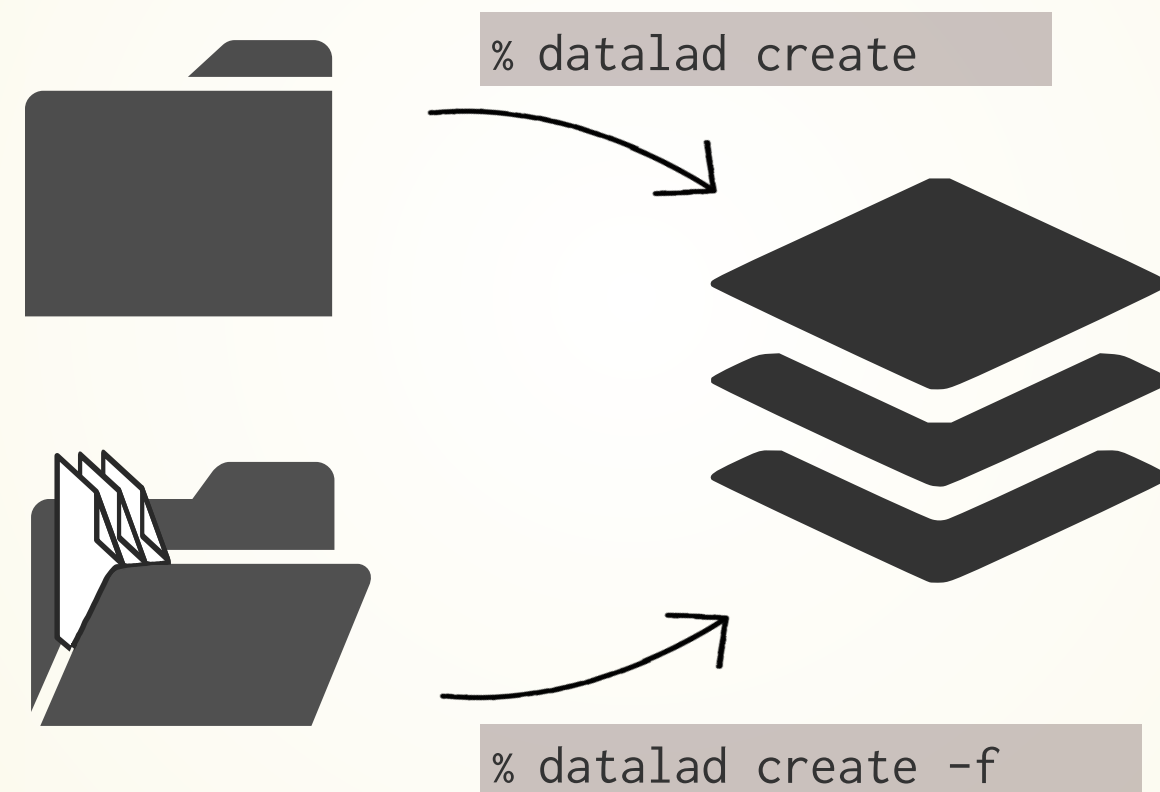


# QUICK RECAP

create new, empty datasets to populate...



.... or transform existing directories into datasets

# QUICK RECAP



## Git

- dataset history (commit messages, run records)
- All files + content committed into Git (useful with code, text, ...)
- File identity information of all annexed files (file name, identity hash, storage locations where to retrieve it from)

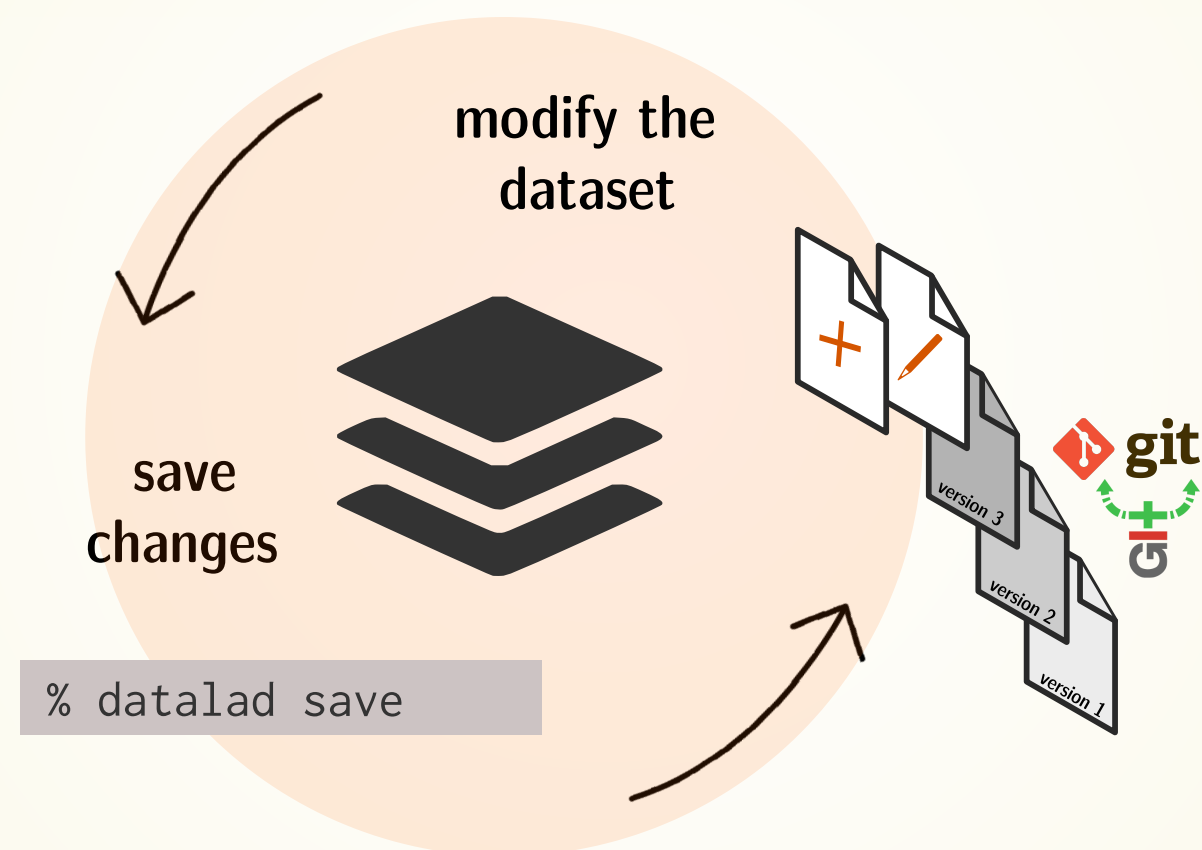


## git-annex

- contents of annexed files
- organized in the "annex" or "object tree" of the dataset



# QUICK RECAP



# QUICK RECAP

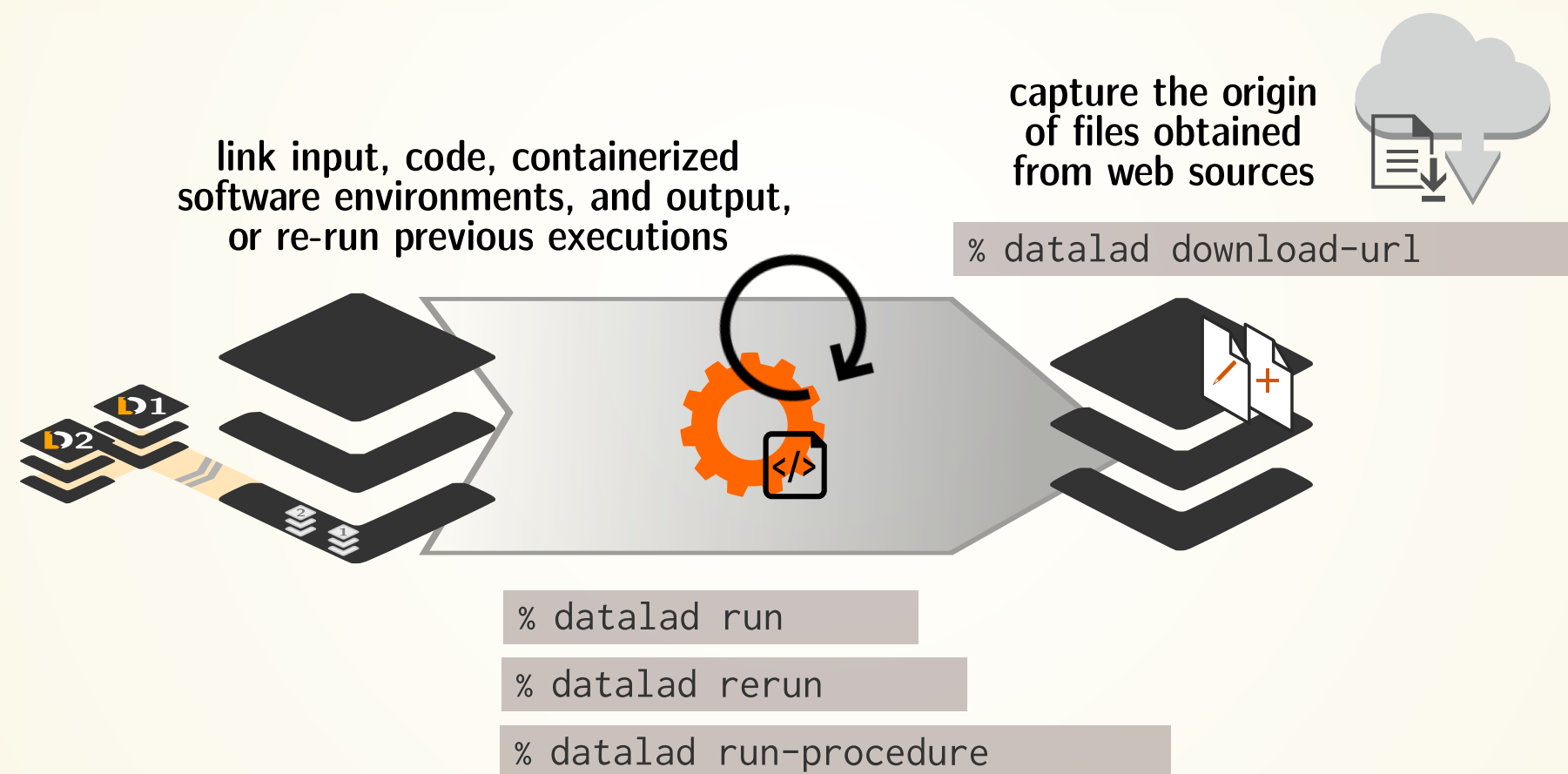
|                  |       |               |   |                                                                                            |
|------------------|-------|---------------|---|--------------------------------------------------------------------------------------------|
| 2020-06-05 10:58 | +0200 | Adina Wagner  | M | [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-d |
| 2020-06-05 08:24 | +0200 | Adina Wagner  | o | [finalround] {upstream/finalround} add results from computing with mean instead of median  |
| 2020-06-05 09:09 | +0200 | Michael Hanke | o | Change wording, clarify comment                                                            |
| 2020-06-05 07:26 | +0200 | Michael Hanke | M | Merge remote-tracking branch 'gh-mine/finalround'                                          |
| 2020-05-28 16:39 | +0200 | Asim H Dar    | o | Added datalad.get() so S2SRMS() pulls data and can run standalone                          |
| 2020-05-18 08:25 | +0200 | Adina Wagner  | o | {gh-asim/finalround} S2SRMS: example implementation of the S2SRMS method suggested by R2   |
| 2020-05-01 17:38 | +0200 | Adina Wagner  | o | Minor edits as suggested by reviewer 2                                                     |
| 2020-05-29 09:04 | +0200 | Adina Wagner  | M | Merge pull request #13 from psychoinformatics-de/adswa-patch-1                             |
| 2020-05-24 09:15 | +0200 | Adina Wagner  | o | {upstream/adswa-patch-1} Fix installation instructions                                     |
| 2020-05-24 09:53 | +0200 | Adina Wagner  | M | Merge pull request #14 from psychoinformatics-de/bf-data                                   |
| 2020-05-24 09:33 | +0200 | Adina Wagner  | o | [bf-data] One-time datalad import                                                          |
| 2020-05-24 09:32 | +0200 | Adina Wagner  | o | install and get relevant subdataset data                                                   |
| 2020-03-18 10:19 | +0100 | Michael Hanke | M | Merge pull request #8 from psychoinformatics-de/adswa-patch-1                              |
| 2019-12-19 10:22 | +0100 | Adina Wagner  | o | {gh-asim/adswa-patch-1} add sklearn to requirements                                        |
| 2020-03-18 10:13 | +0100 | Michael Hanke | o | Tune new figure caption                                                                    |
| 2020-03-18 10:03 | +0100 | Michael Hanke | M | Merge pull request #11 from ElectronicTeaCup/revision_2                                    |
| 2020-03-18 09:59 | +0100 | Adina Wagner  | M | [revision_2] {gh-asim/revision_2} Merge branch 'revision_2' of github.com:ElectronicTe     |
| 2020-03-18 09:58 | +0100 | Michael Hanke | o | last detection                                                                             |
| 2020-03-18 09:59 | +0100 | Adina Wagner  | o | name parameter in caption                                                                  |

