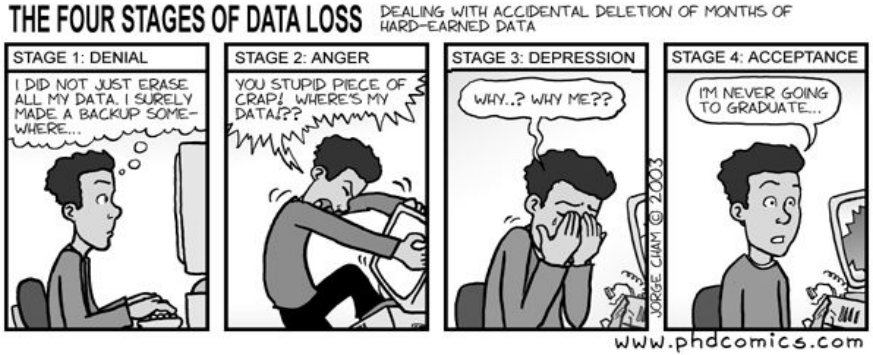
Gestão de Dados: Tudo, em Todos os Lugares, Ao Mesmo Tempo: Caso de uso

Bruno Direito [a] ***bruno.direito@uc.pt,***
João Pereira[b],
Alexandre Sayal[b]

[a] Centro de informática e Sistemas da Universidade de Coimbra, [b] Centro de Imagem Biomédica e Investigação Translacional (CIBIT), Universidade de Coimbra

https://www.themoviedb.org/movie/545611-everything-everywhere-all-at-once/images/posters

- *Everything Everywhere All at once*
  - Evelyn Wang, uma mãe sobrecarregada, embarca numa aventura multiversal onde descobre o seu potencial oculto para salvar todas as realidades.
  - **Há elementos comuns no multiverso**
    - **A gestão e interoperabilidade de dados entre pares (universos) é chave**

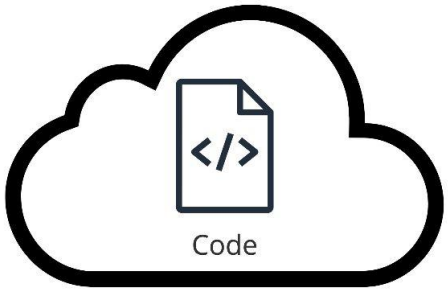


https://phdcomics.com/comics/archive.php?comicid=382

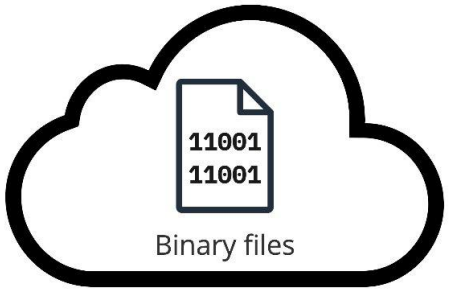
- *Os vários elementos que envolvem o processo científico não são estáticos*
  - *a sua partilha/ trabalho colaborativo é um desafio*

