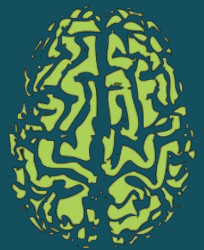
Datasets

...

Sebastian
@s_urchs (slack: @surchs)



BrainHack
School

Before we start:

AWS CLI

cyberduck.io

https://docs.aws.amazon.com/cli/latest/userguide/cli-chap-install.html

```
→ pip install awscli
```

```
~ via aws
→ conda install awscli
```

If you ask yourself...

how

- do I choose a good dataset?
- do I find open datasets?
- do I get the data?
- do I work with the data?

then this is for you!

How do I choose a good dataset?

Ask yourself:

- How easy can I get access?
- How “raw” is it?
- How useful is it?

Copied liberally from Chris Gorgolewski’s slides (@chrisgorgo)

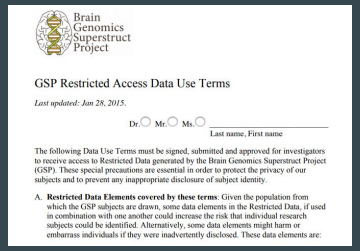
Ease of access

Signed Data Usage Agreement

Access through managed database

“Just get it” Direct download

“It’s right there”



Brain Genomics Superstruct Project

GSP Restricted Access Data Use Terms

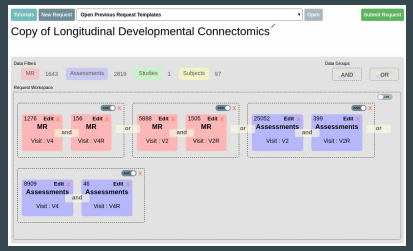
Last updated: Jan 28, 2015.

Dr.◯ Mr.◯ Ms.◯ _____

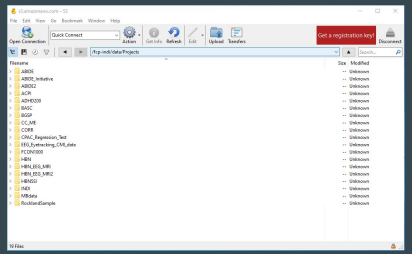
Last name, First name

The following Data Use Terms must be signed, submitted and approved for investigators to receive access to Restricted Data generated by the Brain Genomics Superstruct Project (GSP). These special precautions are essential in order to protect the privacy of our subjects and to prevent any inappropriate disclosure of subject identity.

A. **Restricted Data Elements covered by these terms**: Given the population from which the GSP subjects are drawn, some data elements in the Restricted Data, if used in combination with one another could increase the risk that individual research subjects could be identified. Alternatively, some data elements might harm or embarrass individuals if they were inadvertently disclosed. These data elements are:



Copy of Longitudinal Developmental Connectomics



Hard

Easy

How raw is the data set

More control

Less work

Organized

Preprocessed

Derivatives

- DICOM
- Idiosyncratic organization
- Must be converted

- Ideally standard organization (BIDS)
- Text-based metadata
- Must be preprocessed

- Minimally or fully preprocessed
- Standardized organization
- Data Quality metrics
- Must be analyzed

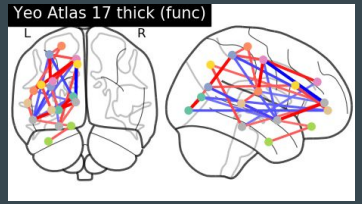
- Statistical maps
- Summary metrics

```
total 62M
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 1
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 10
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 100
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 101
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 102
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 103
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 104
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 105
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 106
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 107
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 108
-rwxrwxr-x 1 cmoreau def-jacquese 181K Feb 10  2018 109
```

```
sub-control01/
    anat/
        sub-control01_T1w.nii.gz
        sub-control01_T1w.json
        sub-control01_T2w.nii.gz
        sub-control01_T2w.json
    func/
        sub-control01_task-nback_bold.nii.gz
        sub-control01_task-nback_bold.json
        sub-control01_task-nback_events.tsv
        sub-control01_task-nback_physio.tsv.gz
        sub-control01_task-nback_physio.json
        sub-control01_task-nback_sbref.nii.gz
    dwi/
        sub-control01_dwi.nii.gz
        sub-control01_dwi.bval
        sub-control01_dwi.bvec
    fmap/
        sub-control01_phasediff.nii.gz
        sub-control01_phasediff.json
        sub-control01_magnitude1.nii.gz
        sub-control01_scans.tsv

    code/
        deface.py
    derivatives/
    README
    participants.tsv
    dataset_description.json
    CHANGES
```

```
total 7.1G
-rw-r--r-- 1 surchs cisl  18K Apr  8 13:38 fmri_sub0028852_session1_rest1_extra.mat
-rw-r--r-- 1 surchs cisl 125M Apr  8 13:38 fmri_sub0028852_session1_rest1.nii.gz
-rw-r--r-- 1 surchs cisl  18K Apr  8 12:21 fmri_sub0028853_session1_rest1_extra.mat
-rw-r--r-- 1 surchs cisl 125M Apr  8 12:21 fmri_sub0028853_session1_rest1.nii.gz
-rw-r--r-- 1 surchs cisl  21K Apr  8 13:57 fmri_sub0028854_session1_rest1_extra.mat
-rw-r--r-- 1 surchs cisl 126M Apr  8 13:57 fmri_sub0028854_session1_rest1.nii.gz
-rw-r--r-- 1 surchs cisl  23K Apr  8 12:28 fmri_sub0028855_session1_rest1_extra.mat
-rw-r--r-- 1 surchs cisl 124M Apr  8 12:27 fmri_sub0028855_session1_rest1.nii.gz
-rw-r--r-- 1 surchs cisl  22K Apr  8 12:42 fmri_sub0028856_session1_rest1_extra.mat
-rw-r--r-- 1 surchs cisl 125M Apr  8 12:41 fmri_sub0028856_session1_rest1.nii.gz
-rw-r--r-- 1 surchs cisl  13K Apr  8 12:22 fmri_sub0028857_session1_rest1_extra.mat
-rw-r--r-- 1 surchs cisl 126M Apr  8 12:21 fmri_sub0028857_session1_rest1.nii.gz
-rw-r--r-- 1 surchs cisl  15K Apr  8 14:11 fmri_sub0028858_session1_rest1_extra.mat
```