Date

Author

Change summary (commit message)

# QUICK RECAP



# **BEFORE WE BEGIN...**

Any left-over questions from yesterday?

# PUBLISHING DATASETS

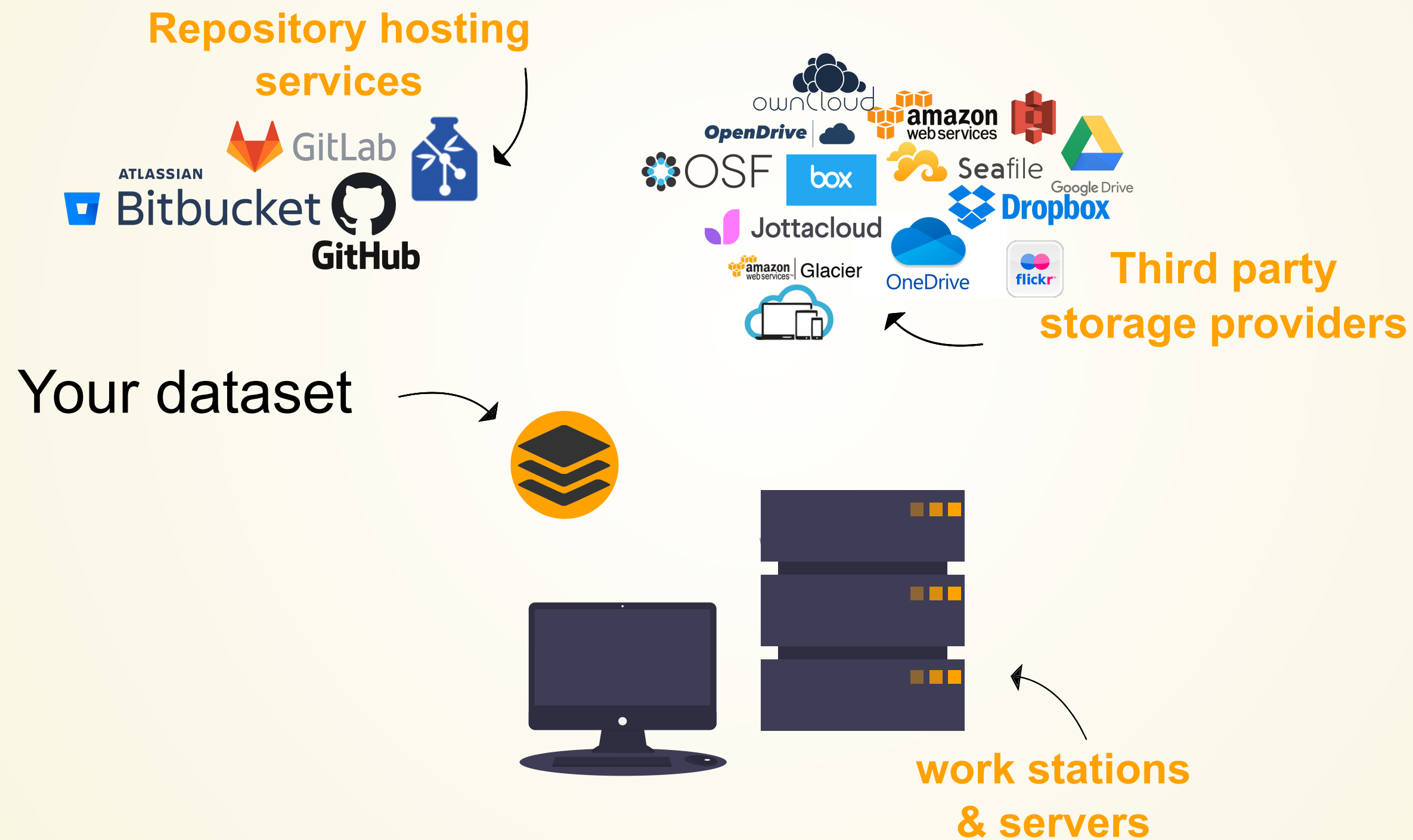
How to share your work with others

---

Repository hosting services, siblings, and `data-lad push`

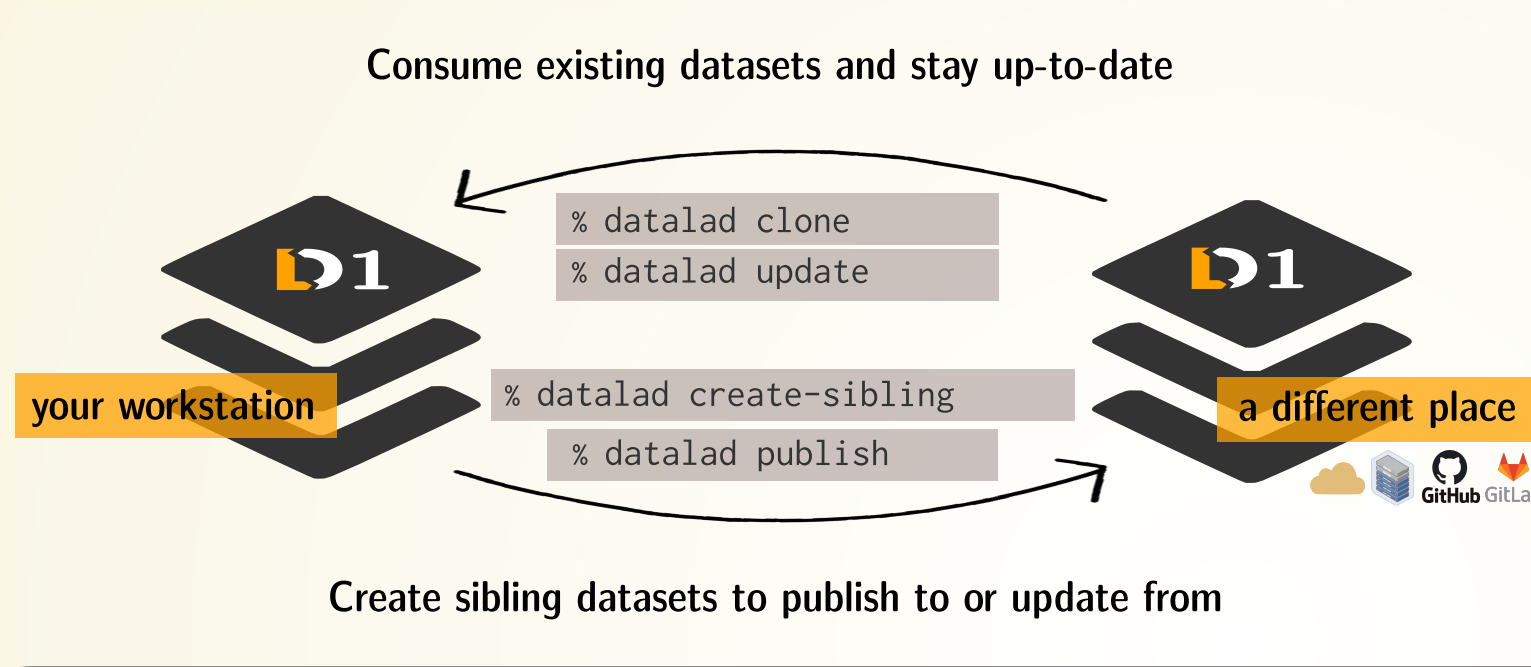
# PUBLISHING DATASETS

I have a dataset on my computer. How can I share it, or collaborate on it?



