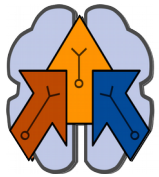


Brain Communication Pathways
Sinergia Consortium
Swiss National Science Foundation

A (brief) Introduction to Open and Reproducible Science in Computational Brain Imaging

Sébastien Tourbier[1]

Connectomics Lab, Radiology Research Center, CHUV

ReproNim Training Fellow 2019-2020

June 28th 2019, TRABIT Summer School, EPFL, Lausanne

Purpose

- To introduce you **why we should care about reproducibility in research**
- To give you an **overview of a set of solutions in computational brain imaging**

  that **have proven** to be **capable of effective large-scale collaboration**

  to make your research more open, reproducible, and reusable and your **collaborations more efficient**

What is reproducibility at the publication level?

- **Re-executability** (publication-level replication):

  Using the **same data** and **executing** the **same code** in the **same computing environment,** can you **recreate** the **same results** ?

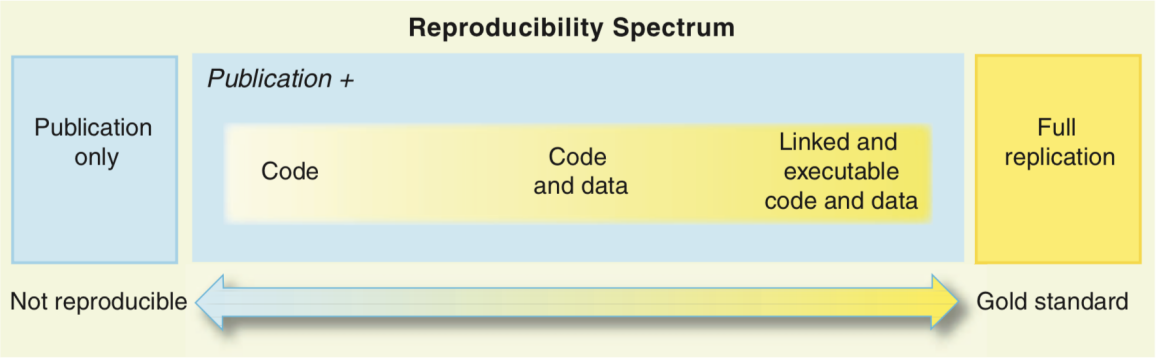


Illustration from Peng R.D., *Reproducible Research in Computational Science*, Science 334 (6060), pp. 1226-1227 doi: 10.1126/science.1213847

The reproducibility problem

- A number of studies have brought the reproducibility into question

Reproducible Research in Computational Science

Roger D. Peng

DOI: 10.1126/science.1213847

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

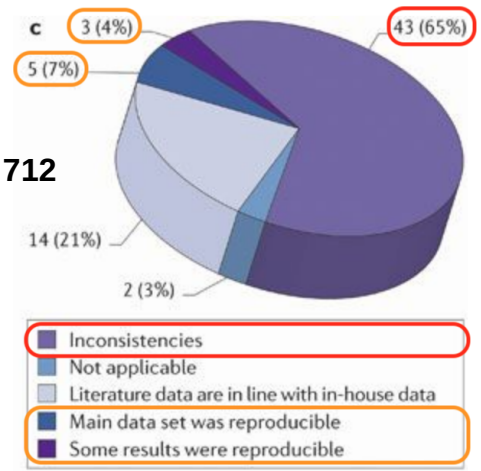


Chart from Prinz et al.
Nature Reviews Drug Discovery 10, 712
(September 2011)

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

https://www.economist.com/leaders/2013/10/21/how-science-goes-wrong

Issues that can affect reproducibility

- **Ineffective data sharing** (heterogeneous data organization, data not shared,...)
- **Low power** (small sample size limited to be extended due to heterogeneous open data organization)
- **Methodological variance** (Incomplete description of method/workflow, code not shared,...)
- **Computing environment variance** (Operating system, Tool version,...)
- **Mistakes**

→ **Addressed** with **Open Science**

What is Open Science?

“Open Science is the **practice of science** in such a way that **others can collaborate and contribute**, where **research data, lab notes and other research processes are freely available**, under terms that **enable reuse, redistribution and reproduction** of the research and its underlying data and methods.”

From https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition

Push to Open Science – Open Data

"**Research data** are the **evidence that underpins the answer to the research question**, and can be **used to validate findings** regardless of its form (e.g. print, digital, or physical)." UK Concordat on Open Research Data, 28 July 2016

"The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the **direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community**. Such results include peer-reviewed publications and digital data." US office of Science and Technology Policy, *Increasing access to the results of federally funded scientific research*, Memo, 2013

"During the course of the research work and after its completion, **grantees are obliged to make available to the public in an appropriate manner the research results obtained with the help of SNSF funding,** thereby explicitly mentioning the support obtained from the SNSF." SNSF policy on Open Research Data, Art 47 of *Funding Regulations, 2015*

"Since October 2017, researchers **have to include a data management plan (DMP) in their funding application** for most of the funding schemes. At the same time, the **SNSF expects that data generated by funded projects are publicly accessible in digital databases** provided there are no legal, ethical, copyright or other issues."

http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx#Guidelines%20and%20Regulations

"My goal is to make EPFL an institution that bets on open science and shared knowledge."

Martin Vetterli, President of EPFL, 2017

Keys for open and reproducible computational brain imaging

- **Data standardization**

  Version-controlled **structured** dataset → **same data**

- **Data processing standardization**

  Version-controlled data analysis workflow with **data provenance,** encapsulated in **container**

  → **Re-execute** the **same code** on the **same data** using the **same computing environment**

Road map

- How to **control** the **version** of **code**
- How to **control** the **version of data**
- How to **manage neuroimaging data**
- How to **release code with its computing environment**
- How to **better control** the **execution** and **re-usability** of **data analysis workflows**

