TOWARDS COMPUTATIONAL REPRODUCIBILITY WHEN WORKING WITH VERY LARGE DATASETS

Adina Wagner
mas.to/@adswa

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

Slides: DOI 10.5281/zenodo.7835784

ACKNOWLEDGEMENTS

Co-authors

- **Laura K. Waite***
- **Małgorzata Wierzba***
- Felix Hoffstaedter
- Alexander Q. Waite
- Benjamin Poldrack
- Simon B. Eickhoff
- Michael Hanke

DataLad software & ecosystem

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

Funders

NSF 1429999

germany-usa

SPONSORED BY THE Federal Ministry of Education and Research

BMBF 01GQ1411

JÜLICH Forschungszentrum

DFG

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

OpenNEURO

eBRAIN Health

MOTOR SFB 1451

brainlife.io

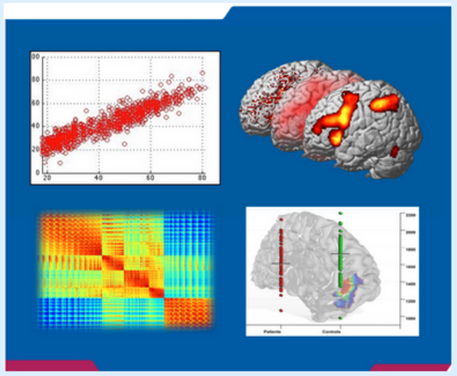
cbrain

VirtualBrainCloud

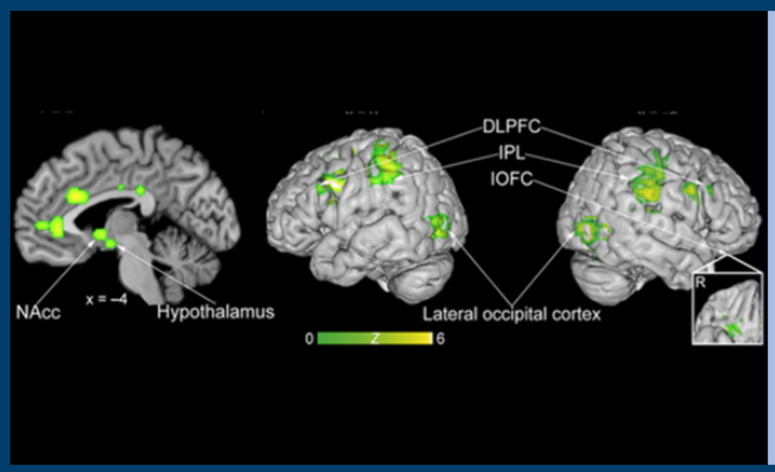
Welcome to the INM-7 Homepage

Our institute is specialized in integrating multi-modal neuroimaging data and using this information to develop machine learning models for predicting complex phenotypes.