# "SHARE DATA LIKE SOURCE CODE"

- Datasets can be cloned, pushed, and updated from and to local and remote paths, remote hosting services, external special remotes



- Examples:  
Local path

```
../my-projects/experiment_data
```

Remote path

```
myuser@myinstitutes.hcp.system:/home/myuser/my-projects/experiment_data
```

Hosting service

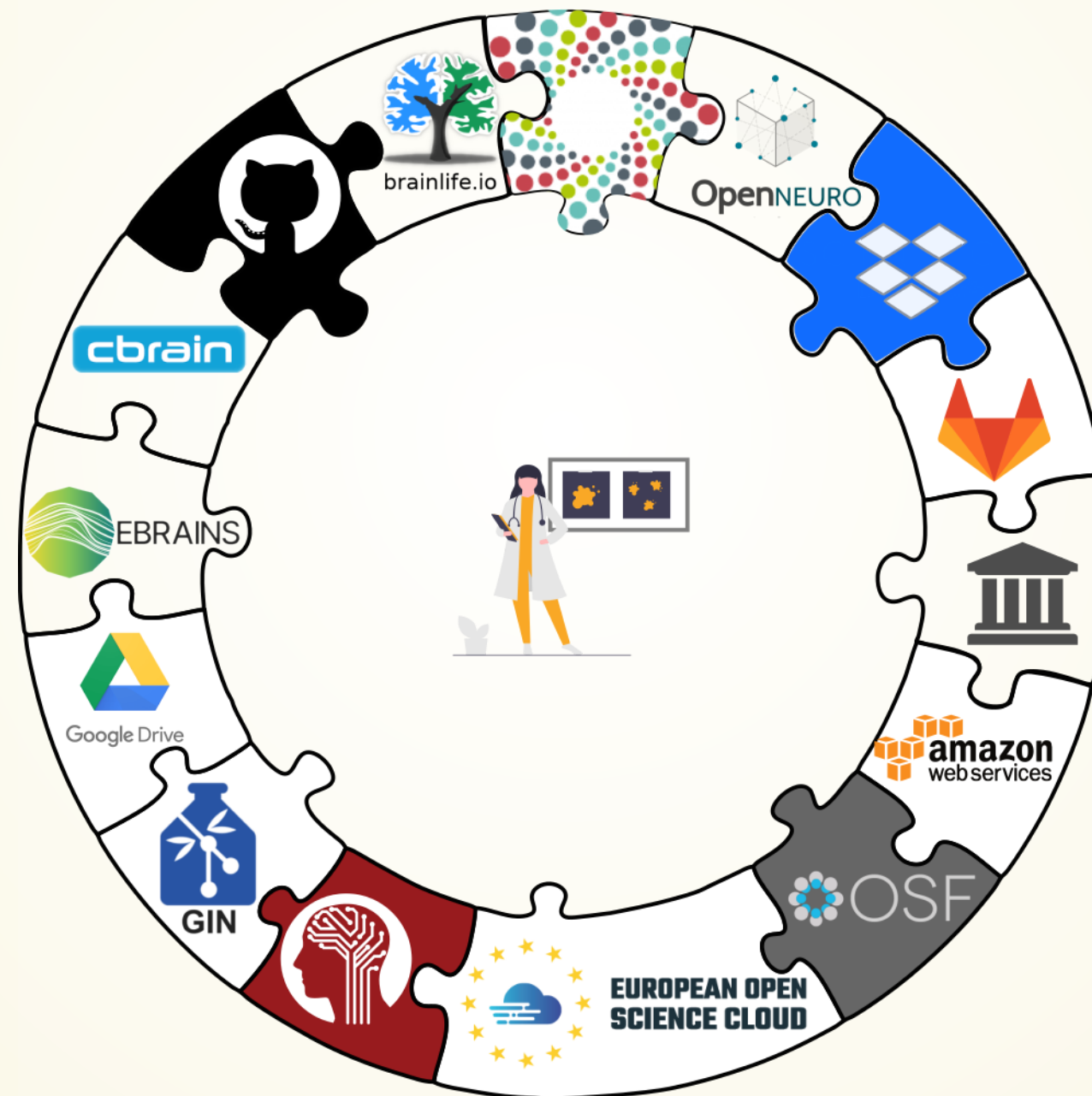
```
git.github.com:myuser/experiment_data.git
```

External special remotes

```
osf://my-osf-project-id
```

# INTEROPERABILITY

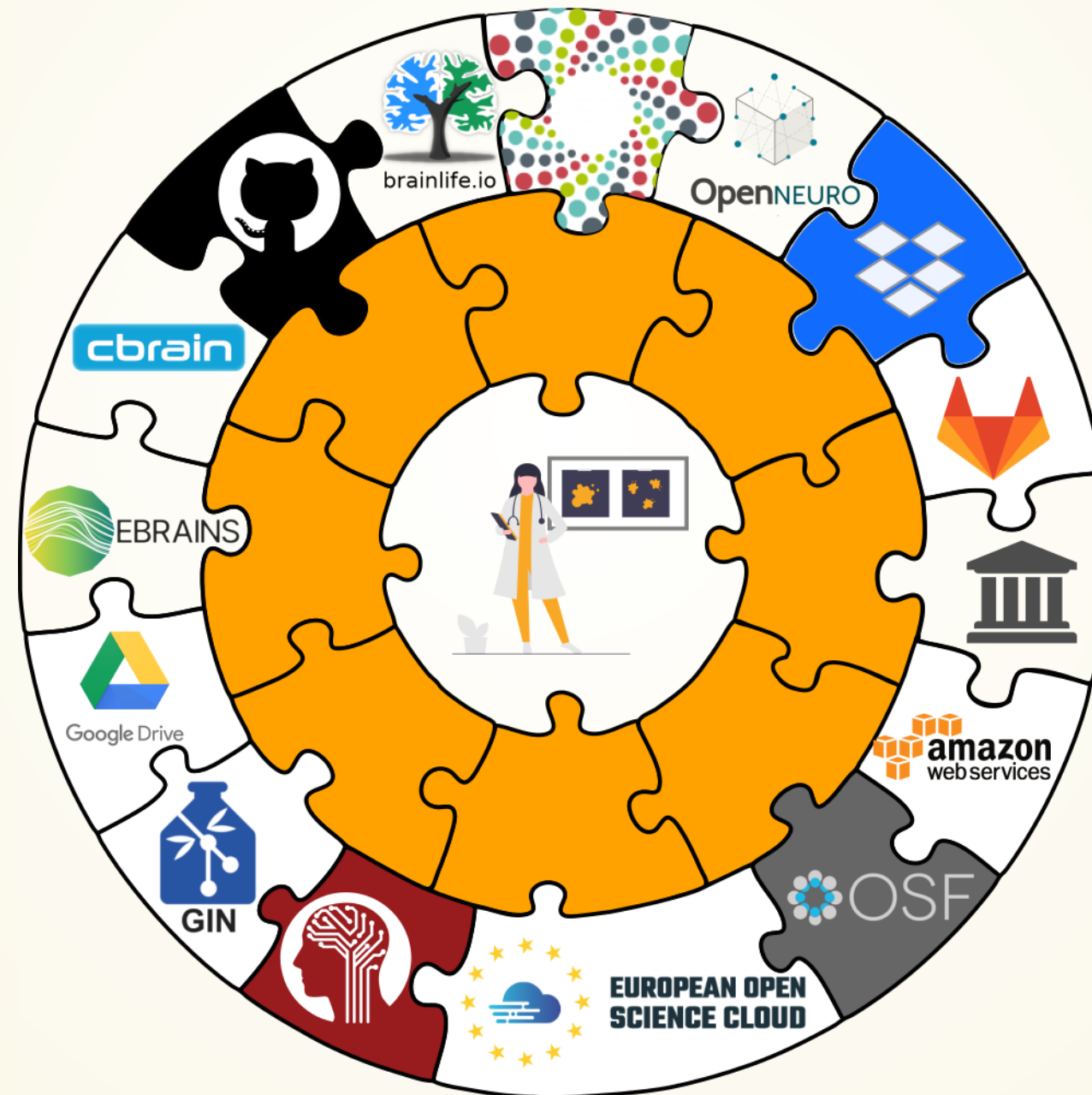
- DataLad is built to maximize interoperability and use with hosting and storage technology



See the chapter [Third party infrastructure](#) for walk-throughs for different services

# INTEROPERABILITY

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



See the chapter [Third party infrastructure](#) for walk-throughs for different services

# GLOSSARY

## **Sibling (remote)**

Linked clones of a dataset. You can usually update (from) siblings to keep all your siblings in sync (e.g., ongoing data acquisition stored on experiment compute and backed up on cluster and external hard-drive)

## **Repository hosting service**

Webservices to host Git repositories, such as GitHub, GitLab, Bitbucket, Gin, ...

## **Third-party storage**

Infrastructure (private/commercial/free/...) that can host data. A "special remote" protocol is used to publish or pull data to and from it

