DECENTRALIZED MANAGEMENT OF DIGITAL OBJECTS FOR OPEN SCIENCE

Adina Wagner
@AdinaKrik

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

Debian in Arts & Science

Slides: https://github.com/datalad-handbook/course/

ACKNOWLEDGEMENTS

Software

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

Documentation project

- Michael Hanke
- Laura Waite
- *28 additional 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

Collaborators

Human Brain Project

CONP PCNO

VirtualBrainCloud

OpenNEURO

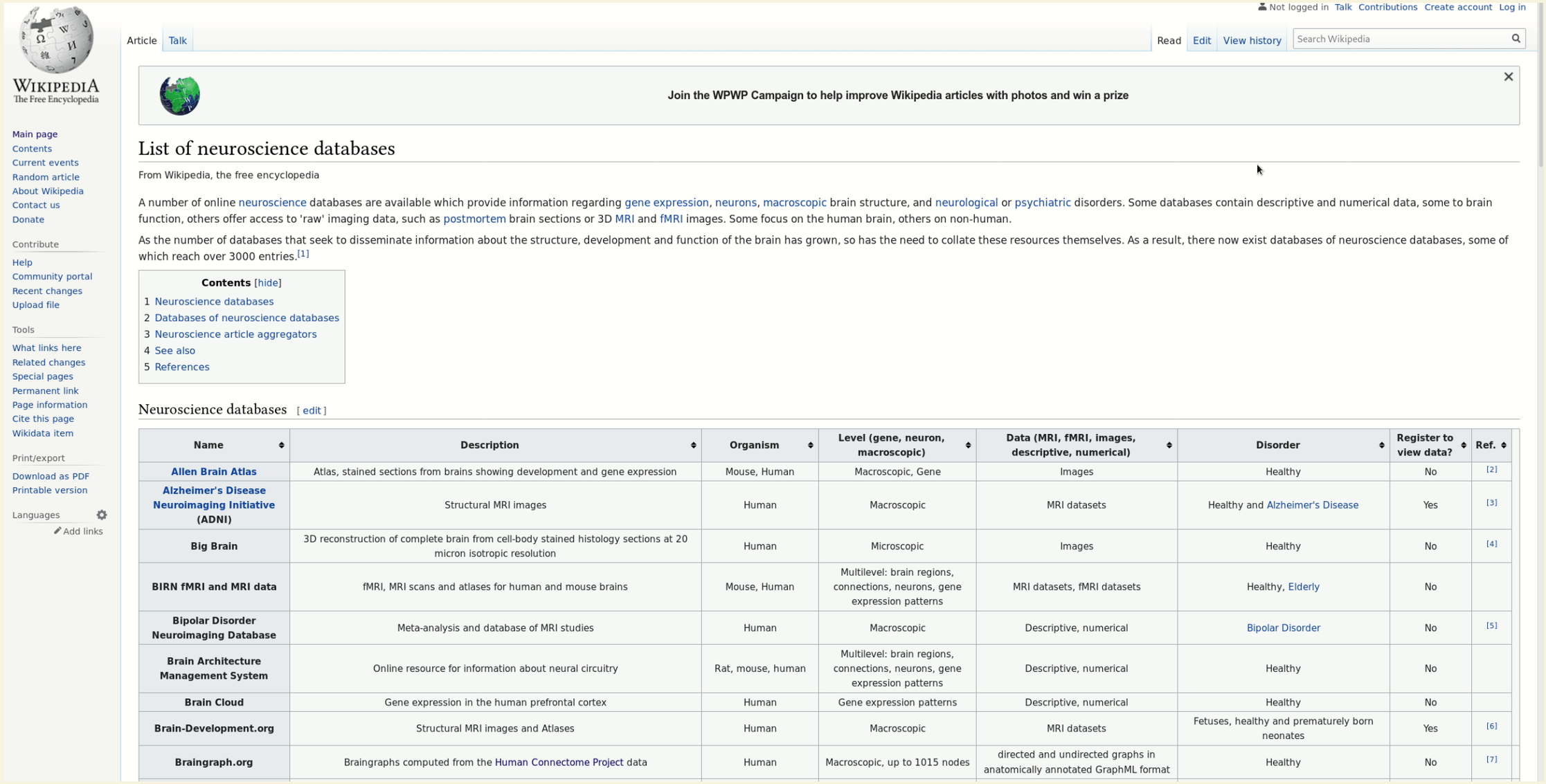
cbrain

brainlife.io

PERKS OF BEING A (NEURO)SCIENTIST...

PERKS OF BEING A (NEURO)SCIENTIST...

A growing culture of open data



Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

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

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

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 (NEURO)SCIENTIST...

A large and growing amount of open source software

... and many more!

PERKS OF BEING A (NEURO)SCIENTIST...

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

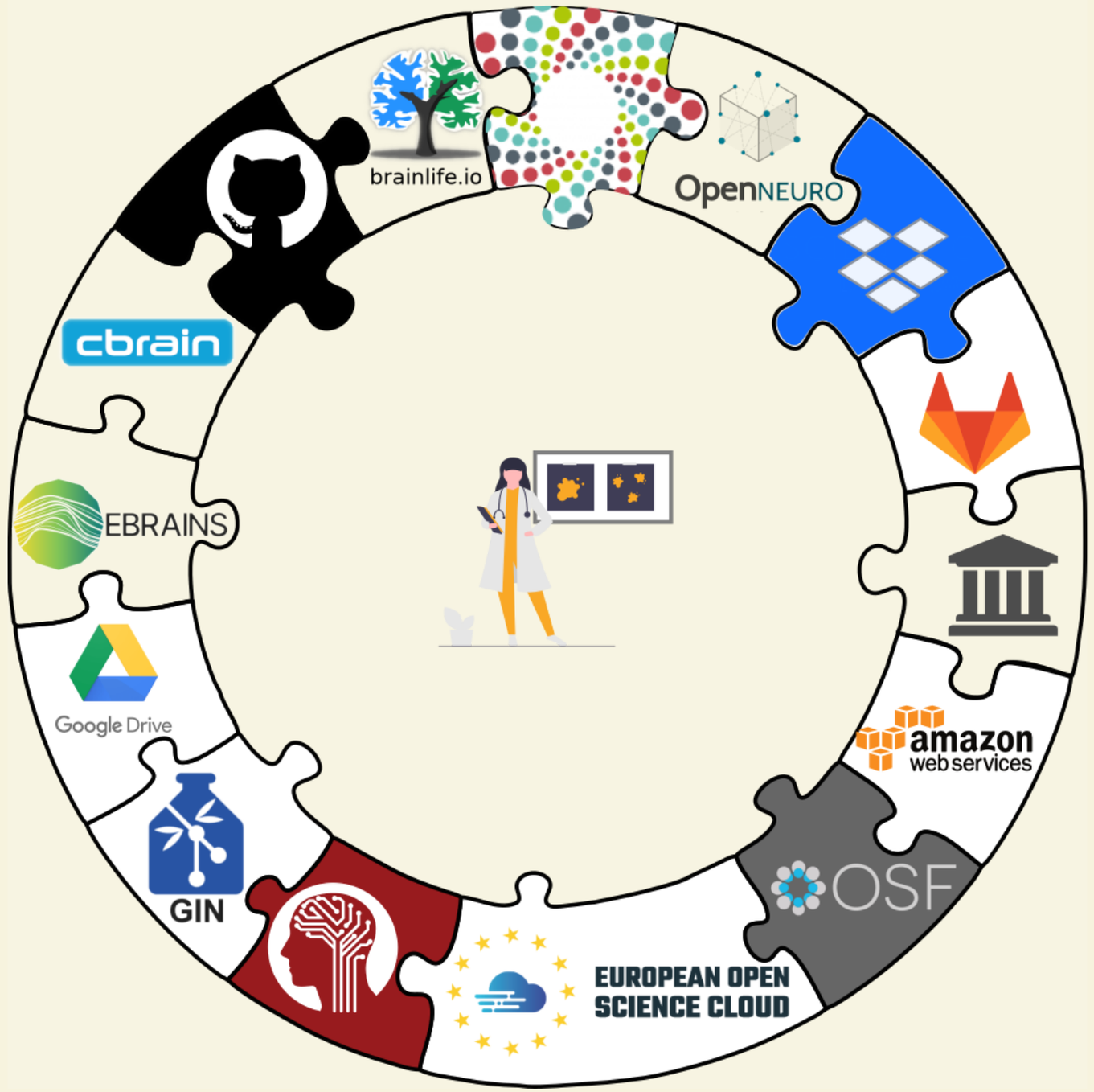


Image credit: https://www.incf.org/resources/sbps

PERKS OF BEING A (NEURO)SCIENTIST...