How useful is the data set

Data Quality

- Is there a publication describing the data
- Did other studies re-use the data
- Ask around

Meta Data Quality

- Are the data well described (inclusion, acquisition, processing)
- Are the metrics available that you are interested in?
- How much missing data are there?

Data Cost

- How large is the data set (storage space, download)
- Are the data preprocessed?
- Could you preprocess them (time, resources, knowledge) ?

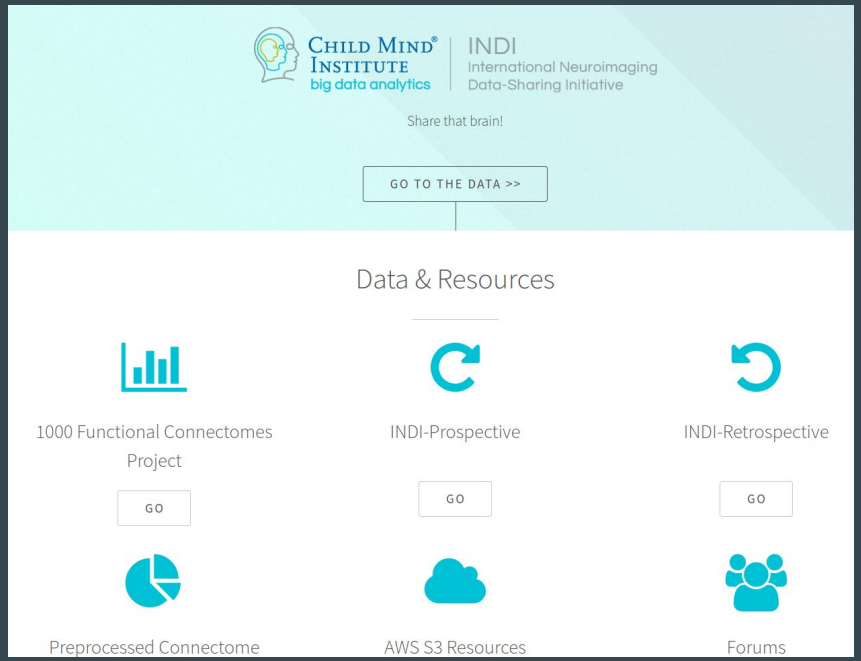
How do I find open datasets

https://openneuro.org/

http://fcon_1000.projects.nitrc.org/