Director: Prof. Simon Eickhoff



RESEARCH FOCUS



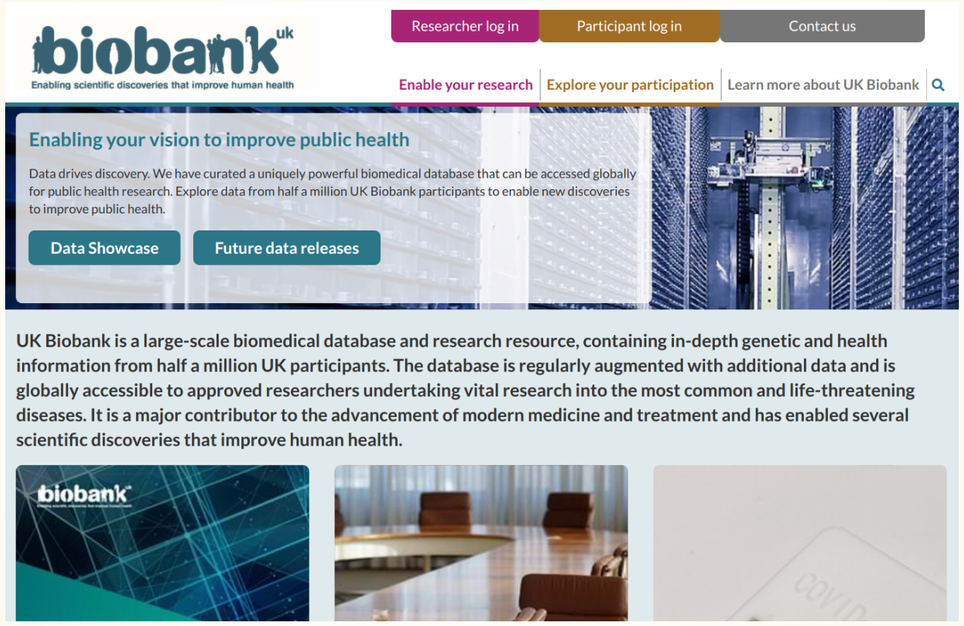
Understanding human brain organization

We develop and apply novel methods for mapping the regional organization of the human brain into cortical areas as well as the topography of large-scale, distributed networks recruited by cognitive, socio-affective or sensorimotor tasks.

www.fz-juelich.de/en/inm/inm-7

FAIRLY BIG: SCALING UP

Objective: Process the UK Biobank (imaging data)



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

CHALLENGES

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

FAIRLY BIG RESULTS

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

- Rendered from provenance records, automatically captured in the output dataset. Full video: https://youtube.com/datalad
- Two full (re-)computations, programmatically comparable, verifiable, reproducible -- on any system with data access

www.nature.com/scientificdata

scientific data

Check for updates

OPEN

ARTICLE

FAIRly big: A framework for computationally reproducible processing of large-scale data

Adina S. Wagner[1,4]✉, Laura K. Waite[1,4], Małgorzata Wierzba[1,2,4], Felix Hoffstaedter[1], Alexander Q. Waite[1], Benjamin Poldrack[1], Simon B. Eickhoff[1,3] & Michael Hanke[1,3]

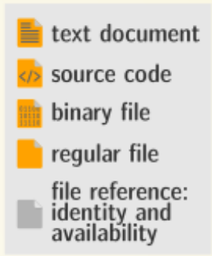
Large-scale datasets present unique opportunities to perform scientific investigations with unprecedented breadth. However, they also pose considerable challenges for the findability, accessibility, interoperability, and reusability (FAIR) of research outcomes due to infrastructure limitations, data usage constraints, or software license restrictions. Here we introduce a DataLad-based, domain-agnostic framework suitable for reproducible data processing in compliance with open science mandates. The framework attempts to minimize platform idiosyncrasies and performance-related complexities. It affords the capture of machine-actionable computational provenance records that can be used to retrace and verify the origins of research outcomes, as well as be re-executed independent of the original computing infrastructure. We demonstrate the framework's performance using two showcases: one highlighting data sharing and transparency (using the studyforrest.org dataset) and another highlighting scalability (using the largest public brain imaging dataset available: the UK Biobank dataset).

https://www.nature.com/articles/s41597-022-01163-2

Datalad

http://datalad.org (Click here to try it in your browser)

EXHAUSTIVE TRACKING OF RESEARCH COMPONENTS



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

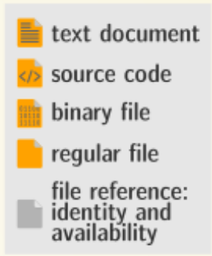
```
# turn any directory into a dataset
# with version control


% datalad create <directory>
```

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


% datalad save
```

CAPTURE COMPUTATIONAL PROVENANCE



Which data was needed at which version, as input into which code, running with what parameterization in which computional 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 for 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>
```

