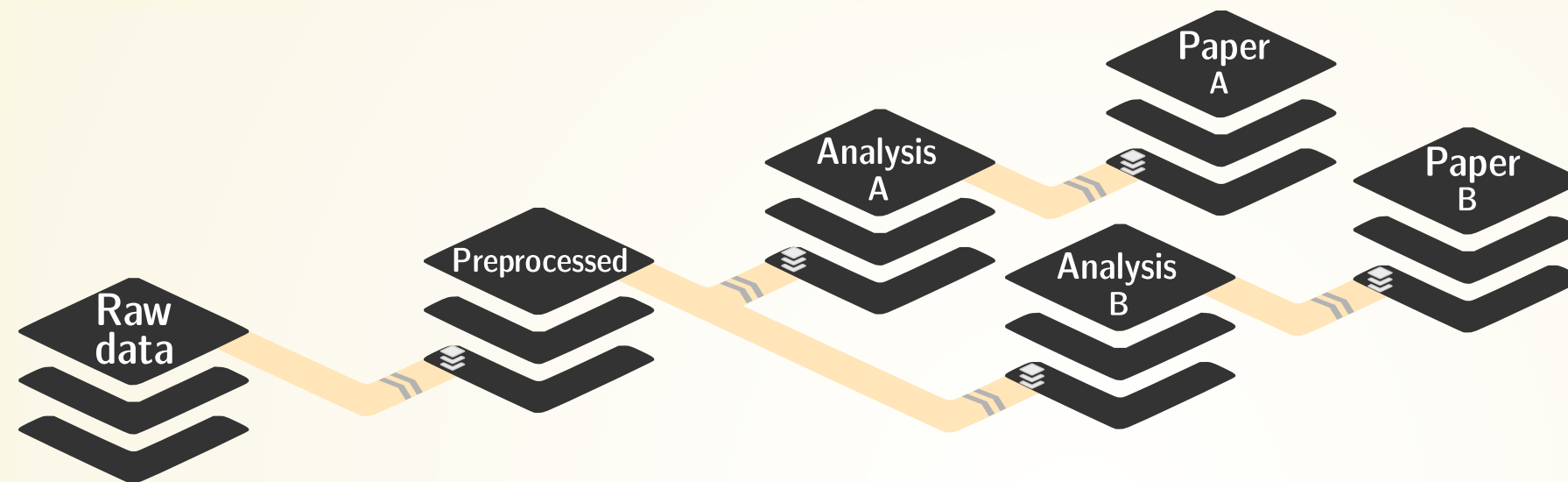
- hierarchies of datasets in super-/sub-dataset relationships
- based on Git submodules, but more seamless: Mono-repo feel thanks to recursive operations
- Overcomes scaling issues with large amounts of files

```
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)

DATASET 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: Mono-repo feel thanks to recursive operations
- Overcomes scaling issues with large amounts of files

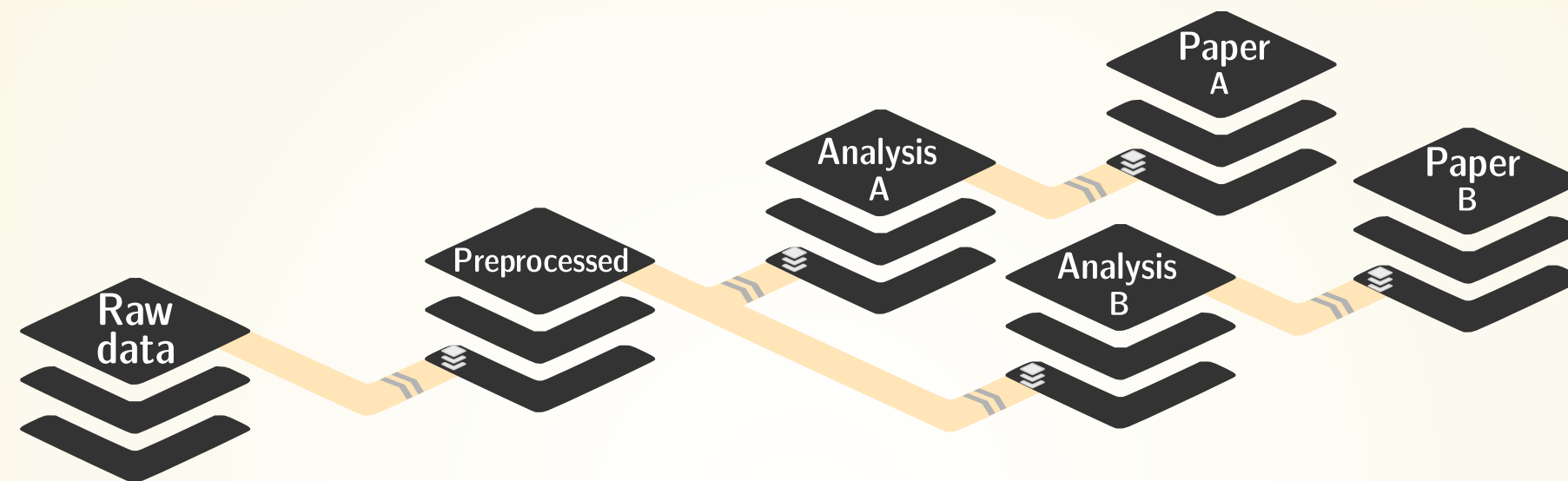
```
adina@bulk1 in /ds/hcp/super on git:master> datalad status --annex -r
15530572 annex'd files (77.9 TB recorded total size)
nothing to save, working tree clean
```

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

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

INTUITIVE DATA ANALYSIS STRUCTURE

- You can link datasets together in superdataset-subdataset hierarchies:



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

```
1 $ cd myanalysis
2 # we can install analysis input data as a subdataset to the dataset
3 $ datalad clone -d . https://github.com/datalad-handbook/iris_data.git input/
4 [INFO ] Scanning for unlocked files (this may take some time)
5 [INFO ] Remote origin not usable by git-annex; setting annex-ignore
6 install(ok): input (dataset)
7 add(ok): input (file)
8 add(ok): .gitmodules (file)
9 save(ok): . (dataset)
10 action summary:
11   add (ok: 2)
12   install (ok: 1)
13   save (ok: 1)
```

copy

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?"

Image credit: CC-BY Scriberia and The Turing Way



LEAVING A TRACE

LEAVING A TRACE

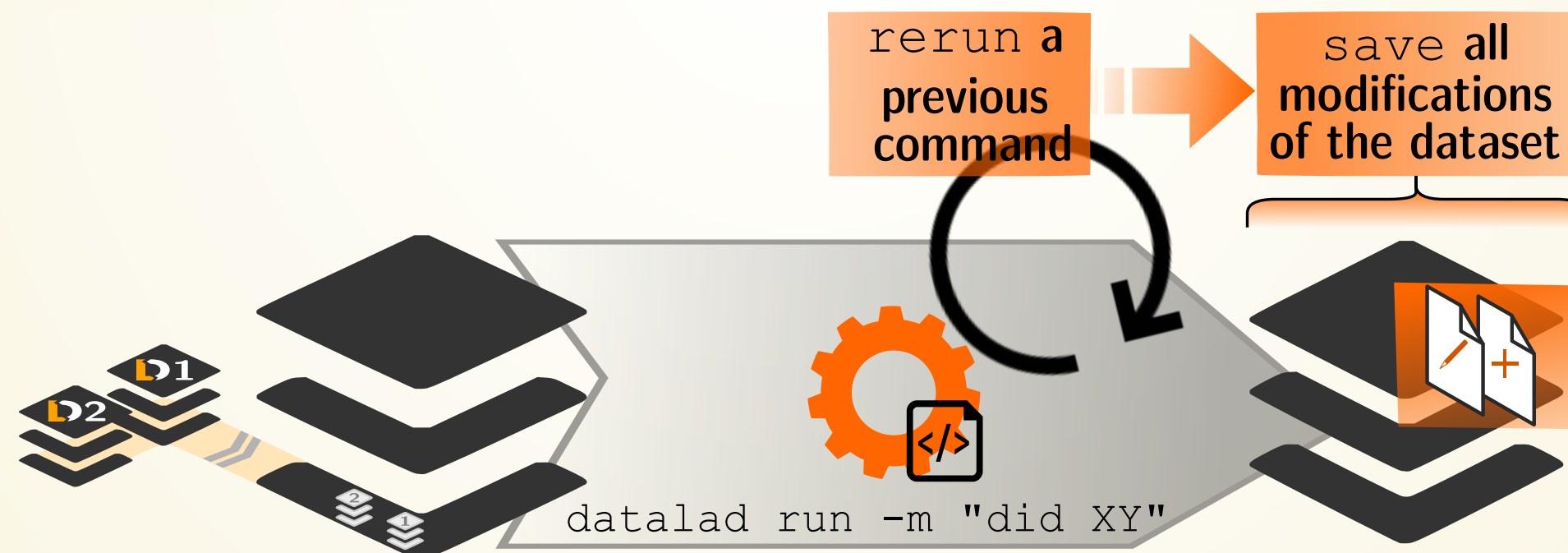
datalad run wraps around anything expressed in a command line call and saves the dataset modifications resulting from the execution.



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.



