

RDM IN SYSTEMS NEUROSCIENCE: CHALLENGES AND BEST PRACTICES

Adina Wagner

 mas.to/@adswa



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



Institute for Experimental Psychology,
HHU Düsseldorf

Slides: files.inm7.de/adina/talks/html/fdm_nrw_hhu.html
DOI; 10.5281/zenodo.10122803

PERKS OF BEING A NEUROSCIENTIST...

A growing culture of open data



WIKIPEDIA

The Free Encyclopedia

Main page

Contents

Current events

Random article

About Wikipedia

Contact us

Donate

Contribute

Help

Community portal

Recent changes

Upload file

Tools

What links here

Related changes

Special pages

Permanent link

Page information

Cite this page

Wikidata item

Print/export

Download as PDF

Printable version

Languages

Add links

Article

Talk

Read

Edit

View history

Search Wikipedia



Join the WPWP Campaign to help improve Wikipedia articles with photos and win a prize

X

List of neuroscience databases

From Wikipedia, the free encyclopedia

A number of online [neuroscience](#) databases are available which provide information regarding [gene expression](#), [neurons](#), [macroscopic](#) brain structure, and [neurological](#) or [psychiatric](#) disorders. Some databases contain descriptive and numerical data, some to brain function, others offer access to 'raw' imaging data, such as [postmortem](#) brain sections or 3D [MRI](#) and [fMRI](#) images. Some focus on the human brain, others on non-human.

As the number of databases that seek to disseminate information about the structure, development and function of the brain has grown, so has the need to collate these resources themselves. As a result, there now exist databases of neuroscience databases, some of which reach over 3000 entries.^[1]

Contents [hide]

1

Neuroscience databases

2

Databases of neuroscience databases

3

Neuroscience article aggregators

4

See also

5

References

Neuroscience databases

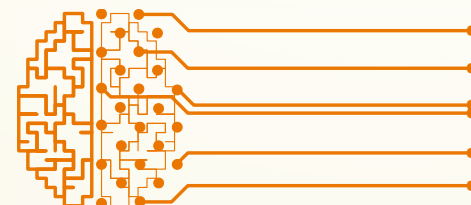
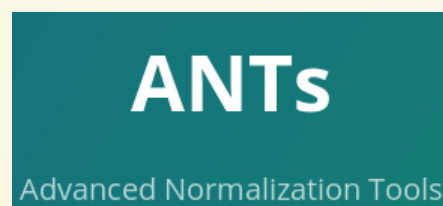
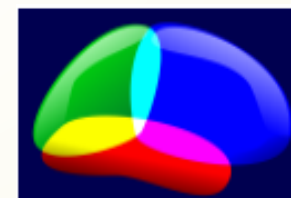
[edit]

Name	Description	Organism	Level (gene, neuron, macroscopic)	Data (MRI, fMRI, images, descriptive, numerical)	Disorder	Register to view data?	Ref.
Allen Brain Atlas	Atlas, stained sections from brains showing development and gene expression	Mouse, Human	Macroscopic, Gene	Images	Healthy	No	[2]
Alzheimer's Disease Neuroimaging Initiative (ADNI)	Structural MRI images	Human	Macroscopic	MRI datasets	Healthy and Alzheimer's Disease	Yes	[3]
Big Brain	3D reconstruction of complete brain from cell-body stained histology sections at 20 micron isotropic resolution	Human	Microscopic	Images	Healthy	No	[4]
BIRN fMRI and MRI data	fMRI, MRI scans and atlases for human and mouse brains	Mouse, Human	Multilevel: brain regions, connections, neurons, gene expression patterns	MRI datasets, fMRI datasets	Healthy, Elderly	No	
Bipolar Disorder Neuroimaging Database	Meta-analysis and database of MRI studies	Human	Macroscopic	Descriptive, numerical	Bipolar Disorder	No	[5]
Brain Architecture Management System	Online resource for information about neural circuitry	Rat, mouse, human	Multilevel: brain regions, connections, neurons, gene expression patterns	Descriptive, numerical	Healthy	No	
Brain Cloud	Gene expression in the human prefrontal cortex	Human	Gene expression patterns	Descriptive, numerical	Healthy	No	
Brain-Development.org	Structural MRI images and Atlases	Human	Macroscopic	MRI datasets	Fetuses, healthy and prematurely born neonates	Yes	[6]
Braingraph.org	Braingraphs computed from the Human Connectome Project data	Human	Macroscopic, up to 1015 nodes	directed and undirected graphs in anatomically annotated GraphML format	Healthy	No	[7]

Image credit: https://en.wikipedia.org/wiki/List_of_neuroscience_databases

PERKS OF BEING A NEUROSCIENTIST...