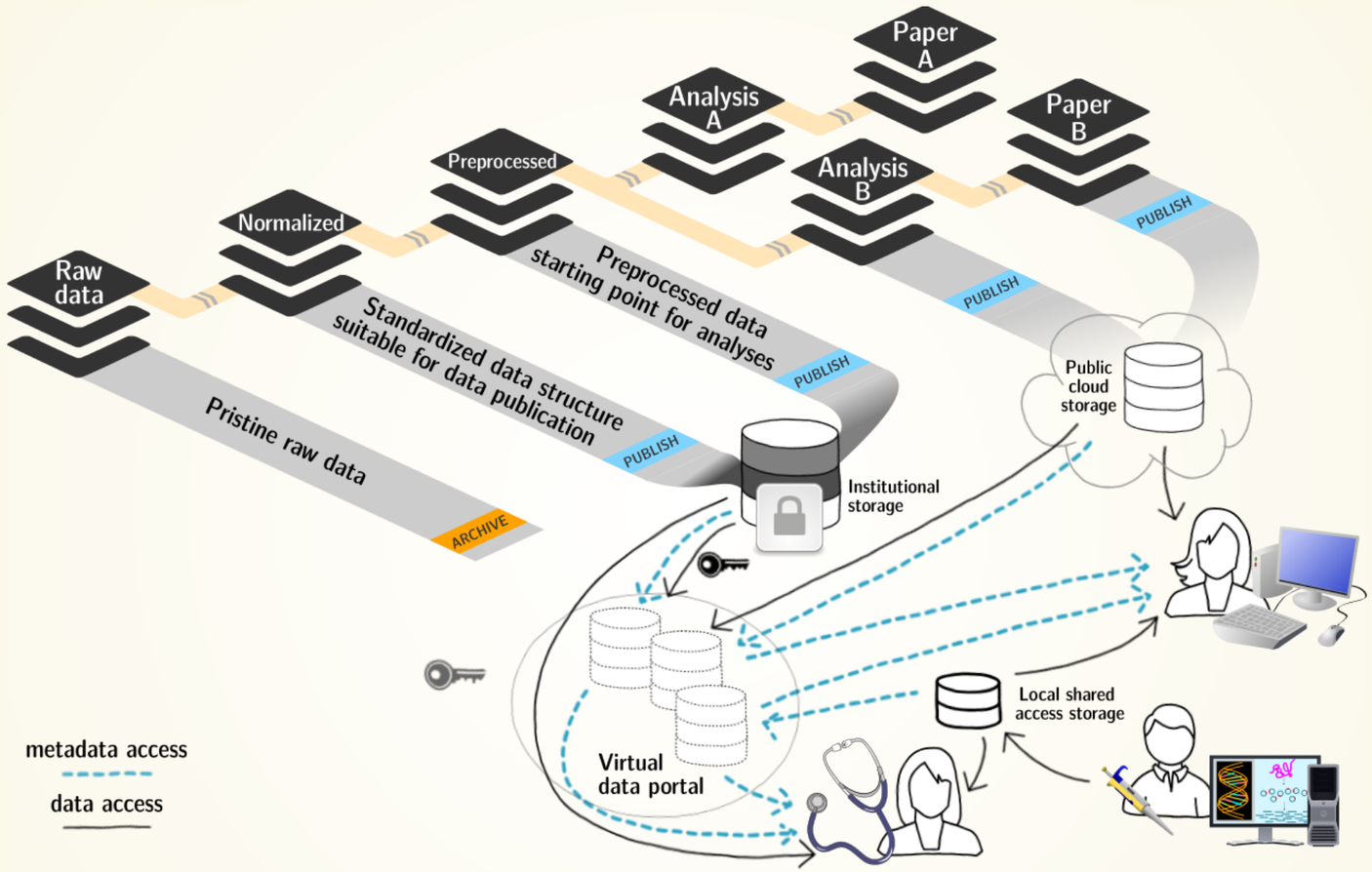
ULTIMATE GOAL: (RE-)USABILITY

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>
```

DATALAD: MANAGE (CO-)EVOLUTION OF DIGITAL OBJECTS

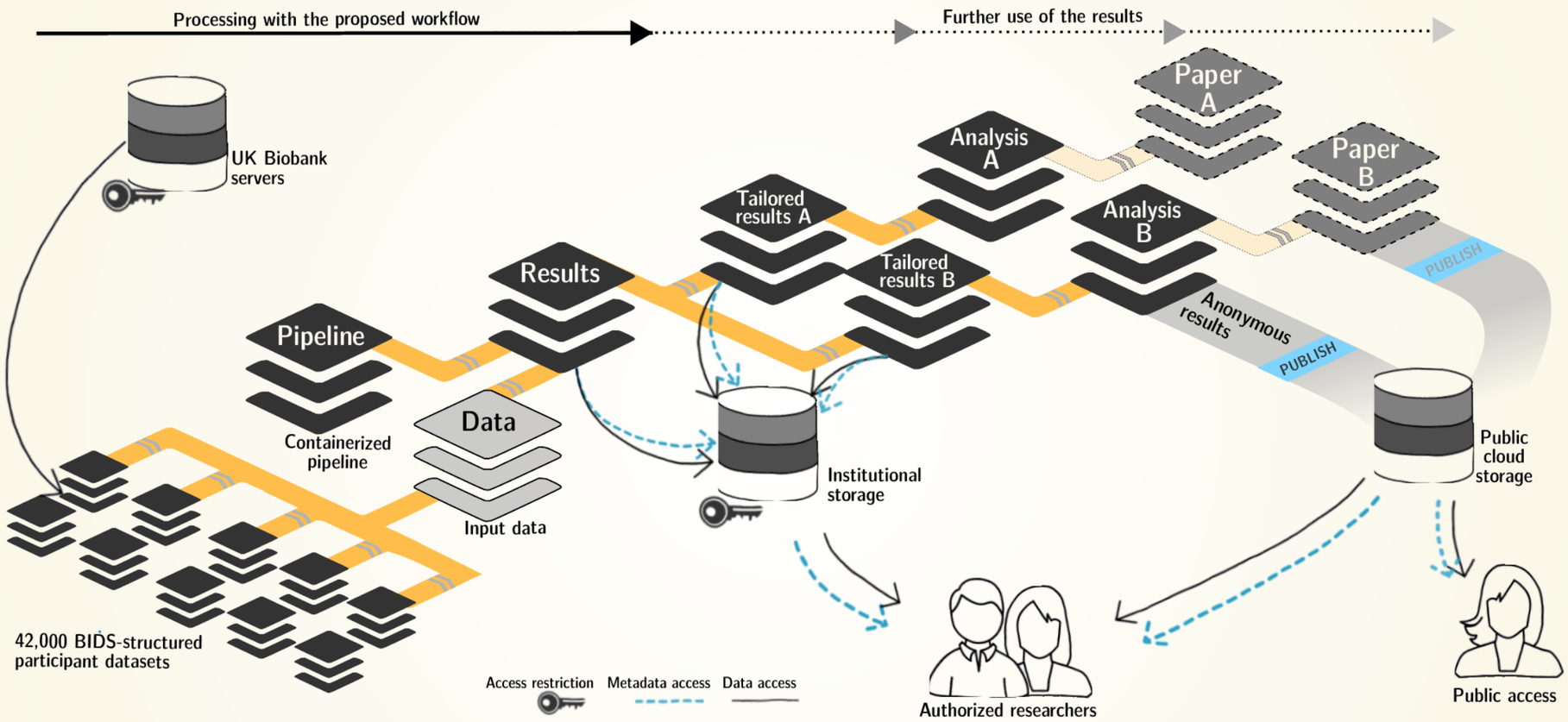


Consume, create, curate, analyze, publish, and query data with full provenance capture and "universal" metadata support.

DataLad is free and open source (MIT-licensed). http://datalad.org