How to control the version of code

Git

- **Git** is the most commonly **adopted version control system for code repository**

  based on Eclipse survey 2014

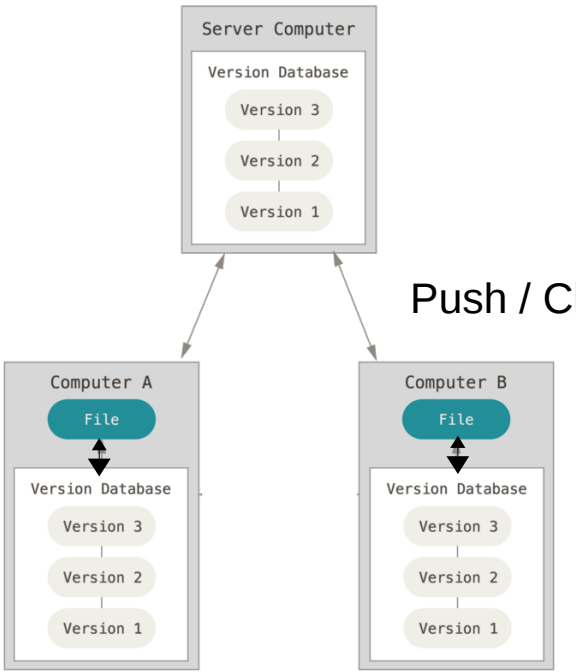
- **Distributed way** to record changes/versions of files (such as codes, documents, datasets)

  → **Make collaboration much more efficient**

- Each project team member has a **local working directory** and a **local repository** where they can commit changes

- **Changes** are accessible to other members only if they are **pushed to the central/remote repository**

- For **students and teachers** (see https://education.github.com/)**:**
  - **GitHub Pro plan for free**



Push / Clone or Pull

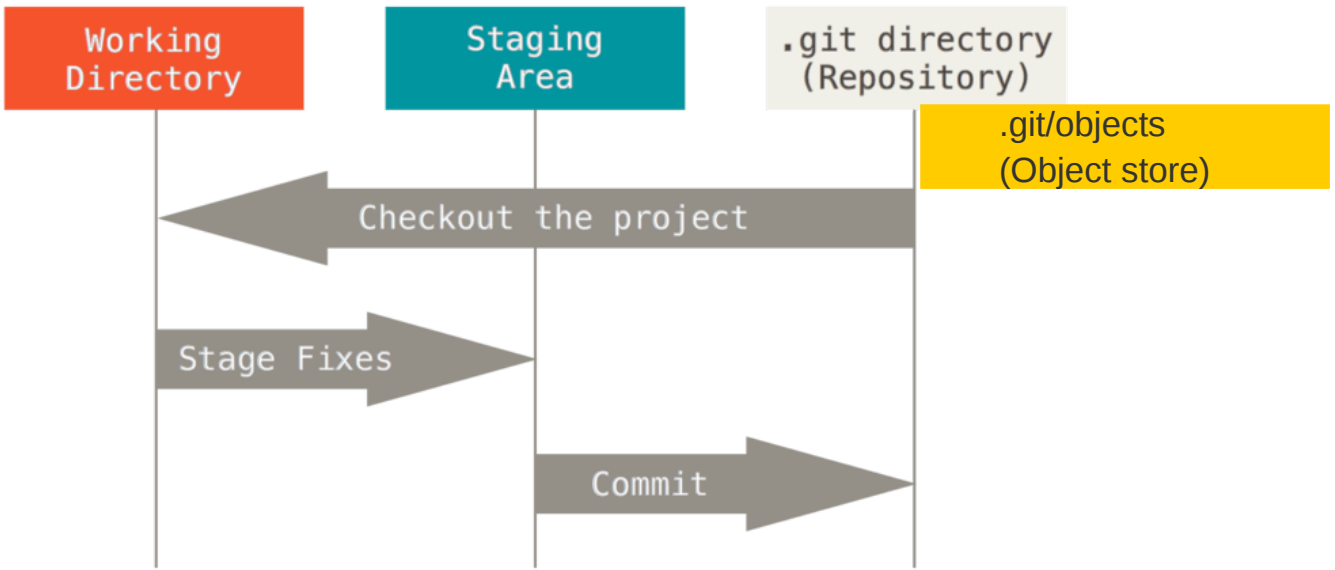
Commit / Checkout

Source:

https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

Git Basics

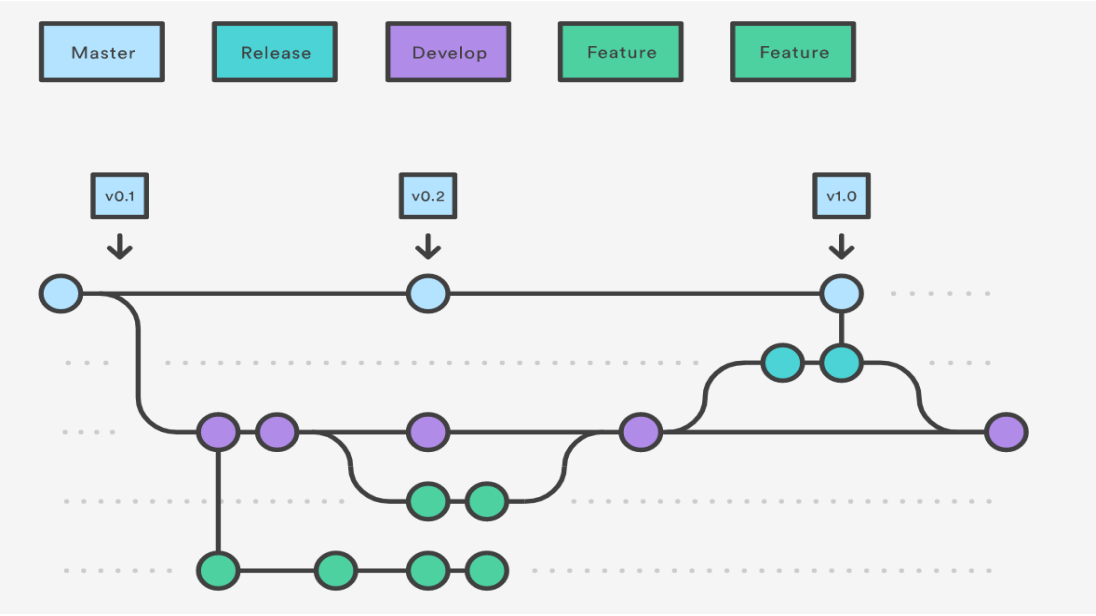
- A **git project** is composed of 1) the git repository with object store (*.git/objects*), 2) the working directory, and 3) the staging area



Source: https://git-scm.com/book/en/v2/Getting-Started-Basics

Git basics

- Branches and version tags



Source:
https://www.atlassian.com/git/tutorials/comparing-workflows/gitflow-workflow

Git is great for code

- Ability to:
  - Tracks content (files, file trees, commits)
  - Stores all known content in object store (*.git/objects*)
  - Efficiently exchange objects and references (branches, version tags) with remote repository