Componentes de um projeto de investigação

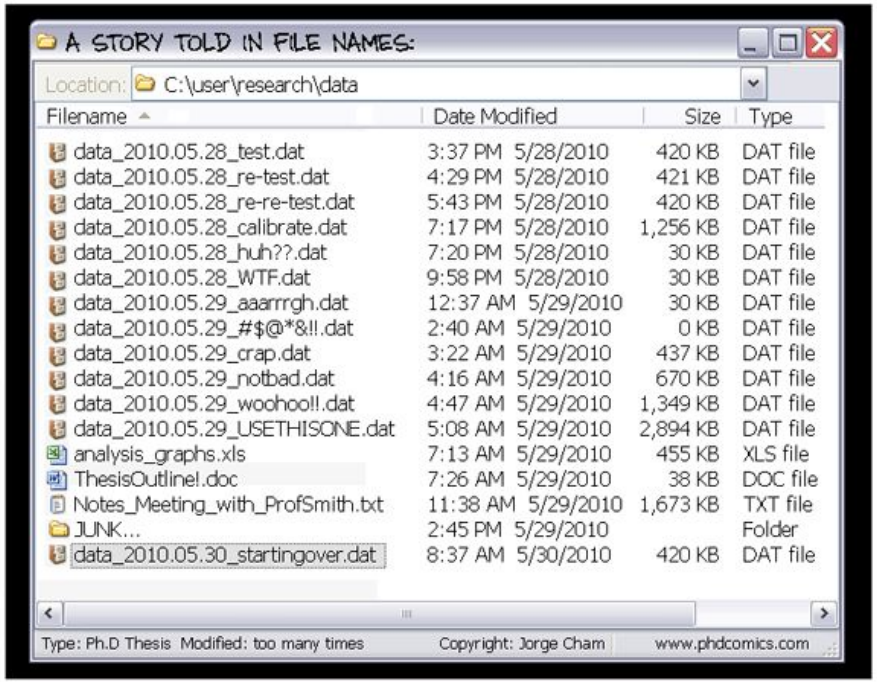
Google drive
Dropbox
OneDrive



GitHub



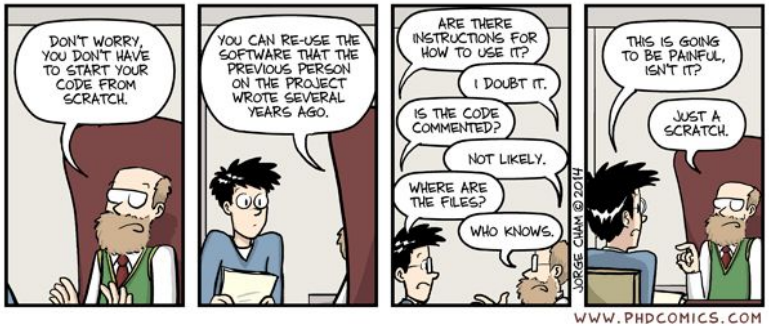
Dropbox
Data repositories



https://phdcomics.com/comics/archive.php?comicid=1323

Desafios do trabalho colaborativo em projetos de investigação

- **Partilha** (métodos de compactação *tarball, zip, rar* e encriptação)
- **Gestão de atualizações**
  - ***controlo de versões dos dados***
  - ***código***
    - processamento e análise de dados
- **Interoperabilidade**
  - Funciona no meu ecossistema (SO, versões de packages, etc.) e no dos meus colegas
- **Reprodutibilidade**
  - Temos os mesmos resultados



PI - research funds

Data manager (research assistant?, PI?)
(responsible for acquisition)