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

```
1 $ datalad containers-run \  
2   --message "Time series extraction from Locus Coeruleus" \  
3   --container-name nilearn \  
4   --input 'mri/*_bold.nii' \  
5   --output 'sub-*/LC_timeseries_run-*.csv' \  
6   "python3 code/extract_lc_timeseries.py" \  
7 \  
8 -- Git commit -- \  
9   commit 5a7565a640ff6de67e07292a26bf272f1ee4b00e \  
10  Author:      Adina Wagner adina.wagner@t-online.de \  
11  AuthorDate:  Mon Nov 11 16:15:08 2019 +0100 \  
12  Commit:      Adina Wagner adina.wagner@t-online.de \  
13  CommitDate:  Mon Nov 11 16:15:08 2019 +0100 \  
14 \  
15  [DATALAD RUNCMD] Time series extraction from Locus Coeruleus \  
16  === Do not change lines below === \  
17  { \  
18    "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..", \  
19    "dsid": "92ealfaa-632a-11e8-af29-a0369f7c647e", \  
20    "inputs": [ \  
21      "mri/*.bold.nii.gz", \  
22      ".datalad/environments/nilearn.simg" \  
23    ], \  
24    "outputs": ["sub-*/LC_timeseries_run-*.csv"], \  
25    ... \  
26  } \  
27  ^^^ Do not change lines above ^^^ \  
28  --- \  
29  sub-01/LC_timeseries_run-1.csv | 1 + \  
30  ...
```

[copy](#)

DATA ANALYSIS PROVENANCE

```
1 $ datalad containers-run \  
2   --message "Time series extraction from Locus Coeruleus" \  
3   --container-name nilearn \  
4   --input 'mri/*_bold.nii' \  
5   --output 'sub-*/LC_timeseries_run-*.csv' \  
6   "python3 code/extract_lc_timeseries.py" \  
7 \  
8 -- Git commit -- \  
9   commit 5a7565a640ff6de67e07292a26bf272f1ee4b00e \  
10  Author:      Adina Wagner adina.wagner@t-online.de \  
11  AuthorDate:  Mon Nov 11 16:15:08 2019 +0100 \  
12  Commit:      Adina Wagner adina.wagner@t-online.de \  
13  CommitDate:  Mon Nov 11 16:15:08 2019 +0100 \  
14 \  
15  [DATALAD RUNCMD] Time series extraction from Locus Coeruleus \  
16  === Do not change lines below === \  
17  { \  
18    "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..", \  
19    "dsid": "92ealfaa-632a-11e8-af29-a0369f7c647e", \  
20    "inputs": [ \  
21      "mri/*.bold.nii.gz", \  
22      ".datalad/environments/nilearn.simg" \  
23    ], \  
24    "outputs": ["sub-*/LC_timeseries_run-*.csv"], \  
25    ... \  
26  } \  
27  ^^^ Do not change lines above ^^^ \  
28  --- \  
29  sub-01/LC_timeseries_run-1.csv | 1 + \  
30  ...
```

[copy](#)

DATA ANALYSIS PROVENANCE

```
1 $ datalad containers-run \  
2   --message "Time series extraction from Locus Coeruleus" \  
3   --container-name nilearn \  
4   --input 'mri/*_bold.nii' \  
5   --output 'sub-*/LC_timeseries_run-*.csv' \  
6   "python3 code/extract_lc_timeseries.py" \  
7 \  
8 -- Git commit -- \  
9   commit 5a7565a640ff6de67e07292a26bf272f1ee4b00e \  
10  Author:      Adina Wagner adina.wagner@t-online.de \  
11  AuthorDate:  Mon Nov 11 16:15:08 2019 +0100 \  
12  Commit:      Adina Wagner adina.wagner@t-online.de \  
13  CommitDate:  Mon Nov 11 16:15:08 2019 +0100 \  
14 \  
15  [DATALAD RUNCMD] Time series extraction from Locus Coeruleus \  
16  === Do not change lines below === \  
17  { \  
18    "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..", \  
19    "dsid": "92ealfaa-632a-11e8-af29-a0369f7c647e", \  
20    "inputs": [ \  
21      "mri/*.bold.nii.gz", \  
22      ".datalad/environments/nilearn.simg" \  
23    ], \  
24    "outputs": ["sub-*/LC_timeseries_run-*.csv"], \  
25    ... \  
26  } \  
27  ^^^ Do not change lines above ^^^ \  
28  --- \  
29  sub-01/LC_timeseries_run-1.csv | 1 + \  
30  ...
```

[copy](#)

DATA ANALYSIS PROVENANCE

```
1 $ datalad rerun 5a7565a640ff6de67
2 [INFO ] run commit 5a7565a640ff6de67; (Time series extraction from Locus Coeruleus)
3 [INFO ] Making sure inputs are available (this may take some time)
4 get(ok): mri/sub-01_bold.nii (file)
5 get(ok): mri/sub-02_bold.nii (file)
6     [...]
7 [INFO ] == Command start (output follows) =====
8 [INFO ] == Command exit (modification check follows) =====
9 add(ok): sub-01/LC_timeseries_run-*.csv(file)
10 add(ok): sub-02/LC_timeseries_run-*.csv (file)
11     [...]
12 action summary:
13   add (ok: 30)
14   get (ok: 30)
15   save (ok: 2)
16   unlock (ok: 30)
```

[copy](#)

DATA ANALYSIS PROVENANCE

```
1 $ datalad rerun 5a7565a640ff6de67
2 [INFO   ] run commit 5a7565a640ff6de67; (Time series extraction from Locus Coeruleus)
3 [INFO   ] Making sure inputs are available (this may take some time)
4 get(ok): mri/sub-01_bold.nii (file)
5 get(ok): mri/sub-02_bold.nii (file)
6     [...]
7 [INFO   ] == Command start (output follows) =====
8 [INFO   ] == Command exit (modification check follows) =====
9 add(ok): sub-01/LC_timeseries_run-*.csv(file)
10 add(ok): sub-02/LC_timeseries_run-*.csv (file)
11     [...]
12 action summary:
13   add (ok: 30)
14   get (ok: 30)
15   save (ok: 2)
16   unlock (ok: 30)
```

[copy](#)

DATALAD FOR SCIENTIFIC WORKFLOWS?

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.
Version control beyond text

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

Version control beyond text

Science is build from modular units.

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

Version control beyond text

Science is build from modular units.

Nesting

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

Version control beyond text

Science is build from modular units.

Nesting

Science is exploratory, iterative, multi-stepped, and complex.

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

Version control beyond text

Science is build from modular units.

Nesting

Science is exploratory, iterative, multi-stepped, and complex.

Provenance

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

Version control beyond text

Science is build from modular units.

Nesting

Science is exploratory, iterative, multi-stepped, and complex.

Provenance

Science is collaborative.

DATALAD FOR SCIENTIFIC WORKFLOWS?

Scientific building blocks are not static.

Version control beyond text

Science is build from modular units.

Nesting

Science is exploratory, iterative, multi-stepped, and complex.

Provenance

Science is collaborative.