Work on questions targeting understanding of and treatments for mental and neural illness

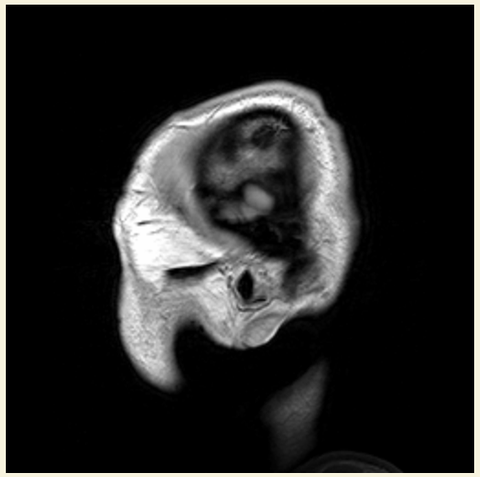


Image credit: Dwayne Reed; https://commons.wikimedia.org/wiki/File:Parasagittal_MRI_of_human_head_in_patient_with_benign_familial_macrocephaly_prior_to

THE GOOD NEWS, THE BAD NEWS

THE GOOD NEWS, THE BAD NEWS

Data sharing

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	... that's sparse & subject to strict data protection, making sharing difficult

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	... that's sparse & subject to strict data protection, making sharing difficult
Data analysis	

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	... that's sparse & subject to strict data protection, making sharing difficult
Data analysis	Reproducibility is threatened by intransparent, multi-stepped analyses & unstable results across software versions[1]

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

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	... that's sparse & subject to strict data protection, making sharing difficult
Data analysis	Reproducibility is threatened by intransparent, multi-stepped analyses & unstable results across software versions[1]
Collaboration	

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

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	... that's sparse & subject to strict data protection, making sharing difficult
Data analysis	Reproducibility is threatened by intransparent, multi-stepped analyses & unstable results across software versions[1]
Collaboration	Few interoperable workflows across institutes, rather: isolated solutions

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

THE GOOD NEWS, THE BAD NEWS

Data sharing	Heterogenous distribution and updating, scientists lack data management skills
Patient data	... that's sparse & subject to strict data protection, making sharing difficult
Data analysis	Reproducibility is threatened by intransparent, multi-stepped analyses & unstable results across software versions[1]
Collaboration	Few interoperable workflows across institutes, rather: isolated solutions

General underlying difficulties (distribution, updates, management, interoperability, reproducibility) also exist(ed) for software -
what can we learn?

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

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

"Share and treat data like software"

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

"Share and treat data like software"

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

"Share and treat data like software"

NeuroDebian

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

"Share and treat data like software"

NeuroDebian

IMPROVE SCIENTIFIC WORKFLOWS, COMING FROM THE PERSPECTIVE OF SOFTWARE DISTRIBUTIONS

"Share and treat data like software"

NeuroDebian

DATALAD

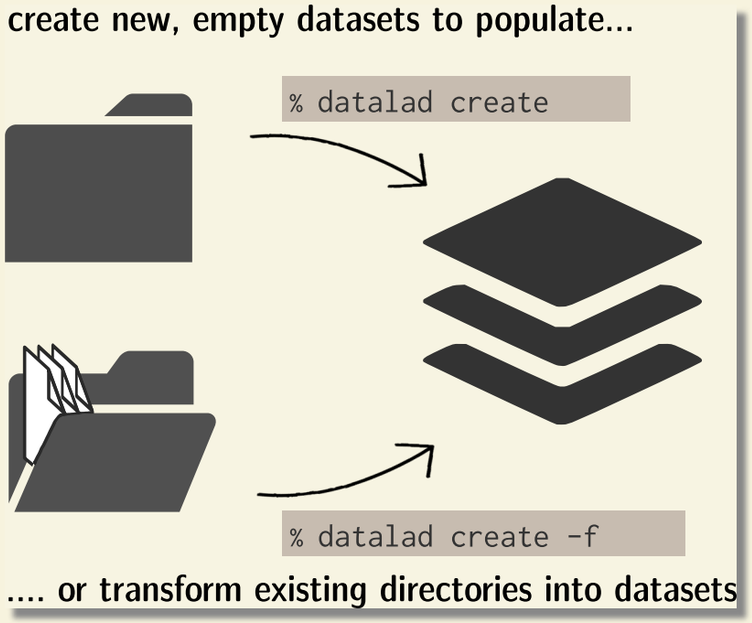
- DataLad: Joint management of digital objects through their entire life cycle
    - Version control
    - Transport logistics
    - Interoperability
    - Provenance capture
- Basics
    - Built on top of Git and git-annex (Joey Hess)
    - Free and open source Python software (MIT license)
    - Python API and command line interface

