

DATA ... LAD

Michael Hanke & Adina Wagner



USE CASES IN DATA STORAGE AND RETRIEVAL

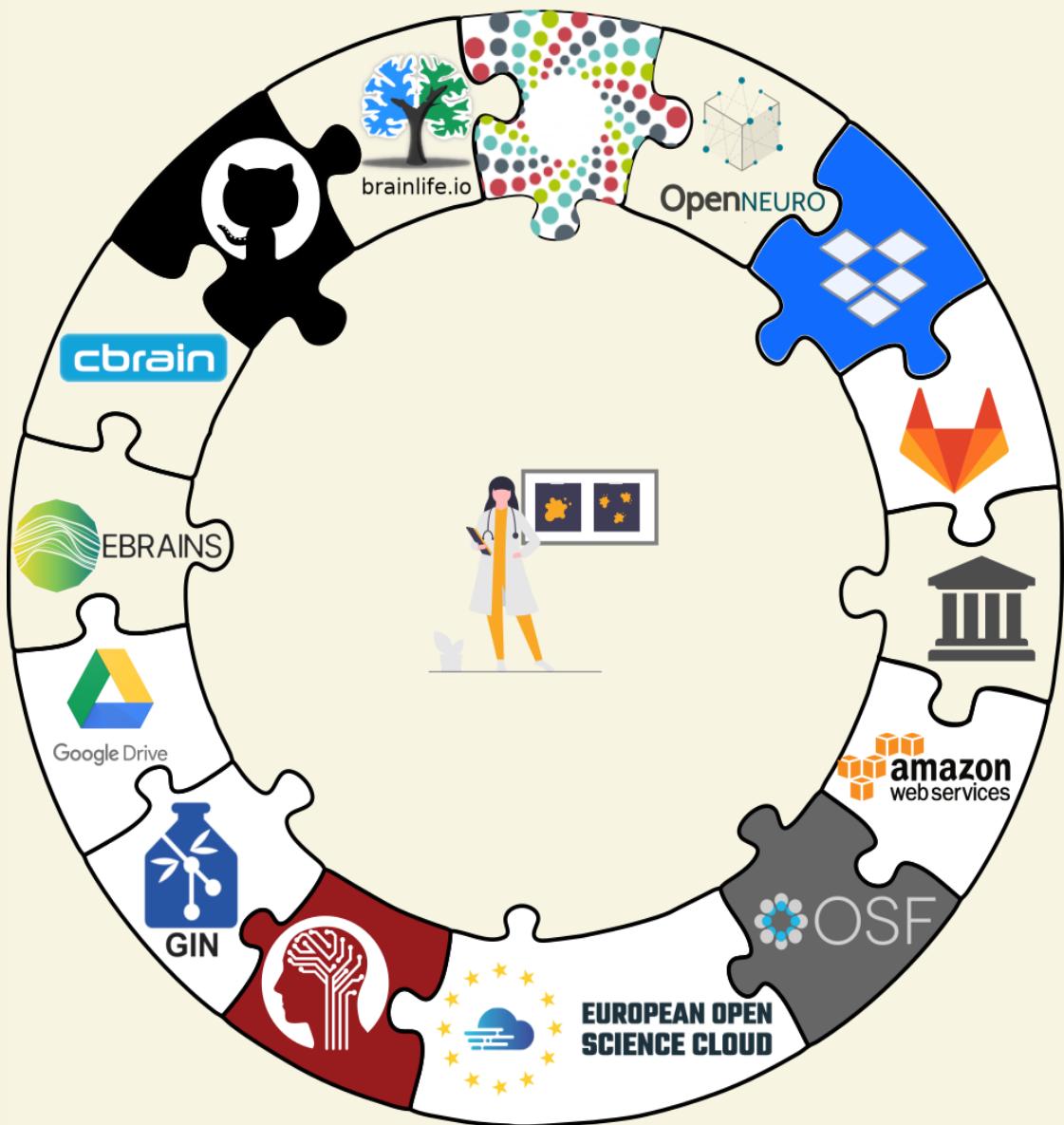
- DataLad: Joint management of digital objects through their entire life cycle

DataLad's features,

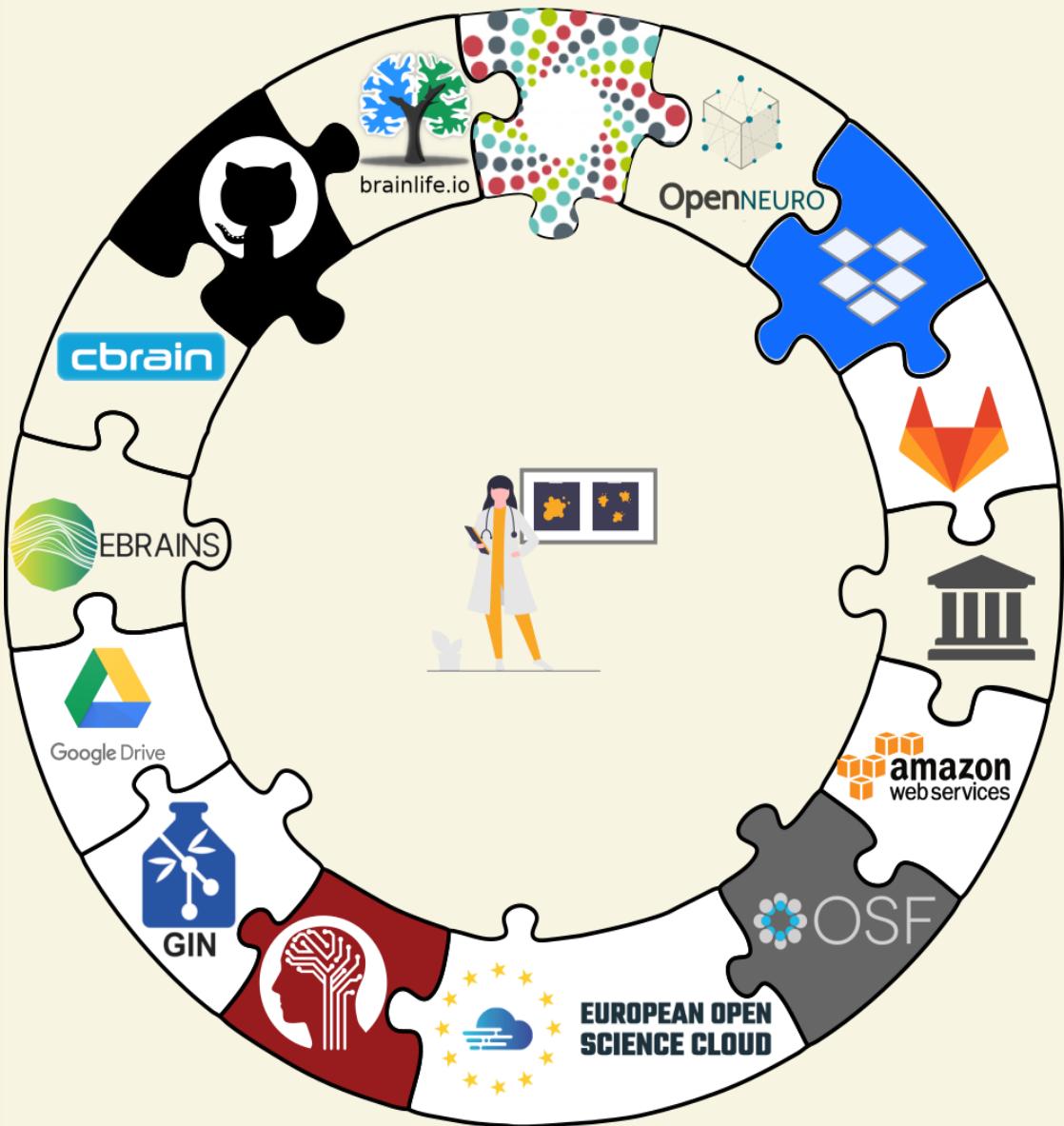
- Version control
- Transport logistics
- Interoperability
- Provenance capture