A large and growing amount of open source software



... and many more!

PERKS OF BEING A NEUROSCIENTIST...

Many readily available, often free, sometimes FOSS, services for data storage and collaboration

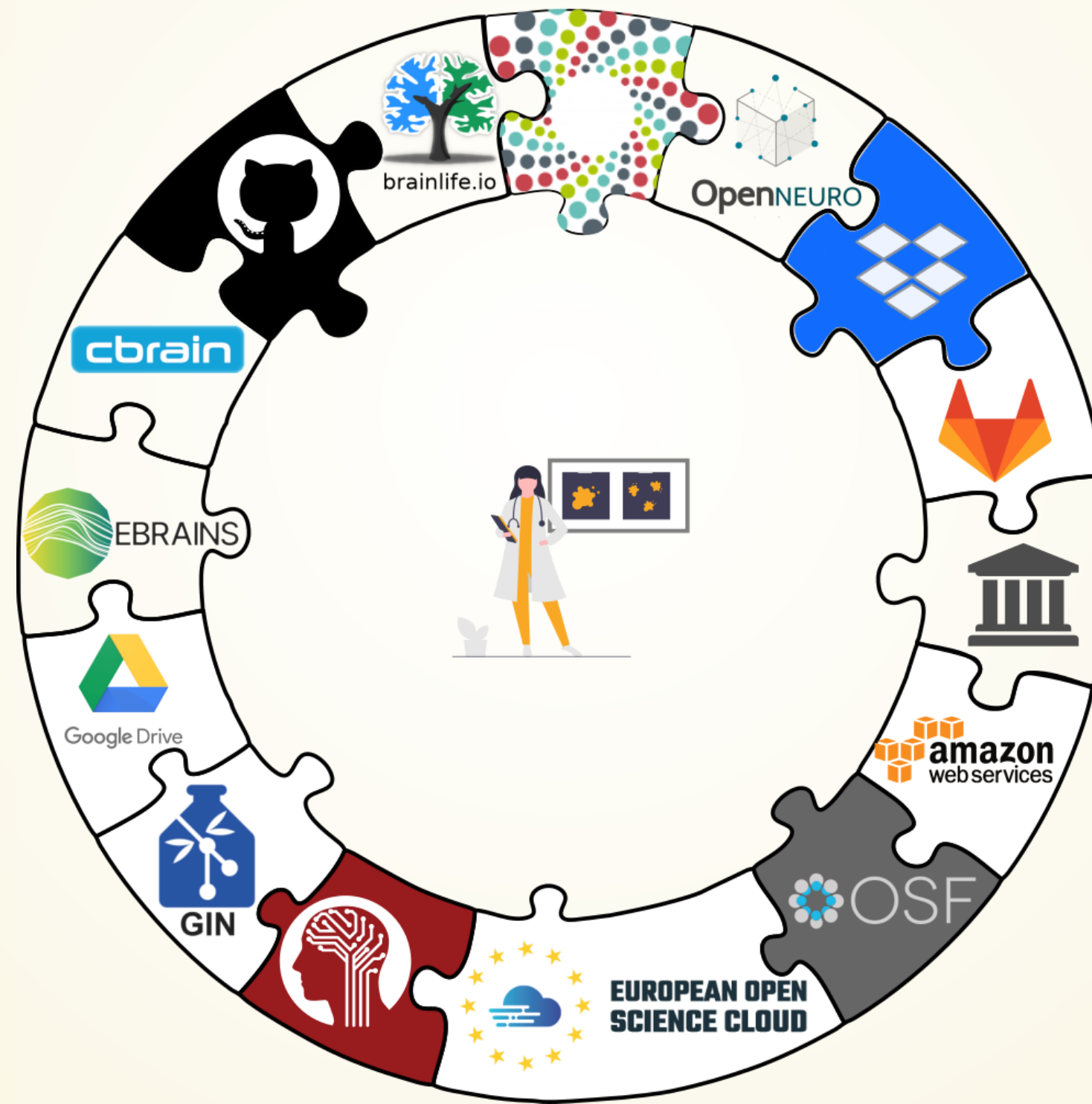
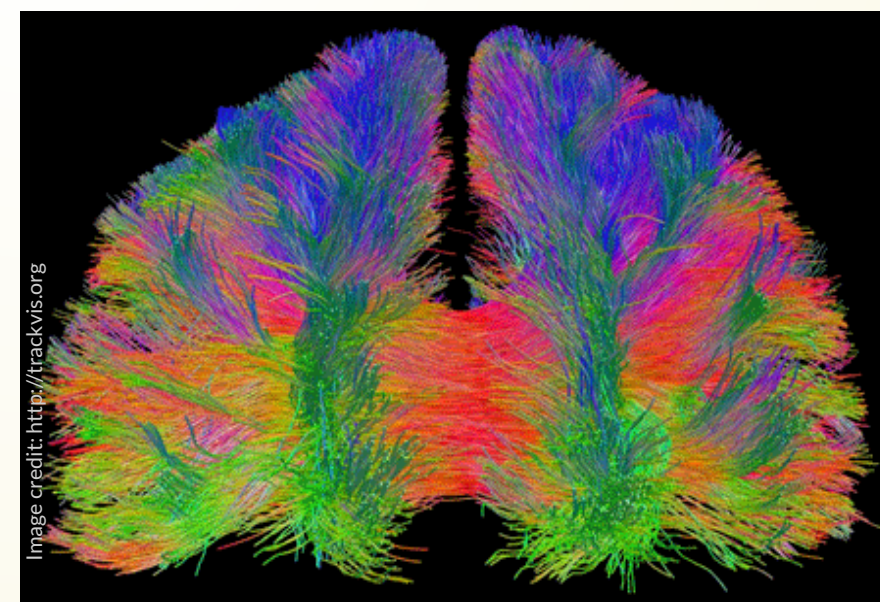
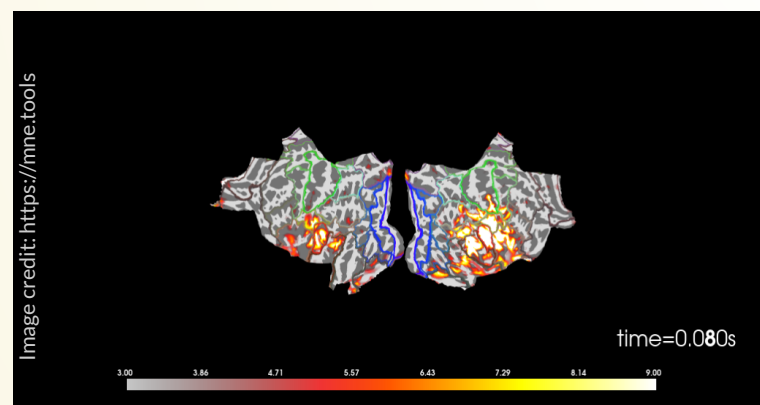
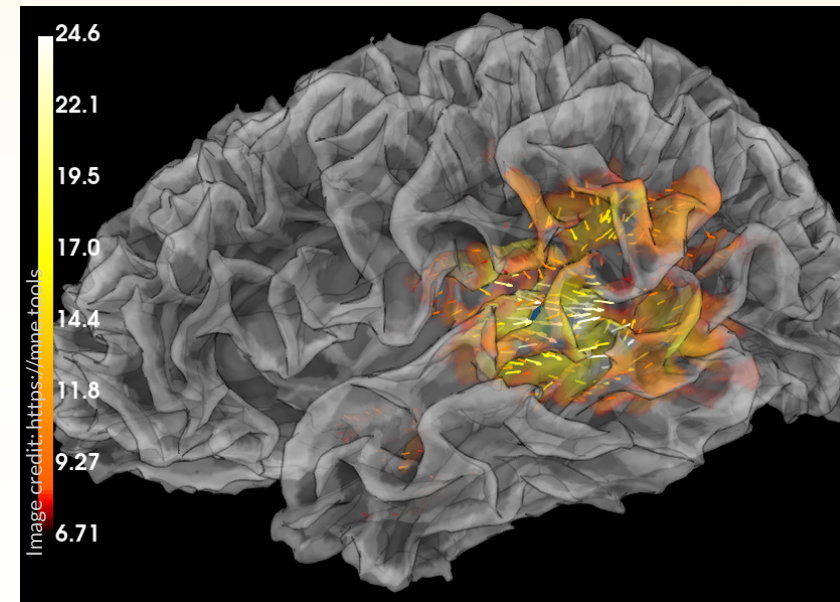
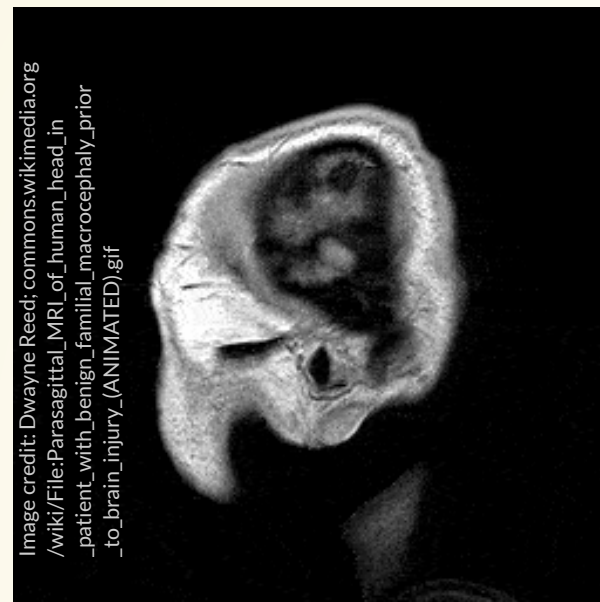