VERSION CONTROL

- DataLad knows two things: Datasets and files

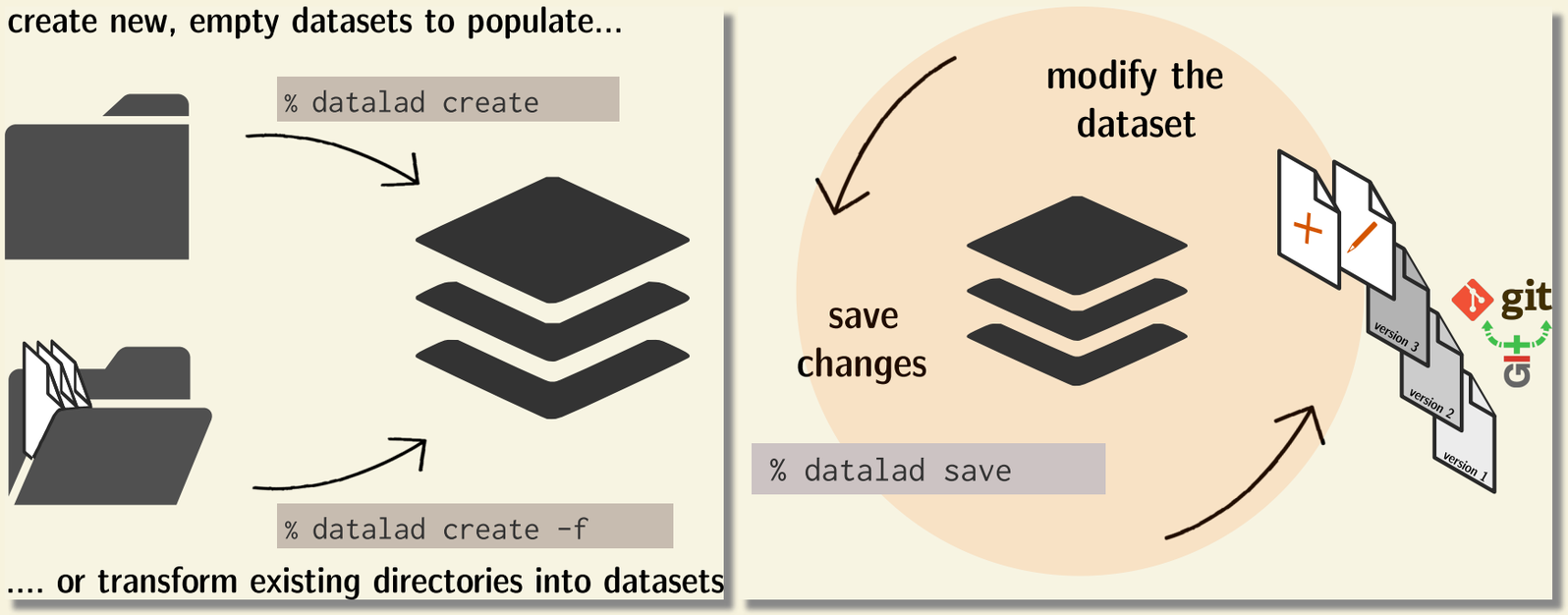
VERSION CONTROL

- DataLad knows two things: Datasets and files



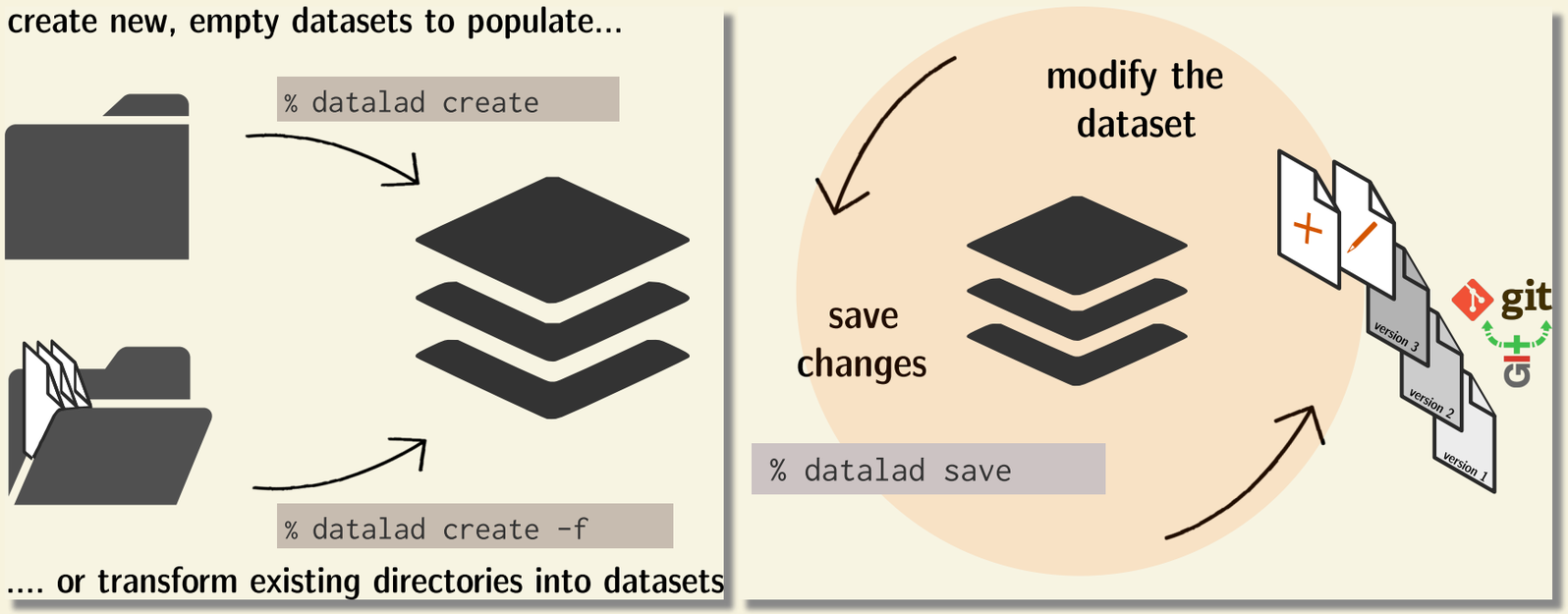
VERSION CONTROL

- DataLad knows two things: Datasets and files



VERSION CONTROL

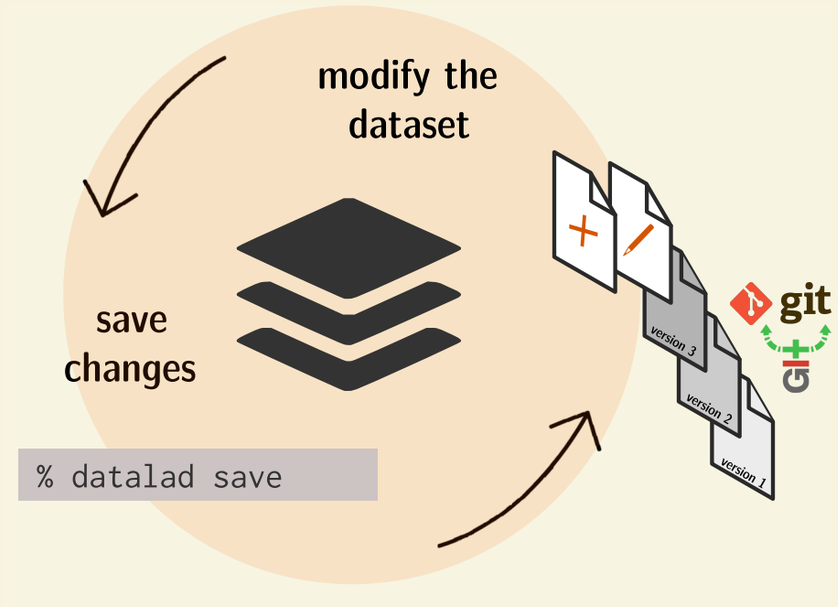
- DataLad knows two things: Datasets and files



- 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**

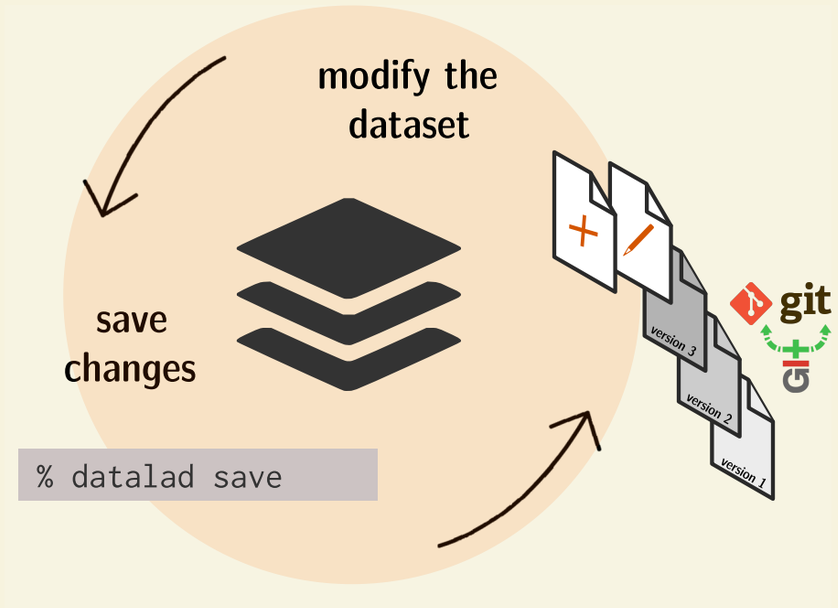
VERSION CONTROL: DATA

VERSION CONTROL: DATA



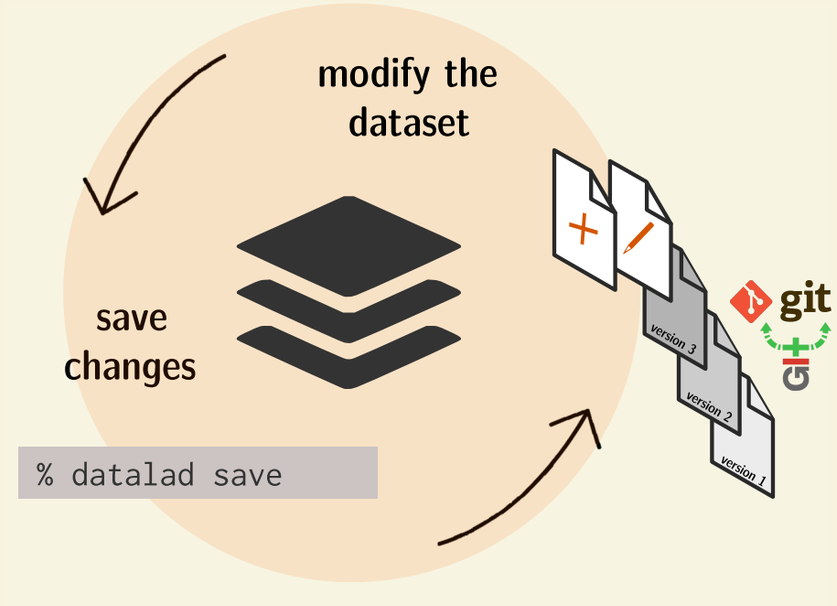
VERSION CONTROL: DATA

- Datasets have an optional annex for (large or sensitive) data (or text/code).