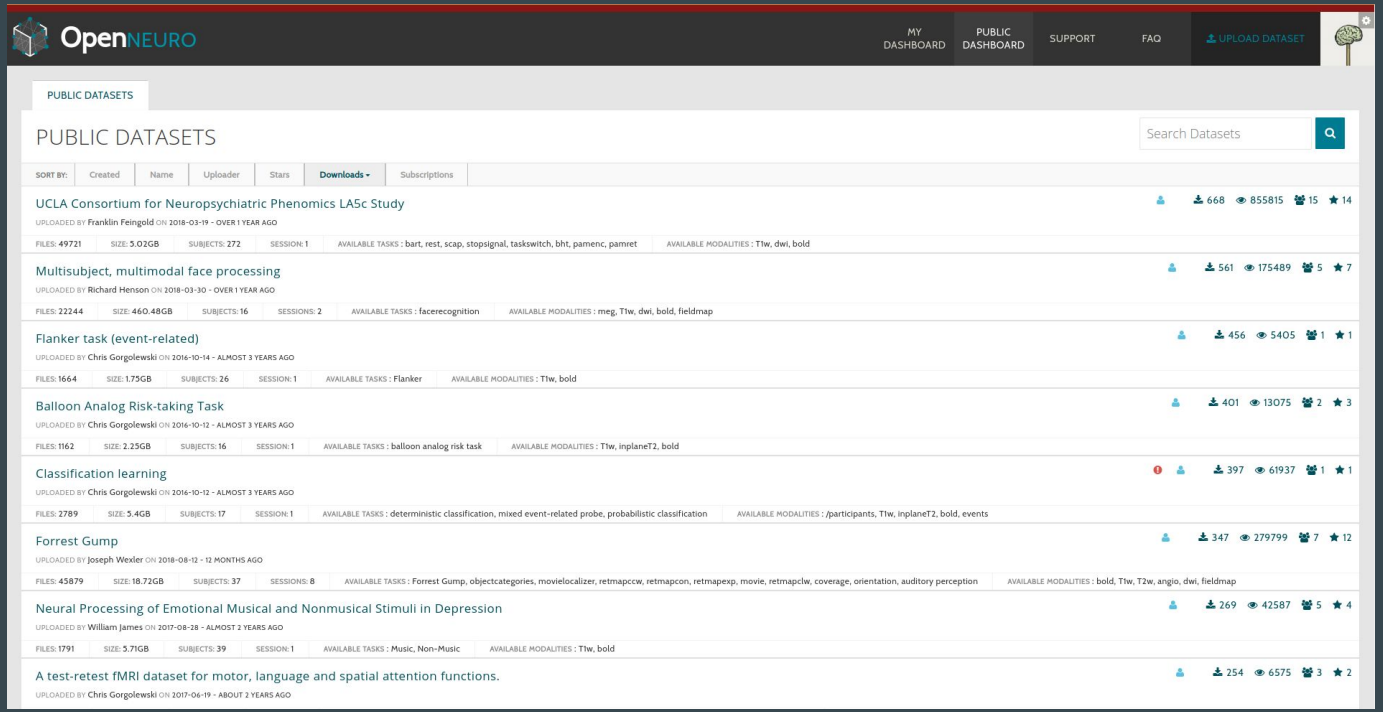
FCP-INDI

http://fcon_1000.projects.nitrc.org/



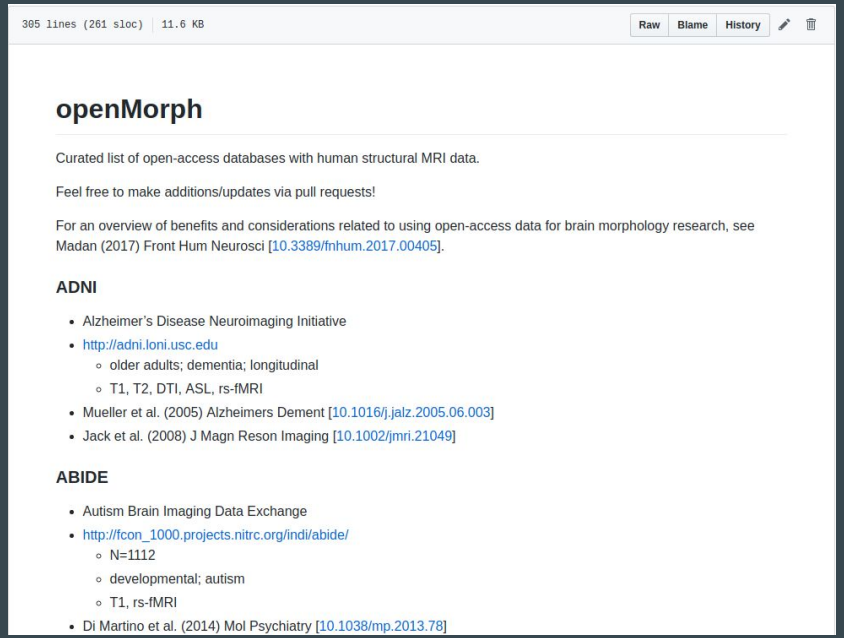
Open-Neuro

https://openneuro.org/



openMorph

https://github.com/cMadan/openMorph



305 lines (261 sloc) | 11.6 KB

Raw | Blame | History

openMorph

Curated list of open-access databases with human structural MRI data.

Feel free to make additions/updates via pull requests!

For an overview of benefits and considerations related to using open-access data for brain morphology research, see Madan (2017) Front Hum Neurosci [10.3389/fnhum.2017.00405].