## **Publishing datasets**

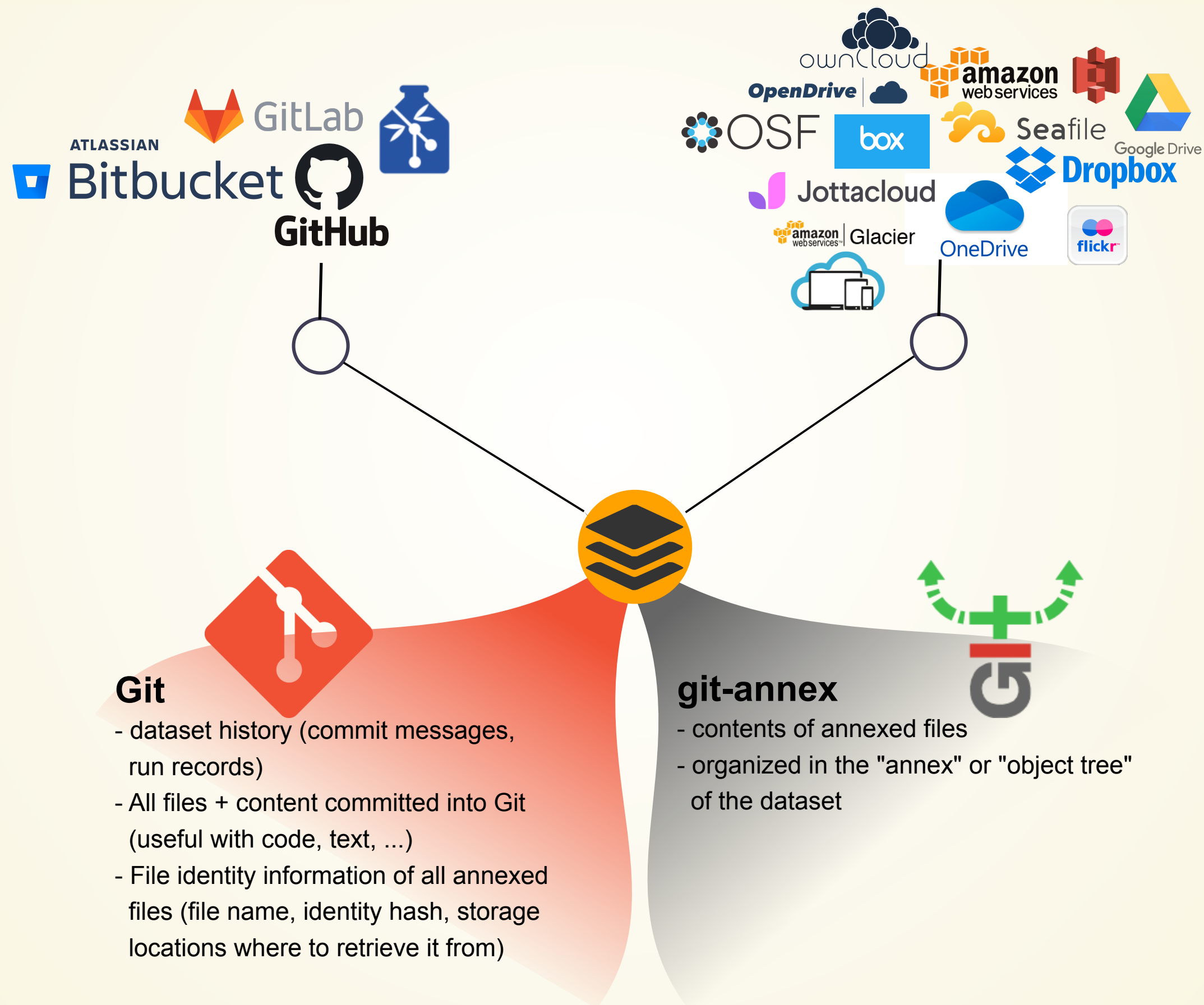
*Pushing* dataset contents (Git and/or annex) to a sibling using **datalad push**

## **Updating datasets**

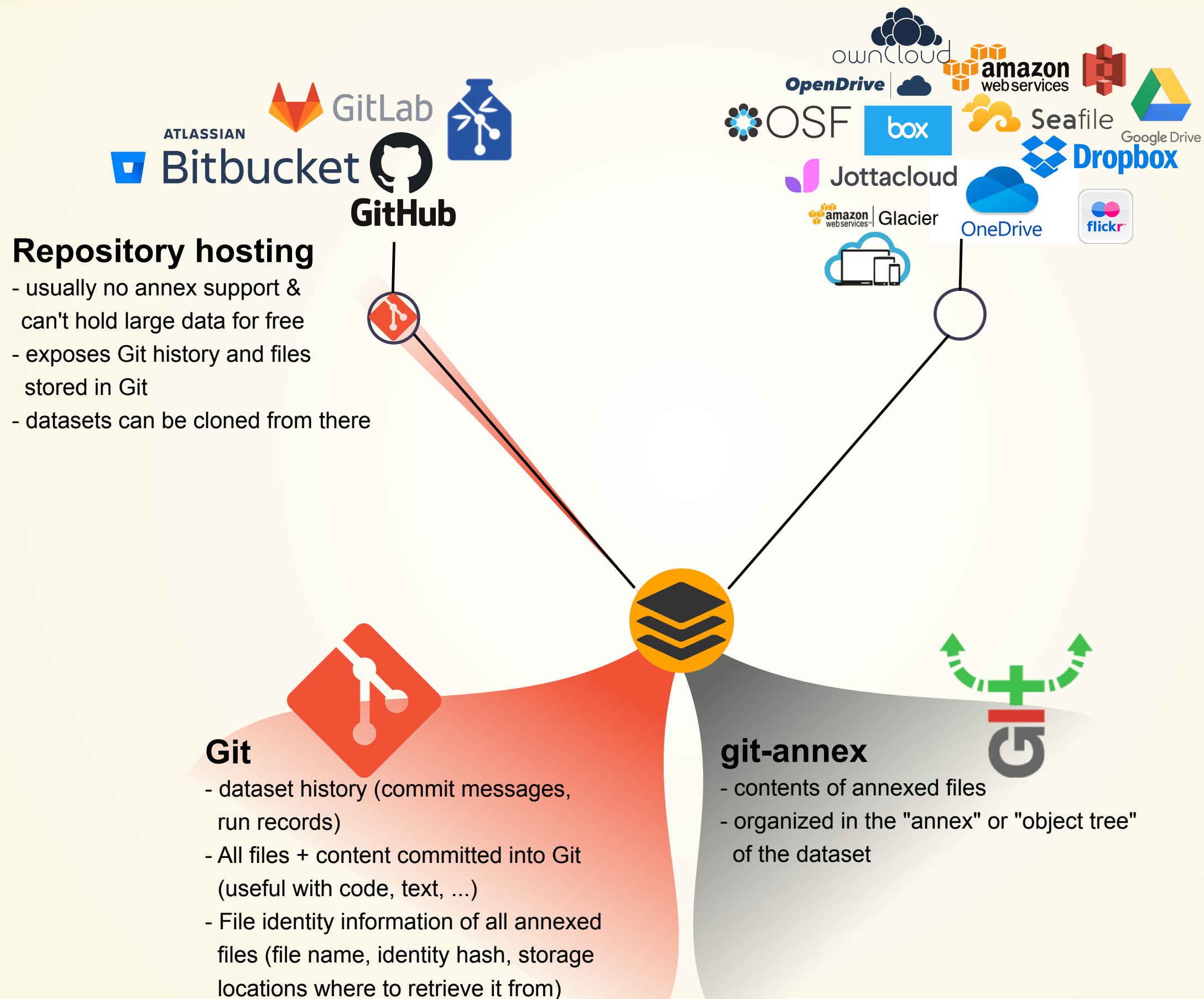
*Pulling* new changes from a sibling using **datalad update --merge**

# PUBLISHING DATASETS

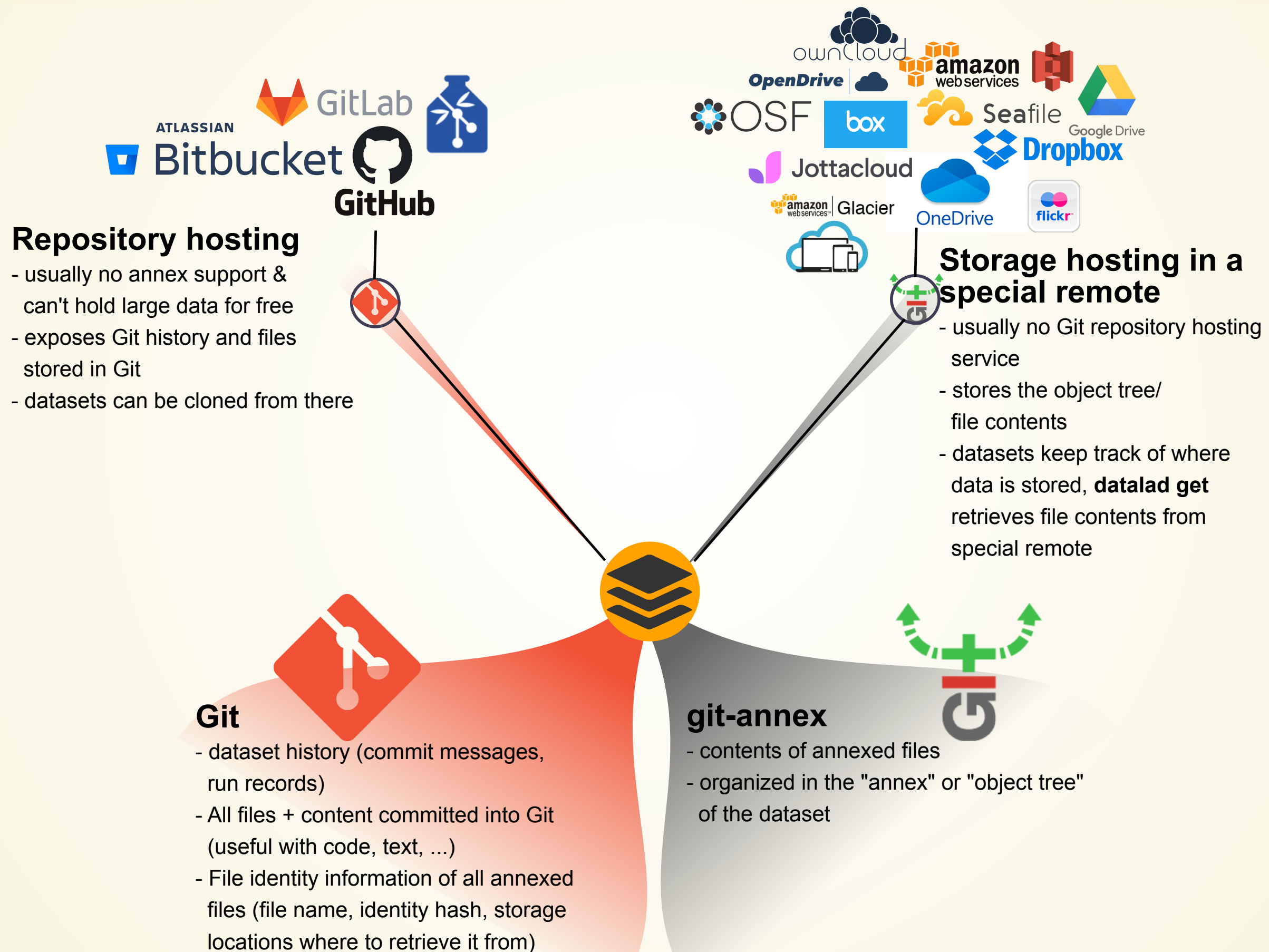
- Most public datasets separate content in Git versus git-annex behind the scenes