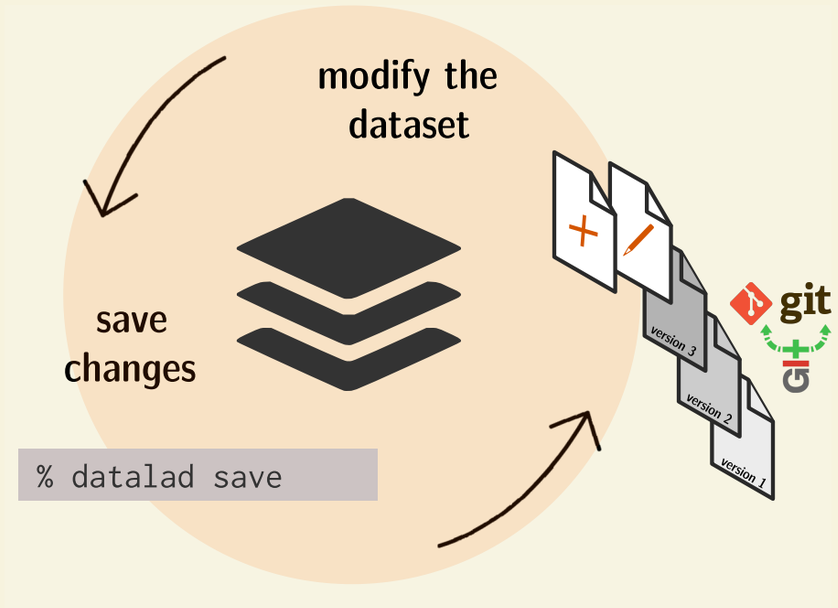
VERSION CONTROL: DATA

- Datasets have an optional annex for (large or sensitive) data (or text/code).
- Identity (hash) and location information is put into Git, rather than file content. The annex, and transport to and from it is managed with **git-annex** (git-annex.branchable.com)
  → decentralized version control for files of any size.



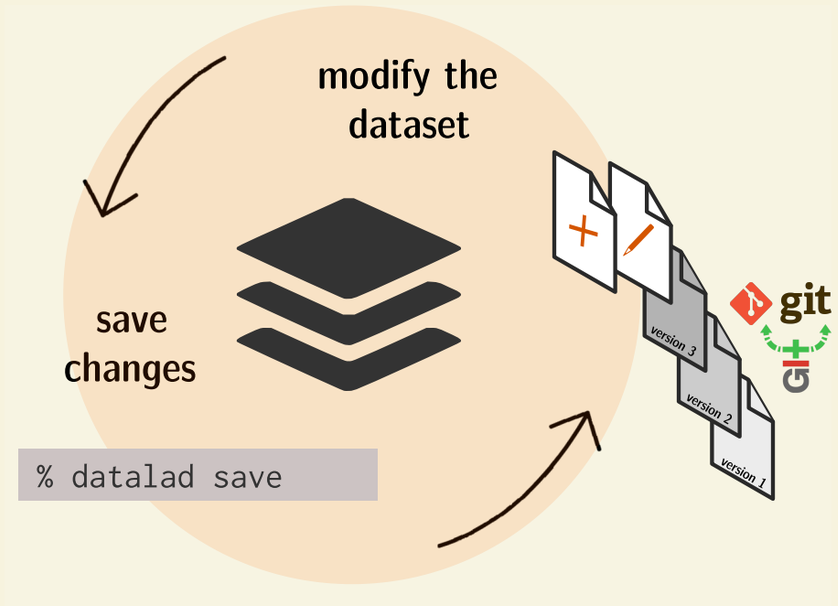
VERSION CONTROL: DATA

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



VERSION CONTROL: DATA

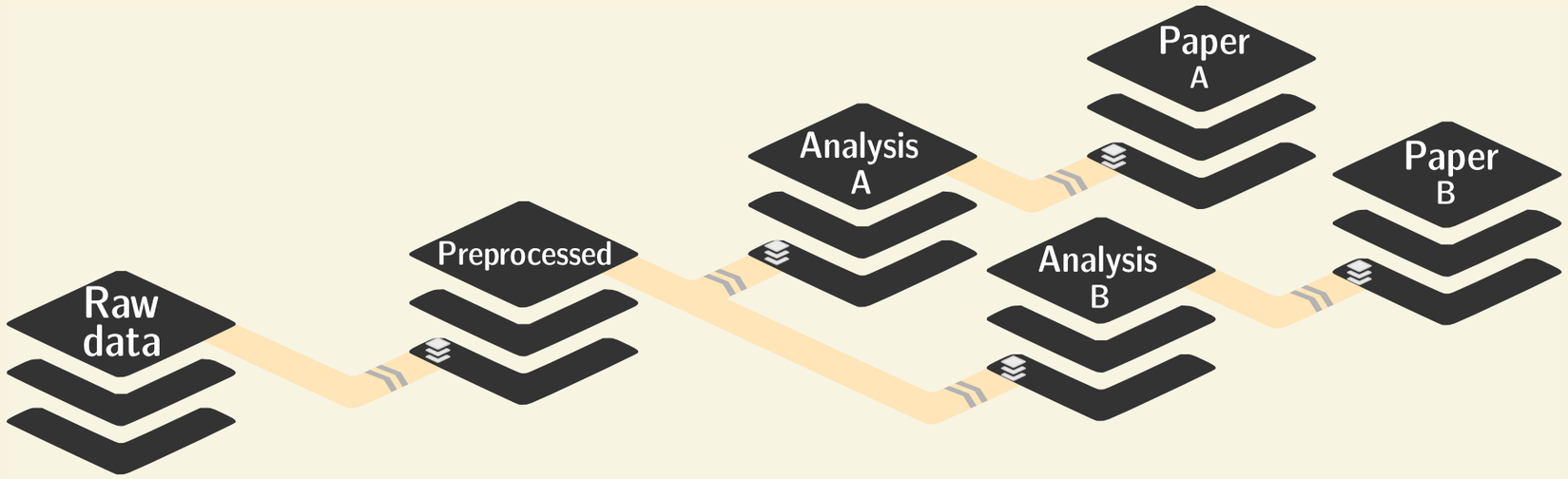
- Datasets have an optional annex for (large or sensitive) data (or text/code).
- Identity (hash) and location information is put into Git, rather than file content. The annex, and transport to and from it is managed with **git-annex** (git-annex.branchable.com)
  → decentralized version control for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



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

VERSION CONTROL: NESTING

- Seamless nesting mechanisms:

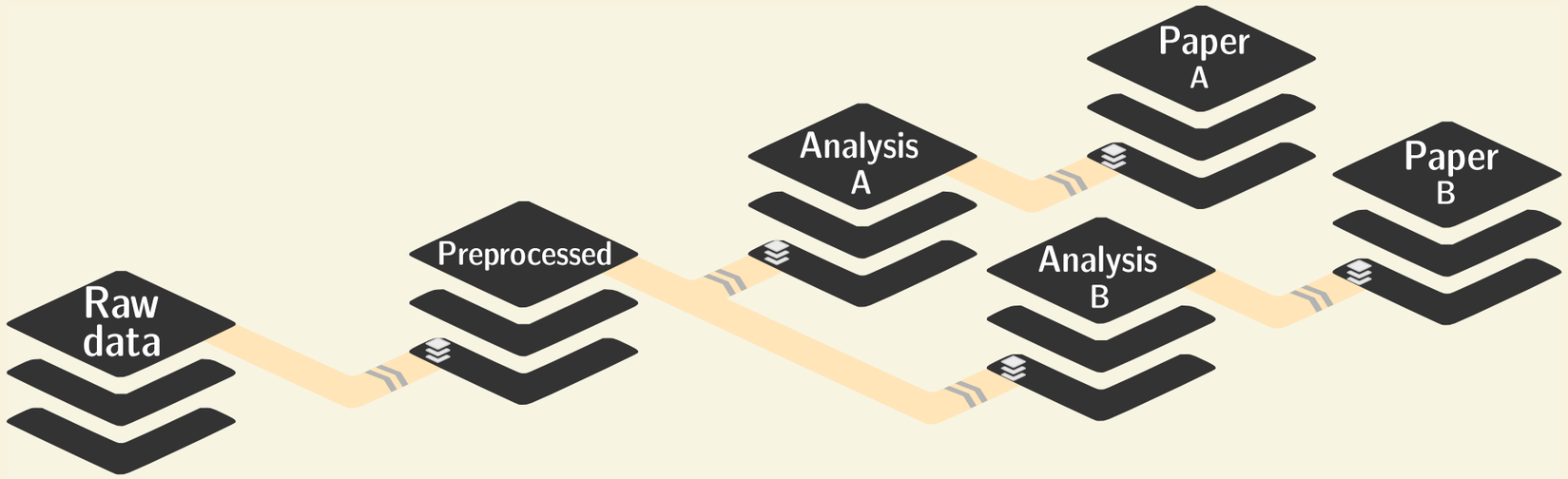


Nest modular datasets to create a linked hierarchy of datasets, and enable recursive operations throughout the hierarchy

  - hierarchies of datasets in super-/sub-dataset relationships
  - based on Git submodules, but more seamless

VERSION CONTROL: NESTING

- Seamless nesting mechanisms:



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

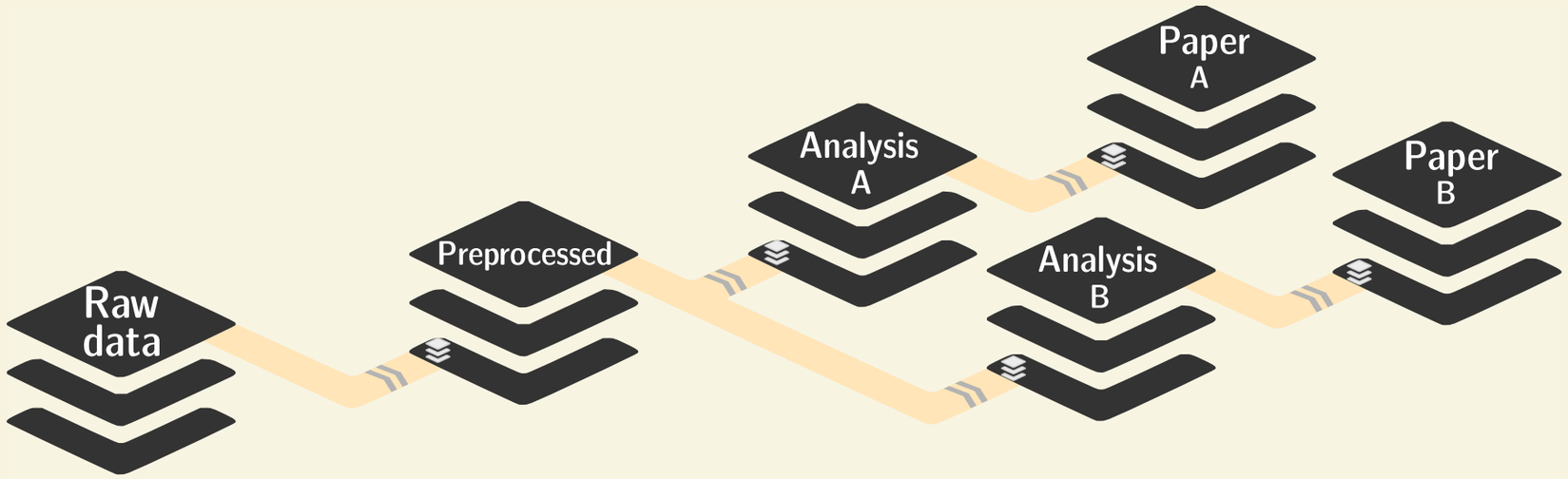
  - hierarchies of datasets in super-/sub-dataset relationships
  - based on Git submodules, but more seamless
- Overcomes scaling issues with large amounts of files

```
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)

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)

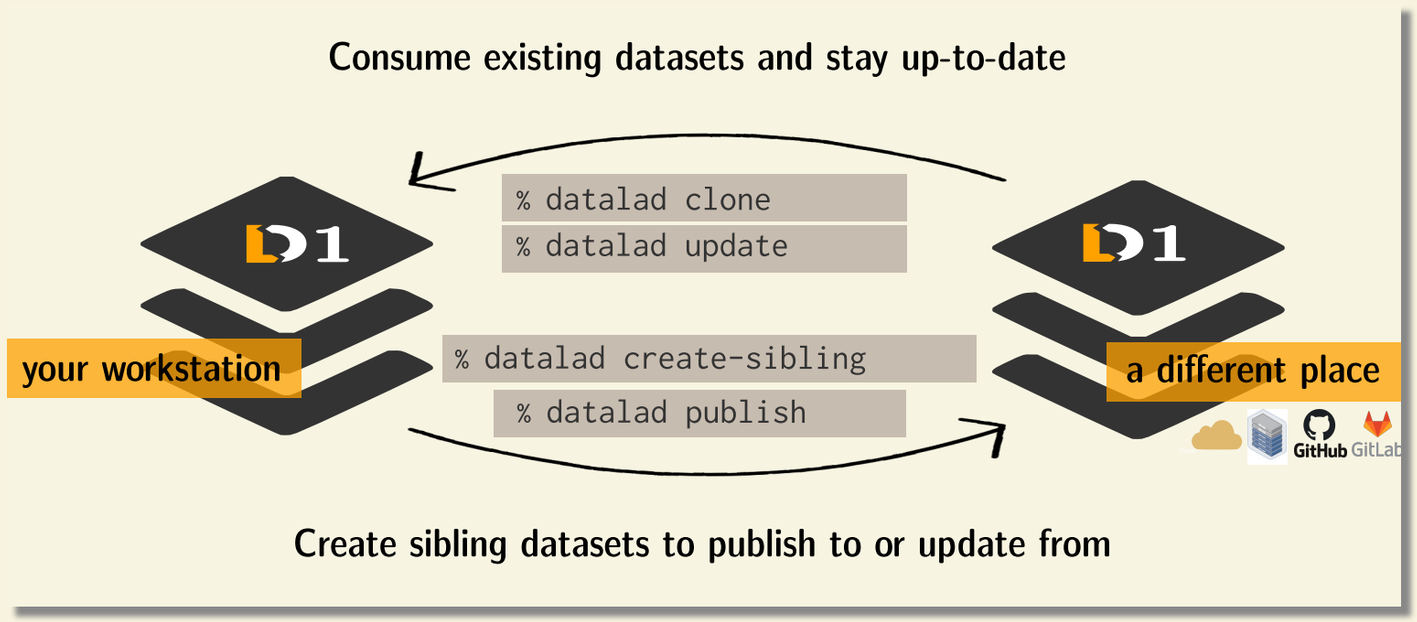
- Modularizes research components for transparency, reuse, and access management

TRANSPORT LOGISTICS

- Share data like source code

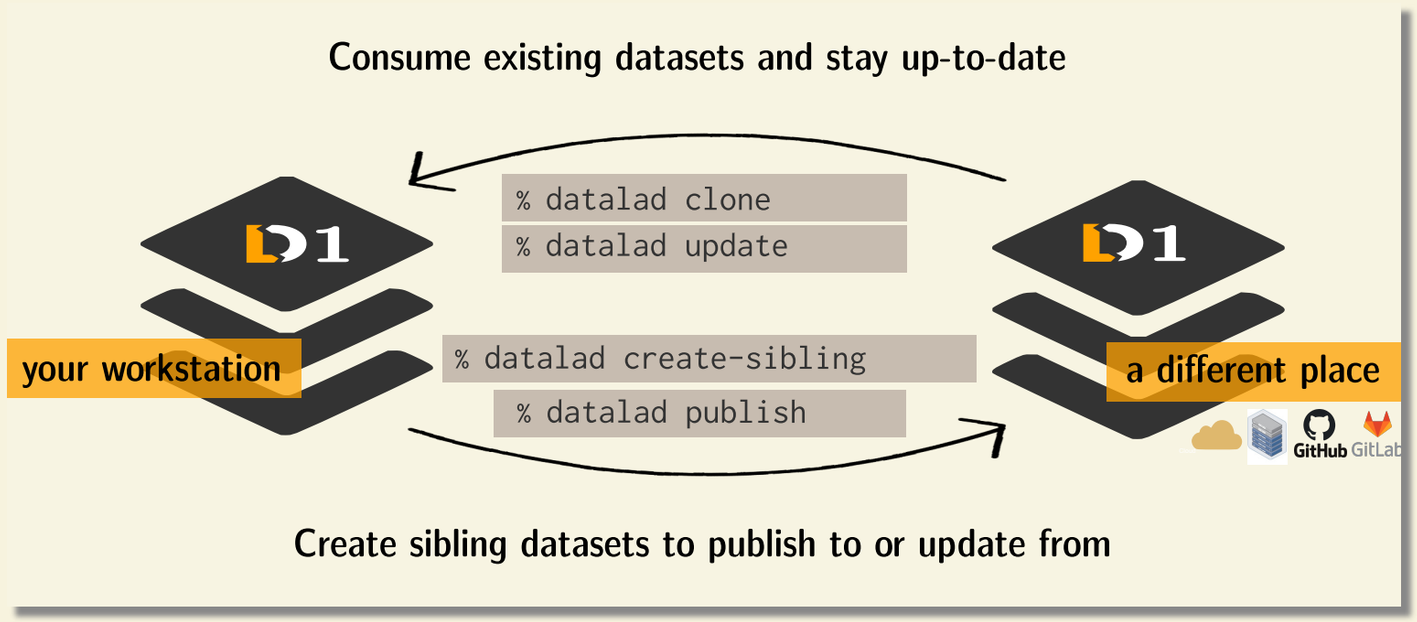
TRANSPORT LOGISTICS

- Share data like source code
- Datasets can be cloned, pushed, and updated from and to local paths, remote hosting services, external special remotes



TRANSPORT LOGISTICS

- Share data like source code
- Datasets can be cloned, pushed, and updated from and to local paths, remote hosting services, external special remotes



- Flexible data access management for annexed file contents based on storage location

TRANSPORT LOGISTICS

TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean (only Git):

TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean (only Git):

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

TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean (only Git):

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

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

TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean (only Git):