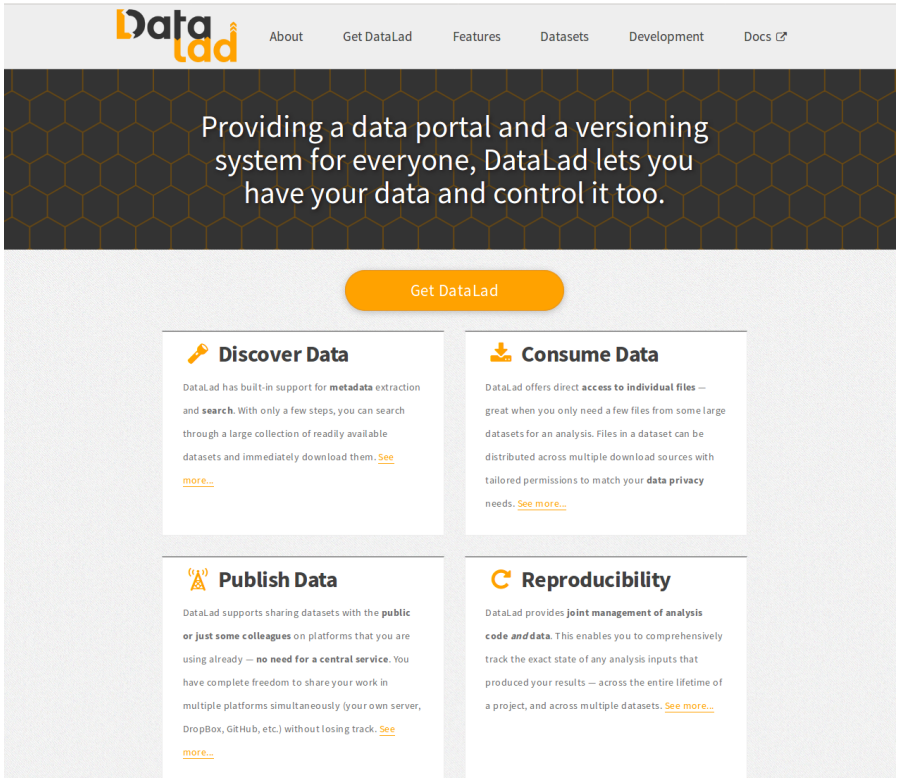
Git is great for code but...

- **Inadequate for data**:
  - Designed for projects that contains files having **relatively small sizes**
  - All known **content** is **distributed across all copies**
  - **Duplicate content** between working directory and the object store of the .git repository (*.git/objects*)

How to control the version of data

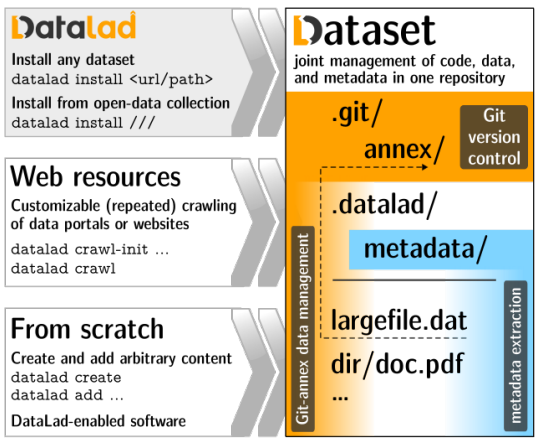
Datalad

https://www.datalad.org/



Datalad principles

- There are only two things in the world: **datasets and files**
- A **dataset is a Git repository**

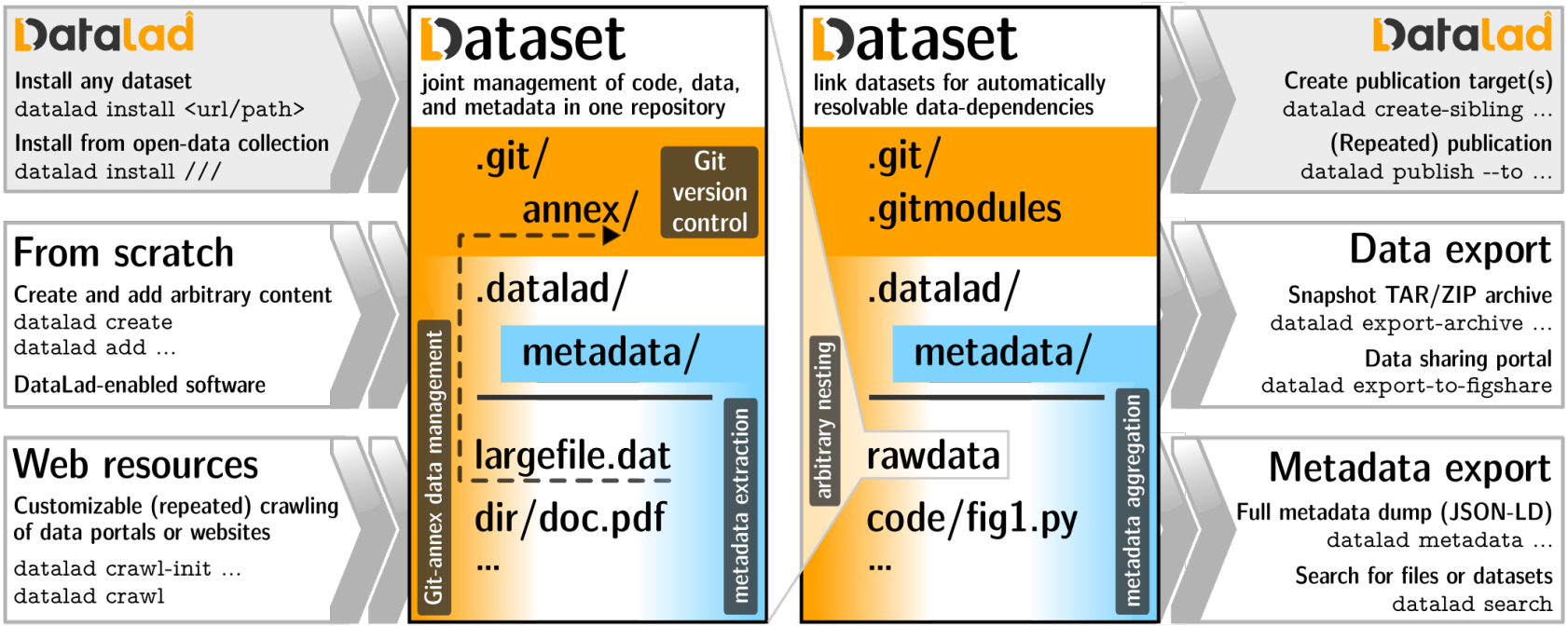


Source: https://github.com/datalad/talk-2019-mila

- A dataset can have an **optional annex** for (large) file content tracking (transport to and from the annex managed with Git-annex, https://git-annex.branchable.com)