Transport logistics

RESEARCH DATA MANAGEMENT IS TIED TO REPRODUCIBILITY

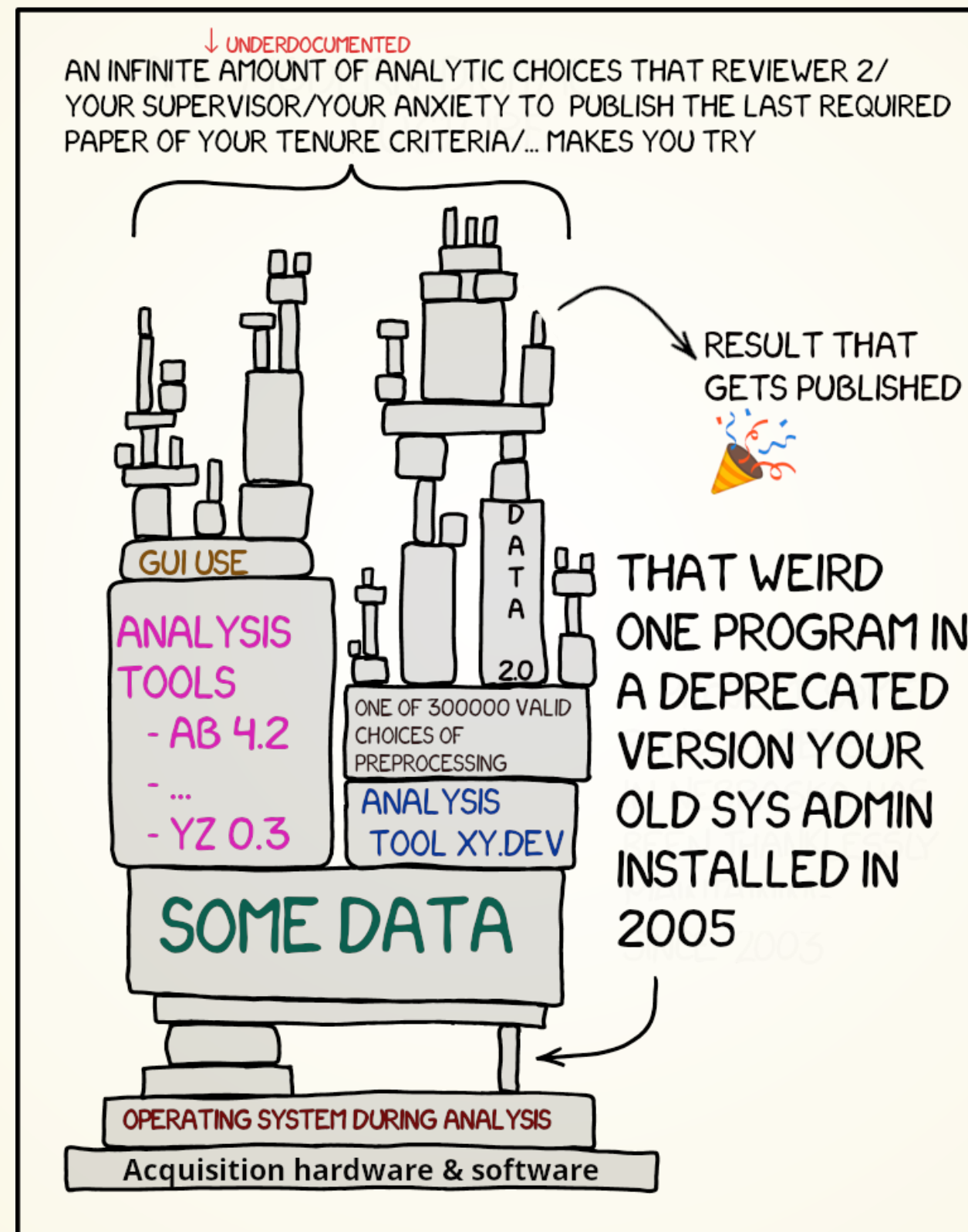
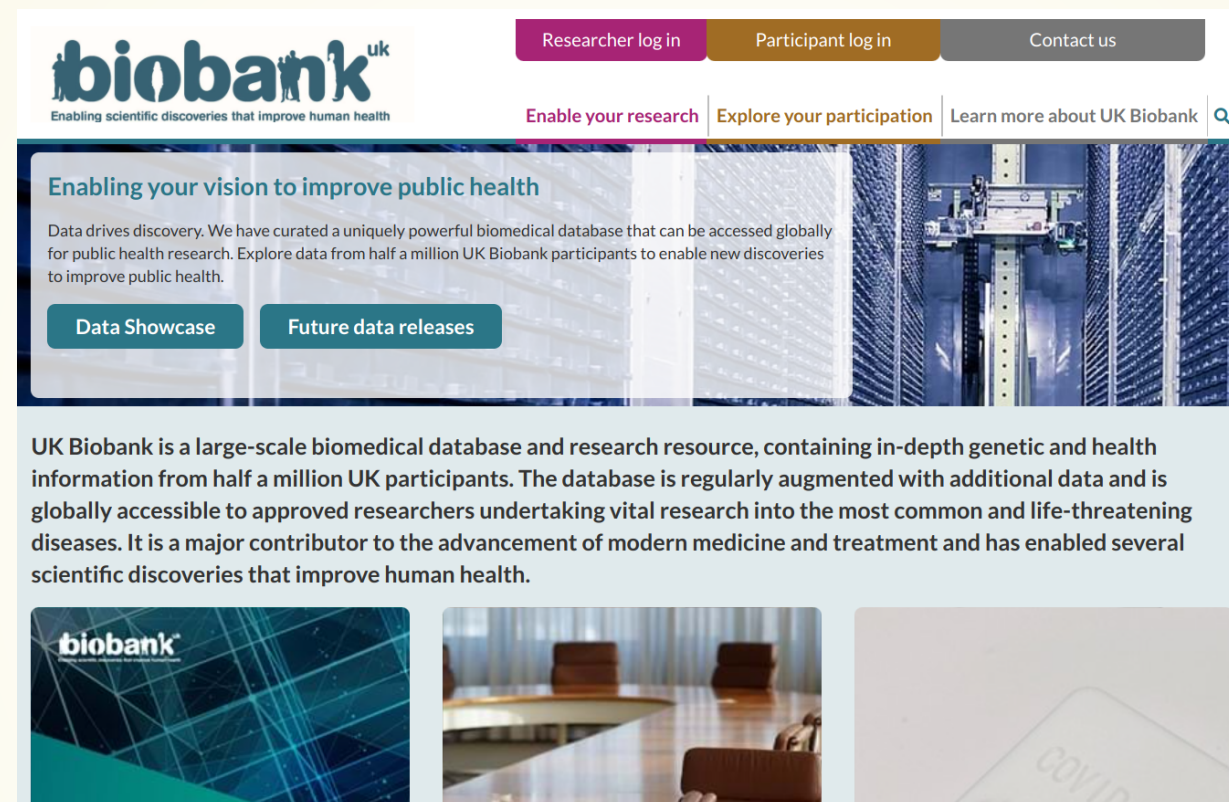


Image credit: Based on xkcd.com/2347/ (CC-BY)

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

FAIRLY BIG: SCALING UP

Objective: Process the UK Biobank (imaging data)

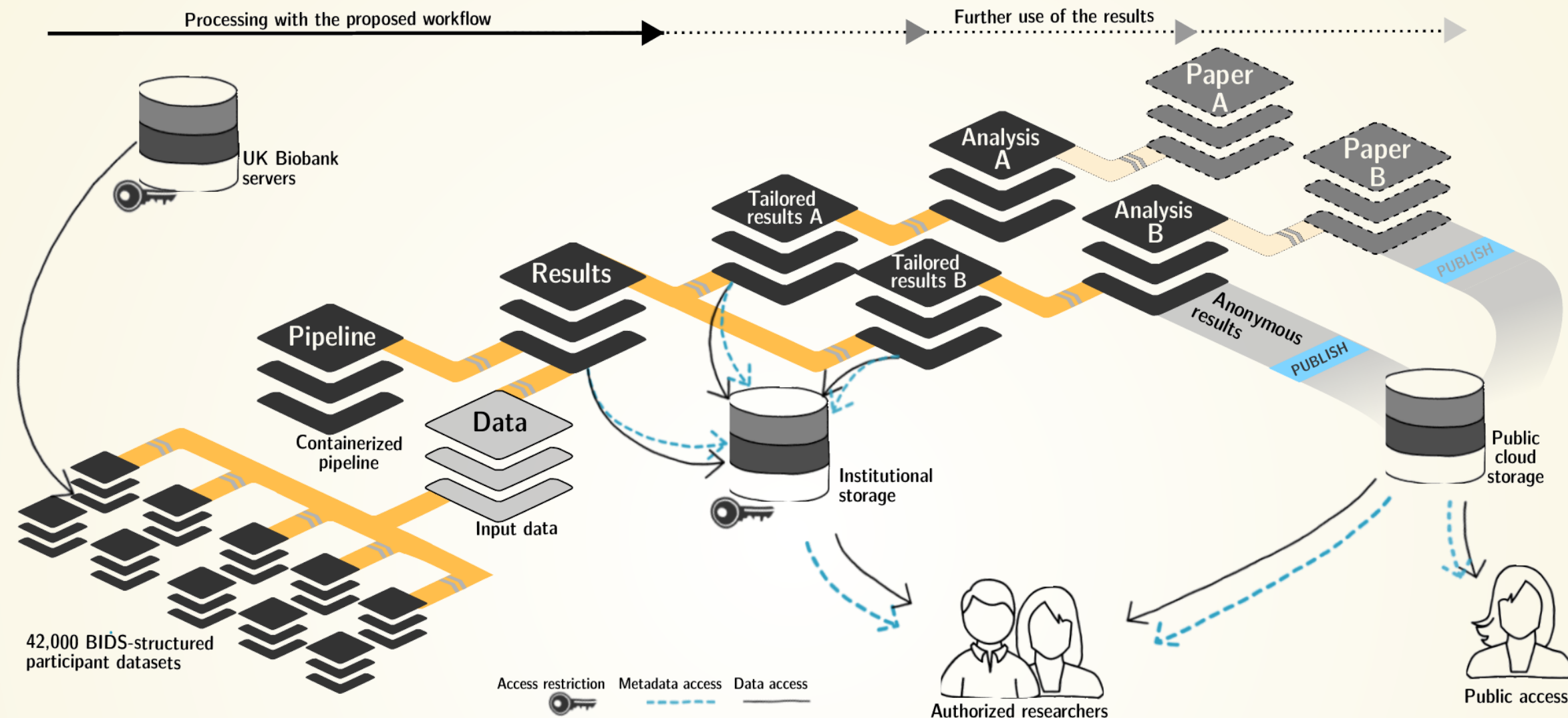


- 76 TB in 43 million files in total
- 42,715 participants contributed personal health data
- Strict DUA
- Custom binary-only downloader
- Most data records offered as (unversioned) ZIP files

CHALLENGES

- Process data such that
 - Results are computationally reproducible (without the original compute infrastructure)
 - There is complete linkage from results to an individual data record download
 - It scales with the amount of available compute resources
- Data processing pipeline
 - Compiled MATLAB blob
 - 1h processing time per image, with 41k images to process
 - 1.2 M output files (30 output files per input file)
 - 1.2 TB total size of outputs

FAIRLY BIG SETUP

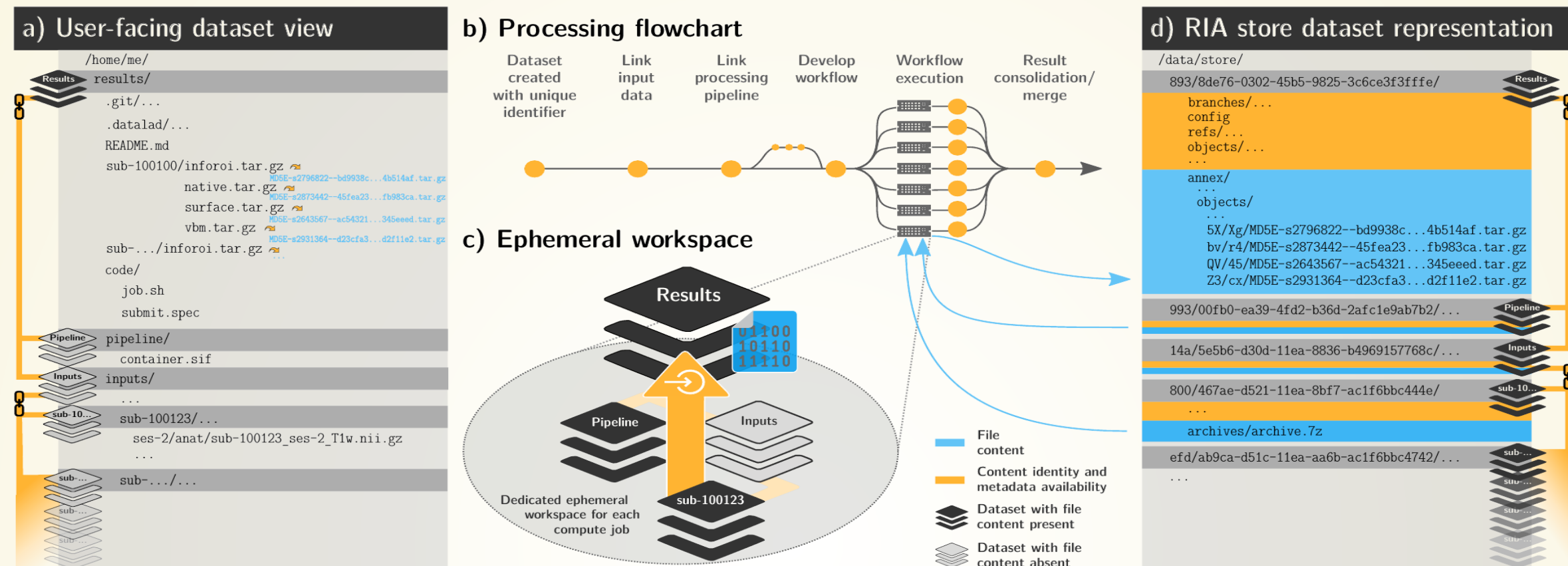


Exhaustive tracking

- **data-lad-ukbiobank** extension downloads, transforms & track the evolution of the complete data release in DataLad datasets
- Native and BIDSified data layout (at no additional disk space usage)
- Structured in 42k individual datasets, combined to one superdataset
- Containerized pipeline in a software container
- Link input data & computational pipeline as dependencies

Wagner, Waite, Wierzbica et al. (2021). FAIRly big: A framework for computationally reproducible processing of large-scale data.

FAIRLY BIG WORKFLOW



portability