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

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

- annexed file's contents can be retrieved & dropped on demand:

TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean (only Git):

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

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

- annexed file's contents can be retrieved & dropped on demand:

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

TRANSPORT LOGISTICS

- Disk-space aware workflows: Cloned datasets are lean (only Git):

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

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

- annexed file's contents can be retrieved & dropped on demand:

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

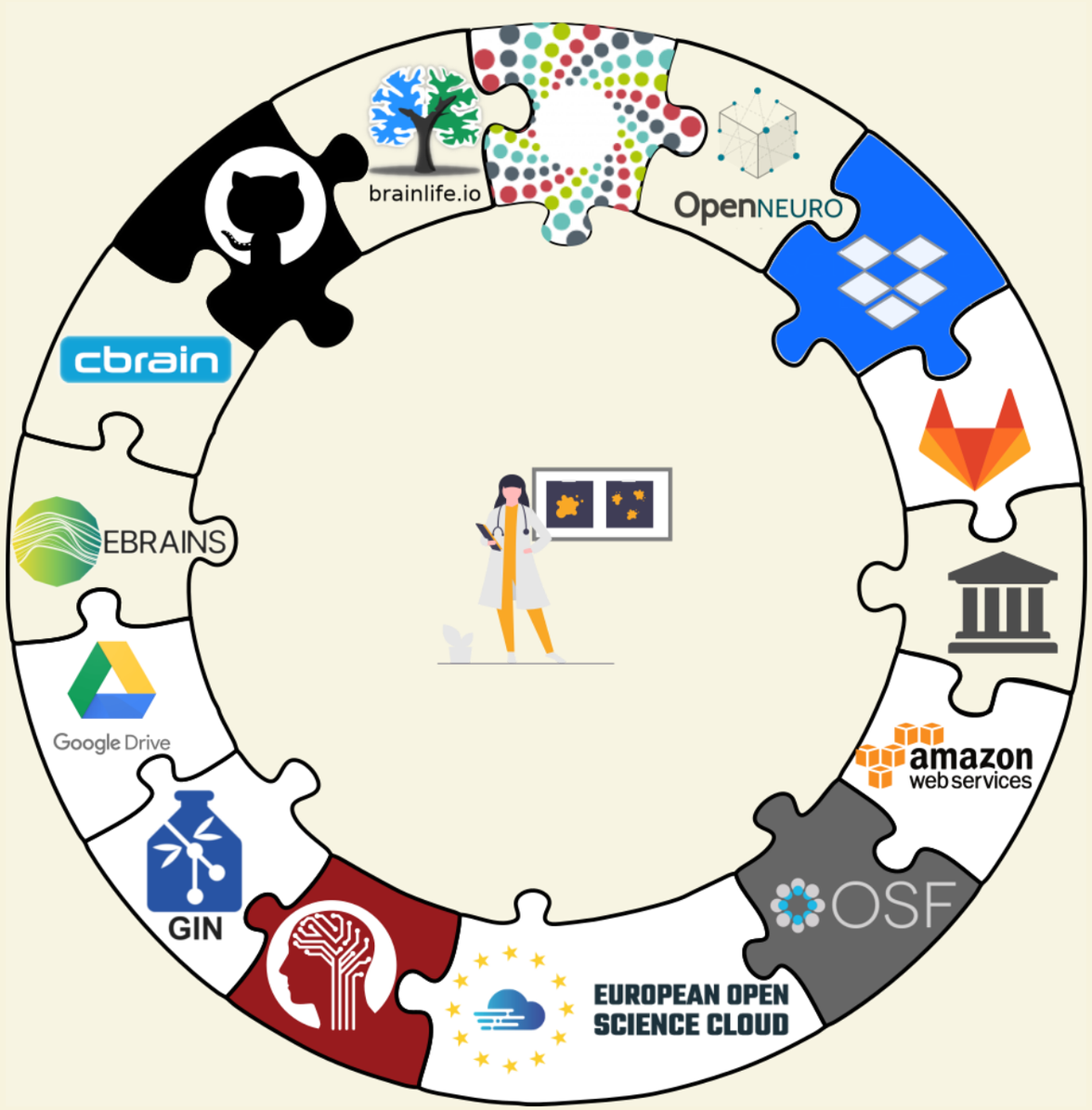
```
$ datalad drop A.Shashua-Introduction_to_Machine_Learning.pdf
drop(ok): /tmp/machinelearning-books/A.Shashua-Introduction_to_Machine_Learning.pdf (file) [checking https
```

INTEROPERABILITY

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

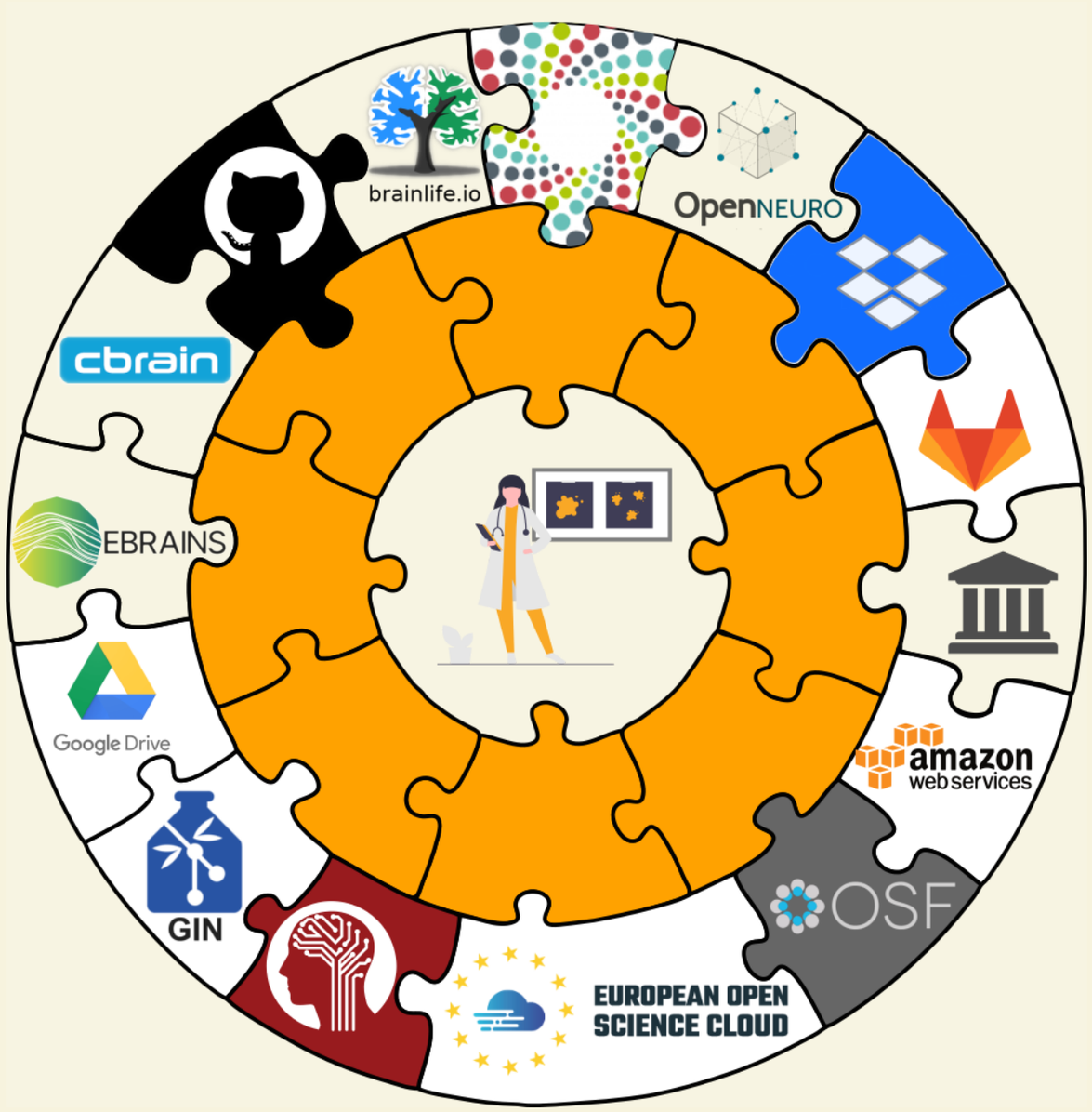
INTEROPERABILITY

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



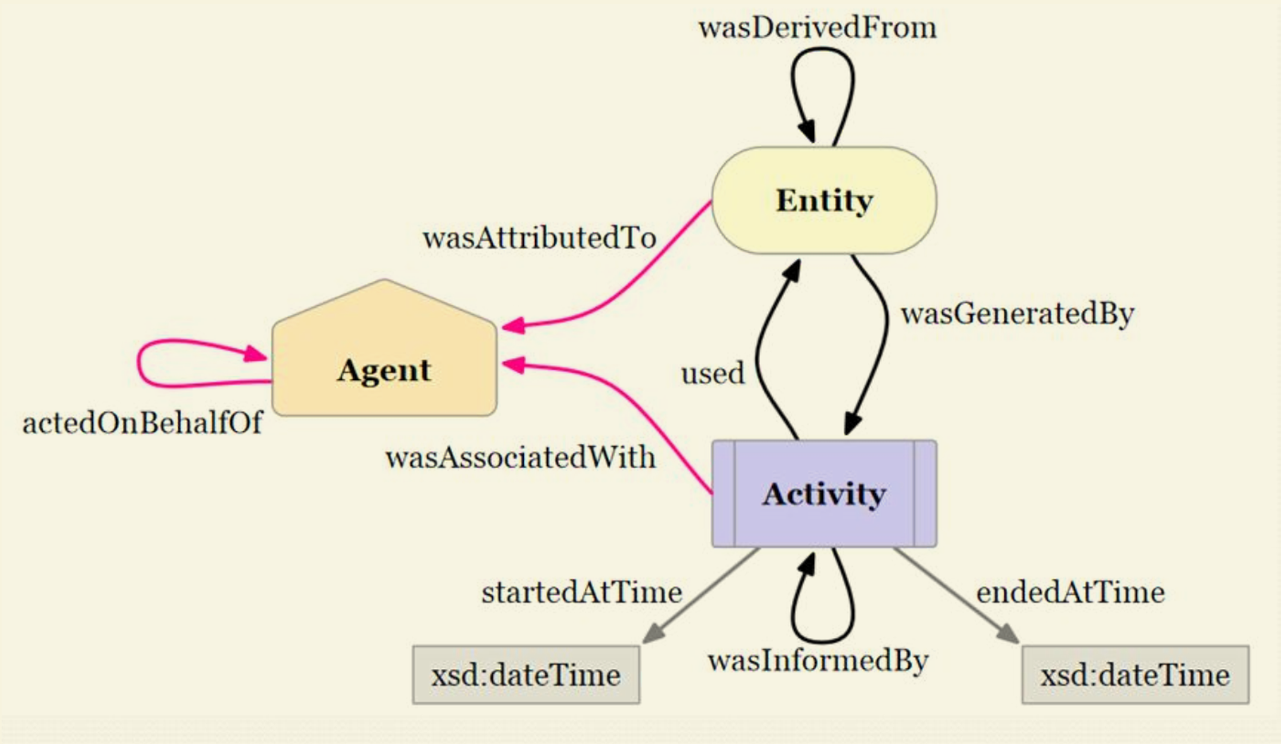
INTEROPERABILITY

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



PROVENANCE CAPTURE

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



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

PROVENANCE CAPTURE

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