enable a range of common use cases for storing and retrieving data

DATA SHARING & CONSUMPTION

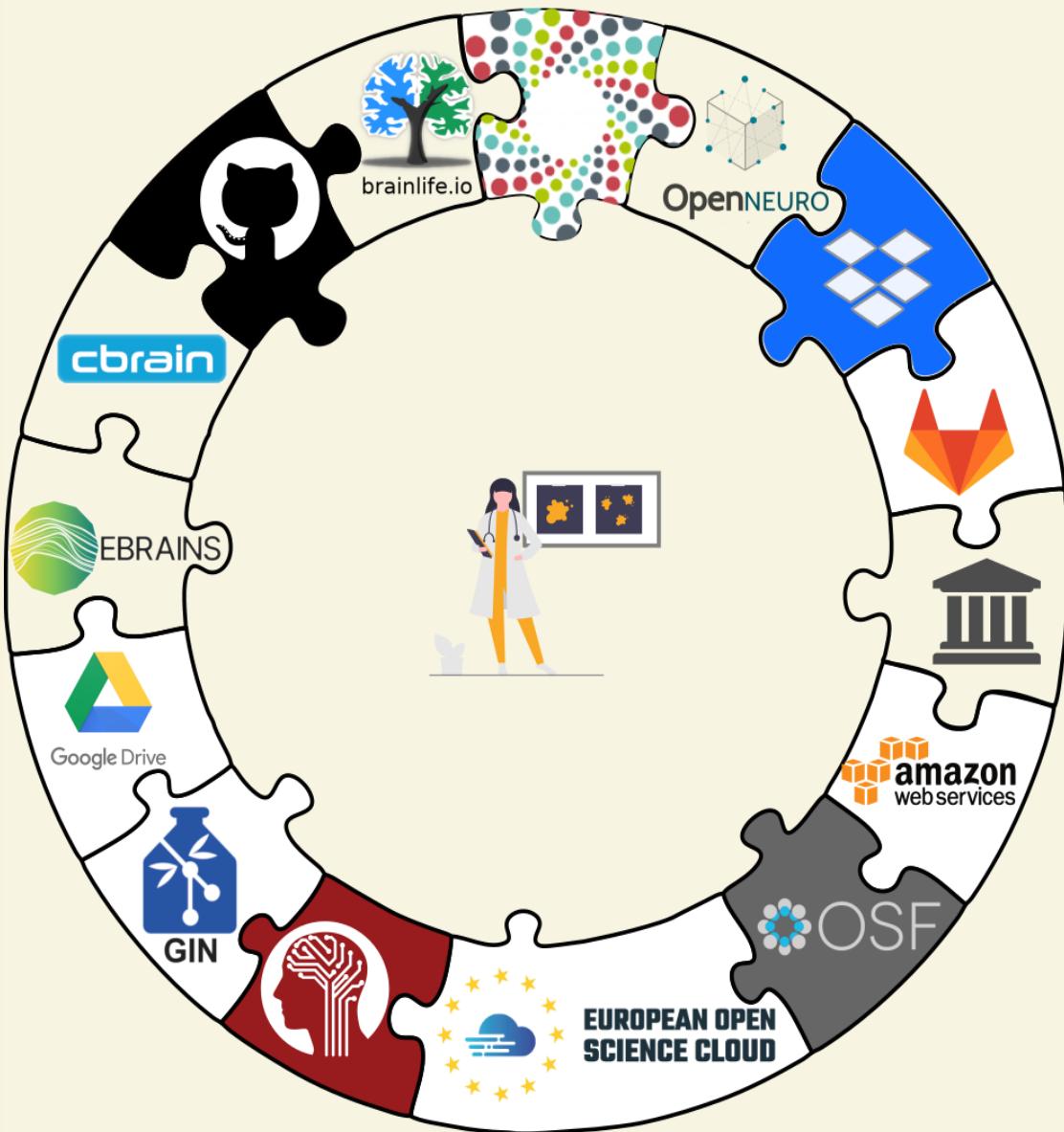


DATA SHARING & CONSUMPTION



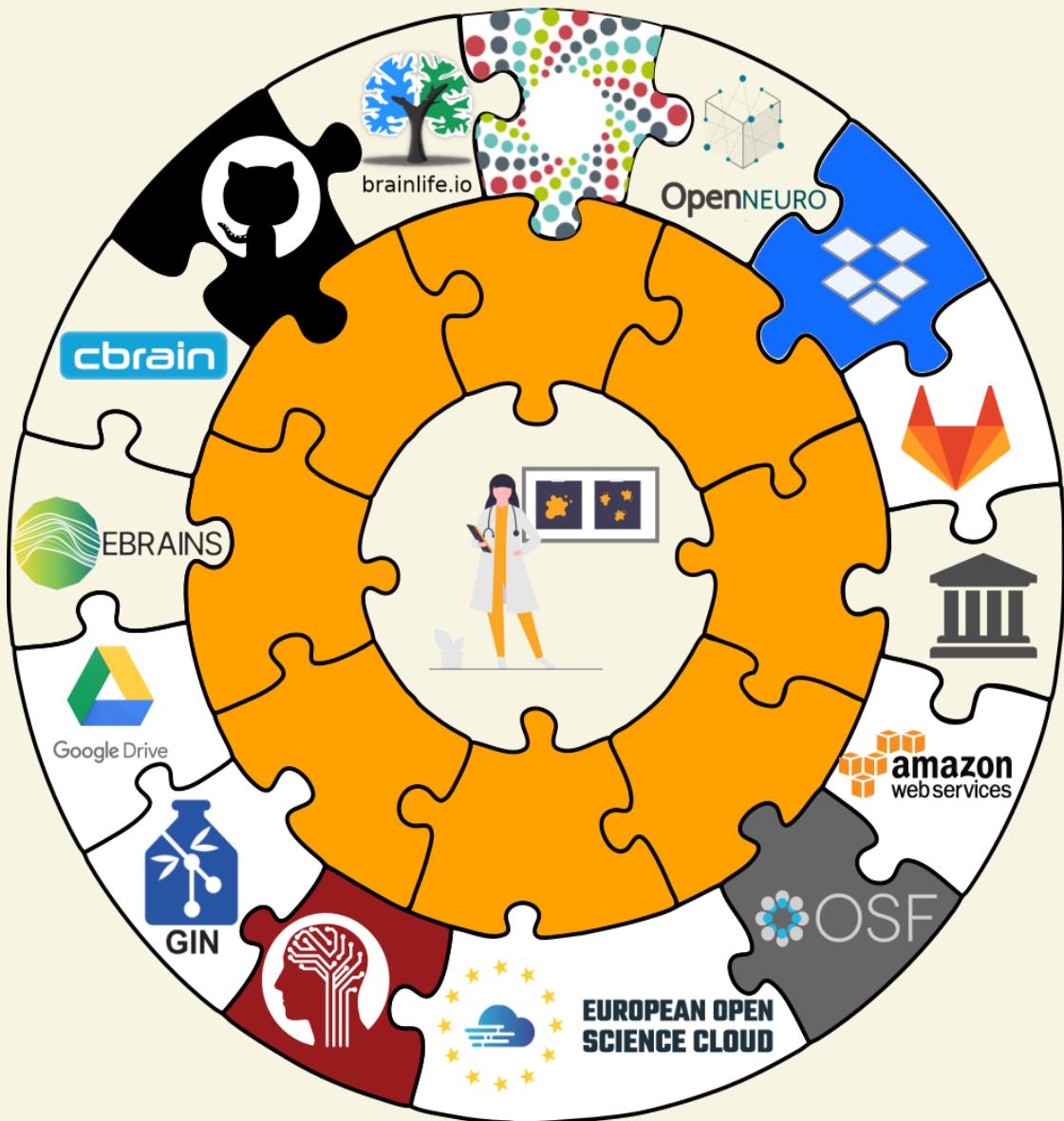