ADNI

- Alzheimer's Disease Neuroimaging Initiative
- http://adni.loni.usc.edu
  - older adults; dementia; longitudinal
  - T1, T2, DTI, ASL, rs-fMRI
- Mueller et al. (2005) Alzheimers Dement [10.1016/j.jalz.2005.06.003]
- Jack et al. (2008) J Magn Reson Imaging [10.1002/jmri.21049]

ABIDE

- Autism Brain Imaging Data Exchange
- http://fcon_1000.projects.nitrc.org/indi/abide/
  - N=1112
  - developmental; autism
  - T1, rs-fMRI
- Di Martino et al. (2014) Mol Psychiatry [10.1038/mp.2013.78]

nilearn

7.2. `nilearn.datasets`: Automatic Dataset Fetching

Helper functions to download NeuroImaging datasets

User guide: See the Fetching open datasets from Internet section for further details.

Functions:

`fetch_atlas_craddock_2012`([data_dir, url, …])	Download and return file names for the Craddock 2012 parcellation
`fetch_atlas_destrieux_2009`([lateralized, …])	Download and load the Destrieux cortical atlas (dated 2009)
`fetch_atlas_harvard_oxford`(atlas_name[, …])	Load Harvard-Oxford parcellations from FSL.
`fetch_atlas_msdl`([data_dir, url, resume, …])	Download and load the MSDL brain atlas.
`fetch_coords_power_2011`()	Download and load the Power et al.
`fetch_atlas_smith_2009`([data_dir, mirror, …])	Download and load the Smith ICA and BrainMap atlas (dated 2009)
`fetch_atlas_yeo_2011`([data_dir, url, …])	Download and return file names for the Yeo 2011 parcellation.
`fetch_atlas_aal`([version, data_dir, url, …])	Downloads and returns the AAL template for SPM 12.
`fetch_atlas_basc_multiscale_2015`([version, …])	Downloads and loads multiscale functional brain parcellations
`fetch_atlas_allen_2011`([data_dir, url, …])	Download and return file names for the Allen and MIALAB ICA atlas (dated 2011).
`fetch_atlas_pauli_2017`([version, data_dir, …])	Download the Pauli et al.
`fetch_coords_dosenbach_2010`([ordered_regions])	Load the Dosenbach et al.
`fetch_abide_pcp`([data_dir, n_subjects, …])	Fetch ABIDE dataset
`fetch_adhd`([n_subjects, data_dir, url, …])	Download and load the ADHD resting-state dataset.
`fetch_haxby`([data_dir, n_subjects, …])	Download and loads complete haxby dataset
`fetch_icbm152_2009`([data_dir, url, resume, …])	Download and load the ICBM152 template (dated 2009)
`fetch_icbm152_brain_gm_mask`([data_dir, …])	Downloads ICBM152 template first, then loads ‘gm’ mask image.
`fetch_localizer_button_task`([data_dir, url, …])	Fetch left vs right button press contrast maps from the localizer.
`fetch_localizer_contrasts`(contrasts[, …])	Download and load Brainomics/Localizer dataset (94 subjects)