- Minimization of custom procedures and data structures: **Users must not loose data or data access**, if DataLad would vanish
- **Complete decentralization**, no required central server or service
- **Maximization of existing 3rd-party data resources and infrastructure re-use**

Datalad is more than a version control system



Source: https://github.com/datalad/talk-2019-mila

How to manage neuroimaging datasets

Getting lost in data

Heterogeneity in data description practices causes:

- Unnecessary manual metadata input
- Problems in sharing, reusing and combining datasets
- No way to automatically validate/curate datasets

→ Need for a **standard** that is **commonly adopted**

Version-Controlled Structured Datasets

- **One dataset formal** → **share-ability** and **re-usability**
  - Studies sharing data have higher statistical quality
  - Sharing data related to higher citation rate

- **Data versioning** → guaranteed to use the **same data**

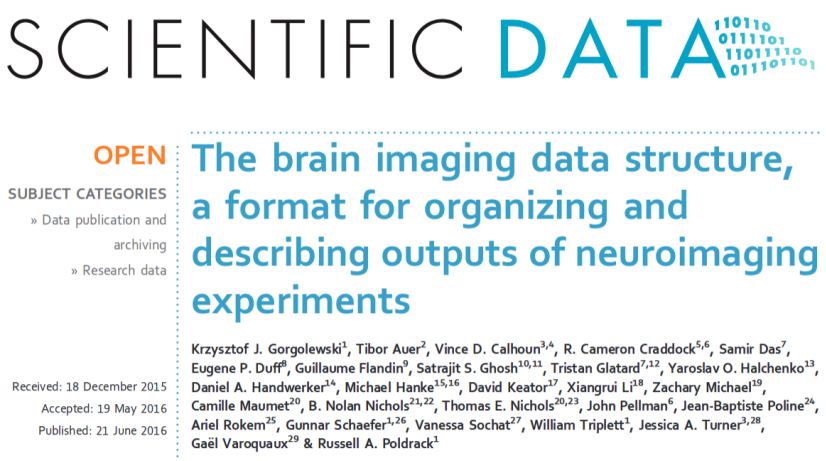
Rising-star solution in neuroscience:

Brain Imaging Data Structure (BIDS) Datasets + Datalad

→ No more “Getting lost in your data”

BIDS: Brain Imaging Data Structure

www.nature.com/scientificdata



SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Data publication and archiving
» Research data

Received: 18 December 2015
Accepted: 19 May 2016
Published: 21 June 2016

The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments

Krzysztof J. Gorgolewski[1], Tibor Auer[2], Vince D. Calhoun[3,4], R. Cameron Craddock[5,6], Samir Das[7], Eugene P. Duff[8], Guillaume Flandin[9], Satrajit S. Ghosh[10,11], Tristan Glatard[7,12], Yaroslav O. Halchenko[13], Daniel A. Handwerker[14], Michael Hanke[15,16], David Keator[17], Xiangrui Li[18], Zachary Michael[19], Camille Maumet[20], B. Nolan Nichols[21,22], Thomas E. Nichols[20,23], John Pellman[6], Jean-Baptiste Poline[24], Ariel Rokem[25], Gunnar Schaefer[1,26], Vanessa Sochat[27], William Triplett[1], Jessica A. Turner[3,28], Gaël Varoquaux[29] & Russell A. Poldrack[1]

http://bids.neuroimaging.io/

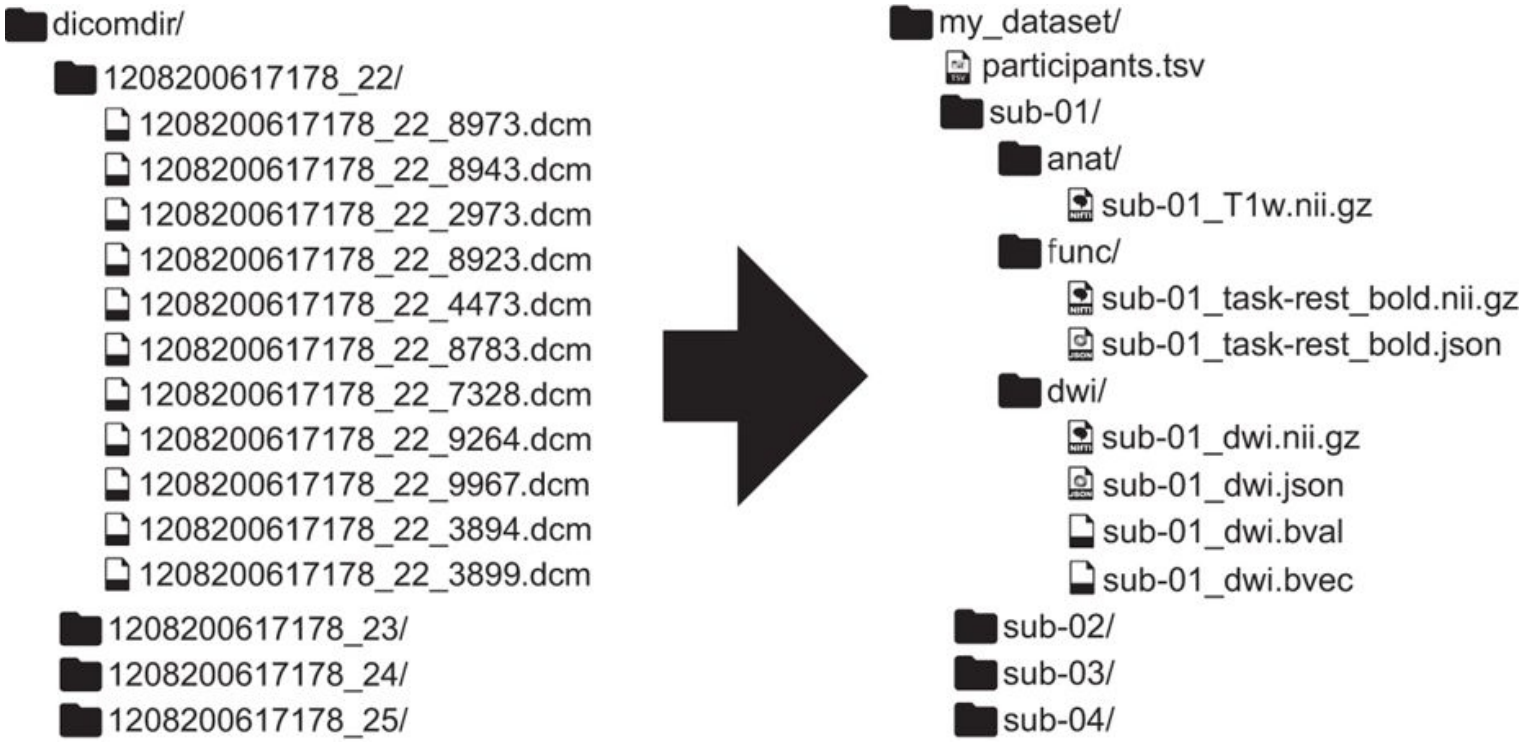
- **What is it?**
  - **Standard** for **organizing and describing brain MRI datasets**
  - Many **on-going extensions**, including **BIDS-EEG** and **BIDS-iEEG**