# PUBLISHING DATASETS



# PUBLISHING DATASETS

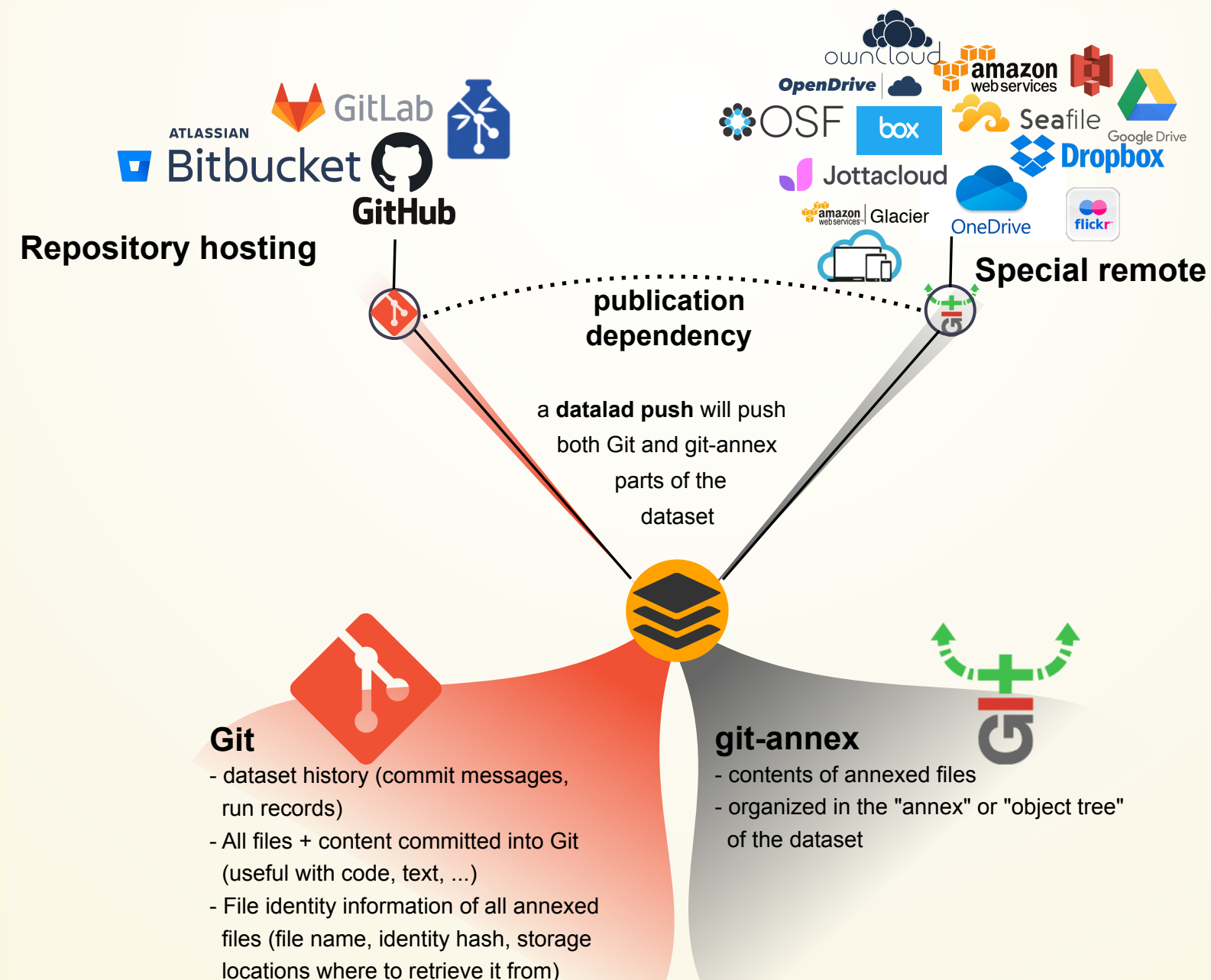


# PUBLISHING DATASETS

Typical case:

- Datasets are exposed via a private or public repository on a repository hosting service
- Data can't be stored in the repository hosting service, but can be kept in almost any third party storage
- Publication dependencies automate pushing to the correct place, e.g.,

```
$ git config --local remote.github.datalad-publish-depends gdrive
# or
$ datalad siblings add --name origin --url git@git.jugit.fzj.de:adswa/experiment-data.git --publish-depends s3
```



# PUBLISHING DATASETS

- Real-life example 1:  
GitHub for repository hosting, data hosting via datapub.fz-juelich.de + GNODE

The screenshot shows the GitHub repository page for `psychoinformatics-de/studyforrest-data-phase2`. The repository has 1 pull request, 4 issues, 6 stars, and 8 forks. The main content area displays a list of files and folders, including `.datalad`, `code`, `src`, `stimuli`, and a series of subdirectories (`sub-01` through `sub-20`). The `sub` directories are associated with the commit message "BF: Re-import respiratory trace after bug fix in converte...". The right sidebar provides additional information, including the repository's description, a link to `studyforrest.org`, a README, a license, a release history (showing the first public release on Mar 26, 2016), and a list of contributors (mih Michael Hanke, dakot Daniel Kottke, and adswa Adina Wagner).

psychoinformatics-de / studyforrest-data-phase2

Unwatch 1 Unstar 6 Fork 8

Code Issues 4 Pull requests 1 Actions Projects Security Insights

master 2 branches 1 tag Go to file Add file Code

mih Merge pull request #15 from adswa/ENH/README b5306e2 on May 7 77 commits

|          |                                                              |             |
|----------|--------------------------------------------------------------|-------------|
| .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 bug fix in converte... | 3 years ago |

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

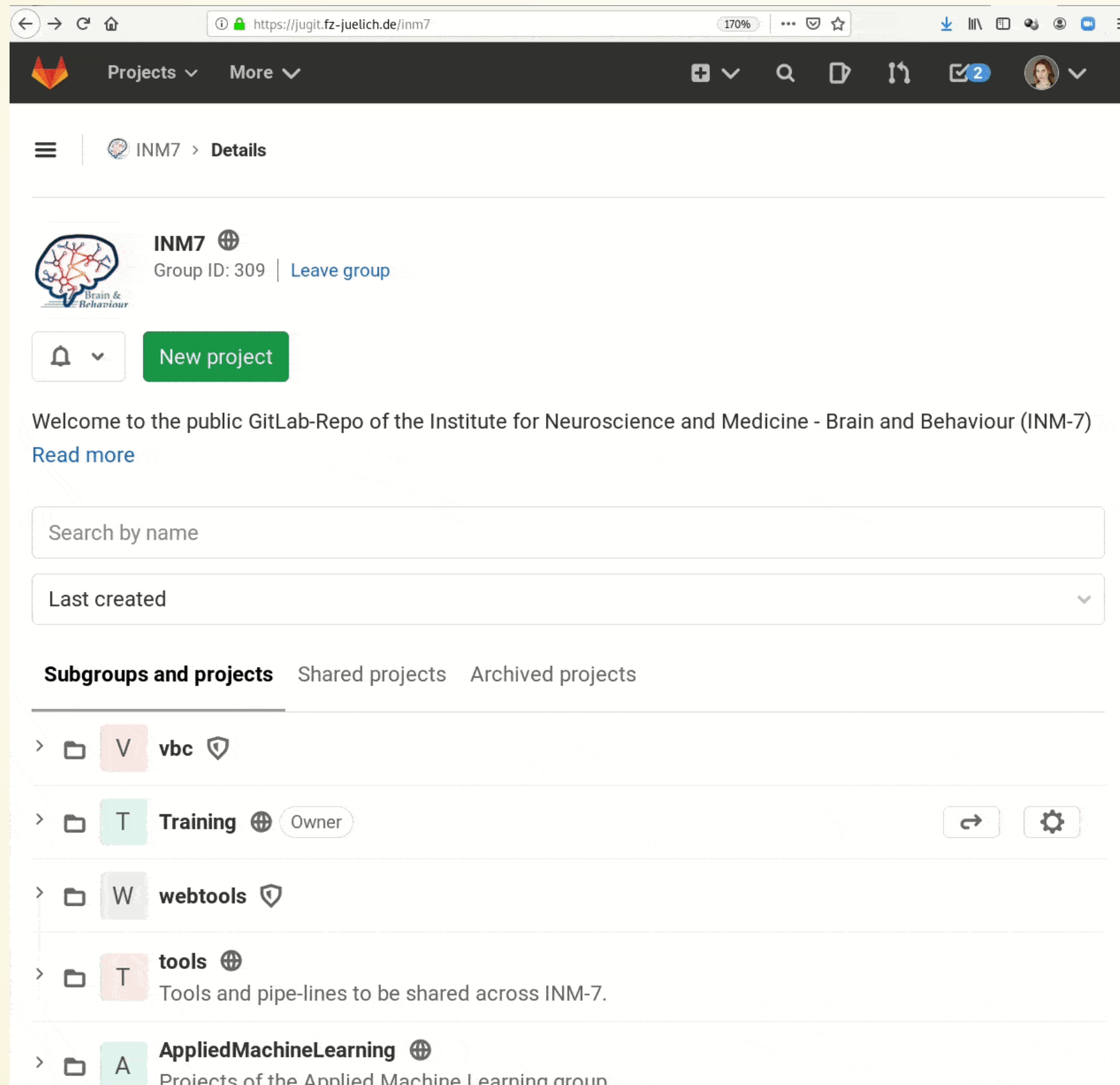
dakot Daniel Kottke

adswa Adina Wagner

Languages

# PUBLISHING DATASETS

- Real-life example 2:  
GitLab for repository hosting, data hosting via internal webserver (access restricted)



# PUBLISHING DATASETS

- Real-life example 3:  
GitHub for repository hosting, data hosting via Amazon S3 (requires DUA)

The screenshot shows the GitHub repository page for 'datalad-datasets/human-connectome-project-openaccess'. The repository is owned by 'datalad-datasets' and has 7 stars, 11 forks, and 1 pull request. The main branch is 'master' with 2 branches and 0 tags. The repository contains several files and folders, including '.datalad', 'HCP1200', '.gitmodules', '.noannex', 'DATA\_USE\_AGREEMENT...', and 'README.md'. The 'README.md' file is open, showing the title 'Get data from the Human Connectome Project Open Access dataset with DataLad' and a description of the dataset. The 'About' section on the right provides more details about the dataset, including the WU-Minn HCP1200 Data and a link to the Human Connectome Project website. The 'Releases' and 'Packages' sections are empty, and the 'Contributors' section lists three contributors: adswa, mih, and yarikoptio.