- Parallel processing: 1 job = 1 subject (number of concurrent jobs capped at the capacity of the compute cluster)
- Each job is computed in a ephemeral (short-lived) dataset clone, results are pushed back: Ensure exhaustive tracking & portability during computation
- Content-agnostic persistent (encrypted) storage (minimizing storage and inodes)
- Common data representation in secure environments

Wagner, Waite, Wierzba et al. (2021). FAIRly big: A framework for computationally reproducible processing of large-scale data.

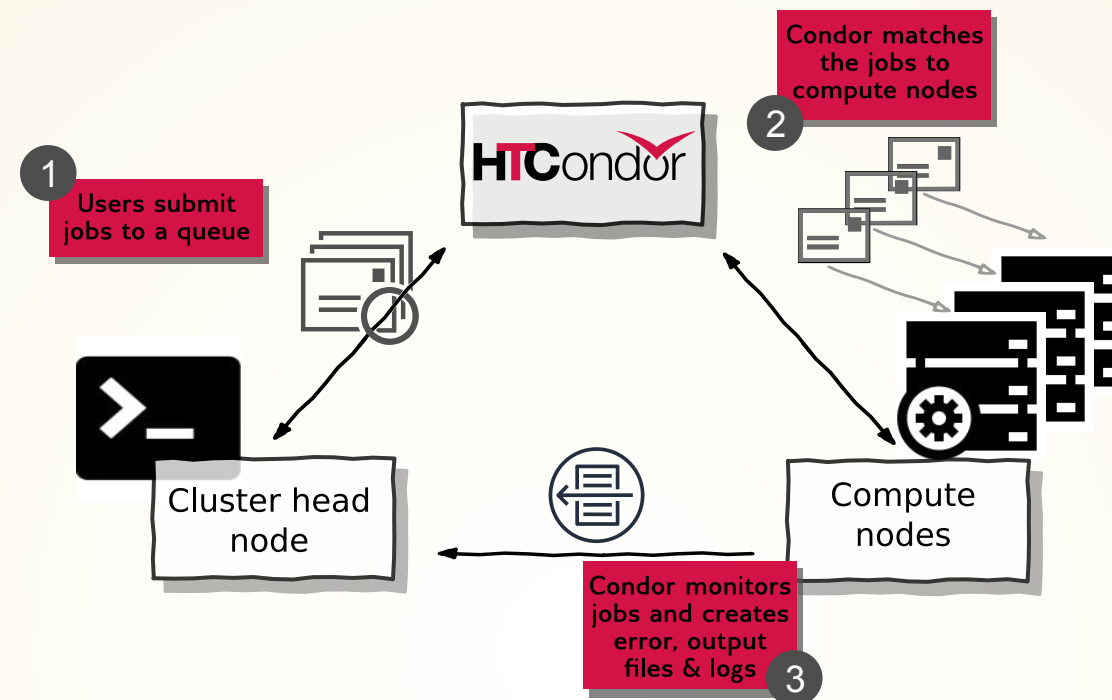
FAIRLY BIG WORKFLOW

portability

- Parallel processing: 1 job = 1 subject (number of concurrent jobs capped at the capacity of the compute cluster)
- Each job is computed in a ephemeral (short-lived) dataset clone, results are pushed back: Ensure exhaustive tracking & portability during computation
- Content-agnostic persistent (encrypted) storage (minimizing storage and inodes)
- Common data representation in secure environments

Wagner, Waite, Wierzba et al. (2021). FAIRly big: A framework for computationally reproducible processing of large-scale data.

FAIRLY BIG WORKFLOW

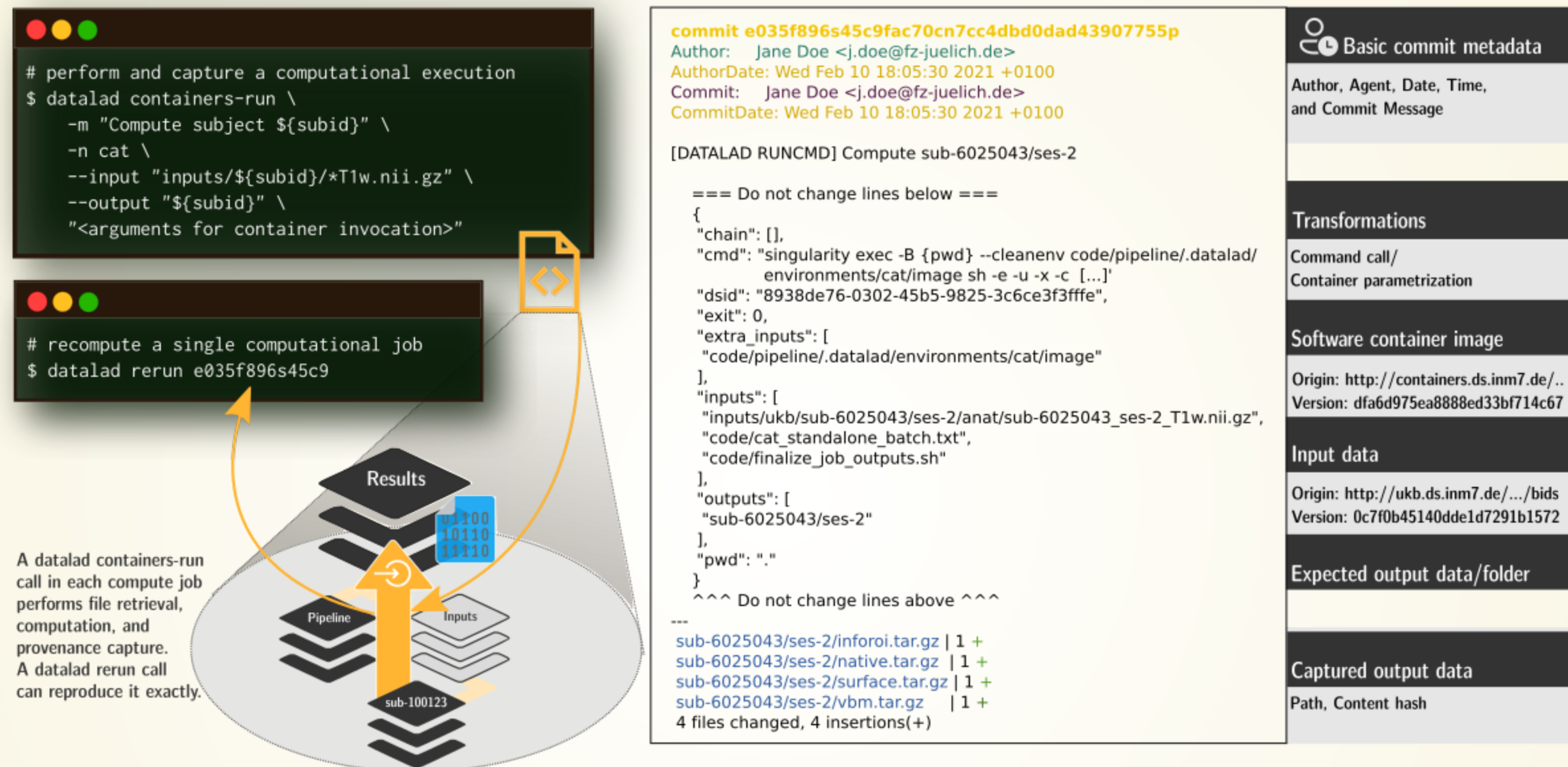


portability

- Parallel processing: 1 job = 1 subject (number of concurrent jobs capped at the capacity of the compute cluster)
- Each job is computed in a ephemeral (short-lived) dataset clone, results are pushed back: Ensure exhaustive tracking & portability during computation
- Content-agnostic persistent (encrypted) storage (minimizing storage and inodes)
- Common data representation in secure environments

Wagner, Waite, Wierzba et al. (2021). FAIRly big: A framework for computationally reproducible processing of large-scale data.