    http://bids.neuroimaging.io/#get_involved
- **What is it for?**
  - **Make handling easy over one dataset** from one person to another one
  - **Make easier** to write pipelines and to inter-operate between softwares
  - **Make curation easier** by accepting **one dataset formal**

    Online automatic validator of BIDS structured datasets:

    https://bids-standard.github.io/bids-validator/
- **Community outreach**
  - Reached **over 5000 researchers**
  - **Adopted by open databases** such as *OpenfMRI, SchizConnect, Developing Human Connectome Project, FCP-INDI,* and *openNeuro*, as well as **dozens of labs** around the world
  - About **40 example dataset, rapidly growing number**

A BIDS structured dataset



From: The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments

A growing number of softwares supporting BIDS

- BIDS Apps (a growing set of portable containerized data processing pipelines that understand BIDS datasets)
- Converters
  - AFNI BIDS-tools
  - BIDS2ISATab
  - BIDSto3col
  - BIDS2NDA
  - bidskit
  - Dcm2Bids
  - DCM2NIIx
  - DICM2NII
  - HeuDiConv
  - OpenfMRI2BIDS
  - ReproIn (HeuDiConv-based turnkey solution)
  - bids2xar (for XNAT import)
  - XNAT2BIDS
  - Horos (Osirix) export plugin
  - BIDS2NIDM
  - BIDScoin
  - MNE-BIDS (MEG/EEG/iEEG)
- Institution specific data management/conversion tools
  - dac2bids
    Conversion tool for the Donders Institute
  - Autobids from the Centre for Functional and Metabolic Mapping (CFMM) at Western's Robarts Research Institute
- Other Tools
  - Automatic Analysis (fMRI processing toolbox)
  - Brainstorm (MEG/EEG analysis package)
  - C-PAC (Configurable Pipeline for the Analysing Connectomes)
  - FMRIPREP (preprocessing workflow)
  - OpenNeuro (repository)
  - PyBIDS (Python module to harmonize access and manipulation)
- Quality Assessment
  - MRIQC
  - QAP

http://bids.neuroimaging.io/#benefits

The most useful BIDS Softwares

- **Heudiconv:** a flexible DICOM converter for organizing brain imaging data into datalad repositories with BIDS structured working directory layouts

  http://nipy.org/heudiconv

- **BIDS Validator**: validator for the Brain Imaging Data Structure which helps you to make sure that your BIDS dataset is compliant to the BIDS specifications

  http://bids-standard.github.io/bids-validator/

- **Pybids**: python tools for querying and manipulating BIDS datasets.

  https://bids-standard.github.io/pybids/

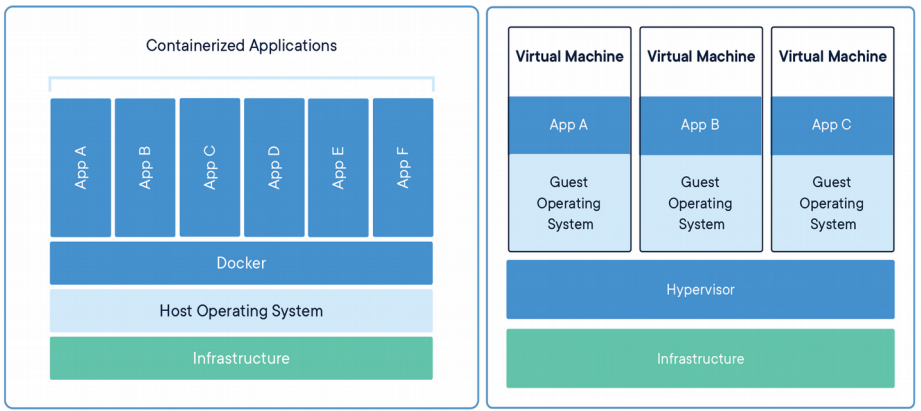
Repositories for BIDS datasets

- Curated:
  - OpenNeuro.org
  - FCP-INDI
  - NITRC
  -
- Uncurated:
  - Data Dryad
  - Figshare
  - Harvard Dataverse
  - Zenodo

How to release code
with
its required computing environment

Software container technology

- **Docker and Singularity** Software Container: much more **lightweight** and **use far fewer resources** than virtual machines



From https://www.docker.com/resources/what-container

- Container images can be run on Linux, Mac, and Windows platforms
  → **High-level of portability**