Data analyst - software A

Data analyst - software b

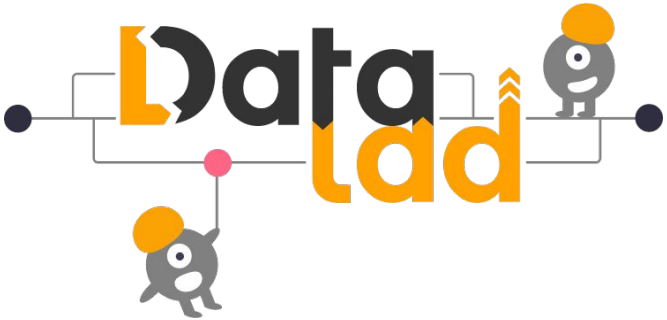
PI - another idea for the same data

Original RAW

RAW Copy 1

RAW Copy 2

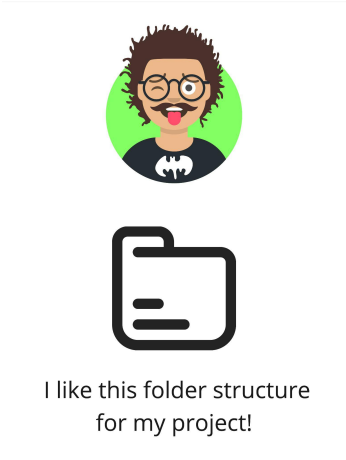
RAW Copy 3 - from RAW Copy 1

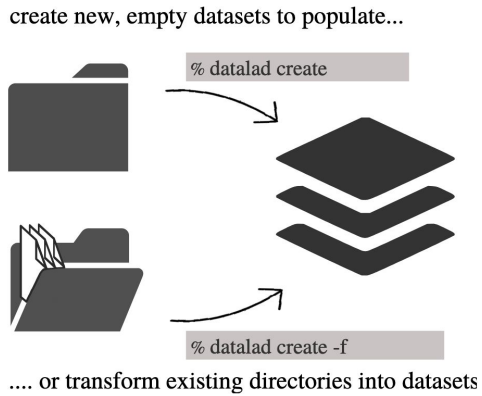


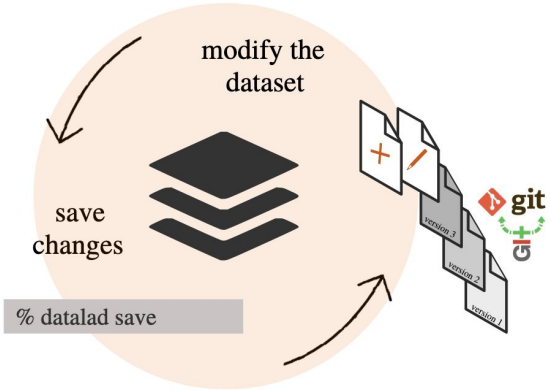
- *Under the hood* *(Git is a distributed version control system that tracks versions of files)*
  - Git-Annex
    - Sistema de sincronização distribuída de ficheiros
    - Permite a gestão de ficheiros e controlo de versões, fazer cópias sem adicionar o conteúdo dos ficheiros ao repositório
    - Integra com várias soluções de *storage* (hooks)

Ciclo de gestão de dados - **Planeamento**

- Uma estrutura para o projeto
  - Criar um ***dataset***
    - *Operações sobre o dataset são guardadas*
      - *Repositório git*



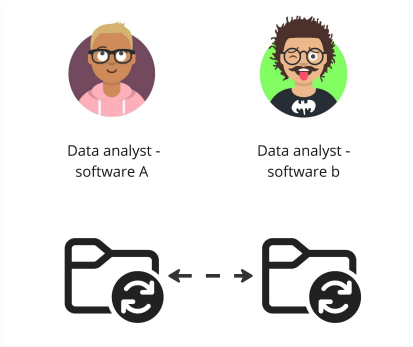


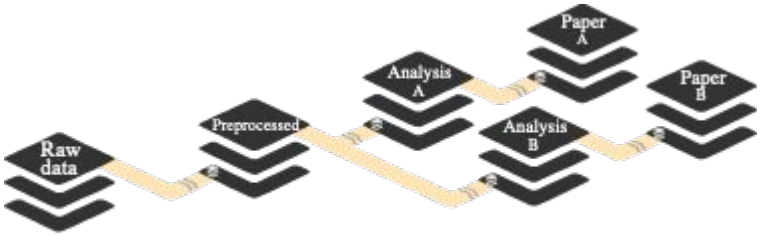


https://handbook.datalad.org/en/latest/_images/dataset.svg

Ciclo de gestão de dados - **Aquisição de dados**

- Uma estrutura modular dentro do dataset
  - (Sub)datasets dentro de datasets

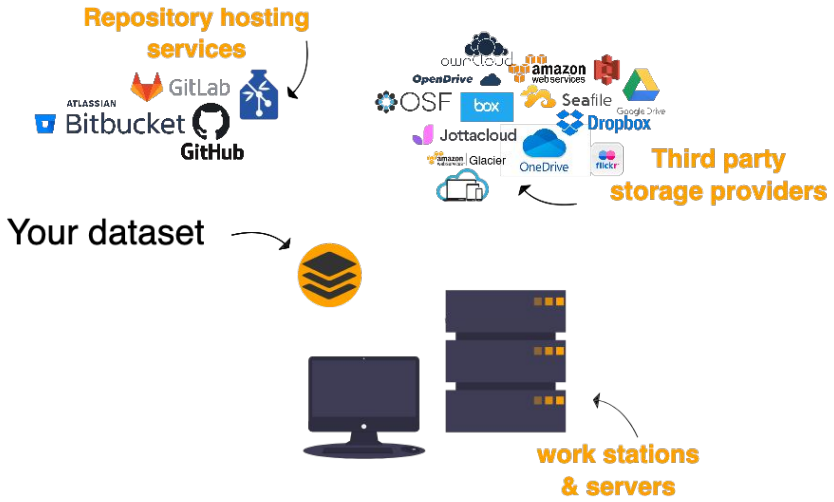




Nest modular datasets to create a linked hierarchy of datasets, and enable recursive operations throughout the hierarchy