Search or jump to... Pull requests Issues Marketplace Explore

datalad-datasets / human-connectome-project-openaccess Unwatch 7 Unstar 11 Fork 1

<> Code Issues 5 Pull requests 1 Actions Projects Wiki Security Insights

master 2 branches 0 tags Go to file Add file Code

adswa Merge pull request #18 from datalad-datasets/RES... c71d3b6 28 days ago 22 commits

| File                  | Description                                                | Time          |
|-----------------------|------------------------------------------------------------|---------------|
| .datalad              | Let DataLad find subdatasets in the main dataset store ... | 14 months ago |
| HCP1200               | add a range of MNINonLinear/Results/ files for subjects... | 28 days ago   |
| .gitmodules           | Replace participants broken participant sub(sub)datasets   | 14 months ago |
| .noannex              | Turn into an actual DataLad dataset (no annex, though)     | 14 months ago |
| DATA_USE_AGREEMENT... | add copy of Data Usage Agreement from the HCP              | 14 months ago |
| README.md             | Adjust README with note on data from ConnectomeDB          | 7 months ago  |

README.md

## Get data from the Human Connectome Project Open Access dataset with DataLad

Made with DataLad

This dataset enables data retrieval with DataLad (0.12.2 or later) from the [HCP Open Access dataset](#) for users that accepted the WU-Minn HCP Consortium Open Access Data Use Terms and obtained valid AWS credentials via [db.humanconnectome.org](#).

### Human Connectome Project

The Human [Connectome Project](#) (HCP) aims to construct a map of the complete structural and functional neural connections in vivo within and across individuals.

Its 'WU-Minn HCP Open Access Data' data release includes high-resolution 3T MR scans from young healthy adult twins and non-twin siblings (ages 22-35) using four imaging modalities: structural images (T1w and T2w), resting-state fMRI (rfMRI), task-fMRI (tfMRI), and high angular resolution diffusion imaging (dMRI). It further includes behavioral and other individual subject measure data for all, and MEG data and 3T MR data for a subset of subjects (twins).

WU-Minn HCP1200 Data:  
3T/7T MR scans from young healthy adults twins and non-twin siblings (ages 22-35)  
[T1w, T2w, resting-state and task fMRI, high angular resolution dMRI]

[db.humanconnectome.org/...](#)

Readme

### Releases

No releases published  
[Create a new release](#)

### Packages

No packages published  
[Publish your first package](#)

### Contributors 3

- adswa Adina Wagner
- mih Michael Hanke
- yarikoptio Yaroslav Halche...

# PUBLISHING DATASETS

Special case 1: repositories with annex support



## Repositories with annex support

- examples: GIN (gin.g-node.org), GitLab instances with enabled annex support
- can hold large data for free
- exposes Git history and all files + content
- datasets can be cloned from there

### Git

- dataset history (commit messages, run records)
- All files + content committed into Git (useful with code, text, ...)
- File identity information of all annexed files (file name, identity hash, storage locations where to retrieve it from)

### git-annex

- contents of annexed files
- organized in the "annex" or "object tree" of the dataset

# PUBLISHING DATASETS

## Special case 2: Special remotes with repositories



# PUBLISHING DATASETS

Special case 1: repositories with annex support

[LIVE DEMO GIN]

# PUBLISHING DATASETS

Special case 2: Special remotes with repositories

Requires the DataLad extension `datalad-osf`

[LIVE DEMO OSF]

# PUBLISHING DATASETS

## Special case 2: Special remotes with repositories

Requires the DataLad extension `datalad-next`

### [DEMO WEBDAV]

Deposit and retrieve DataLad datasets with WebDAV services

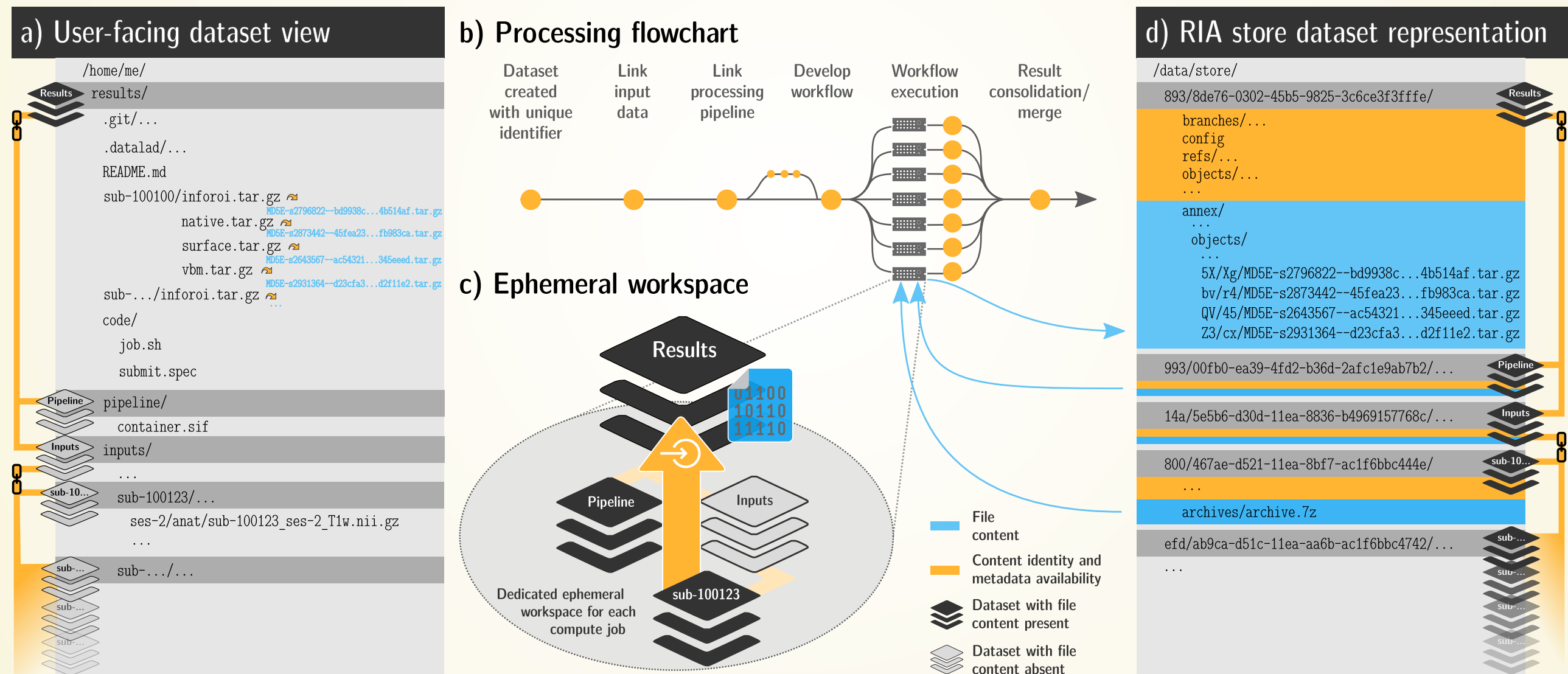


source: [youtube.com/watch?v=XkcwpqPQHQY](https://youtube.com/watch?v=XkcwpqPQHQY)

# PUBLISHING DATASETS

## Special case 3: RIA stores for dataset hosting/backup

Tutorial for large scale, reproducible computation: [github.com/psychoinformatics-de/fairly-big-processing-workflow](https://github.com/psychoinformatics-de/fairly-big-processing-workflow)



# PUBLISHING DATASETS

DataLad can create siblings from the command line for the following services:

## **GitHub**

```
datalad create-sibling-github
```

## **GitLab**

```
datalad create-sibling-gitlab
```

## **Gin**

```
datalad create-sibling-gin
```

## **Gogs**

```
datalad create-sibling-gogs
```

## **local or remote paths**

```
datalad create-sibling
```

## **RIA stores**

```
datalad create-sibling-ria
```

## **Open Science Framework (needs datalad-osf)**

```
datalad create-sibling-osf
```

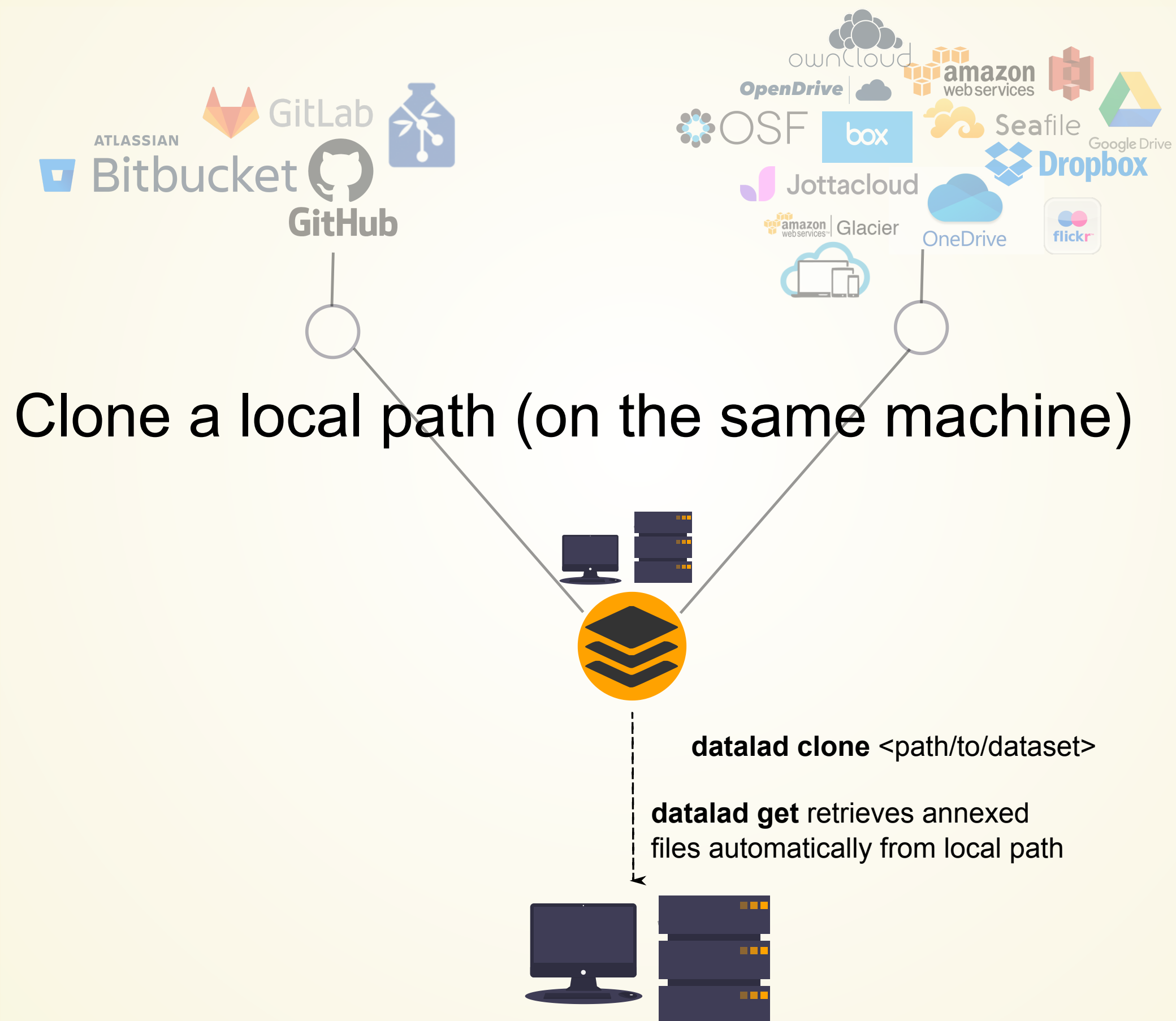
## **WebDAV-based hosting (e.g., Sciebo, EOSC; needs datalad-next)**

```
datalad create-sibling-webdav
```