- Users do not need to install any extra software dependencies

  → **easy-to-install**

- Container image version. Easy to switch between versions

  → **High-level of reproducibility**

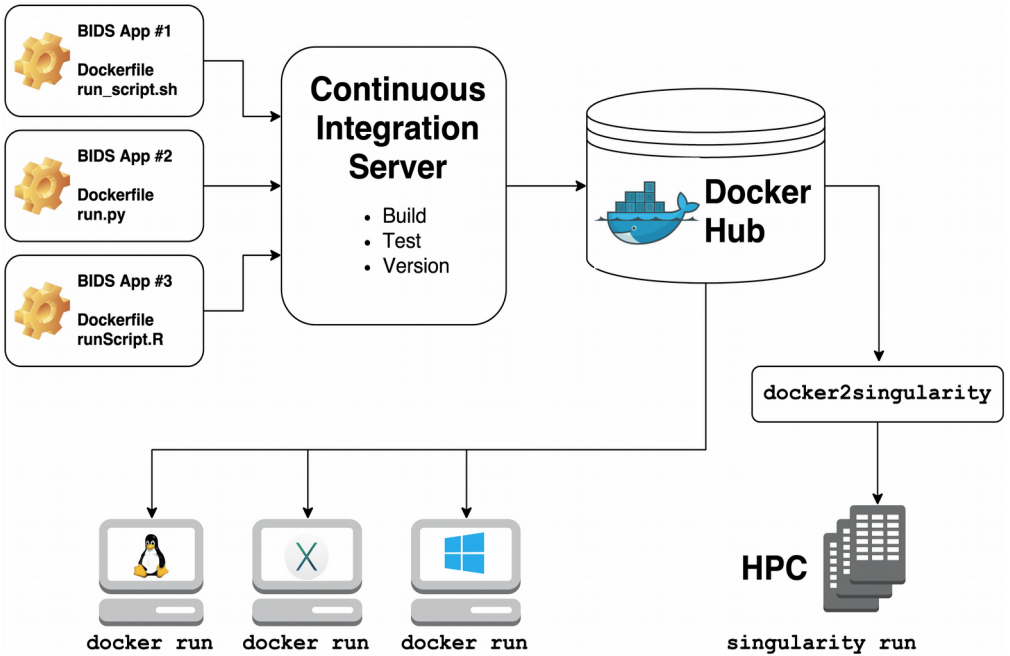
- Fixed version of third party softwares

  → **No need to worry about incompatible updates of dependencies**

The BIDS App framework

http://bids-apps.neuroimaging.io/

- **Portable neuroimaging pipeline** that takes **BIDS datasets** as inputs



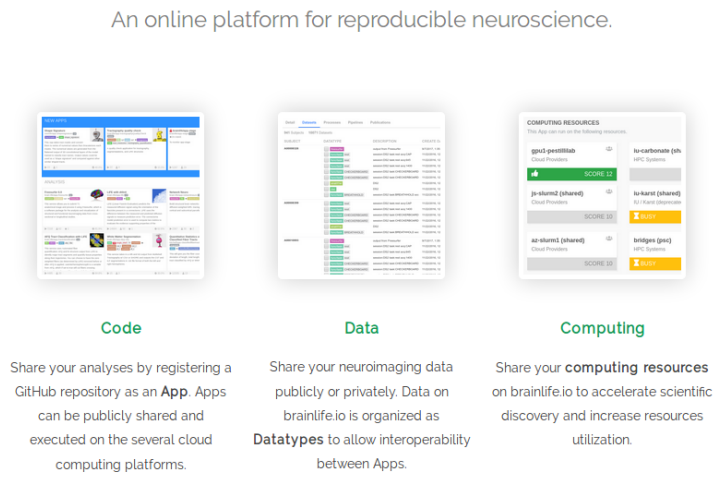
Source: Gorgolewski K. J. et al., PLOS Computational Biology 13(3): e1005209. https://doi.org/10.1371/journal.pcbi.1005209

- Based on two software container technologies:
  - **Docker** - for building, hosting as well as running containers on **local hardware** (running Windows, Mac OS X or Linux)
  - **Singularity** - for running containers on **HPCs**.
- Each BIDS App has the **same core set of command line arguments**

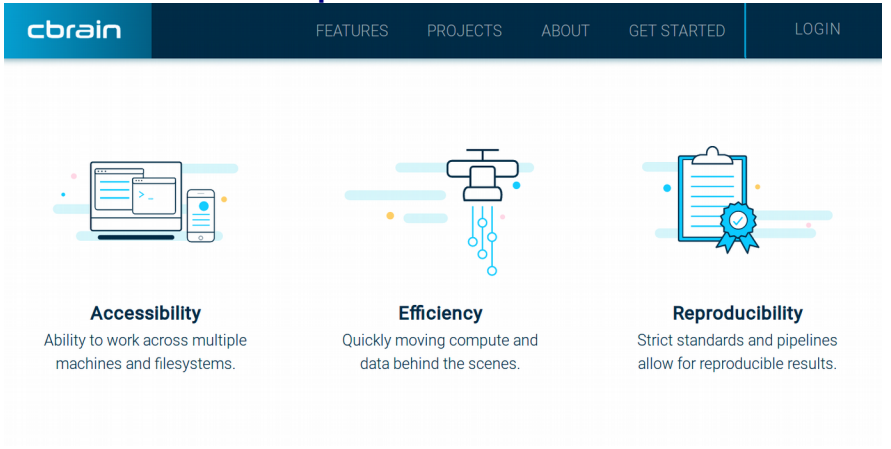
  → **Easy** to run and integrate into **automated platforms**