FAIRLY BIG PROVENANCE CAPTURE



Provenance

- Every single pipeline execution is tracked
- Execution in ephemeral workspaces ensures results individually reproducible without HPC access

Wagner, Waite, Wierzba et al. (2021). FAIRly big: A framework for computationally reproducible processing of large-scale data.

FAIRLY BIG MOVIE

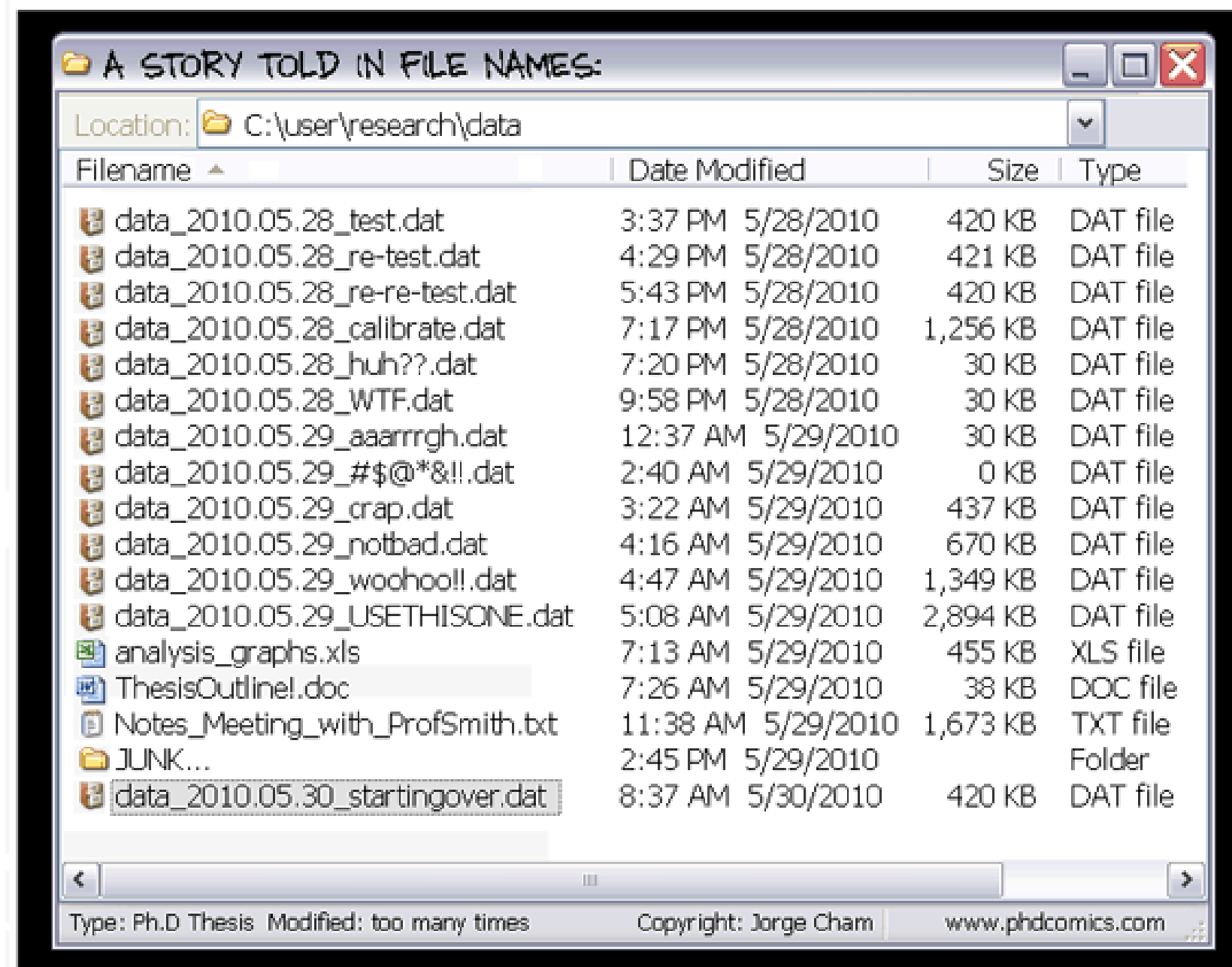
Reproducible processing of 41,180 brain images from the UK Biobank with DataLad



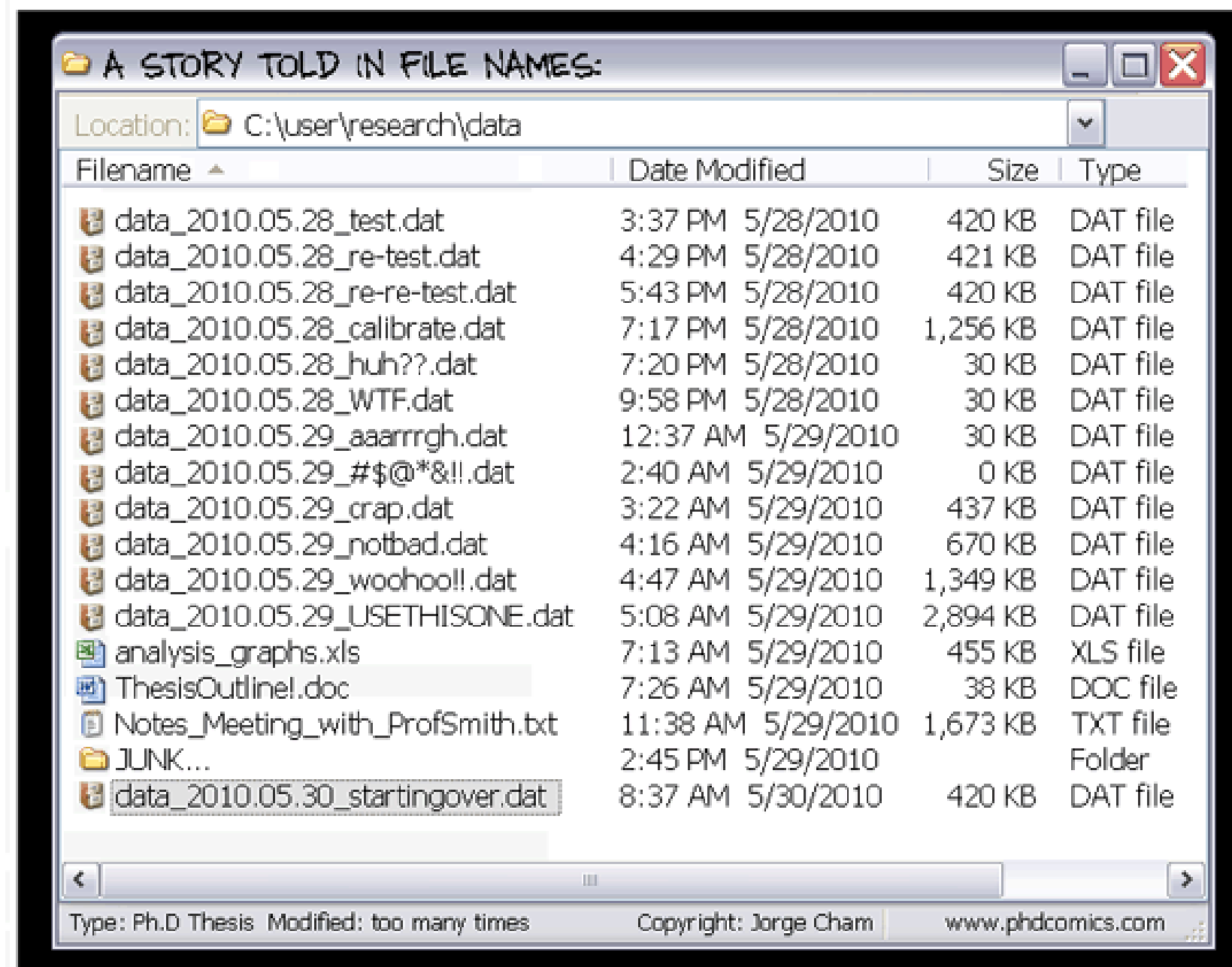
- Two computations on clusters of different scale (small cluster, supercomputer).