- Many options for sharing and storing data

DATA SHARING & CONSUMPTION



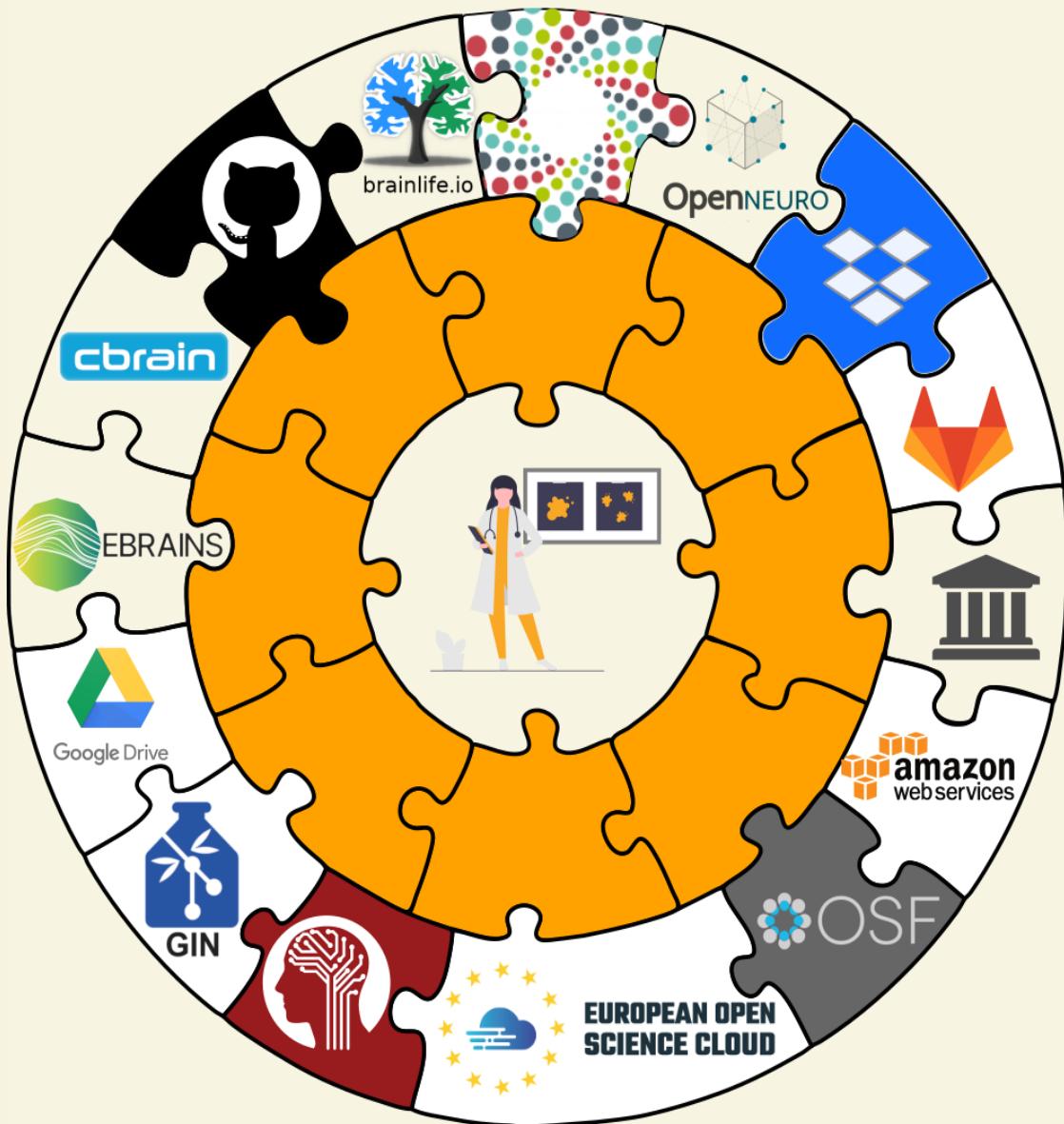
- Many options for sharing and storing data
- Often university-/lab-wide standards, but lack of generic and interoperable workflows