Image credit: <https://www.incf.org/resources/sbps>

PERKS OF BEING A NEUROSCIENTIST...

Work on fascinating questions with fascinating data



THE GOOD NEWS, THE BAD NEWS

Interesting data	... that's subject to GDPR, making storage, analysis, and sharing difficult
Data analysis	Reproducibility is threatened by intransparent, multi-stepped analyses ¹ & unstable results across software versions ² . Dataset sizes exceed computational capabilities (e.g., HCP project: ~100TB, 1.2k people).
Data sharing	Heterogenous distribution and updating, many scientists lack data management skills
Collaboration	Few interoperable workflows across institutes, rather: isolated solutions

¹ Botvinik-Nezer et al., 2020: Variability in the analysis of a single neuroimaging dataset by many teams

² Kiar et al., 2020: Comparing perturbation models for evaluating stability of neuroimaging pipelines



- Domain-agnostic data management tool (command-line + graphical user interface), built on top of [Git](#) & [Git-annex](#)
- 10+ year open source project (100+ contributors), available for all major OS
- Born from rethinking data:
 - Just like code, **data is not static**.
 - Just like code, **data is subject to collaboration**. Stream-lined workflows for sharing and collaborating should be possible, mirroring those in software development.
 - **Provenance** of data is essential for reproducible, trustworthy, and FAIR science
 - Flexibility and **interoperability with existing tools** is the key to sustainability and ease of use



- 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)**

EXHAUSTIVE TRACKING OF RESEARCH COMPONENTS



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

Well-structured datasets (using community standards), and portable computational environments — and their evolution — are the precondition for reproducibility

```
# 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
```

CAPTURE COMPUTATIONAL PROVENANCE



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

Which data was needed at which version, as input into which code, running with what parameterization in which computational environment, to generate an outcome?