Full video: <https://youtube.com/datalad>
- Two full (re-)computations, programmatically comparable, verifiable, reproducible -- on any system with data access

CURRENT AND FUTURE DEVELOPMENTS

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE ... SOMETIMES UNRELIABLE OR THREATENED



THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE ... SOMETIMES UNRELIABLE OR THREATENED



THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE ... SOMETIMES UNRELIABLE OR THREATENED

Mar 2019

ABCD Data Release 2.0 available now on the NIMH Data Archive!

The second annual curated ABCD **data release 2.0** is available now on the **NIMH Data Archive**. Data Release 2.0 includes baseline data on the full participant cohort, ages 9-10 years. This release contains much of the same type of data as in Data Release 1.1, and also includes genotypic data for the first time. Smokescreen genotyping array data are available on almost 11,000 participants. These include common variations, as well as variations associated with addiction, smoking behavior and nicotine metabolism.

ABCD Study data will be released annually. **The next data release will be in early summer 2020** and will include the first longitudinal data from the 6-month and 1-year follow-up assessments.

THE BUILDING BLOCKS OF A SCIENTIFIC RESULT ARE ... SOMETIMES UNRELIABLE OR THREATENED

Mar 2025

www.404media.co/nih-archives-repositories-marked-for-review-for-potential-modification

nda.nih.gov

NIMH Data Archive

NDA | ABCD | CCF | OAI | NIAAA_{DA} | AMP SCZ | HEAL | NBB | PEC

 NDA

[Contribute Data](#)

[Get Data](#)

[Tools](#)

[Webinars & Tutorials](#)

[About Us](#)

[Filter Cart \(0\)](#)

[LOGIN](#)

This repository is under review for potential modification in compliance with Administration directives.



Welcome to the NIMH Data Archive

The National Institute of Mental Health Data Archive (NDA) makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science and discovery. De-identified human subjects data, harmonized to a common standard, are available to qualified researchers. Summary data are available to all.

The NDA mission is to accelerate scientific research and discovery through data sharing, data harmonization, and the reporting of research results.

NIMH common data elements now available:
Go to [NIMH Common Data Elements](#)

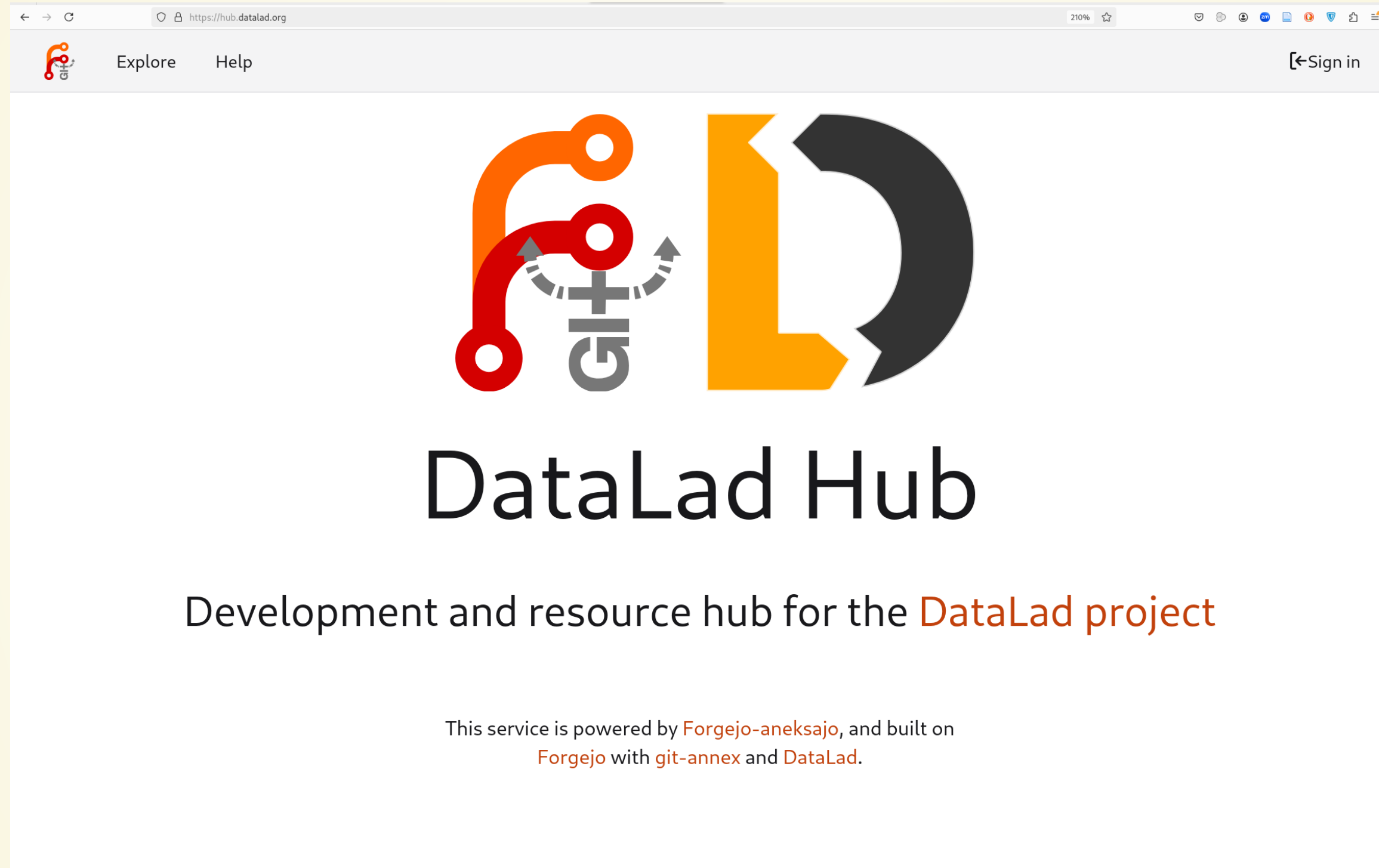
FREEDOM? CHOSE DECENTRALIZATION

- Infrastructure is ephemeral:
 - Change of institutional contracts
 - Change of affiliations
 - Geopolitical developments?