DATA SHARING & CONSUMPTION



- Generic interface to a variety of third party services & storage providers

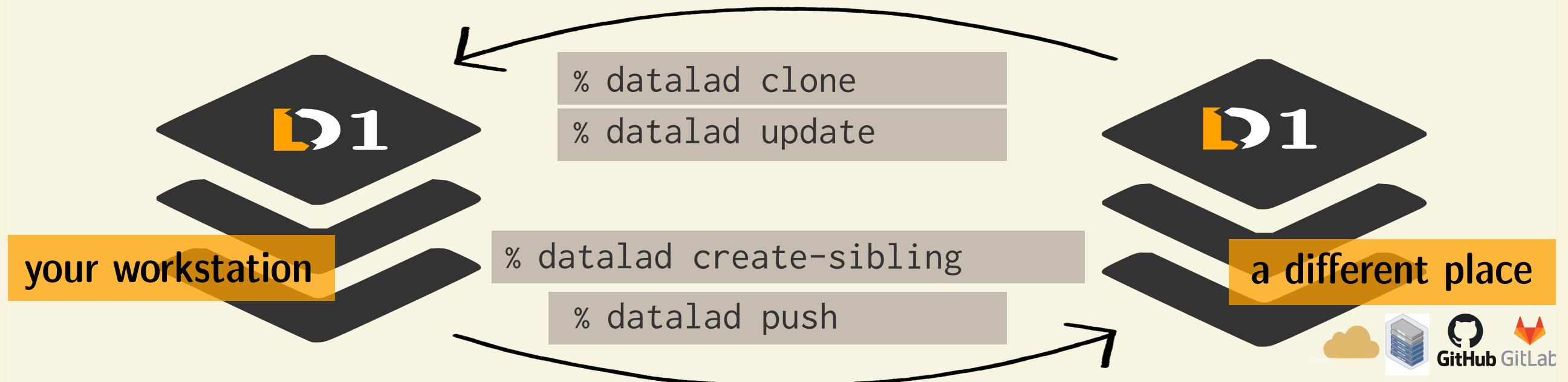
DATA SHARING & CONSUMPTION



- Generic interface to a variety of third party services & storage providers
- Enables interoperable workflows independent of the chosen service

DATA SHARING & CONSUMPTION

Consume existing datasets and stay up-to-date



Create sibling datasets to publish to or update from

DATA SHARING & CONSUMPTION

DATA SHARING & CONSUMPTION

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

DATA SHARING & CONSUMPTION

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

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

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

JavaScript (99%)

Commit	Date
mih Merge pull request #15 from adswa/ENH/README	May 7, 2016
.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 bua fix in converte...	3 years ago

DATA SHARING & CONSUMPTION

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

 [dkp / sub-219_bids_datalad](#)

Watch 1 Star 0 Fork 0

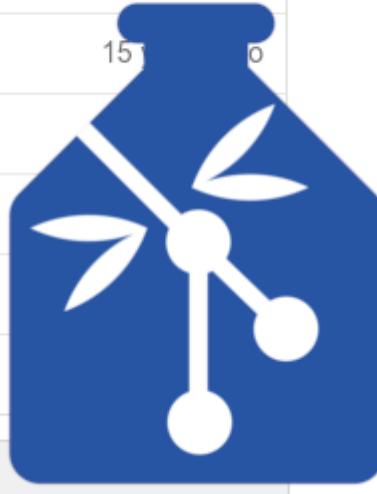
Files Issues 0 Pull Requests 0 Wiki

This is a BIDS MRI dataset generated from sub-219 (me), so no privacy issues. I used heudiconv with the datalad flag to generate it.

7 Commits 4 Branches 0 Releases

Branch: master sub-219_bids... GIN HTTPS SSH git@gin.g-node.org:dkp/sub-219_bids_datalad.zip Download

Commit	Message	Time
root 0ef0b644b0	removed derivatives	3 months ago
.datalad eff691fb27	[DATALAD] new dataset	15 years ago
.heudiconv @ 7cbd6df26 41bf012430	Converted subject 219, session itbs, 1857 dicoms	15 years ago
code eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 years ago
sub-219 41bf012430	Converted subject 219, session itbs, 1857 dicoms	15 years ago
.gitattributes d8d9540de6	Custom .gitattributes	15 years ago
.gitmodules 0ef0b644b0	removed derivatives	3 months ago
CHANGES eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 years ago
README eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 years ago
dataset_description.json eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 years ago
participants.tsv eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 years ago
task-rest_bold.json 41bf012430	Converted subject 219, session itbs, 1857 dicoms	15 years ago
README		

 gin.g-node.org

DATA SHARING & CONSUMPTION

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

The screenshot displays the project page for 'datalad-osf 0.2.0' on the Open Science Framework (OSF). The top navigation bar includes a search bar, help links, and user authentication options. The main content area features the project title, a 'pip install' command, a 'Latest version' button (which is green and indicates the current version), and a release date of July 17, 2020. A brief description of the extension follows. The bottom section contains a navigation menu with 'Project description' selected, and a detailed project description summary.

datalad-osf 0.2.0

pip install datalad-osf

Latest version

Released: Jul 17, 2020

DataLad extension to interface with the Open Science Framework (OSF)

Navigation

Project description

Project description

Release history

DataLad-OSF: Opening up the Open Science Framework for DataLad

DATA SHARING & CONSUMPTION

DATA SHARING & CONSUMPTION

- Special Case: Central data management and archival system

DATA SHARING & CONSUMPTION

- Special Case: Central data management and archival system

The screenshot shows a web-based interface for managing projects, specifically for the Institute for Neuroscience and Medicine - Brain and Behaviour (INM-7). The top navigation bar includes links for 'Projects' and 'More', along with user profile and search icons. The main header displays the group name 'INM7' and its Group ID (309), with an option to 'Leave group'. A prominent green button labeled 'New project' is visible. Below this, a welcome message and a 'Read more' link are present. The interface features a search bar and dropdown menus for filtering projects by 'Last created' or 'Search by name'. A tabbed section at the bottom allows switching between 'Subgroups and projects', 'Shared projects', and 'Archived projects'. The main content area lists several projects under 'Subgroups and projects': 'vbc' (protected), 'Training' (owned by the user), 'webtools' (protected), 'tools' (with a note about sharing across INM-7), and 'AppliedMachineLearning' (with a note about being part of the Applied Machine Learning group).