How do I get the data

Directly from the Amazon S3 bucket

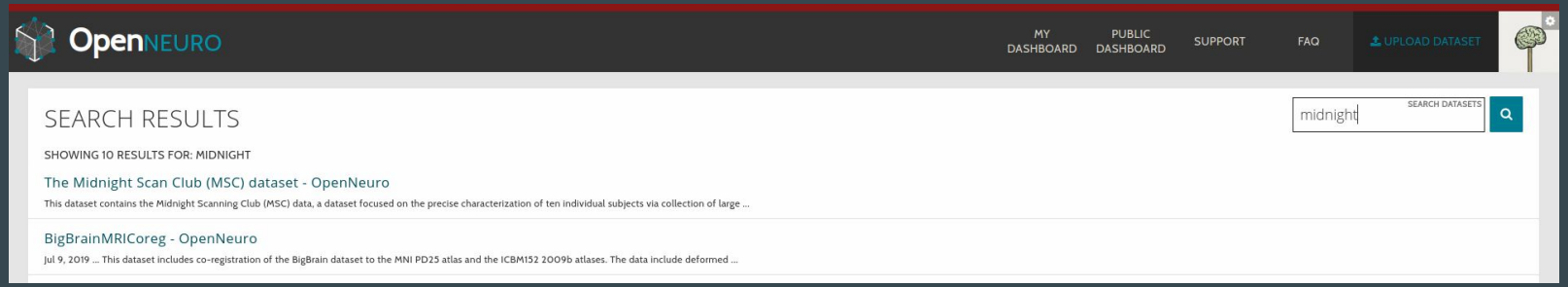
Through the nilearn data grabber

Amazon S3 example

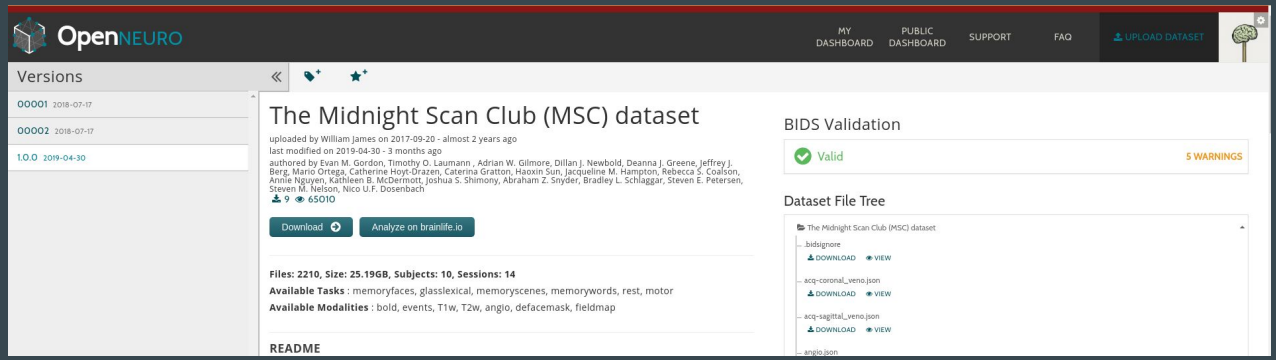
1. Pick a dataset
2. Login to the bucket
3. Download the data

Amazon S3 - pick a dataset

- Go to openneuro.org, search for ‘midnight’

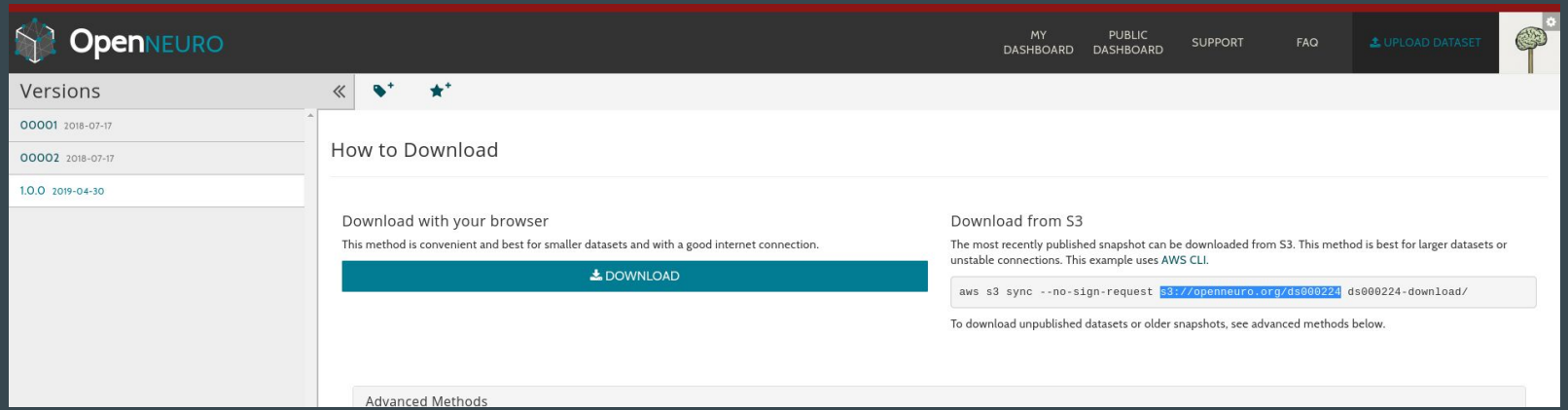


- Click ‘download’



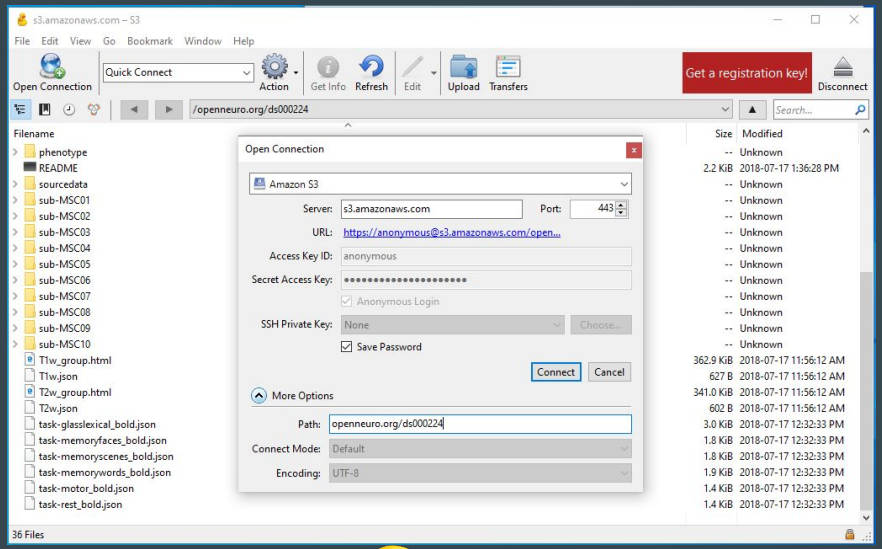
Amazon S3 - pick a dataset

- Copy the AWS S3 path



Amazon S3 - get the data

- Enter the info from openneuro in the “Path” field.