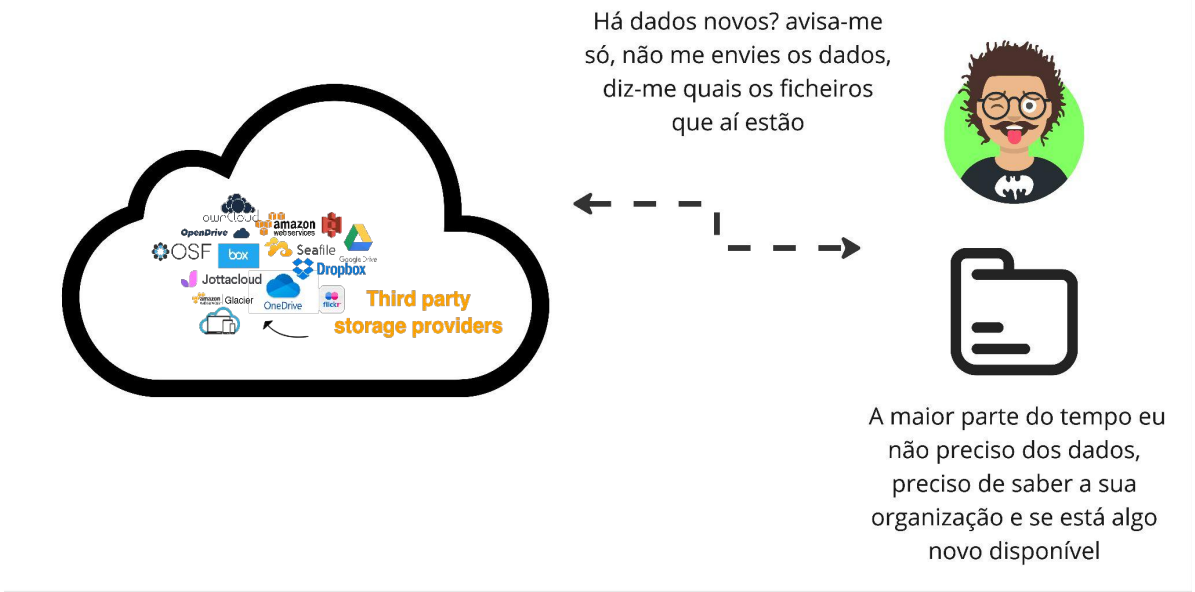
Ciclo de gestão de dados - **Partilha**

- Owncloud (dropit.dei.uc.pt)
  - infraestrutura onde o “**data** datasets” Datalad está alojado
- Github
  - Project dataset
    - código, etc.
    - o Datalad “**data** dataset” está disponível (vemos a estrutura de pastas e ficheiros disponíveis) e linkado (possível aceder aos dados get command) mas não armazenado (por defeito)



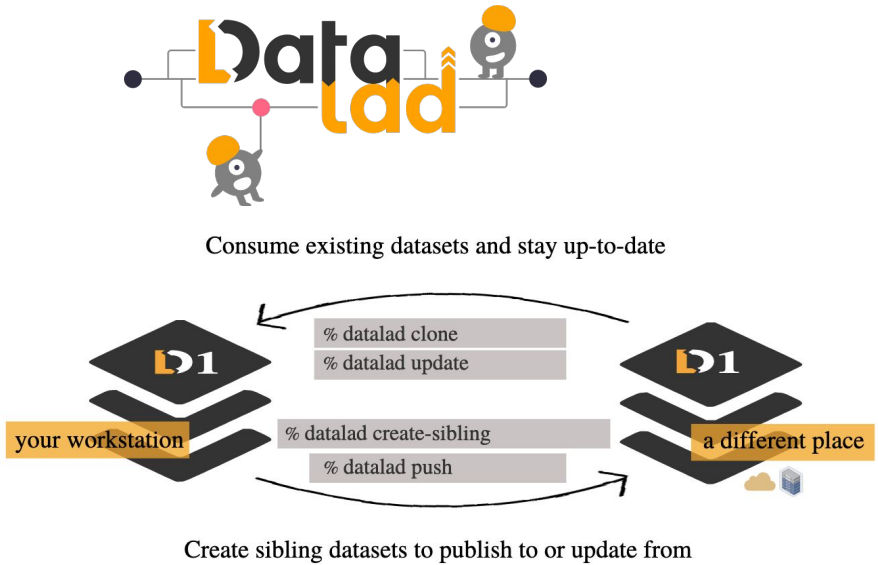
from the DataLad Handbook by Wagner et al. (2022) (CC BY-SA 4.0)