Open platforms supporting BIDS Apps

https://brainlife.io/



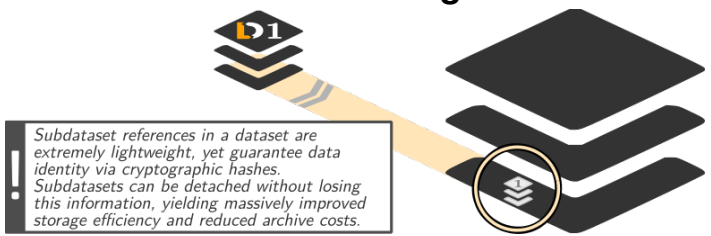
http://www.cbrain.ca/



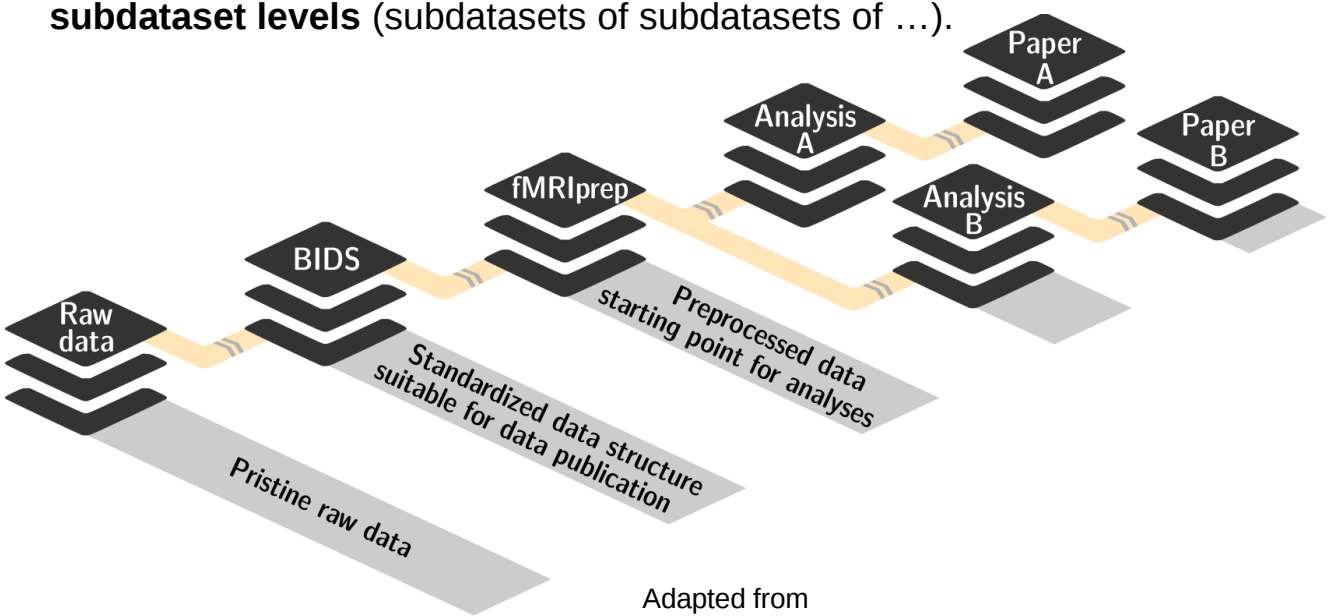
How to better control execution and re-usability of processing workflows

Keep study elements (datasets) modular with Datalad's dataset-subdataset linkage

Datalad's dataset-subdataset linkage



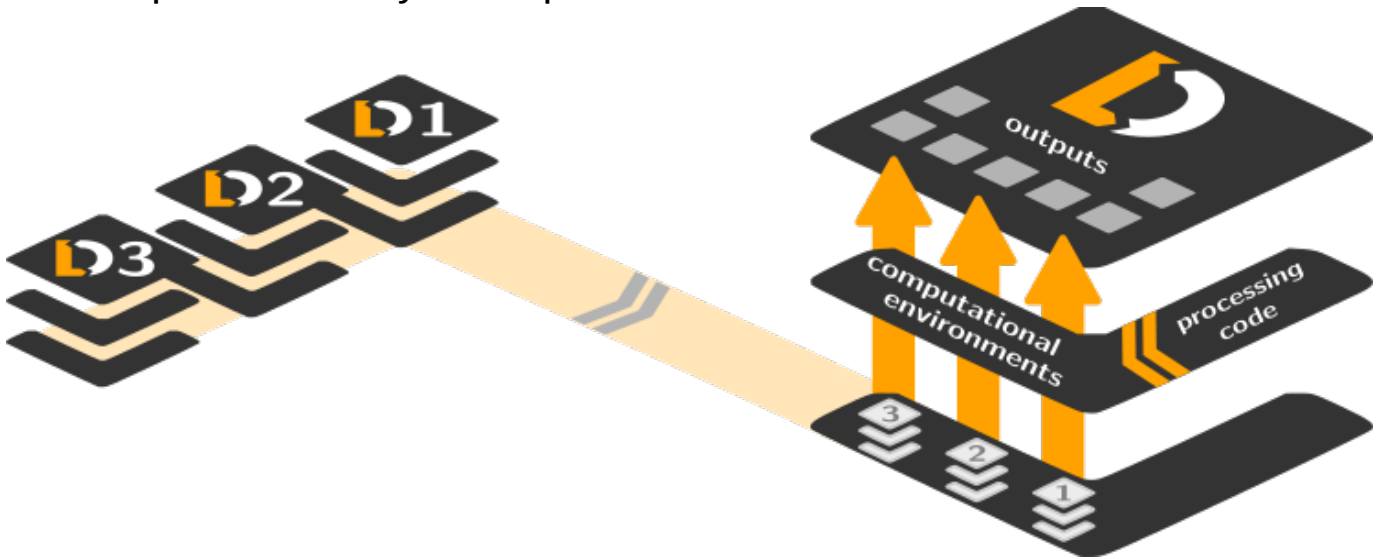
While a single dataset only tracks its immediate inputs, DataLad is capable of resolving any input dependencies recursively across all subdataset levels (subdatasets of subdatasets of …).



Adapted from
https://github.com/datalad/talk-2019-mila

Track Provenance from Data to Results using Datalad's run

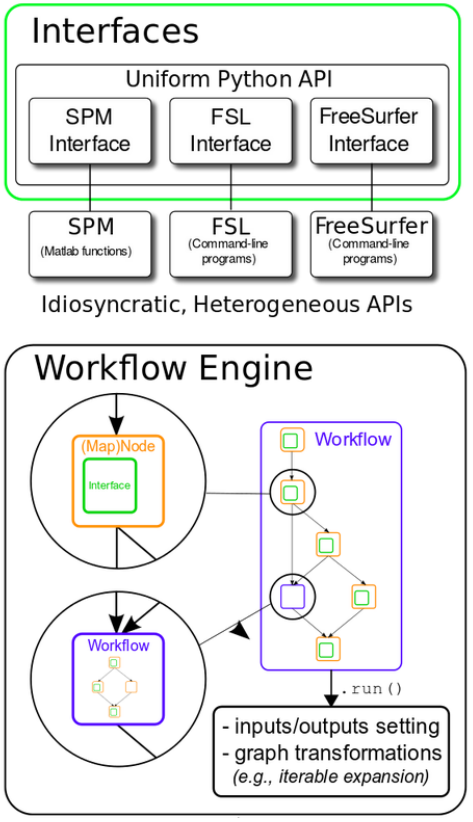
Track all input data, code, and computational environments (e.g. container image) needed to produce analysis outputs



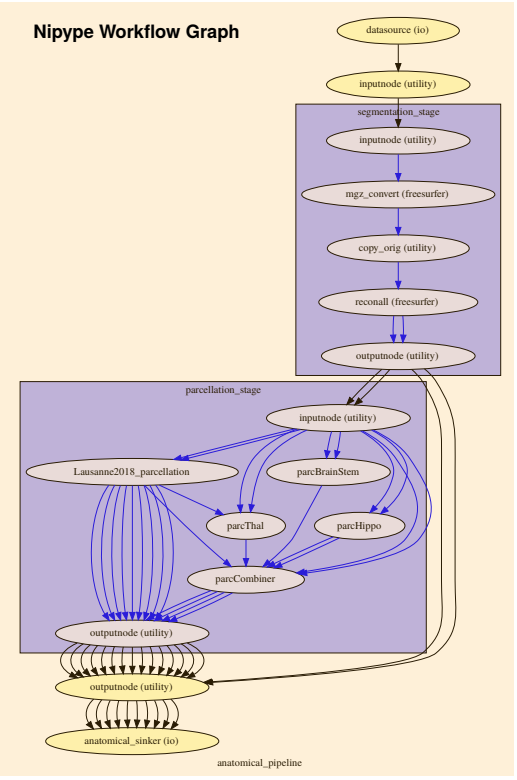
From http://www.repronim.org/coco2019-training/04-02-reproin/