- Copy the line from openneuro
- Replace “sync” with “ls”

```
→ aws s3 ls --no-sign-request s3://openneuro.org/ds000224/
                           PRE .datalad/
                           PRE derivatives/
                           PRE phenotype/
                           PRE sourcedata/
                           PRE sub-MSC01/
                           PRE sub-MSC02/
                           PRE sub-MSC03/
                           PRE sub-MSC04/
                           PRE sub-MSC05/
                           PRE sub-MSC06/
                           PRE sub-MSC07/
```

AWS CLI

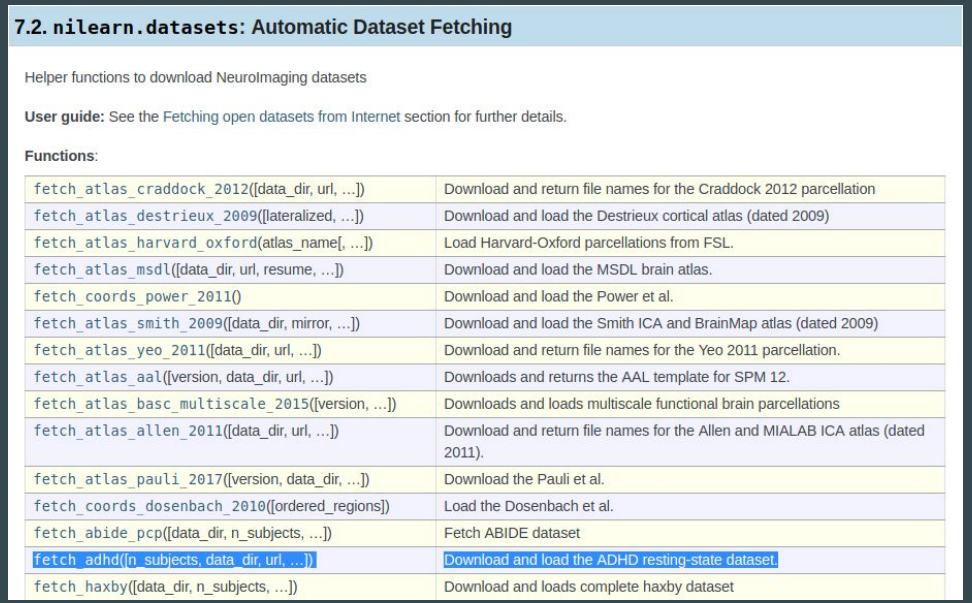
Nilearn example

1. Pick a dataset
2. Ask nilearn to get it
3. Load the data

Nilearn - pick a dataset

https://nilearn.github.io/modules/reference.html#module-nilearn.datasets

Pick the ADHD 200 dataset



7.2. `nilearn.datasets`: Automatic Dataset Fetching

Helper functions to download NeuroImaging datasets

User guide: See the Fetching open datasets from Internet section for further details.

Functions:

`fetch_atlas_craddock_2012`([data_dir, url, ...])	Download and return file names for the Craddock 2012 parcellation
`fetch_atlas_destrieux_2009`([lateralized, ...])	Download and load the Destrieux cortical atlas (dated 2009)
`fetch_atlas_harvard_oxford`(atlas_name[, ...])	Load Harvard-Oxford parcellations from FSL.
`fetch_atlas_msdl`([data_dir, url, resume, ...])	Download and load the MSDL brain atlas.
`fetch_coords_power_2011`()	Download and load the Power et al.
`fetch_atlas_smith_2009`([data_dir, mirror, ...])	Download and load the Smith ICA and BrainMap atlas (dated 2009)
`fetch_atlas_yeo_2011`([data_dir, url, ...])	Download and return file names for the Yeo 2011 parcellation.
`fetch_atlas_aal`([version, data_dir, url, ...])	Downloads and returns the AAL template for SPM 12.
`fetch_atlas_basc_multiscale_2015`([version, ...])	Downloads and loads multiscale functional brain parcellations
`fetch_atlas_allen_2011`([data_dir, url, ...])	Download and return file names for the Allen and MIALAB ICA atlas (dated 2011).
`fetch_atlas_pauli_2017`([version, data_dir, ...])	Download the Pauli et al.
`fetch_coords_dosenbach_2010`([ordered_regions])	Load the Dosenbach et al.
`fetch_abide_pcp`([data_dir, n_subjects, ...])	Fetch ABIDE dataset
`fetch_adhd`([n_subjects, data_dir, url, ...])	Download and load the ADHD resting-state dataset.
`fetch_haxby`([data_dir, n_subjects, ...])	Download and loads complete haxby dataset