INM7

Group ID: 309 | [Leave group](#)

[New project](#)

Welcome to the public GitLab-Repo of the Institute for Neuroscience and Medicine - Brain and Behaviour (INM-7)
[Read more](#)

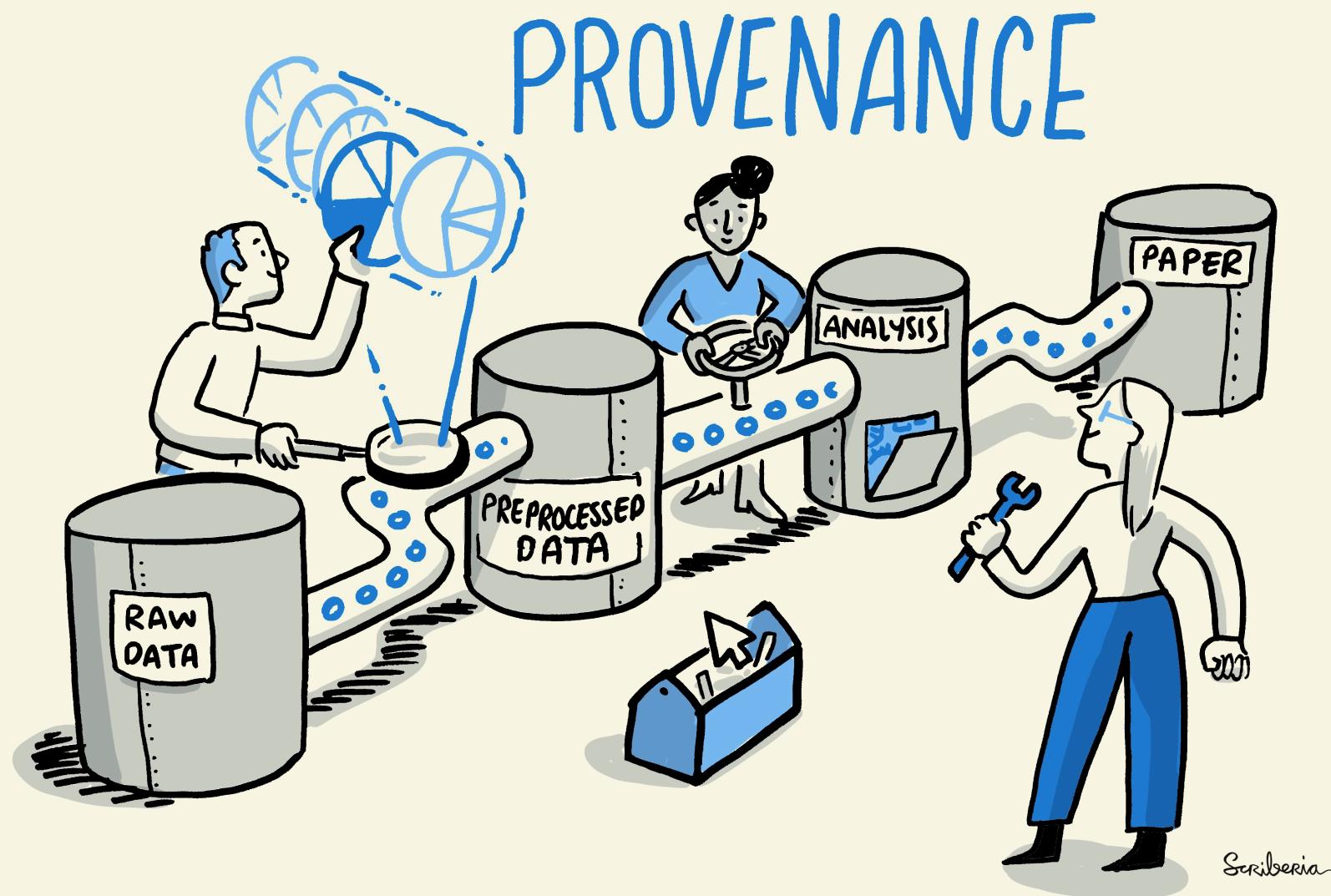
Search by name

Last created

Subgroups and projects Shared projects Archived projects

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

BEYOND DATA: PROVENANCE



BEYOND DATA: PROVENANCE

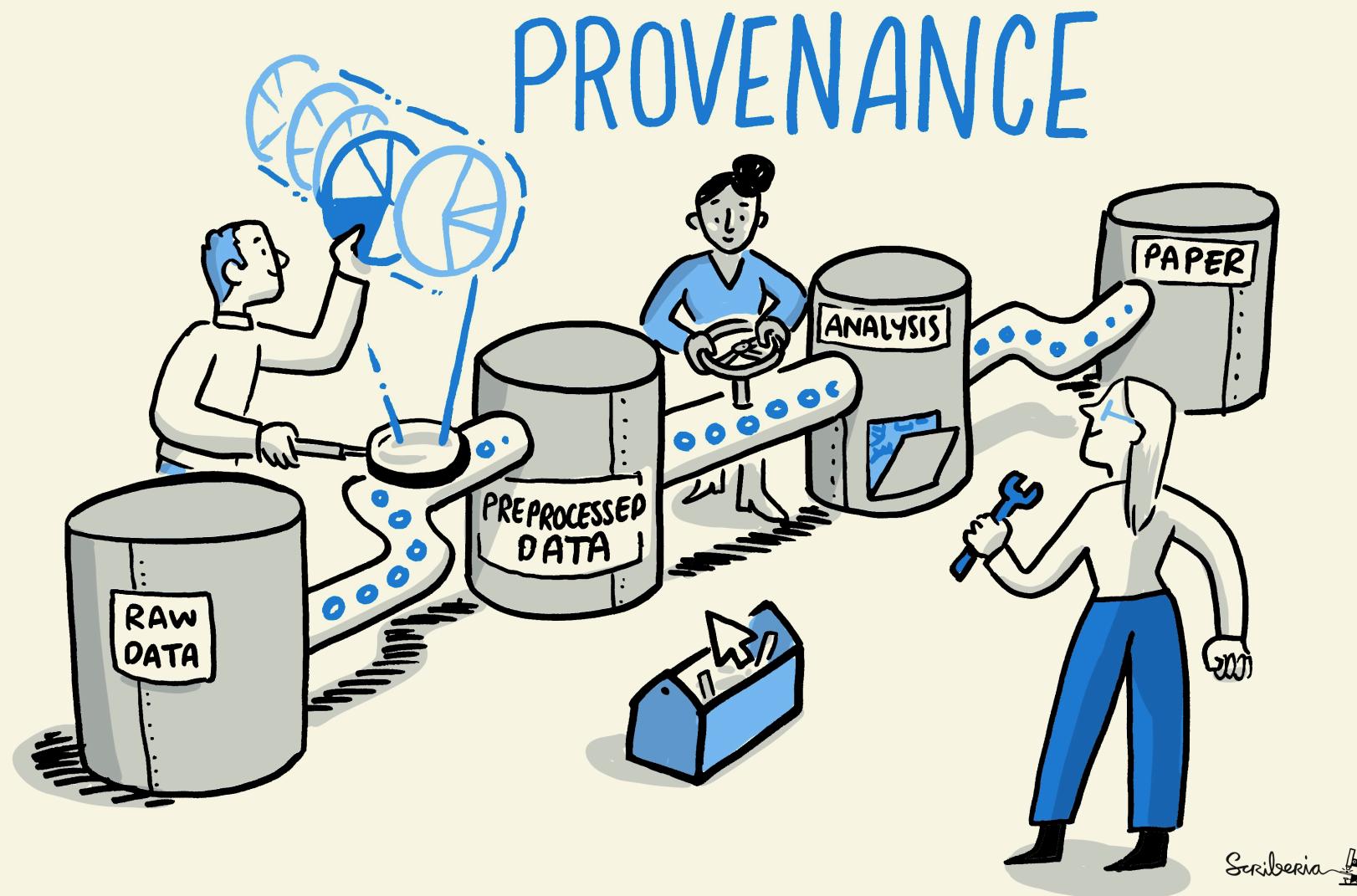


Image credit: CC-BY Scriberia and The Turing Way

- Create, share & use provenance of research objects

BEYOND DATA: PROVENANCE

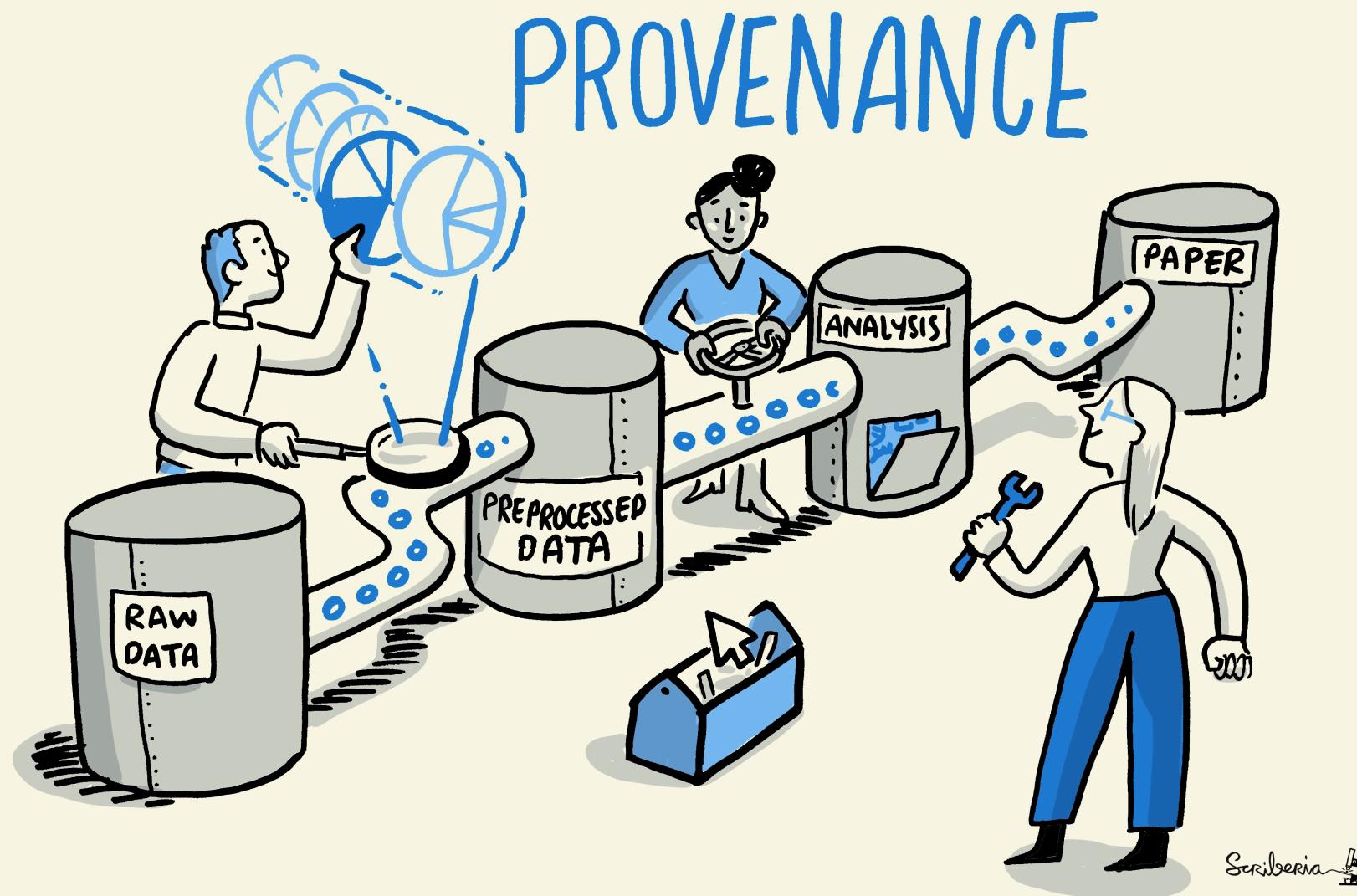


Image credit: CC-BY Scriberia and The Turing Way

- Create, share & use provenance of research objects
- Link code, data, software container and re-executable analysis records

BEYOND DATA: PROVENANCE

BEYOND DATA: PROVENANCE

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

BEYOND DATA: PROVENANCE

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

The screenshot shows a GitHub repository page for 'psychoinformatics-de / paper-remodnav'. The repository has 8 pull requests, 2 stars, and 2 forks. The 'Code' tab is selected, showing the master branch with 7 branches and 2 tags. The commit history lists 429 commits from various contributors, including 'adswa' and 'remodnav @ d289118'. The repository contains files like code, data, img, .gitignore, .gitmodules, COPYING, Makefile, README.md, main.tex, references.bib, results_def.tex, spbasic bst, svglov3.clo, and svjour3.cls. The README.md file describes REMoDNaV: Robust Eye Movement. The 'About' section provides a link to the manuscript at <https://doi.org/10.1101/619254>. The 'Releases' section shows 2 tags. The 'Contributors' section lists Adina Wagner, Michael Hanke, and Asim H ... The 'Languages' section shows TeX at 79.4%, Python at 20.0%, and Makefile at 0.6%.

psychoinformatics-de / **paper-remodnav**

Code Issues Pull requests Actions Projects Wiki Security Insights

master 7 branches 2 tags Go to file Add file Code

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

code Merge pull request #12 from psychoinformatics-de/finalr... 4 months ago

data Point to latest label dataset 17 months ago

img Label panels in flowchart 6 months ago

remodnav @ d289118 Update remodnav with latest test dataset 17 months ago

.gitignore Prevent permanent rebuilds of the figures 17 months ago

.gitmodules Add and use studyforrest raw eyegaze dataset as direct... 17 months ago