Halchenko, Meyer, Poldrack, ... & Hanke, M. (2021). DataLad: distributed system for joint management of code, data, and their relationship. Journal of Open Source Software, 6(63), 3262.

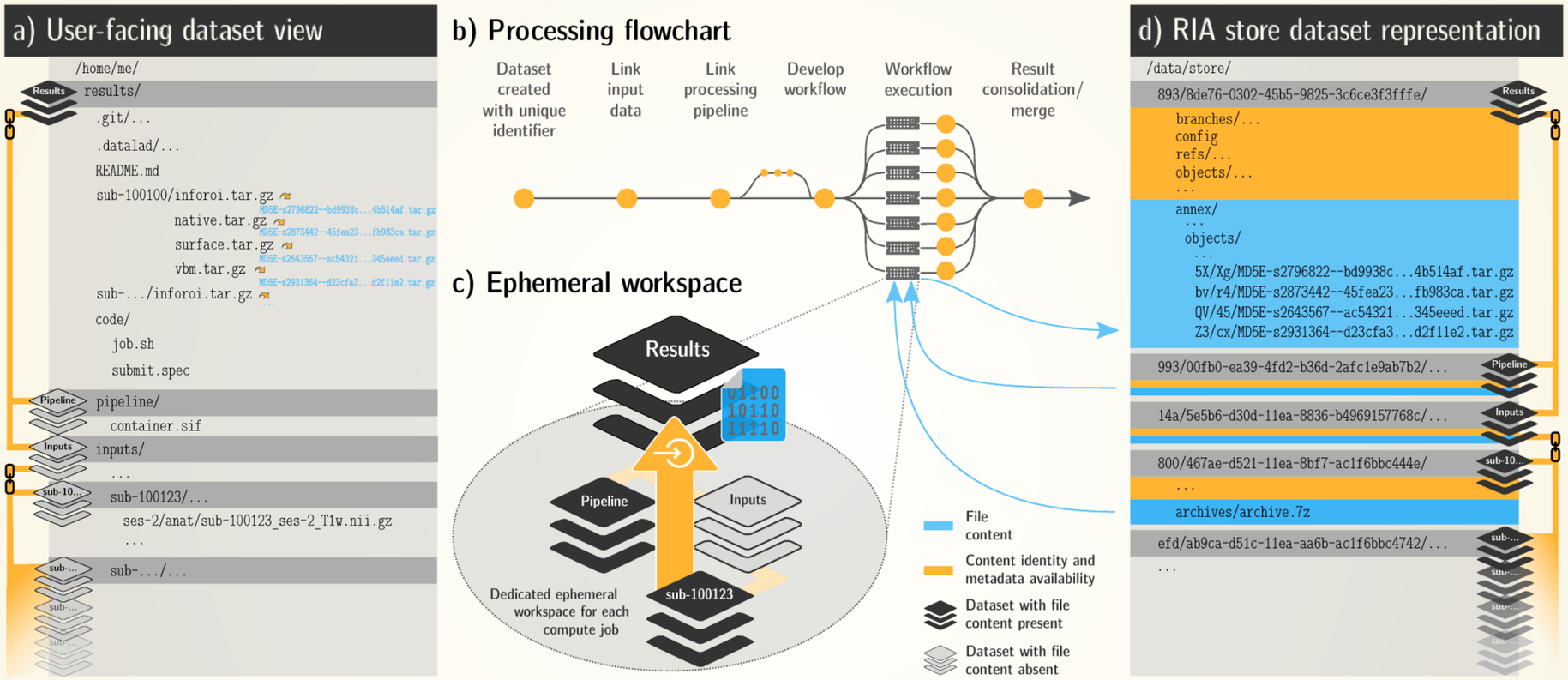
FAIRLY BIG SETUP



- UKB DataLad extension can track the evolution of the complete data release in DataLad datasets
- Full version history
- Native and BIDSified data layout

Wagner, Waite, Wierzba, Hoffstaedter, Waite, Poldrack, Eickhoff, Hanke (2021). FAIRly big: A framework for computationally reproducible processing of large-scale data.