Nilearn - get the data

[1]:
```
%matplotlib inline
```

[2]:
```
# Import some imports
import pandas as pd
import nibabel as nib
import nilearn as nil
from nilearn import plotting as nlp
```

```
/home/surchs/Packages/conda/envs/abide/lib/python3.7/importlib/_bootstrap.py:219: RuntimeWarning: numpy.ufunc size changed, may indicate binary incompatibility. Expected
192 from C header, got 216 from PyObject
  return f(*args, **kwds)
```

[3]:
```
# Get one subject from the ADHD-200 dataset
data = nil.datasets.fetch_adhd(n_subjects=1)
```

```
/home/surchs/Packages/conda/envs/abide/lib/python3.7/site-packages/nilearn/datasets/func.py:503: VisibleDeprecationWarning: Reading unicode strings without specifying the
encoding argument is deprecated. Set the encoding, use None for the system default.
  dtype=None)
```

[4]:
```
# Look inside the data
data.keys()
```

[4]:
```
dict_keys(['func', 'confounds', 'phenotypic', 'description'])
```

[5]:
```
print(data['description'].decode('utf-8'))
```

```
ADHD 200


Notes
-----
Part of the the 1000 Functional Connectome Project. Phenotypic
information includes: diagnostic status, dimensional ADHD symptom measures,
age, sex, intelligence quotient (IQ) and lifetime medication status.
Preliminary quality control assessments (usable vs. questionable) based upon
visual timeseries inspection are included for all resting state fMRI scans.

Includes preprocessed data from 40 participants.

Project was coordinated by Michael P. Milham.

Content
-------
    :'func': Nifti images of the resting-state data
    :'phenotypic': Explanations of preprocessing steps
    :'confounds': CSV files containing the nuisance variables

References
----------
```

Nilearn - access the data

[6]:
```
# Look at the location of the data
data['func']
```

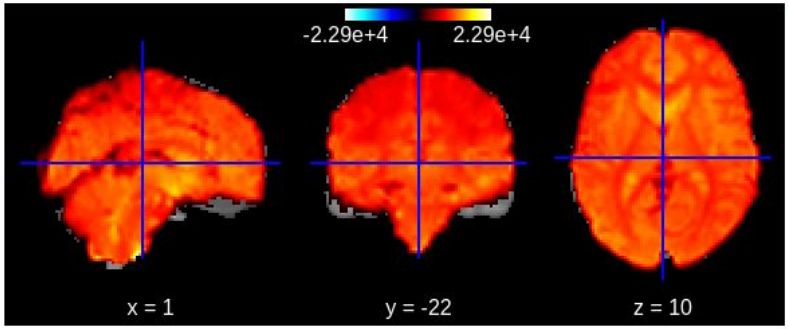
[6]:
```
['/home/surchs/nilearn_data/adhd/data/0010042/0010042_rest_tshift_RPI_voreg_mni.nii.gz']
```

[7]:
```
# Load the downloaded data with nibabel
img = nib.load(data['func'][0])
```

[8]:
```
# Show the first time point of the time series
nlp.view_img(nil.image.index_img(img, 2))
```

```
/home/surchs/Packages/conda/envs/abide/lib/python3.7/site-packages/scipy/ndimage/measurements.py:272: DeprecationWarning: In future, it will be an error for 'np.bool_' sc
alars to be interpreted as an index
  return _nd_image.find_objects(input, max_label)
```

[8]:



Nilearn - access the data

[9]:
```
# Look at the confounds
data['confounds']
```

[9]:
```
['/home/surchs/nilearn_data/adhd/data/0010042/0010042_regressors.csv']
```

[10]:
```
# Get the confounds
confounds = pd.read_csv(data['confounds'][0], sep='\t')
```

[11]:
```
confounds.head()
```

[11]:

	csf	constant	linearTrend	wm	global	motion-pitch	motion-roll	motion-yaw	motion-x	motion-y	motion-z	gm	compcor1	compcor2	compcor3	compcor4	compcor5
0	12140.708282	1.0	0.0	9322.722489	9955.469315	-0.0637	0.1032	-0.1516	-0.0376	-0.0112	0.0840	10617.938409	-0.035058	-0.006713	-0.071532	0.009847	-0.027601
1	12123.146913	1.0	1.0	9314.257684	9947.987176	-0.0708	0.0953	-0.1562	-0.0198	0.0021	0.0722	10611.036827	-0.026949	-0.091152	-0.030126	0.020055	-0.099798
2	12085.963127	1.0	2.0	9319.610045	9945.132852	-0.0795	0.0971	-0.1453	-0.0439	-0.0241	0.0972	10591.877177	0.002552	0.069165	0.090166	-0.016608	-0.071980
3	12109.299348	1.0	3.0	9299.841075	9943.648622	-0.0607	0.0918	-0.1601	-0.0418	-0.0133	0.0877	10592.008336	0.079391	0.029959	-0.098036	0.062493	0.024105
4	12072.330305	1.0	4.0	9297.870869	9925.640852	-0.0706	0.0873	-0.1482	-0.0313	-0.0118	0.0712	10570.445905	0.075471	-0.030123	0.084739	0.088217	0.012996