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

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

EXHAUSTIVE CAPTURE ENABLES PORTABILITY



Precise identification of data and computational environments combined with provenance records form a comprehensive and portable data structure, capturing all aspects of an investigation.

```
# transfer data and metadata to other sites and services
# with fine-grained access control for dataset components

% datalad push --to <site-or-service>
```

REPRODUCIBILITY STRENGTHENS TRUST



Outcomes of computational transformations can be validated by authorized 3rd-parties. This enables audits, promotes accountability, and streamlines automated "upgrades" of outputs

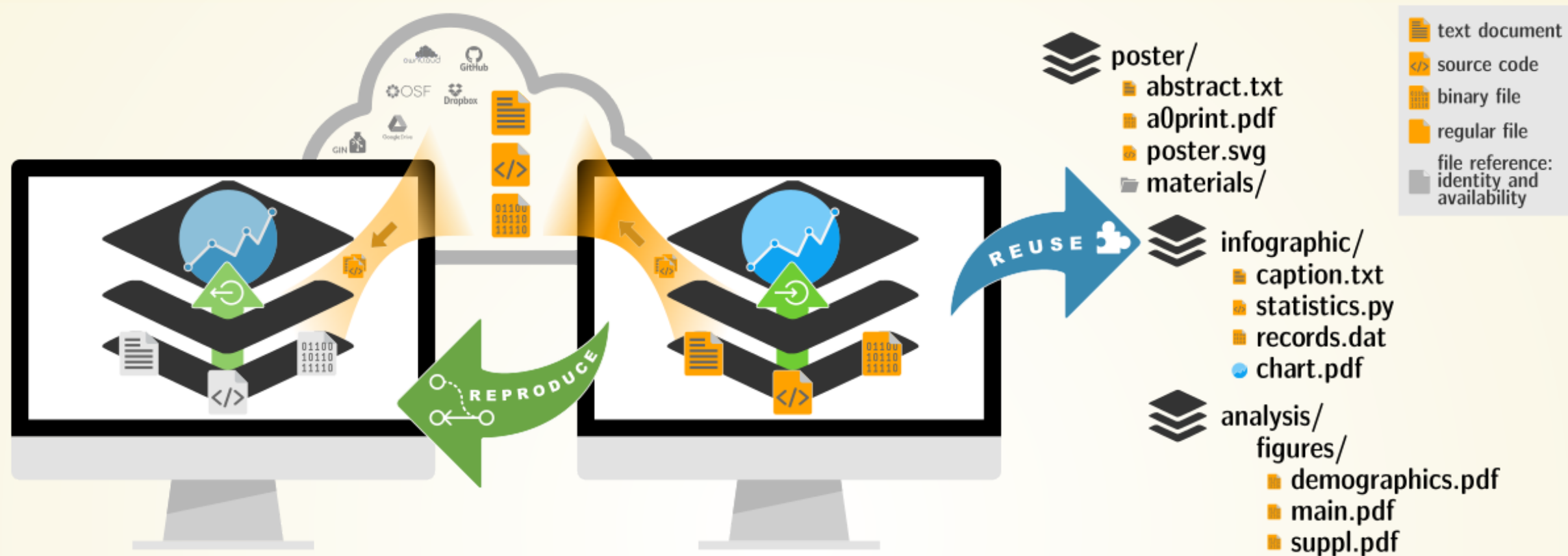
```
# obtain dataset (initially only identity,  
# availability, and provenance metadata)
```

```
% datalad clone <url>
```