- DataLad datasets are portable
 - Effortless migrations to different Git or data hosting
 - Versioning allows for integrity checks

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

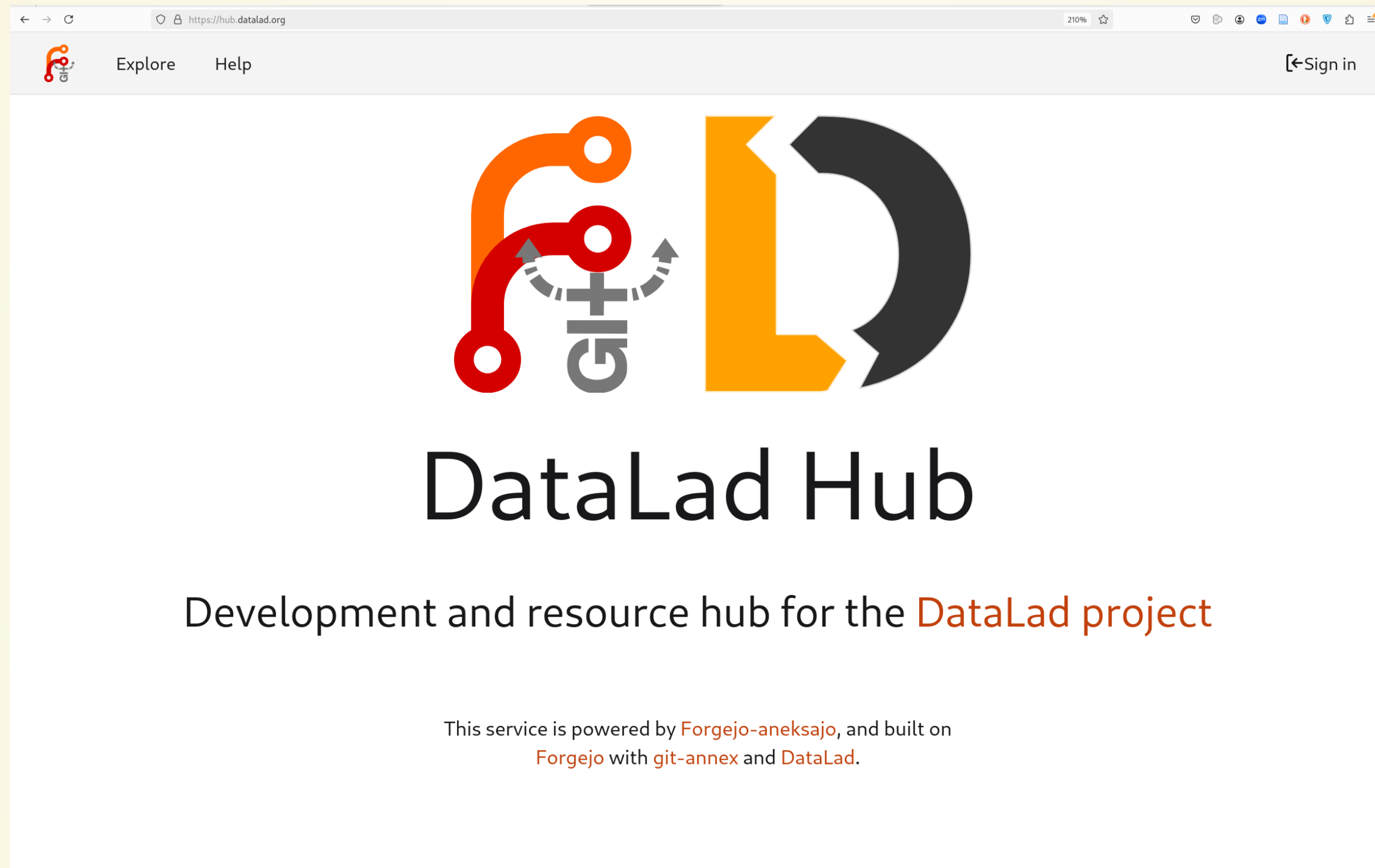
GOING SELF-HOSTED WITH FORGEJO-ANEKSAJO

- Forgejo (forgejo.org): Fork of Gitea



GOING SELF-HOSTED WITH FORGEJO-ANEKSAJO

- Forgejo (forgejo.org): Fork of Gitea
- [Forgejo-aneksajo](#): Forgejo with git-annex support



GOING SELF-HOSTED WITH FORGEJO-ANEKSAJO

- Forgejo (forgejo.org): Fork of Gitea
- [Forgejo-aneksajo](#): Forgejo with git-annex support

← → ↺

https://hub.data-lad.org/naturalistic-imaging

150%

🔔 ⚙️ 👤

🔗

Issues

Pull requests

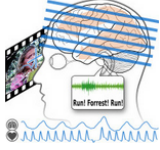
Milestones

Explore

🔔

+

👤



naturalistic-imaging

A collection of datasets on combining naturalistic stimulation paradigms with magnetic resonance tomography and other brain imaging methods. To get a dataset listed here, please get in touch! m.hanke@fz-juelich.de

✉️ m.hanke@fz-juelich.de

🔗

Repositories

26

📁

Projects

📦

Packages

👤

Members

3

👥

Teams

1

⚙️

Settings

🔍

Search repos...

🔍

Filter ▾

Sort ▾

New repository

New migration



narratives

315 healthy participants listened to naturalistic spoken stories during fMRI acquisition. Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C. B., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., Choe, G., Goldstein, A., Halchenko, Y. O., Norman, K. A., & Hasson, U. Narratives: fMRI data for evaluating models of naturalistic language comprehension.

🏷️ audiobook-listening 🏷️ bids 🏷️ functional-mri 🏷️ t1-weighted-mri 🏷️ t2-weighted-mri

Updated 5 days ago

☆ 0 🧑 0



cneuromod-opendata-preprocessed

Preprocessed data derivatives for <https://hub.data-lad.org/naturalistic-imaging/cneuromod-opendata>

🏷️ fmriprip 🏷️ smriprip

Updated 3 weeks ago

☆ 0 🧑 0



Emo-FiLM

ds004892 from OpenNeuro

Updated 2 months ago

🔴 HTML ☆ 0 🧑 0



le-petit-prince-hong-kong

52 healthy Cantonese participants aged 65 years and up listened to excerpts from "The little Prince" in Cantonese during fMRI and EEG acquisition. The dataset also contains detailed information on participants' language history, lifetime experiences, linguistic and cognitive skills, as well as extensive audio and text annotations, such as time-aligned speech segmentation and prosodic features, along with word-by-word predictors from natural language processing (NLP) tools. ds004718 from OpenNeuro

