An automatically and computationally reproducible neuroimaging analysis from scratch

A Tutorial from the DataLad Handbook

Mariana Carmona Baez, Khady Diagne, Heike Schuler

What is DataLad?

A software tool designed to help with anything related to the evolution of digital objects (code/data).

➔ Any field, any experience level, any project!

Allows to share, search and version control data such that it fulfills the **FAIR** principles.

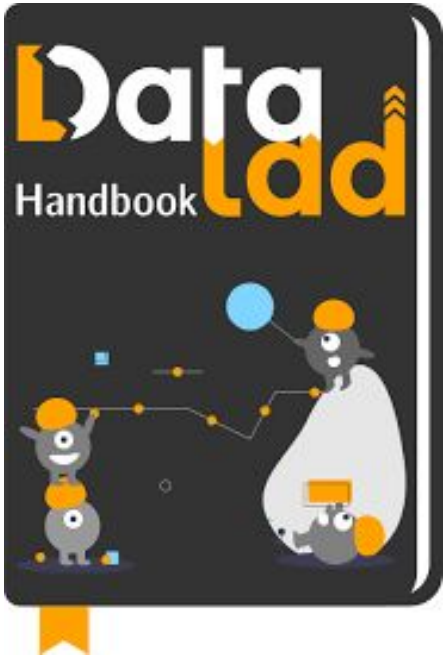
A command line tool built on Python and Git, developed for complete decentralization.

Why DataLad?

Premise…

★ Datalad-containers extension allows for complete *software environments* to be captured in *computational containers* and shared along with datasets

The DataLad Handbook



- A ‘how-to’ guide to use Datalad, intended for any audience.
- Teaches principles of data (and code) management for **FAIR** projects.
- Two parts:
  - The basics - core DataLad functionality
  - Use Cases - concrete DataLad applications

Use Case

Goal: Reproduce a portable analysis starting from the acquisition of a neuroimaging dataset (from MRI scanner) up to complete data analysis results.

We start by creating a home for the raw data:

```
$ datalad create localizer_scans
$ cd localizer_scans
[INFO] Creating a new annex repo at /home/me/usecases/repro2/localizer
[INFO] Scanning for unlocked files (this may take some time)
create(ok): /home/me/usecases/repro2/localizer_scans (dataset)
```

- Step-by-step guide to reproduce the study

Try 1: Khady

1st try: **Error** when trying to load the DICOM converter HeudiConv

```
$ datalad containers-add heudiconv --url shub://ReproNim/
ohbm2018-training:heudiconvn

[ERROR  ] 'GitRepo' object has no attribute
'get_special_remotes' [containers_add.py:_ensur
e_datalad_remote:74] (AttributeError)
```

Try 1: Khady

2nd try: Few days later, started over.

```
(base) Khady:localizer_scans Khady$ datalad containers-run -m "Convert
 sub-02 DICOMs into BIDS"   --container-name heudiconv   heudiconv -f
reproin -s 02 -c dcm2niix -b -l "" --minmeta -a .   -o /tmp/heudiconv.
sub-02 --files inputs/rawdata/dicoms
[INFO   ] Making sure inputs are available (this may take some time)
[INFO   ] == Command start (output follows) =====
/bin/sh: singularity: command not found
[INFO   ] == Command exit (modification check follows) =====
[INFO   ] The command had a non-zero exit code. If this is expected, y
ou can save the changes with 'datalad save -d . -r -F .git/COMMIT_EDIT
MSG'
CommandError: 'singularity exec .datalad/environments/heudiconv/image
heudiconv -f reproin -s 02 -c dcm2niix -b -l '' --minmeta -a . -o /tmp
/heudiconv.sub-02 --files inputs/rawdata/dicoms' failed with exitcode
127 under /Users/Mamie/localizer_scans
```

➢ Now the DICOM converter loads, but getting **Errors** from the 'Singularity' software...

Try 2: Mariana

Wasn't able to meet the requirements for the container installation.

Tried creating another environment using pipenv to switch between different versions of Python.

```
ERROR: Could not find a version that satisfies the requirement datalad>=0.12 (from datalad-neu
roimaging) (from versions: 0.2, 0.2.1.dev1, 0.2.1, 0.2.2, 0.2.3, 0.3, 0.3.1, 0.4, 0.4.1, 0.5.0
, 0.5.1, 0.6.0.dev1, 0.6.0, 0.7.0, 0.8.0, 0.8.1, 0.9.0, 0.9.1, 0.9.2, 0.9.3, 0.10.0rc1, 0.10.0
rc2, 0.10.0rc3, 0.10.0rc4, 0.10.0rc5, 0.10.0, 0.10.1, 0.10.2, 0.10.3.1, 0.11.0, 0.11.1, 0.11.2
, 0.11.3, 0.11.4, 0.11.5, 0.11.6, 0.11.7, 0.11.8, 0.12.0rc1, 0.12.0rc2, 0.12.0rc3, 0.12.0rc4,
0.12.0rc5)
ERROR: No matching distribution found for datalad>=0.12 (from datalad-neuroimaging)
```

Try 3: Heike

Could not run software on Windows (git-archive required).