COPYING Declare CC-BY license 17 months ago

Makefile Simplify reference management 7 months ago

README.md Fix installation instructions 4 months ago

main.tex Minor edits as suggested by reviewer 2 5 months ago

references.bib velocity versus dispersion-based: cite a few algorithms ... 7 months ago

results_def.tex Put generated results back into Git 8 months ago

spbasic bst Basic switch to new layout 2 years ago

svglov3.clo Basic switch to new layout 2 years ago

svjour3.cls Basic switch to new layout 2 years ago

README.md

REMoDNaV: Robust Eye Movement

About

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

Readme

CC-BY-4.0 License

Releases

2 tags

Create a new release

Packages

No packages published

Publish your first package

Contributors 3

adswa Adina Wagner

mih Michael Hanke

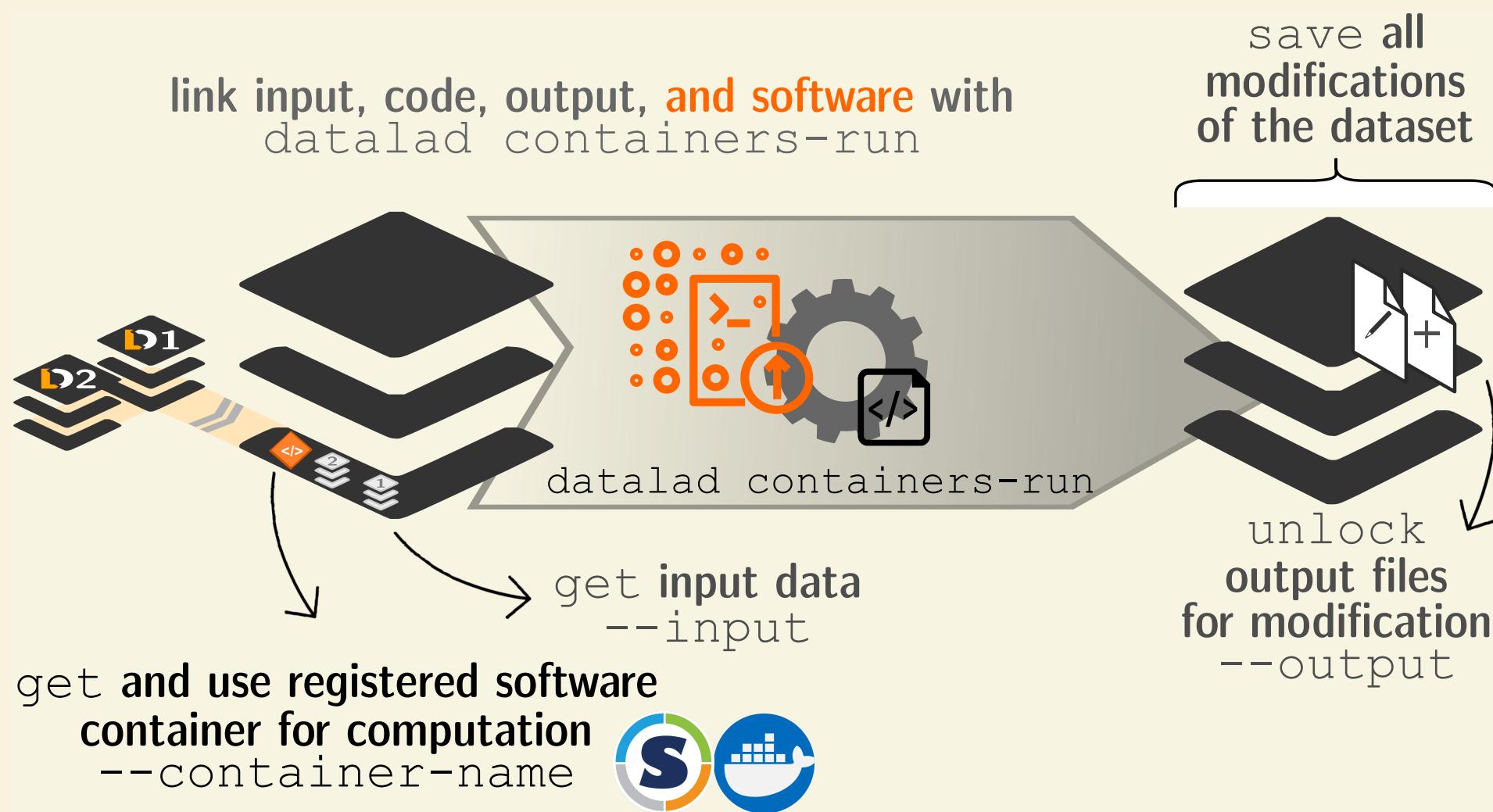
ElectronicTeaCup Asim H ...

Languages

TeX 79.4% Python 20.0%

Makefile 0.6%

CONTAINERIZED WORKFLOWS



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": "92e1faa-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)
...
...
```

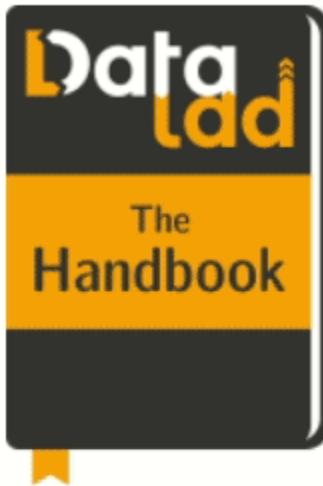
- Aid with the reproducibility of a result and verify it (via content hash)
- Use complete capture and automatic re-computation as alternative to storage and transport

CONTAINERIZED WORKFLOWS