🏷️ 3-tesla-mri 🏷️ 64-channel-eeg 🏷️ audio-annotation 🏷️ audiobook-listening 🏷️ bids 🏷️ eeg 🏷️ flanker-task 🏷️ forward-digit-span-task 🏷️ functional-mri 🏷️ picture-naming-task 🏷️ restingstate-mri 🏷️ t1-weighted-mri 🏷️ text-annotation 🏷️ verbal-fluency-task

Updated 2 months ago

☆ 0 🧑 0



le-petit-prince

49 healthy native English speakers, 35 healthy native Chinese speakers, and 28 healthy native French speakers listened to the English, Chinese or French audiobook of "The little Prince" with auxilliary visual cues. ds003643 from OpenNeuro

🏷️ 3-tesla-mri 🏷️ audiobook-listening 🏷️ bids 🏷️ functional-mri 🏷️ t1-weighted-mri

Updated 2 months ago

🟢 Python ☆ 0 🧑 0

Members

3 >



Teams

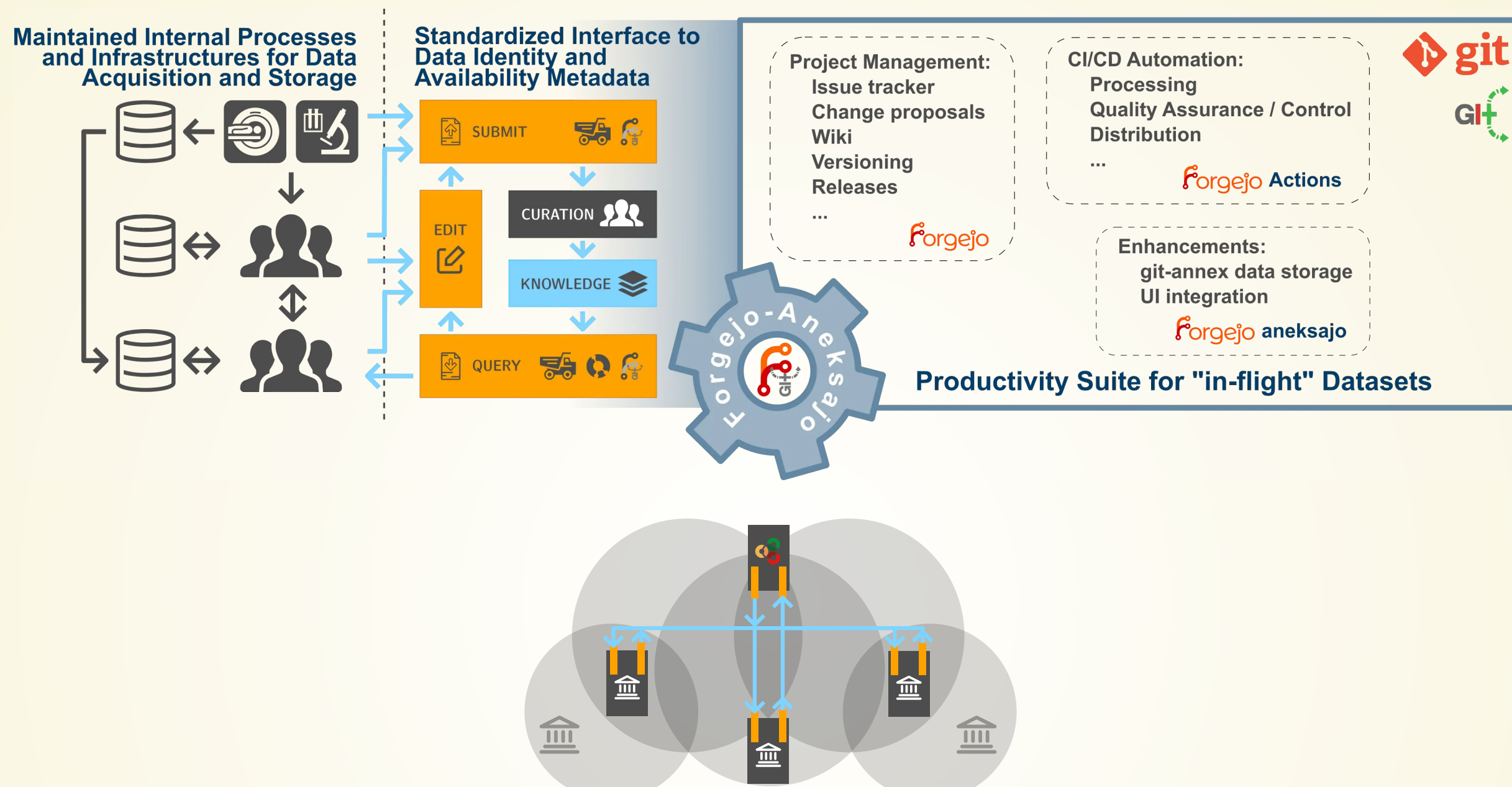
1 >

Owners

3 members · 26 repositories

New team

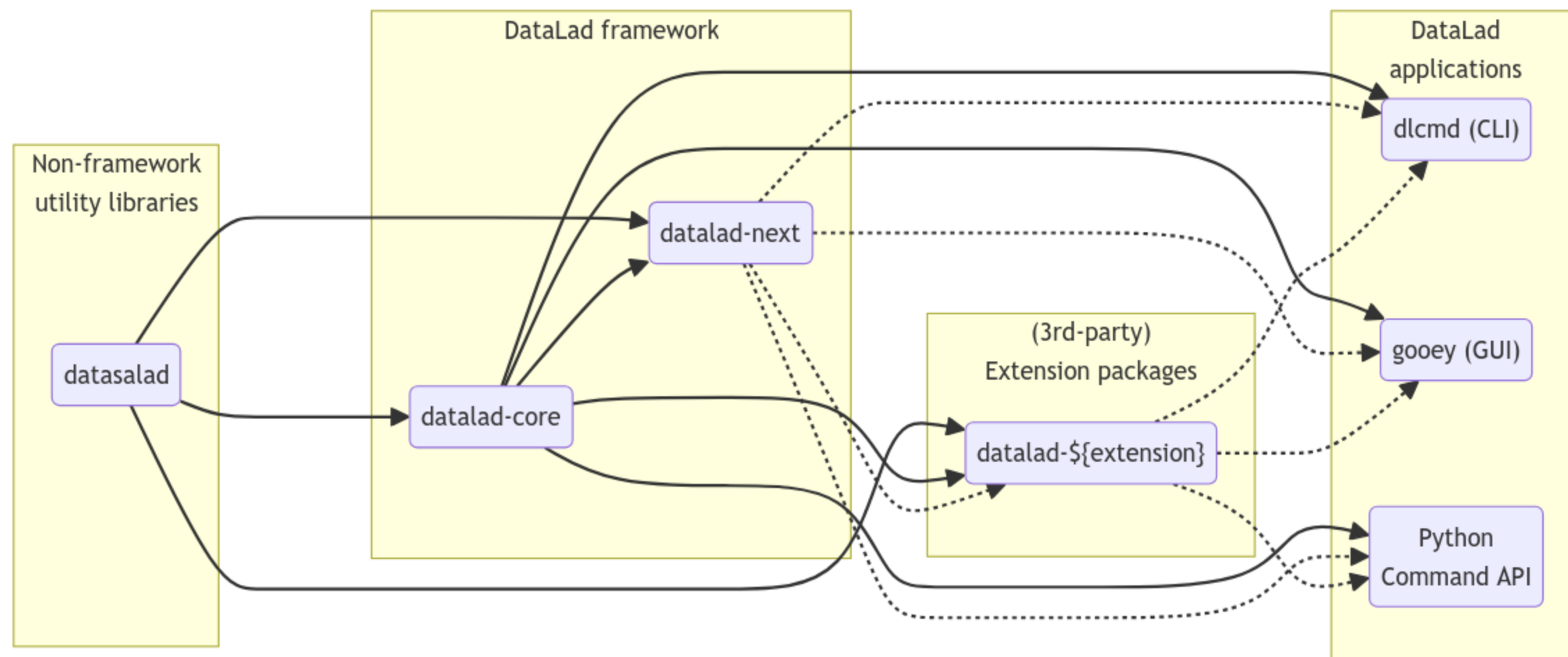
FULL-STACK RDM FOR INDEPENDENT, INTEROPERABLE COLLABORATORS



scale-free organization: consortium, institution, lab, researcher

- maximum contributor benefit
- self-hostable, independently governed solutions, e.g., atris.fz-juelich.de, hub.trr379.de
- minimum contributor cost
- self-contained contributor scopes, not inheriting complexity of others

DEVELOPMENT ROADMAP



JOIN US!

Distribits 2025

- International conference on technologies for distributed data management
- 2 day conference plus single-day Hackathon
- @ Haus der Universität Düsseldorf
- Registration open until May 1st

23.-25. Oct 2025

Düsseldorf

DISTRIBITS.LIVE

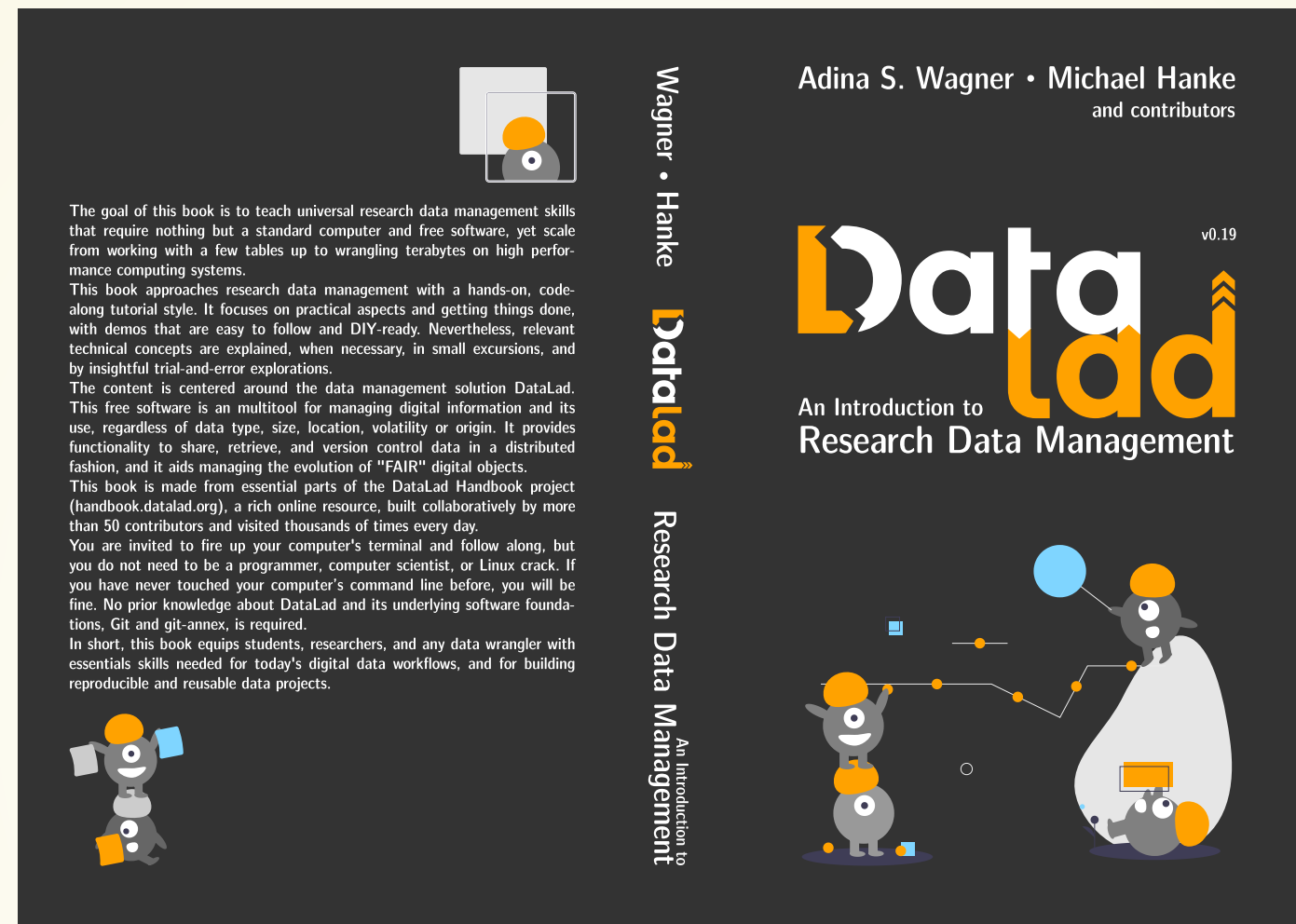
distribits

a6d1c60e7fdf4194320255ae0c8df0ce682c67

DATALAD CONTACT AND MORE INFORMATION

Website + Demos	http://datalad.org
Documentation	http://handbook.datalad.org
Talks and tutorials	https://youtube.com/datalad
Development	http://github.com/datalad
Support	https://matrix.to/#/#datalad:matrix.org
Open data	http://datasets.datalad.org
Mastodon	@datalad@fosstodon.org

EXTENSIVE DOCUMENTATION AND TRAINING MATERIALS



<https://handbook.datalad.org> (or ISBN 979-8857037973)

- **educational materials** on technologies — **targeting researchers**, not developers (executable paper, student supervisor workflow, ...)
- handbook on concepts, workflows, and use cases
- **weekly public (virtual) office hour**

THANKS!



TALK IS CHEAP, SHOW ME THE CODE: GIT VS. DATALAD

DataLad vs Git/Git-annex for modular data management



<https://www.youtube.com/watch?v=Yrg6DgOcbPE>