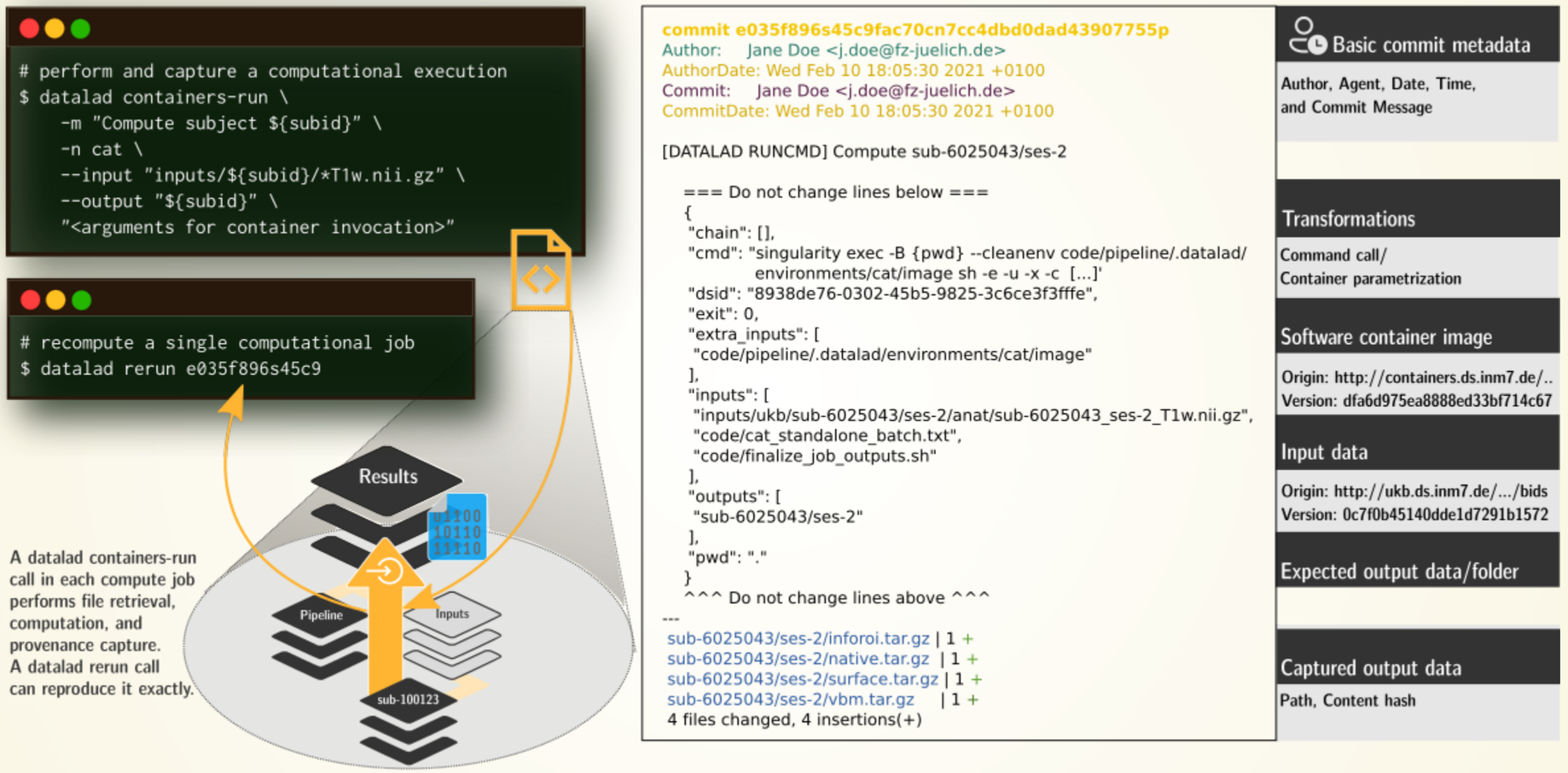
FAIRLY BIG WORKFLOW



- Common data representation in secure environments
- Content-agnostic persistent (encrypted) storage
- All computations in freshly bootstrapped emphemeral environments, only using information from a fully self-contained DataLad dataset

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

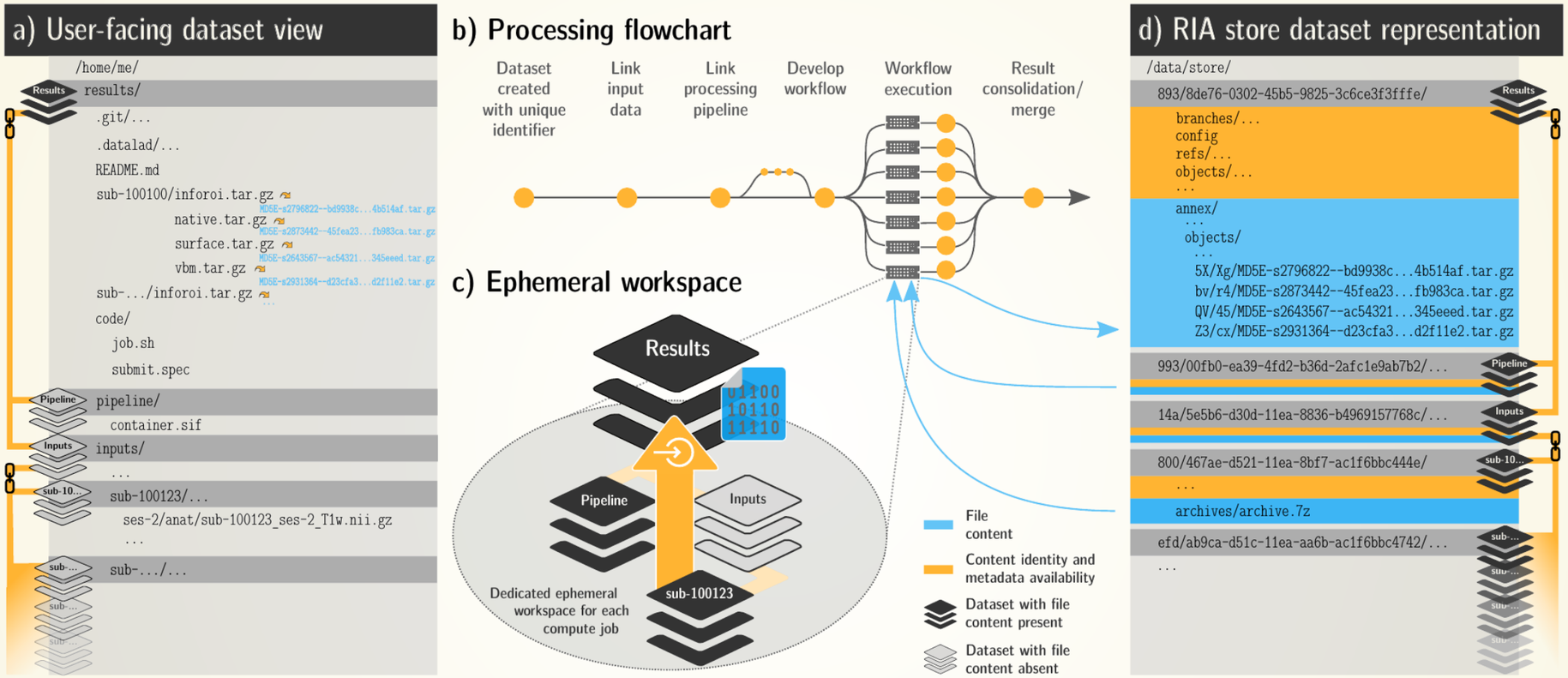
FAIRLY BIG PROVENANCE CAPTURE



- Every single pipeline execution is tracked
- Each execution individually reproducible without HPC access

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

FAIRLY BIG WORKFLOW



- Common data representation in secure environments
- Content-agnostic persistent (encrypted) storage
- All computations in freshly bootstrapped emphemeral environments, only using information from a fully self-contained DataLad dataset

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

CTHULU MERGE