Ciclo de gestão de dados - **Partilha**

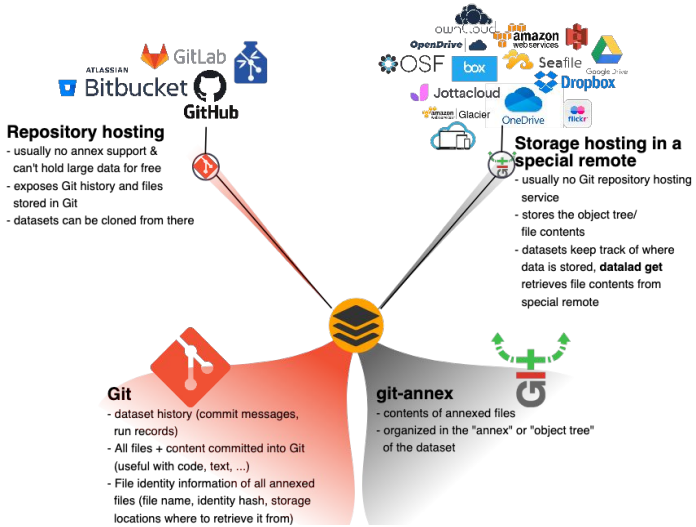


Ciclo de gestão de dados - **Partilha**

- Como funciona?
  - Clone
    - check updates
  - Criação de siblings
    - update siblings, etc



https://handbook.datalad.org/en/latest/_images/collaboration.svg

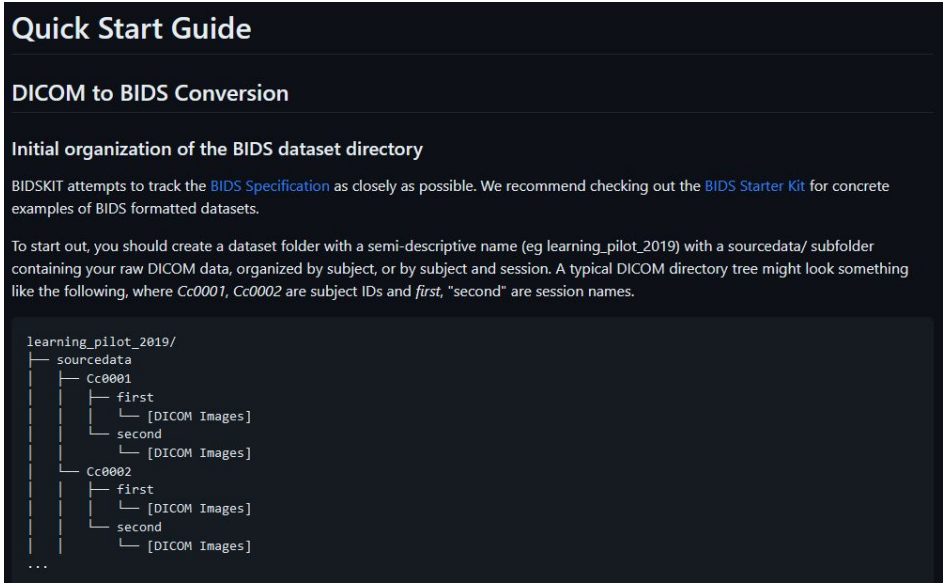


from the DataLad Handbook by Wagner et al. (2022) (CC BY-SA 4.0)

- Permite separar conteúdo Git vs. git-annex
- DataLad datasets são expostos através de repositórios públicos e/ou privados (p.e., GitLab or GitHub)
  - Dados não devem ser armazenados nessas plataformas
    - Third party storage
    - Deve ser clara a dependência entre plataformas

Ciclo de gestão de dados - **Processamento**

- Estrutura de suporte
  - **Formato/standard de dados**
    Brain Imaging Data Structure (*BIDS*)
  - Acesso a ferramentas de apoio



Quick Start Guide

DICOM to BIDS Conversion

Initial organization of the BIDS dataset directory

BIDSKIT attempts to track the BIDS Specification as closely as possible. We recommend checking out the BIDS Starter Kit for concrete examples of BIDS formatted datasets.

To start out, you should create a dataset folder with a semi-descriptive name (eg learning_pilot_2019) with a sourcedata/ subfolder containing your raw DICOM data, organized by subject, or by subject and session. A typical DICOM directory tree might look something like the following, where *Cc0001*, *Cc0002* are subject IDs and *first*, "second" are session names.