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

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

ULTIMATE GOAL: (RE-)USABILITY



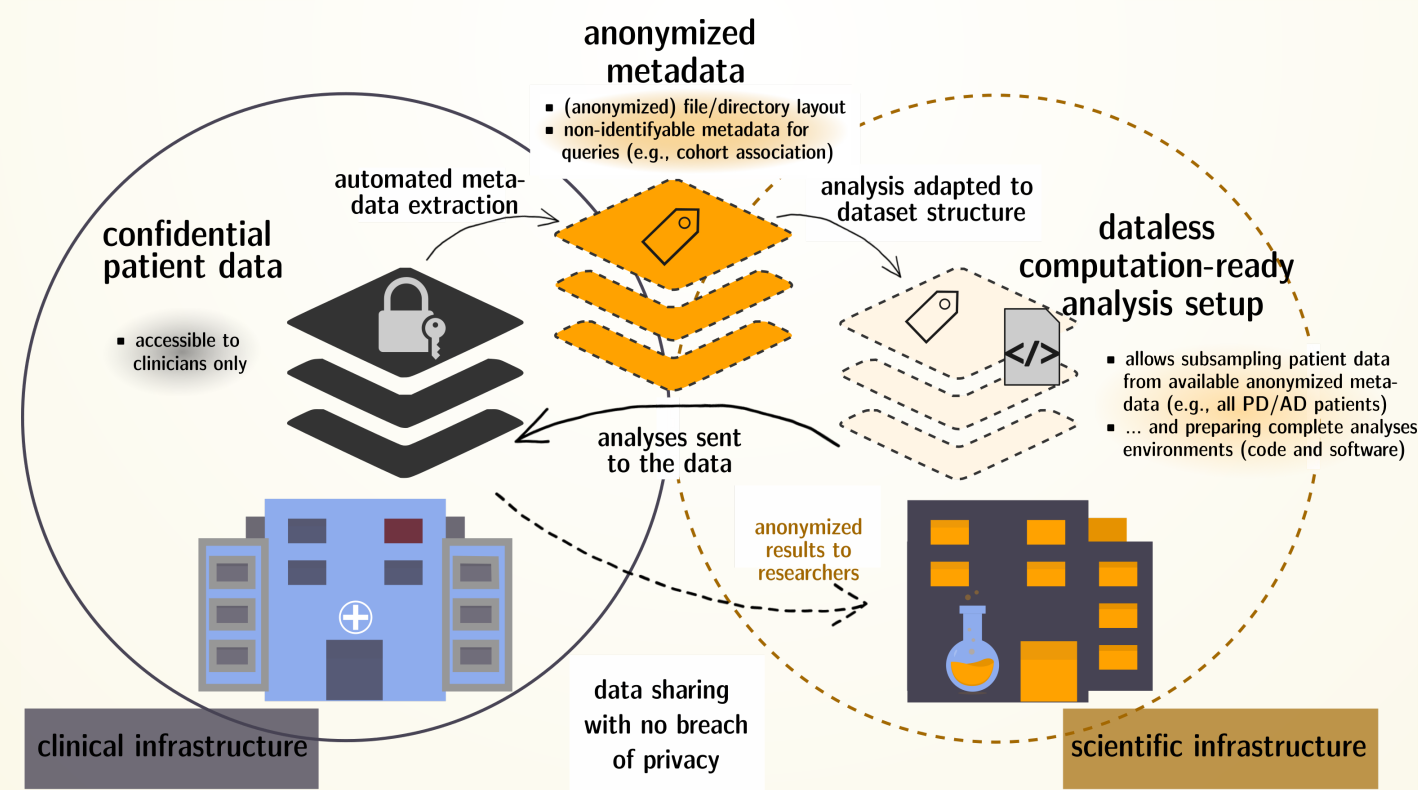
Verifiable, portable, self-contained data structures that track all aspects of an investigation exhaustively can be (re-)used as modular components in larger contexts — propagating their traits

```
# 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>
```

TRACKING & SHARING - WHERE IS THE PRIVACY?