(Additional services being worked on at this moment: Dataverse, ebrains;  
Get in touch with additional service support requests)

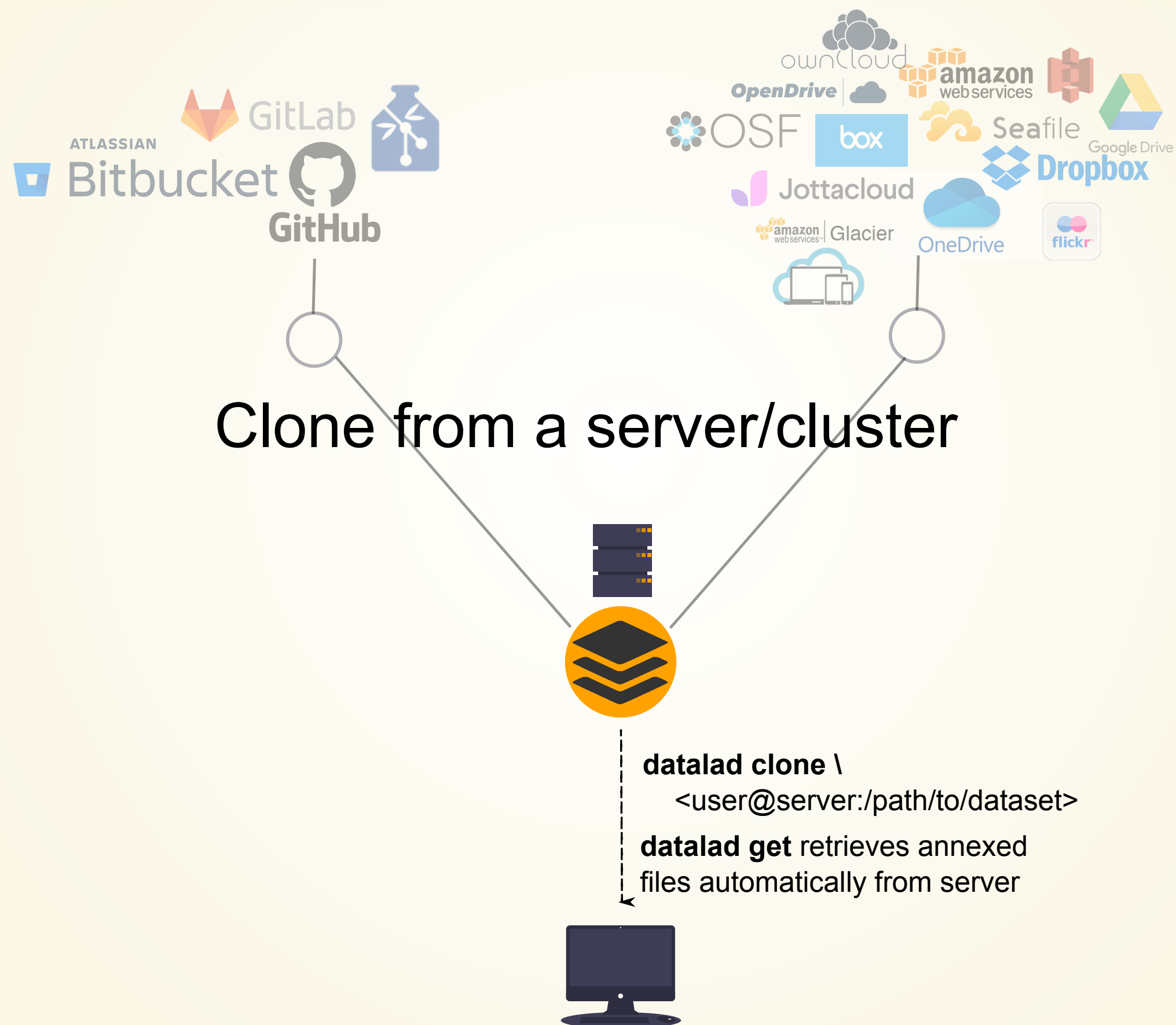
# CLONING DATALAD DATASETS

How does cloning dataset feel like for a consumer?



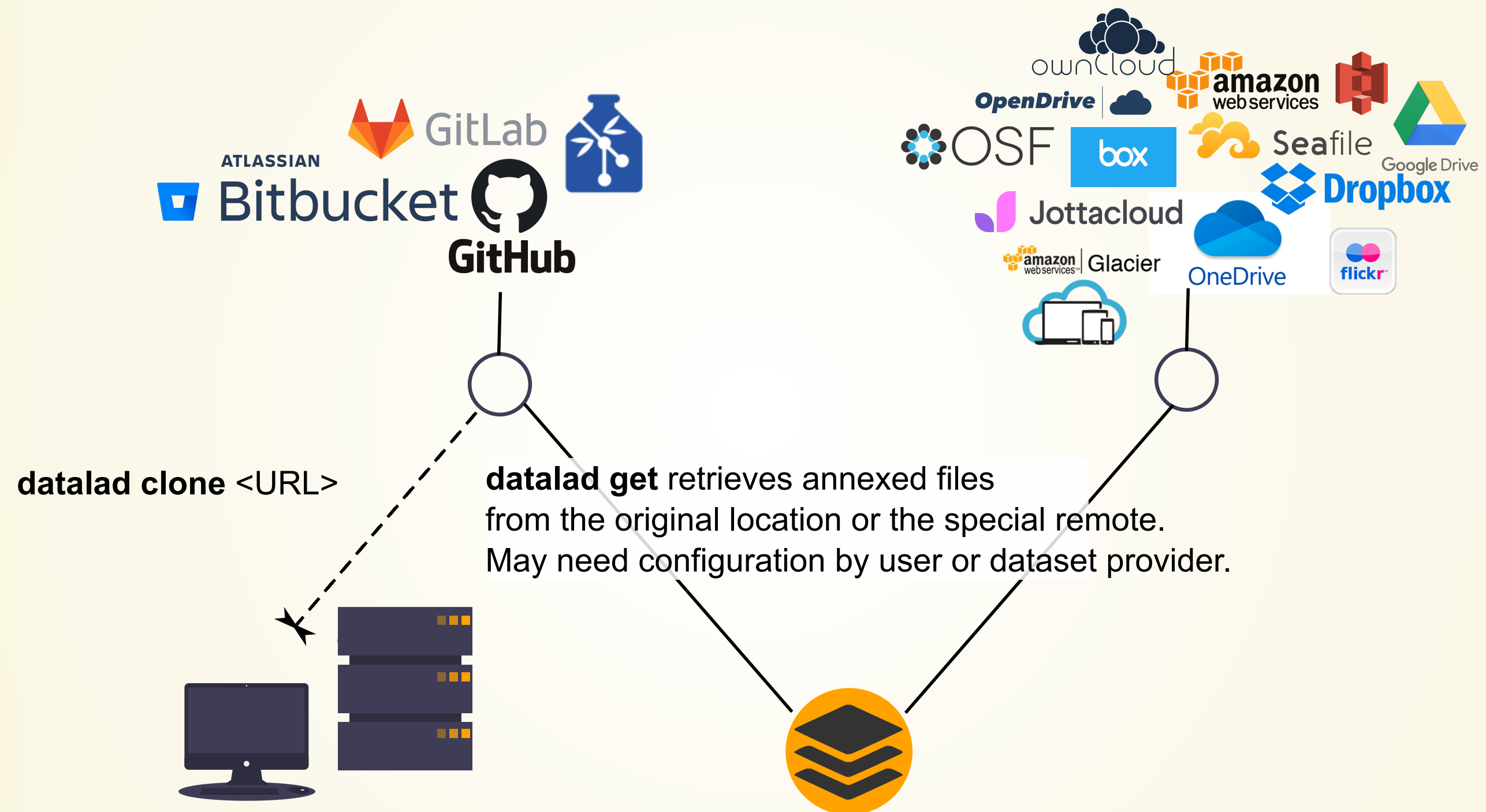
# CLONING DATALAD DATASETS

How does cloning dataset feel like for a consumer?



# CLONING DATALAD DATASETS

How does cloning dataset feel like for a consumer?