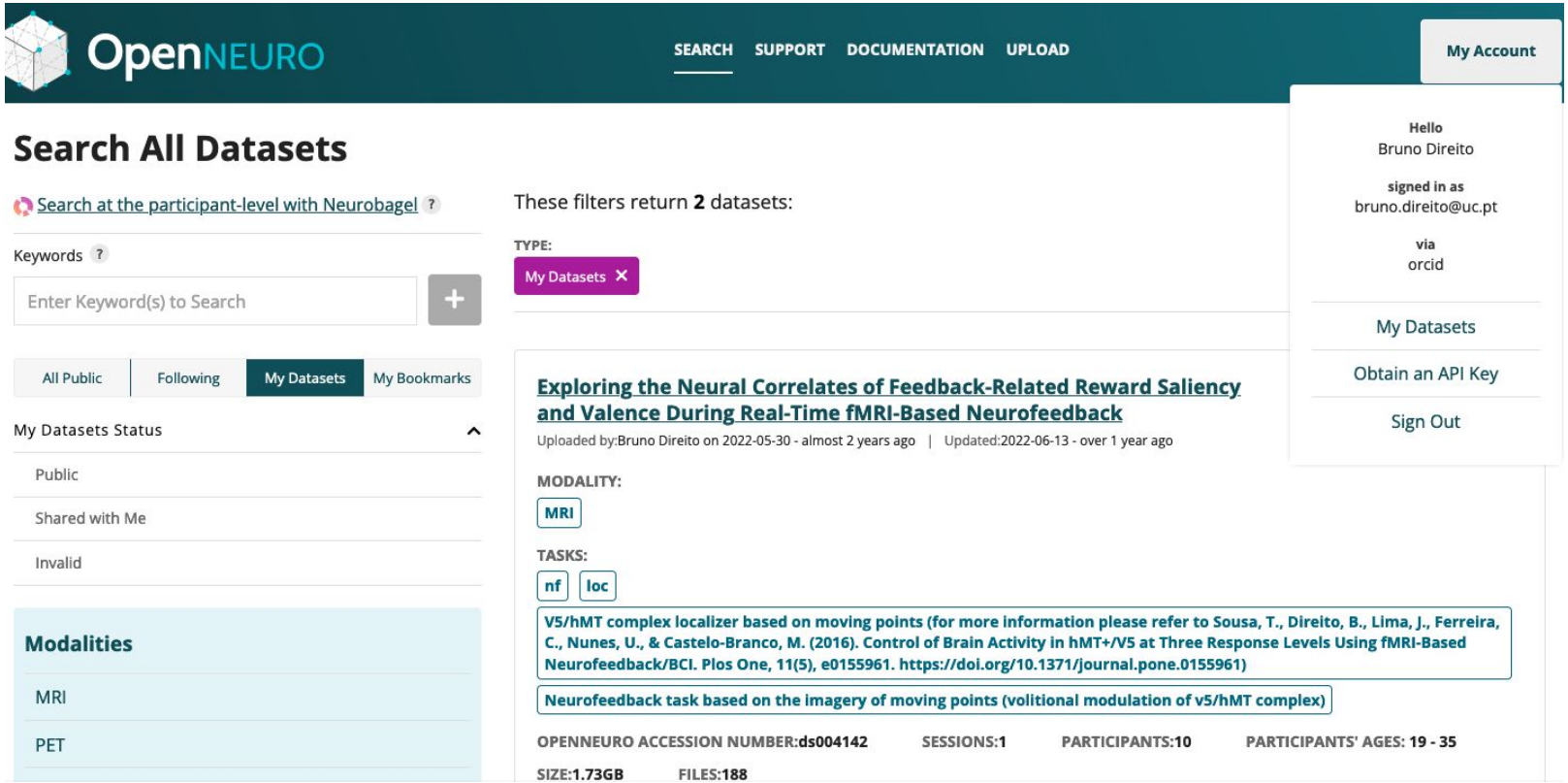
```
learning_pilot_2019/
├── sourcedata
│   ├── Cc0001
│   │   ├── first
│   │   │   └── [DICOM Images]
│   │   └── second
│   │       └── [DICOM Images]
│   └── Cc0002
│   │   ├── first
│   │   │   └── [DICOM Images]
│   │   └── second
│   │       └── [DICOM Images]
...
```