After installing WSL, and installing multiple packages (HeudiConv, dcm2niix, Singularity).

→ Created datalad container and downloaded sample data
→ Created data subset
→ Tried converting DICOM to NifTi, but could not resolve error message

```
[INFO   ] Making sure inputs are available (this may take some time)
[INFO   ] == Command start (output follows) =====
INFO:    Converting SIF file to temporary sandbox...
INFO: Running heudiconv version 0.5.2-dev
Traceback (most recent call last):
  File "/usr/local/bin/heudiconv", line 11, in <module>
    load_entry_point('heudiconv==0.5.2.dev0', 'console_scripts', 'heudiconv')()
  File "/usr/local/lib/python3.5/dist-packages/heudiconv/cli/run.py", line 125, in main
    process_args(args)
  File "/usr/local/lib/python3.5/dist-packages/heudiconv/cli/run.py", line 252, in process_args
    args.subjs, grouping=args.grouping)
  File "/usr/local/lib/python3.5/dist-packages/heudiconv/parser.py", line 164, in get_study_sessions
    for _, files_ex in get_extracted_dicoms(files):
  File "/usr/local/lib/python3.5/dist-packages/heudiconv/parser.py", line 84, in get_extracted_dicoms
    if not tarfile.is_tarfile(t):
  File "/usr/lib/python3.5/tarfile.py", line 2450, in is_tarfile
    t = open(name)
  File "/usr/lib/python3.5/tarfile.py", line 1559, in open
    return func(name, "r", fileobj, **kwargs)
  File "/usr/lib/python3.5/tarfile.py", line 1685, in xzopen
    fileobj = lzma.LZMAFile(fileobj or name, mode, preset=preset)
  File "/usr/lib/python3.5/lzma.py", line 118, in __init__
    self._fp = builtins.open(filename, mode)
FileNotFoundError: [Errno 2] No such file or directory: 'inputs/rawdata/dicoms'
```

```
(mod8) heike@DESKTOP-TKF1DOI:/mnt/c/Users/heike/OneDrive - McGill University/PhD/Courses/QLSC600/Module8/localizer_scans/inputs/rawdata/dicoms$
```

Discussion

- Narrative clear and well structured, with extensive demonstration of how to run code and what output to expect

Missing Information & Points of Improvement

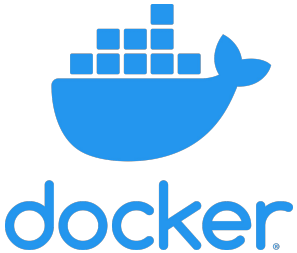
- Requires Git knowledge and a GitHub account
- Assumes knowledge of terminal/command line
- OS specifications missing
  - Does not work on Windows
- Dependencies (and version of dependencies) missing

FAIR principles

	Data*	DataLad Software*	DataLad Handbook / Use Case (Narrative)
Findable	Yes	Yes	Yes
Accessible	Yes	No	Yes
Interoperable	Unclear	Unclear	Yes
Reusable	Unclear	Unclear	Yes
Citeable	No	Partly	No

* when accessed/used through article

Solution: Docker



Use a different tool to manage containers.

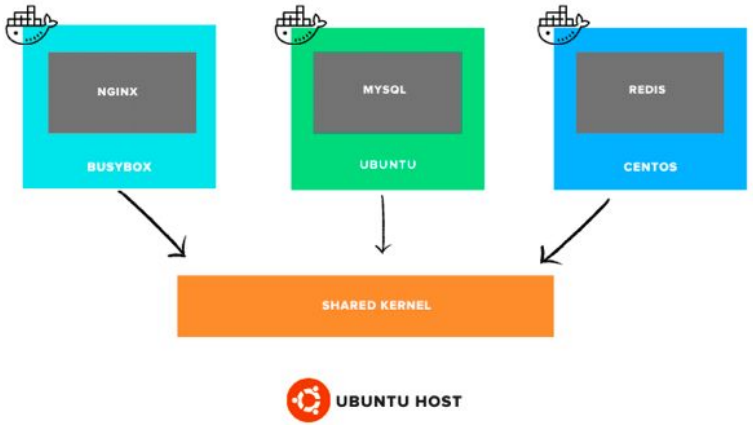
What is a container?

It is a process or group of processes with enough isolation so that it seems like a separate operating system.

Docker is an open-source project that makes setting up a container easier.

Solution: Docker

The main difference between a virtual machine and a container is that the first has its own kernel, while the second only the required files for the processes and uses a shared host kernel.



Solution: Docker

Advantages of using Docker:

- Portability
- Consistent operation
- More efficiency
- Less overhead

Projects like The Rocker Project provide a collection of containers suited for different needs when coding in R.

Questions?