Use a dataflow tool such as Nipype

- **Niype: python-based framework** for **writing workflows**
- **Provide** access to **many brain imaging algorithms** and **integration of your own algorithms** using **a consistent API**
  → **Inter-operability**
- Provide **isolation of data** being analyzed
  → W**orkflows** written using such abstractions **can be reused** on **different datasets**
- **Workflows represented as graphs**
  → **Facilitate execution** and **re-execution**
  → **Track data provenance between the different algorithm interfaces**



Nipype Architecture
Adapted from
https://nipype.readthedocs.io/en/latest/index.html



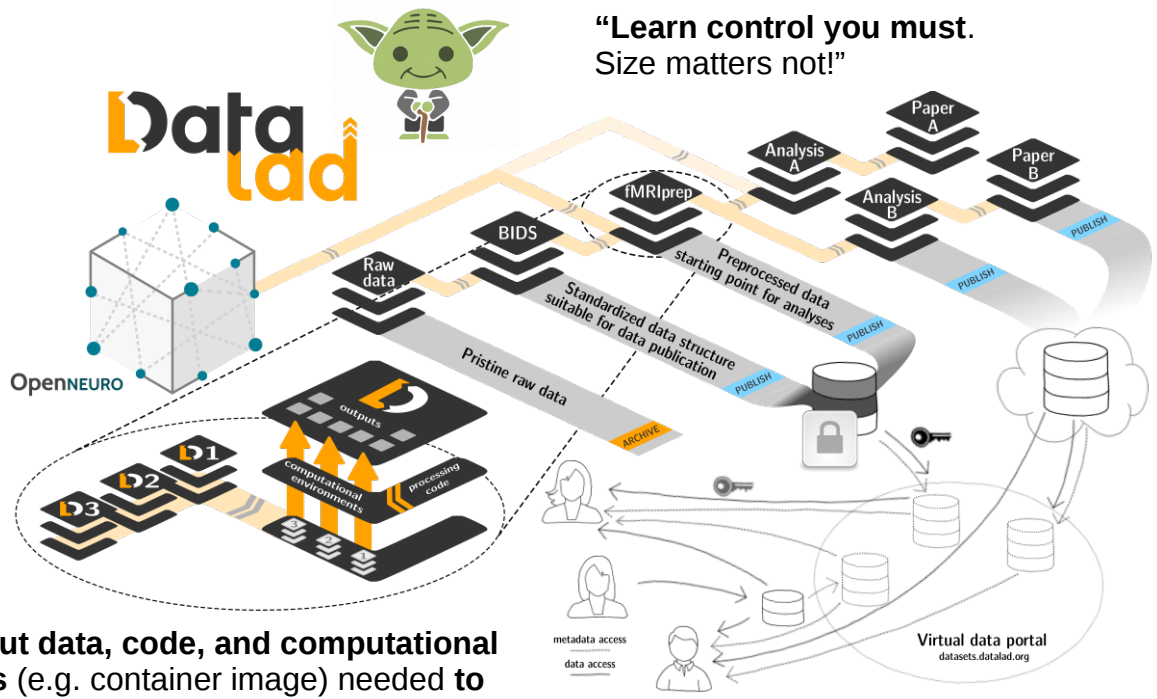
Example of a Nipype execution graph with our own algorithms
(From Tourbier et al., Poster OHBM 2019)

In summary

The YODA principles

Adapted from: Hanke M. et al., YODA: YODA's Organigram on Data Analysis, OHBM18 (https://f1000research.com/posters/7-1965)

“Learn control you must.
Size matters not!”



“Track all input data, code, and computational environments (e.g. container image) needed **to produce analysis outputs** in version controlled datasets — and **reproducibility you will achieve**!”

“Treat study elements as modular, reusable components that are tracked in individual, connected datasets — or the dark side will ravage your productivity!”

“Worry you need not about storage demand duplication and access permissions.

Data/metadata access separation allows you to safely collaborate on individual aspects, far beyond Tatooine's borders.”

Take-home message

- **Achieving complete reproducibility is hard**

- **Requires learning a set of new skills**, which might be **large depending on your IT background** but **a lot of training resources are available**

- This **takes time on the short term…** But **on the long term**:
  - **No more “lost in data”**
  - **No more “what did I do?” or “what did he do?”**

- **Be easy on yourself: learn and adopt** each of the new skills **step by step** and **do not hesitate to ask the open community** if needed (No one has all the answers)

Links to learn more

Git

- **https://education.github.community/**
- **https://www.atlassian.com/git/tutorials**
- **https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control**
- **http://swcarpentry.github.io/git-novice/**

Datalad

- **https://rawgit.com/psychoinformatics-de/talk-datalad-gofair/master/index.html**
- **http://www.repronim.org/coco2019-training/**

Nipype

- **https://nipype.readthedocs.io/en/latest/**
- **https://miykael.github.io/nipype_tutorial/**

Continuous Integration Testing System (not covered)

- **Travis CI: https://travis-ci.org/**
- **Circle CI: https://circleci.com/**
- **ReproNim training module: http://www.repronim.org/module-dataprocessing/06-testing/**

Open Brain Consent (not covered)

- **https://open-brain-consent.readthedocs.io**

Licenses (not covered)

- **Code: https://choosealicense.com/**
- **Data: https://creativecommons.org/licenses/**

More training resources

- **https://software-carpentry.org/lessons/**
- **http://www.repronim.org/5steps**
- **http://www.repronim.org/teach.html**
- **https://github.com/neurohackademy/2018_materials**
- **https://www.ewi-psy.fu-berlin.de/en/v/ccnb/content/Open-Science-Workshop/03_Chris_Gorgolewski.pdf**

brainhack.org

Attending BrainHack/Hackathon events is also a **great opportunity** for **learning through** the **open neuroscience community**

Brain Communication Pathways
Sinergia Consortium
Swiss National Science Foundation

Thank you for your attention!

“Think locally; Act globally”
“Think reproducibly; Act re-executably”
ReproNim Team