```
(master) adina@juseless in /data/group/psyinf/HCP_structural on git:master
> datalad containers-run -n .tools/fmriprep \
  --input '.source/sub-170631' \
  --output '.' \
  ".source . participant --participant-label 170631 --fs-license-file .tools/license.txt --skip-bids-validation --anat-only -w /tmp" !
```

LARGE-SCALE, CONTAINERIZED WORKFLOWS

A screenshot of a web browser window displaying the DataLad Handbook. The title bar shows several tabs, including "The Handbook", "Data storage an", "DataLad for ML", "Reproducible", "3.3. Walk", "datalad-data", "Manuscript:", "favicon.png (PN)", "Search | unD", and a "+" icon. Below the title bar are standard browser controls for back, forward, search, and refresh. The main content area features the DataLad logo, a "Table of Contents" sidebar with a "Star" button and a "53" badge, and the main article content.



[3.3. Walkthrough: Parallel ENKI preprocessing with fMRIprep](#)

- [3.3.1. The analysis](#)
 - [3.3.1.1. Starting point: Datasets for software and input data](#)
 - [3.3.1.2. Analysis dataset setup](#)
 - [3.3.1.3. Workflow script](#)
 - [3.3.1.4. Job submission](#)
 - [3.3.1.5. Merging results](#)
 - [3.3.1.6. Summary](#)

3.3. Walkthrough: Parallel ENKI preprocessing with fMRIprep

The previous section has been an overview on parallel, provenance-tracked computations in DataLad datasets. While the general workflow entails a complete setup, its usually easier to understand it by seeing it applied to a concrete usecase. Its even more informative if that usecase includes some complexities that do not exist in the “picture-perfect” example but are likely to arise in real life. Therefore, the following walkthrough in this section is a write-up of an existing and successfully executed analysis.

3.3.1. The analysis

The analysis goal was standard data preprocessing using [fMRIprep](#) on neuroimaging data of 1300 subjects in the [eNKI](#) dataset. This computational task is ideal for parallelization: Each subject can be preprocessed individually, each preprocessing takes between 6 and 8 hours per subject, resulting in $1300 \times 7\text{h}$ of serial computing, but only about 7 hours of computing time when executed completely in parallel, and fMRIprep is a containerized pipeline that can be pointed to a specific subject to preprocess.

CONCEPTS: DATA SHARING WITHOUT PRIVACY BREACH

- Expose (sensitive) file content versus anonymized metadata on a per-file basis



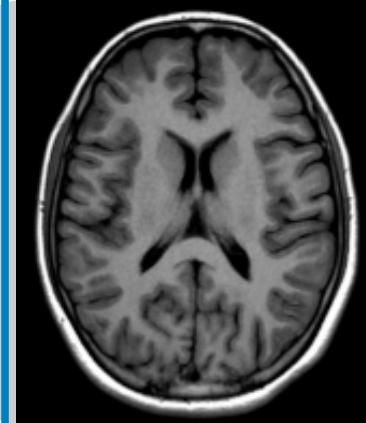
Exposed (anonymized)

VERSION HISTORY	FILE NAMES
MD5E-s3016749--64364d53b5f41fbe5bccfe5e.jpg	patient_record/ photo.jpg
MD5E-s3016749--8ccc1989f2ab223874428092.json	insurance.json
SALTED FILE CONTENT HASHES	visit/
MD5E-s932874--ad9e369d8c908feb6d3fe84e.json	1.json
MD5E-s762273--1ec751b8cb432d16f6a39e64.json	2.json
IMAGING	imaging/
MD5E-s30983443--30dc89ecfd5dd546f8ccc19.img	brain/20031.img
MD5E-s73826423--9432d12b1581f220c8b75953.img	thorax/20201.img

METADATA

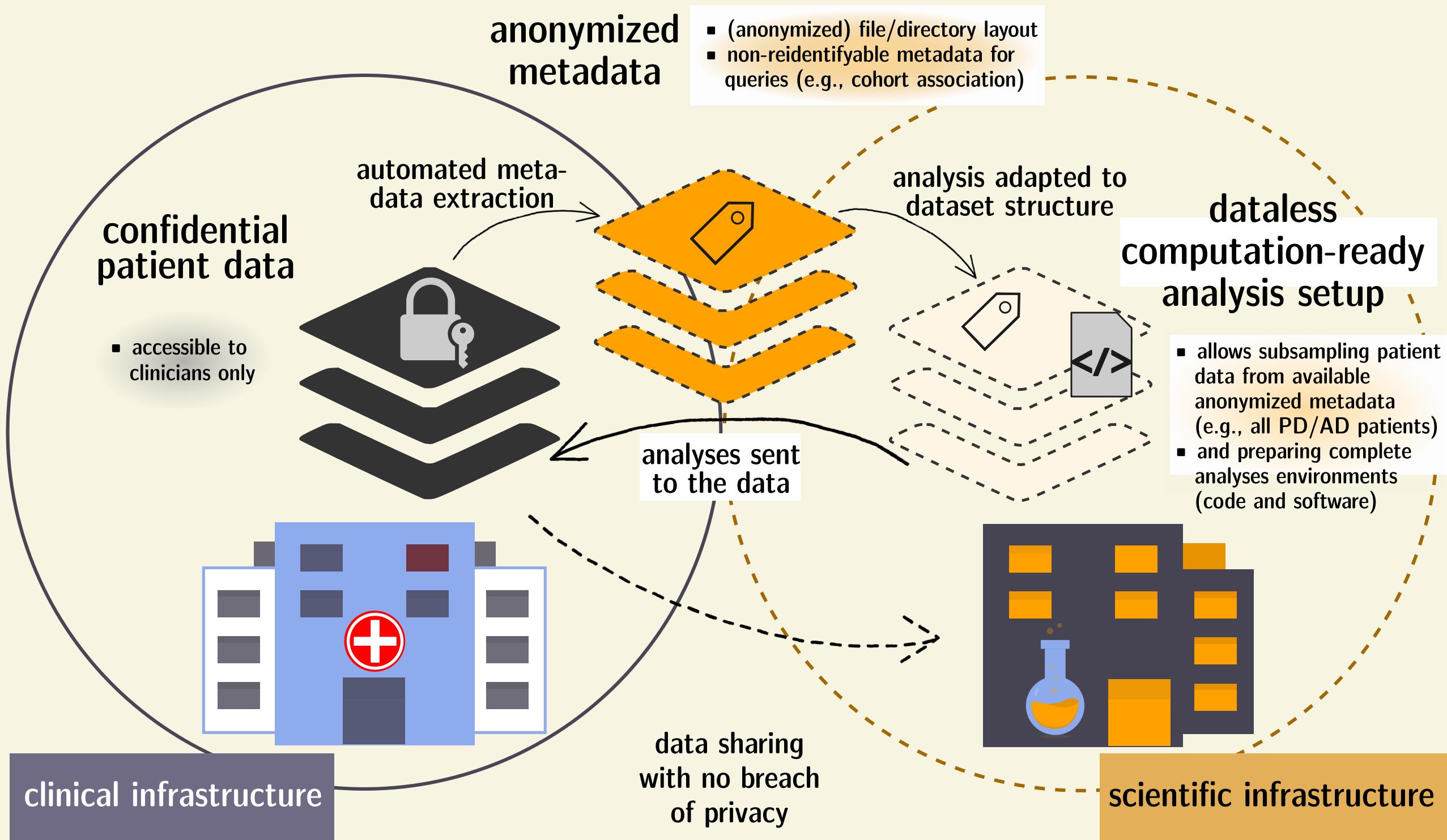
```
{"type": "patient", "id": "122bf-342-422", "ANON IDS": "imaging": ["brain", "thorax"], "sex": "male", "diagnosis": ["F51", "G12", "G52"], "age": 52...}
```

Tracked

PRISTINE INFORMATION
 <pre>{"account": "293924", "id": "a399b30033d"...}</pre>
<pre>{"date": "2020-04-21", "diagnosis": "...", "notes": "..." ...}</pre>
 
 Authorized/local access only

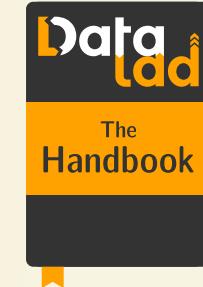
CONCEPTS: COMPUTATION TO DATA

- Expose sufficient anonymized metadata to allow code development, but don't share raw data, only results



FIND OUT MORE

More use cases & 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