```
List:       linux-kernel
Subject:    Re: [GIT PULL] regulator updates for v3.13-rc1
From:       Linus Torvalds <torvalds () linux-foundation ! org>
Date:       2014-01-21 19:16:57

Christ. When you start doing octopus merges, you don't do it by half
measures, do you?

I just pulled the sound updates from Takashi, and as a result got your
merge commit 2cde51fbd0f3. That one has 66 parents.

That kind of merge either needs to be split up, or gitk needs to be
made better about visualizing it, because it ends up being *so* wide
that the history is hard to read.

I think you'll find that having that many parents also breaks old
versions of git.

Anyway, I'd suggest you try to limit octopus merges to ~15 parents or
less to make the visualization tools not go crazy.  Maybe aim for just
10 or so in most cases.

It's pulled, and it's fine, but there's clearly a balance between
"octopus merges are fine" and "Christ, that's not an octopus, that's a
Cthulhu merge".

    Linus
```

42k-way octopus merge -- broke GitLab (JuGit)

READY FOR REPEATED RE-USE

Outcome: Actionable metadata record (in the form of a Git repository).
Orthogonalizes information on content identify and availability from actual data access

DATALAD CONTACT AND MORE INFORMATION

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

THANK YOU FOR YOUR ATTENTION!

Slides: DOI 10.5281/zenodo.7835784 (Scan the QR code)

Women in Neuroscience

Women neuroscientists are underrepresented in neuroscience. You can use the Repository for Women in Neuroscience to find and recommend neuroscientists for conferences, symposia or collaborations, and help making neuroscience more open & divers.