```
$ datalad run -m "Perform eye movement event detection"\
  --input 'raw_data/*.tsv.gz' --output 'sub-*' \
  bash code/compute_all.sh

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

PROVENANCE CAPTURE

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

```
$ datalad containers-run -n neuroimaging-container \
  --input 'mri/*_bold.nii --output 'sub-*/LC_timeseries_run-*.csv' \
  "bash -c 'for sub in sub-*; do for run in run-1 ... run-8;
     do python3 code/extract_lc_timeseries.py \$sub \$run; done; done'"

-- Git commit -- Michael Hanke < ... @gmail.com>; Fri Jul 6 11:02:28 2019
    [DATALAD RUNCMD] singularity exec --bind {pwd} .datalad/e...
    === Do not change lines below ===
    {
     "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..",
     "dsid": "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 +
...
```

PROVENANCE CAPTURE

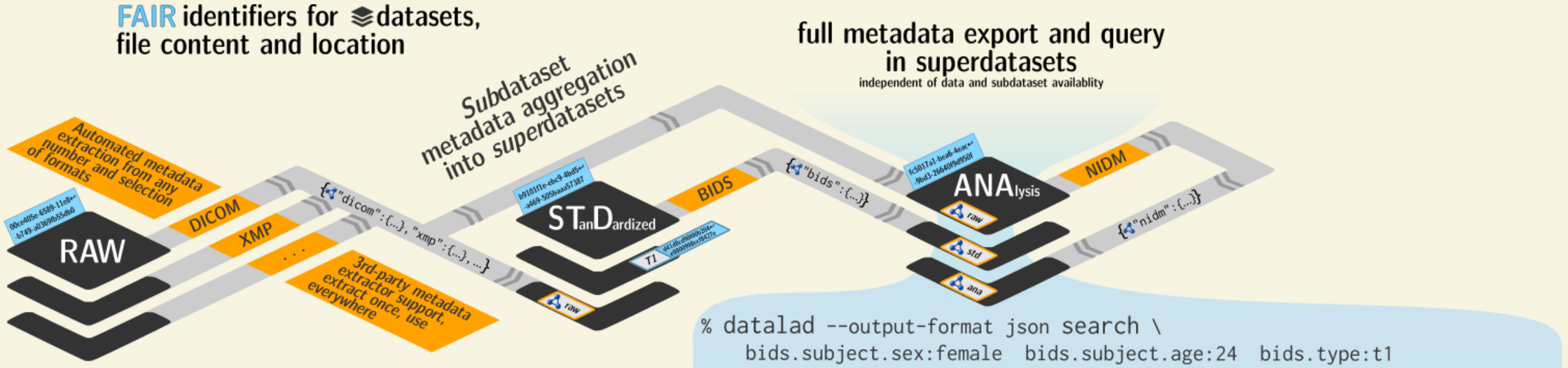
- All recorded transformations can be re-computed automatically

```
$ 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

AND THERE IS MORE...

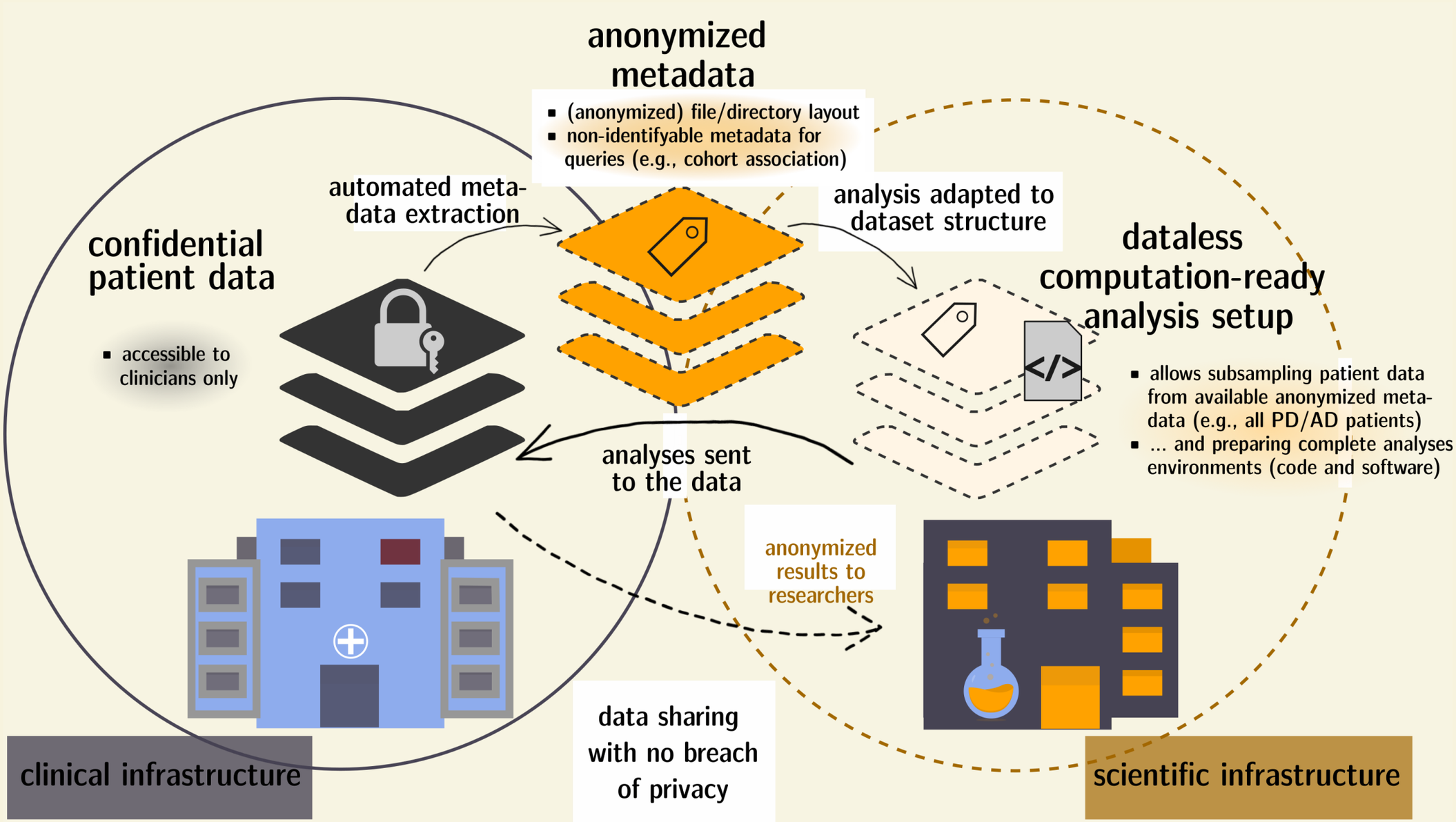
SCALABLE AND ACTIONABLE (META)DATA REPRESENTATIONS



- meta data logistics to generate, store, share, and search arbitrary meta data
- Sharing/retrieving only dataset meta data enables data discovery that doesn't require retrieving (potentially large) file contents first
- facilitates building metadata-driven applications

DEALING WITH SENSITIVE DATA

- Flexible data access management - file contents can be made available to none or selected few
- Share anonymized meta data not subject to privacy concerns
- Viable solution to bring the computation to the data



STILL MORE...

"RIA stores" for central data management, archival, or backup:

self-contained, plain file system storage for datasets, tuned for maintainability and use on systems with inode limitations (any-sized dataset ≈ 25 inodes). More: handbook.datalad.org/r.html?RIA

Extensions with additional or domain-specific functionality:

Available as pip-installable Python packages. More: handbook.datalad.org/r.html?extensions

Open Data collection:

As of August 2020, about 250TB of open data: datasets.datalad.org/

{OPEN,TRANSPARENT,REPRODUCIBLE} SCIENCE

{OPEN,TRANSPARENT,REPRODUCIBLE} SCIENCE

- Treat data like software: obtain, version, share, and update data

{OPEN,TRANSPARENT,REPRODUCIBLE} SCIENCE

- Treat data like software: obtain, version, share, and update data
- Simplified data management, disk-space aware storage & computing

{OPEN,TRANSPARENT,REPRODUCIBLE} SCIENCE

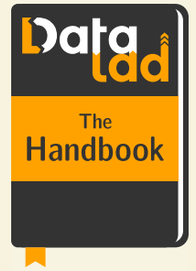
- Treat data like software: obtain, version, share, and update data
- Simplified data management, disk-space aware storage & computing
- Transparent and reproducible science: link code, data, software, and execution in a human- and machine-readable way

{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, interoperabality with established tools & services

FIND OUT MORE

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, like how to make a reproducible paper

FURTHER INFORMATION

- Source code: github.com/datalad/datalad
- Technical docs: docs.datalad.org
- Video tutorials: Youtube channel "DataLad"
- Matrix channel: DataLad

Slides: https://github.com/datalad-handbook/course/

THANKS!