How do I work with the data

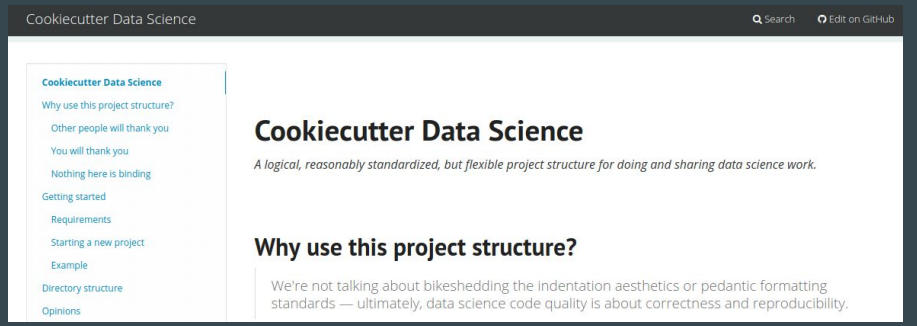
1. Respect the data
    a. Sign and follow data usage agreement
    b. Securely store identifiable information
    c. Cite the data source

How do I work with the data

1. Respect the data
    a. Sign and follow data usage agreement
    b. Securely store identifiable information
    c. Cite the data source
2. Document what you do
    a. Where did you download from?
    b. What command did you use (script if possible)
    c. How did you select data?
    d. What processing did you use?

How do I work with the data

1. Respect the data
    a. Sign and follow data usage agreement
    b. Securely store identifiable information
    c. Cite the data source
2. Document what you do
    a. Where did you download from?
    b. What command did you use (script if possible)
    c. How did you select data?
    d. What processing did you use?
3. Organize your project
    a. Use version control!
    b. You will thank yourself later
    c. https://drivendata.github.io/cookiecutter-data-science/



Some additional resources

zenodo

https://zenodo.org/
Repository for data associated with publication

Digital object identifier

Any license

Hosted by CERN (cool)

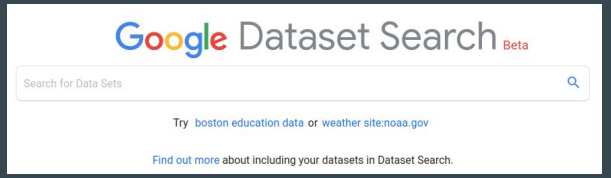
figshare
credit for all your research

https://figshare.com/
Repository for data

Digital object identifier

Creative Commons license

Has commercial side



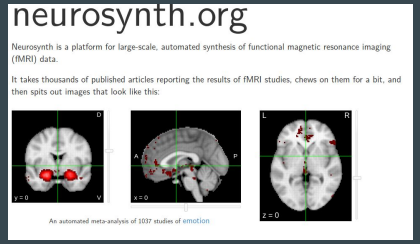
https://toolbox.google.com/datasetsearch
Dataset search engine

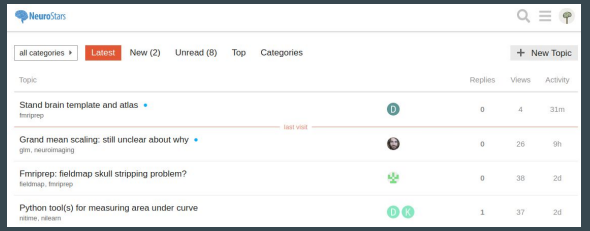
Doesn't store anything

Let's you search other databases

Some additional resources

NEUROVAULT





https://neurovault.org/
Repository of statistical maps of completed studies

http://neurosynth.org/
Aggregated activation data maps with keyword search

Neurosynth-genes has the gene expression data from the Allen brain institute

https://neurostars.org/
Great resource for asking question and getting feedback

Some additional data sets

https://db.humanconnectome.org/

Very high resolution data

Publicly available data for 1200 healthy individuals

Long, repeat imaging data (task and resting state)

Deep meta data

https://www.ukbiobank.ac.uk/

> 100.000 individuals

Deep meta data

Genetics (prospective whole genome sequencing)

Extensive imaging data

Medical records

http://portal.brain-map.org/

Human brain gene expression maps

Histological and developmental atlases

Extensive mouse data