## Clone from a remote URL

# CLONING DATALAD DATASETS

Let's take a look at the special cases:  
[LIVE DEMO CLONING GIN]

# CLONING DATALAD DATASETS

Let's take a look at the special cases:

Requires the DataLad extension `datalad-osf`

[LIVE DEMO CLONING OSF]

# CLONING DATALAD DATASETS

Let's take a look at the special cases:

Requires the DataLad extension `datalad-next`

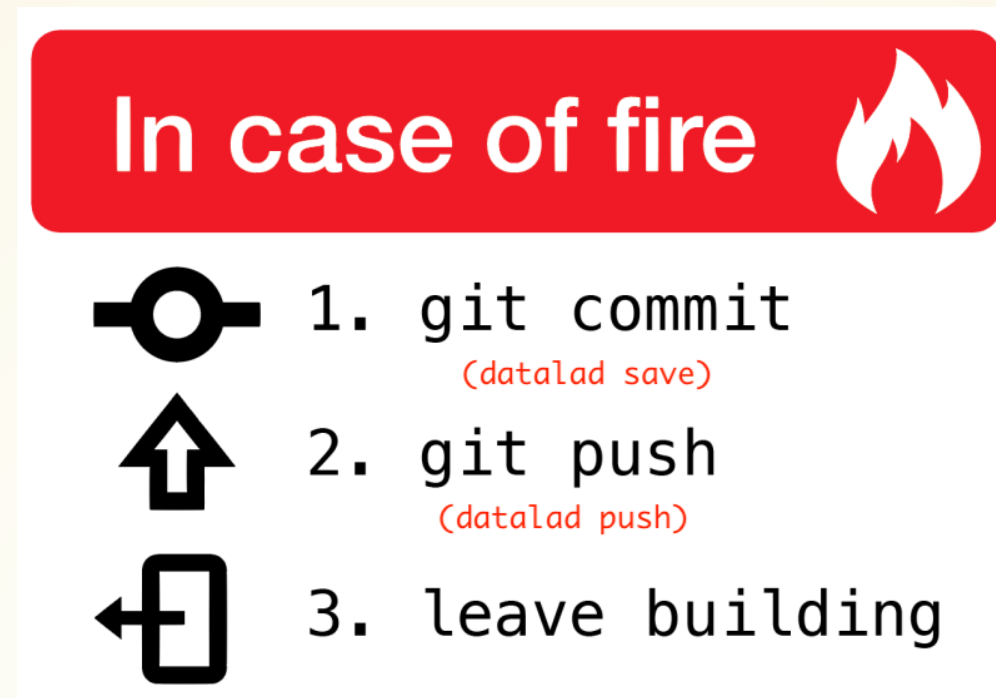
[DEMO CLONING WebDAV]

Deposit and retrieve DataLad datasets with WebDAV services



source: [youtube.com/watch?v=XkcwpqPQHQY](https://youtube.com/watch?v=XkcwpqPQHQY)

# SUMMARY: DATA PUBLICATION



datasets can have "siblings", linked clones in other places

Those can be local or remote, on commercial, free, or personal infrastructure

**Typical repository hosting services do not host annexed contents**

A notable exception is Gin

**Typical storage providers do not host Git repositories**

but datalad extensions can make it possible for certain services, such as the OSF

**Despite the different possible services, operations are streamlined**

**clone** installs datasets, **get** retrieves data, **push** publishes (new changes in) datasets, **update** pulls dataset updates. This remains the case even if underlying data hosting changes.

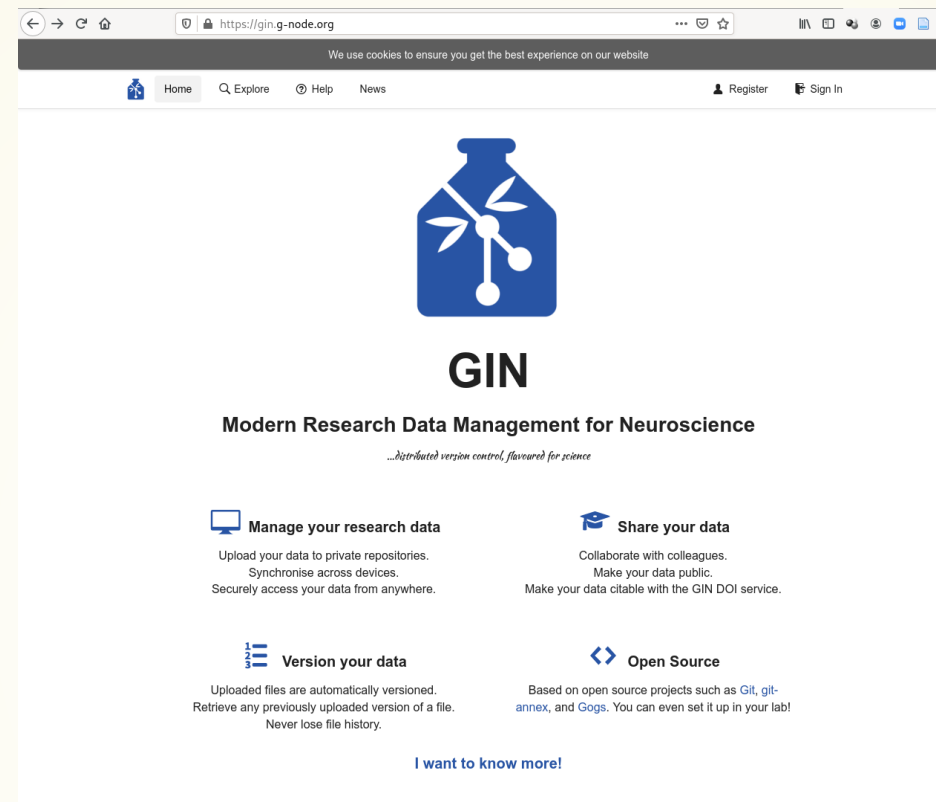
**Siblings serve multiple purposes:**

Personal back-up that's easy to sync; Publicly or privately exposed files to share with (selected) others; Entrypoints for collaborations or others' contributions; ...

# PUBLISH YOUR OWN DATASET

Code: [psychoinformatics-de.github.io/rdm-course/03-remote-collaboration/index.html#publishing-datasets-to-gin](https://psychoinformatics-de.github.io/rdm-course/03-remote-collaboration/index.html#publishing-datasets-to-gin)

# USING GIN FOR DATA PUBLICATION



## Gin has a few advantages for publishing data

- DataLad Integration: Convenience commands to create siblings
- Annex support: Easiest possible publication, preview and individual download of annexed contents in the webinterface
- Open Science support: Archive datasets to obtain a DOI; ensures minimal metadata and a license
- Private or Public repositories
- Runs on European infrastructure (some data protection officers like this)
- Free, and with yet unlimited storage

# USING GIN FOR DATA PUBLICATION

←

→

↺

https://gin.g-node.org/Innrwttkhn/highspeed-bids

110%

 Dashboard

 Issues

 Pull Requests

 Explore

 Help

 News

+

▼



▼

 Innrwttkhn

/

highspeed-bids

DOI

10.12751/g-node.4ivuv8

 Watch

1

 Star

1

 Fork

1

 Files

 Issues 0

 Pull Requests 0

DataLad dataset containing BIDS-converted MRI data used in Wittkuhn & Schuck, 2020, Nature Communications. For details, see: <https://wittkuhn.mpib.berlin/highspeed/>

 60 Commits

 2 Branches

 0 Releases

 Branch: master ▼

highspeed-bids

GIN

HTTPS

SSH

gin get Innrwttkhn/highspeed





|                                                                                                    |            |                                                                 |              |
|----------------------------------------------------------------------------------------------------|------------|-----------------------------------------------------------------|--------------|
|  Lennart Wittkuhn | 0f424f45da | Merge remote-tracking branch 'origin/master'                    | 5 months ago |
|  .datalad       | a4652dfd2b | Configure metadata type(s)                                      | 1 year ago   |
|  .heudiconv     | a2b905b439 | add defaced BIDS-converted MRI data after running heudiconv ... | 1 year ago   |
|  code           | 36bb8a3e68 | update documentation                                            | 1 year ago   |
|  input          | f4b832b93c | [DATALAD] Recorded changes                                      | 1 year ago   |
|  logs           | 4223e4692d | add empty logs directory                                        | 1 year ago   |
|  sub-01         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-02         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-03         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-04         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-05         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-06         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-07         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-08         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-09         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-10         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-11         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-12         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-13         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |
|  sub-14         | 5dd8eb86d8 | [DATALAD RUNCMD] create IntendedFor fields in fieldmap .json... | 1 year ago   |

# USING GIN FOR DATA PUBLICATION

- Step 1: Create a Gin account (requires an email address)
- Step 2: Generate and upload an SSH key
- Step 3: Create and register a sibling repository
- Step 4: Publish your dataset
- Step 5: Update your dataset

## SUMMARY: PUBLISHING AND UPDATING DATA (GIN)

**Gin is a free repository hosting service**

To publish datasets to Gin, you need an account and an SSH key

**DataLad has built-in integration with `datalad create-sibling-gin`**

This requires generating an access token

**Gin has annex support**

`datalad push` published all dataset contents and the Git history

**The dataset can be cloned from Gin by others**

If the dataset is public, this does not even require a Gin account

**You can still publish your dataset to (your lab's) GitHub/GitLab/other places**

and use Gin only for data hosting. Walkthrough: [handbook.datalad.org/basics/101-139-gin.html#ginbts](https://handbook.datalad.org/basics/101-139-gin.html#ginbts)

**Next: Let's collaborate!**

# PUBLICATION AND COLLABORATION EXERCISE

|                                                                | Local                                                                                                                                                                                                                                                                 | Published                                                                                                                                                                                                                                                             | Collaborator                                                                                                                                                                  |                                                                                                                                                                               |
|----------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                |                                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                       | Local                                                                                                                                                                         | Published                                                                                                                                                                     |
| Clone dataset                                                  |                                                                                                                                                                                    |                                                                                                                                                                                                                                                                       |                                                                                            |                                                                                                                                                                               |
| Save dataset modification                                      |  +                                                                                              |                                                                                                                                                                                                                                                                       |  +      |                                                                                                                                                                               |
| Create dataset sibling on GIN                                  |  +                                                                                              |                                                                                                                                                                                                                                                                       |  +      |                                                                                                                                                                               |
| Push dataset to GIN                                            |  +                                                                                              |  +                                                                                              |  +      |  +      |
| Safely remove local dataset                                    |                                                                                                                                                                                                                                                                       |  +                                                                                              |                                                                                                                                                                               |  +      |
| Clone own dataset from GIN again                               |  +                                                                                            |  +                                                                                            |  +    |  +    |
| Save another dataset modification                              |  +                                                                                          |  +                                                                                          |  +  |  +  |
| Push dataset update to GIN                                     |  +                                                                                          |  +                                                                                          |  +  |  +  |
| Add a collaborator's GIN project as a sibling to local dataset |  +                                                                                          |  +                                                                                          |  +  |  +  |
| Merge collaborator's work                                      |  +  +  |  +                                                                                          |  +  |  +  |
| Push dataset update to GIN                                     |  +  +  |  +  +  |  +  |  +  |