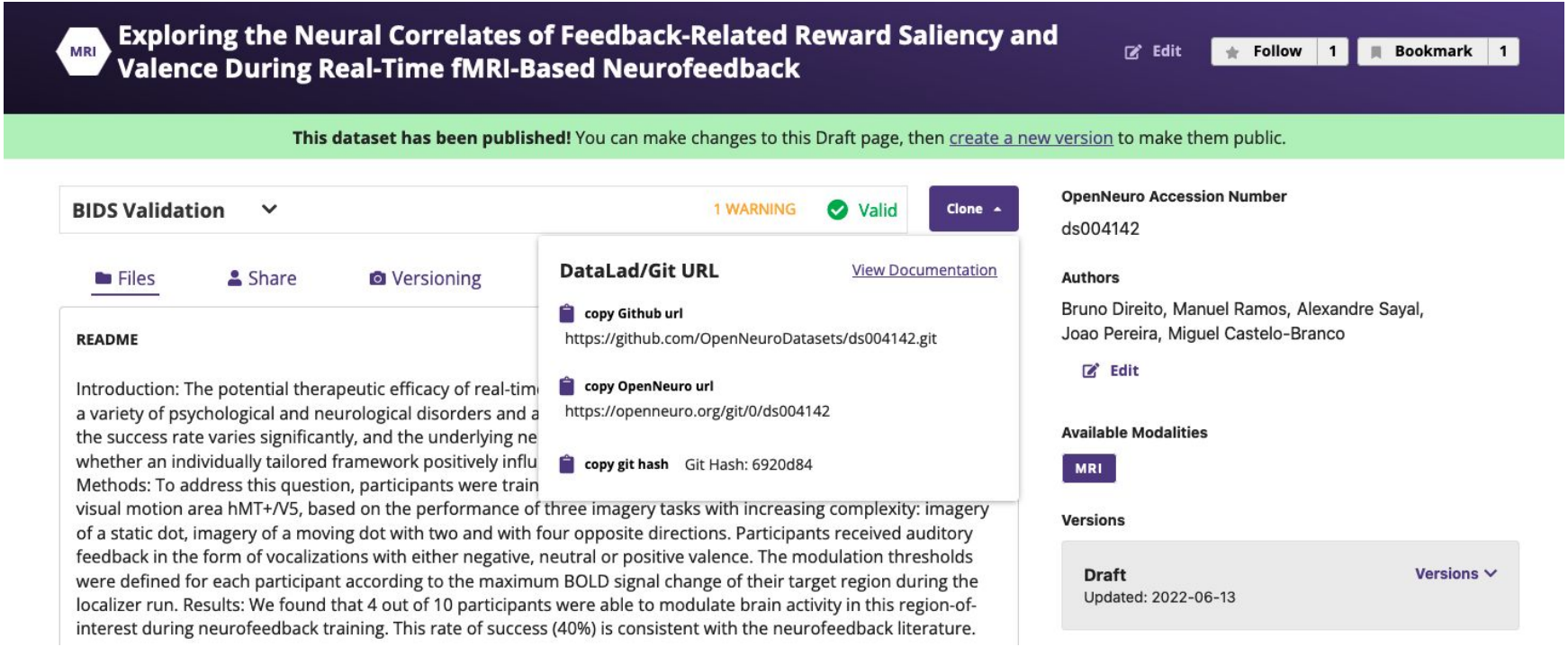
Ciclo de gestão de dados - **Análise**

- Análise de dados
  - Exploratório *mas* replicável
  - Os passos são muitas vezes repetitivos entre colaboradores
  - A análise segue muitas vezes processos / pipelines automatizadas