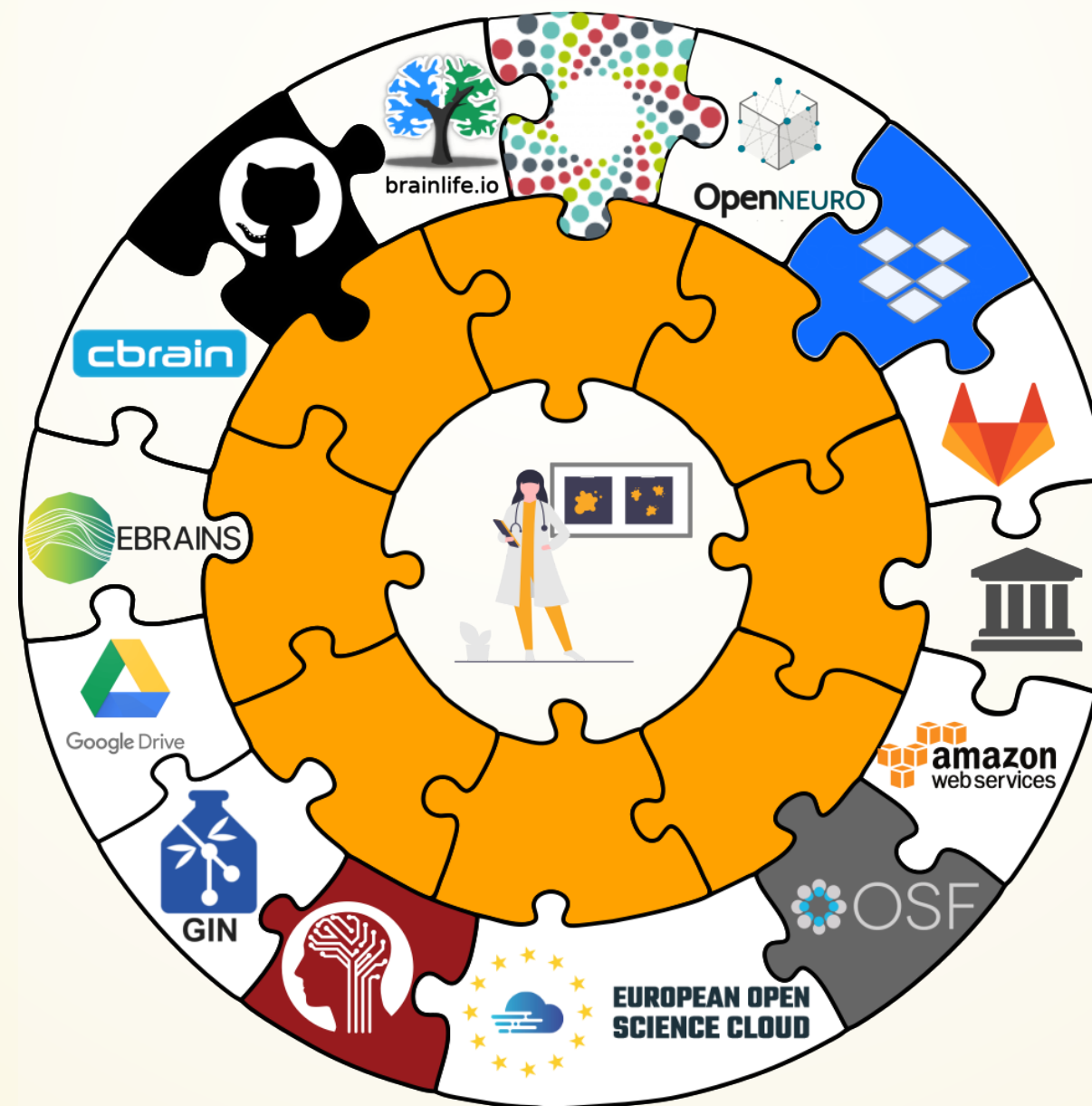
- Datasets have an optional annex for (large or sensitive) data.
- Rather than file content, identity (hash) and location information is tracked. Users have fine-grained control over transport and access; encryption is possible.
- Datasets can separate data access from meta data access



FINE-GRAINED FILE TRANSPORT

- Cloned datasets are lean, because file content is retrieved on demand:

```
1 $ datalad clone \  
2   https://github.com/datalad-datasets/human-connectome-project-openaccess.git  
3 $ cd human-connectome-project-openaccess && du -sh  
4   5.0M .  
5 $ datalad get HCP1200/102513/T1w/T1w_acpc_dc.nii.gz  
6   get(ok): HCP1200/102513/T1w/T1w_acpc_dc.nii.gz (file) [from datalad...]  
7 $ datalad drop HCP1200/102513/T1w/T1w_acpc_dc.nii.gz  
8   drop(ok)
```



BIG DATA

FAIRLY BIG: SCALING UP



Objective: Process the UK Biobank (imaging data)
with 76 TB in 43 million files from 42,715 participants

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

FAIRLY BIG MOVIE

- 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

WHERE CAN I FIND OUT MORE?

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 data1ad 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)

{OPEN,TRANSPARENT,REPRODUCIBLE} SCIENCE

- Treat data like software: obtain, version, share, and update data
- Simplified data management, disk-space aware storage & computing
- Transparent and reproducible science: link code, data, software, and execution in a human- and machine-readable way
- Collaborate: Generic workflows, interoperability with established tools & services

HAUS DER UNIVERSITÄT

DISTRIBITS.LIVE

- First conference on technologies for distributed data management
- 2 day conference plus single-day Hackathon
- @ Haus der Universität Düsseldorf

0. Apr 2024

Düsse

i s t r i b i t

60c7fd54104220245000c8df0c0

ACKNOWLEDGEMENTS

Software

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

THANKS!

Questions?

Funders



Collaborators