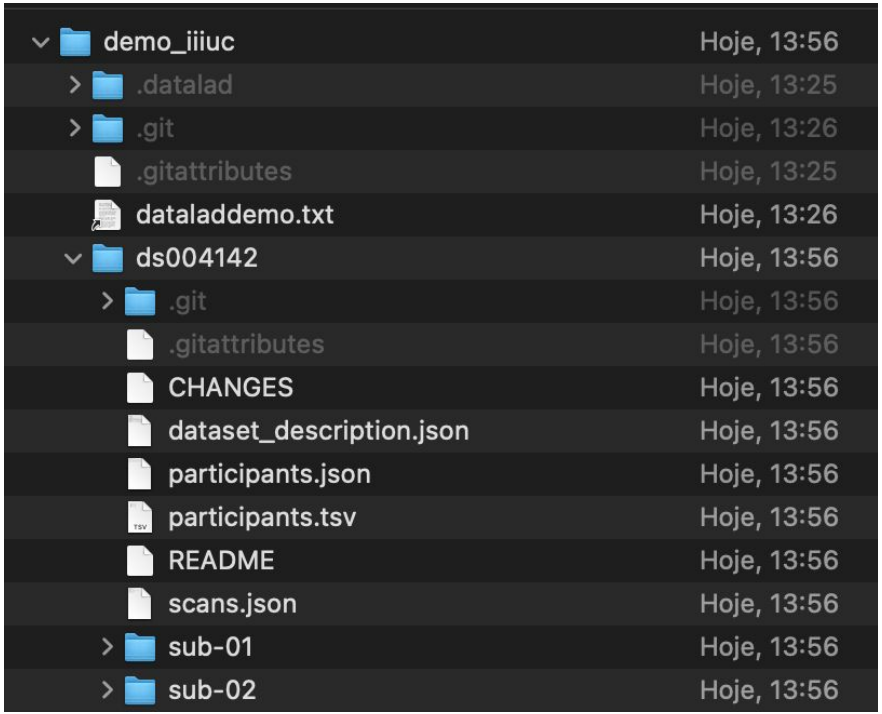
Ciclo de gestão de dados - **Análise**

- Scripts
  - MATLAB scripts
  - Python scripts
- Jupyter notebooks
  - The *Jupyter Notebook*, uma plataforma computacional / interface web-based interactiva (ou podemos montar o sistema localmente em ambientes como o Visual Code)
  - Pode ser usado
    - para criar pipelines e relatórios
    - É possível associar um dataset de dados (*nested*) de forma a que a proveniência dos dados esteja sempre assegurada
      - e referência a fonte original
    - Reutilização
      - neste ou em qualquer outro contexto
      - mudando o dataset associado


OpenNEURO
SEARCH
SUPPORT
DOCUMENTATION
UPLOAD
My Account
Hello
Bruno Direito
signed in as
bruno.direito@uc.pt
via
orcid
My Datasets
Obtain an API Key
Sign Out
Search All Datasets
Search at the participant-level with Neurobagel
Keywords
Enter Keyword(s) to Search
All Public
Following
My Datasets
My Bookmarks
My Datasets Status
Public
Shared with Me
Invalid
Modalities
MRI
PET
These filters return 2 datasets:
TYPE:
My Datasets
Exploring the Neural Correlates of Feedback-Related Reward Saliency and Valence During Real-Time fMRI-Based Neurofeedback
Uploaded by:Bruno Direito on 2022-05-30 - almost 2 years ago | Updated:2022-06-13 - over 1 year ago
MODALITY:
MRI
TASKS:
nf
loc
V5/hMT complex localizer based on moving points (for more information please refer to Sousa, T., Direito, B., Lima, J., Ferreira, C., Nunes, U., & Castelo-Branco, M. (2016). Control of Brain Activity in hMT+/V5 at Three Response Levels Using fMRI-Based Neurofeedback/BCI. Plos One, 11(5), e0155961. https://doi.org/10.1371/journal.pone.0155961)
Neurofeedback task based on the imagery of moving points (volitional modulation of v5/hMT complex)
OPENNEURO ACCESSION NUMBER:ds004142
SESSIONS:1
PARTICIPANTS:10
PARTICIPANTS' AGES: 19 - 35
SIZE:1.73GB
FILES:188


MRI
Exploring the Neural Correlates of Feedback-Related Reward Saliency and Valence During Real-Time fMRI-Based Neurofeedback
Edit
Follow 1
Bookmark 1
This dataset has been published! You can make changes to this Draft page, then create a new version to make them public.
BIDS Validation
1 WARNING
Valid
Clone
Files
Share
Versioning
README
Introduction: The potential therapeutic efficacy of real-tim
a variety of psychological and neurological disorders and a
the success rate varies significantly, and the underlying ne
whether an individually tailored framework positively influ
Methods: To address this question, participants were train
visual motion area hMT+/V5, based on the performance of three imagery tasks with increasing complexity: imagery of a static dot, imagery of a moving dot with two and with four opposite directions. Participants received auditory feedback in the form of vocalizations with either negative, neutral or positive valence. The modulation thresholds were defined for each participant according to the maximum BOLD signal change of their target region during the localizer run. Results: We found that 4 out of 10 participants were able to modulate brain activity in this region-of-interest during neurofeedback training. This rate of success (40%) is consistent with the neurofeedback literature.
DataLad/Git URL
View Documentation
copy Github url
https://github.com/OpenNeuroDatasets/ds004142.git
copy OpenNeuro url
https://openneuro.org/git/0/ds004142
copy git hash
Git Hash: 6920d84
OpenNeuro Accession Number
ds004142
Authors
Bruno Direito, Manuel Ramos, Alexandre Sayal, Joao Pereira, Miguel Castelo-Branco
Edit
Available Modalities
MRI
Versions
Draft
Updated: 2022-06-13
Versions

$ datalad clone
https://github.com/OpenNeuroDatasets/ds004142



Q&A!

or later at bruno.direito@uc.pt