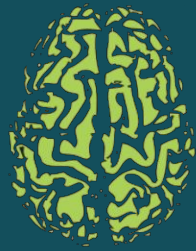


# Datasets

...

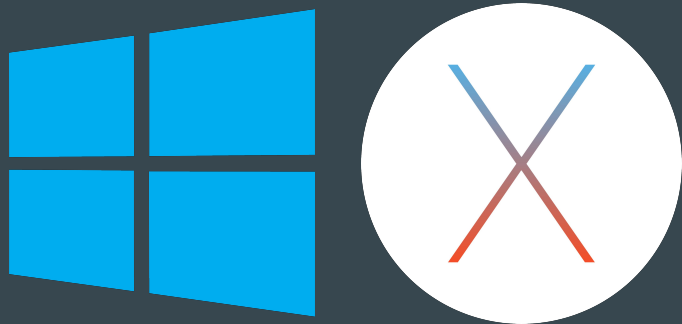
Sebastian

@s\_urchs (slack: @surchs)



BrainHack  
School

# Before we start:




[cyberduck.io](https://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?

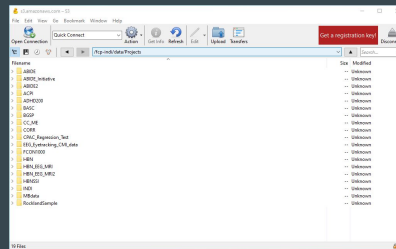
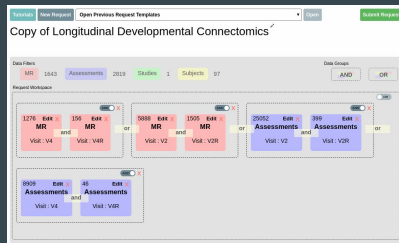
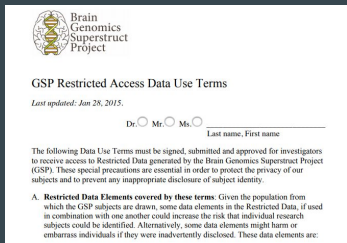
# Ease of access

# Signed Data Usage Agreement

## Access through managed database

“Just get it”  
Direct download

“It’s right there”



# Hard

# Easy

# How raw is the data set

More control



Organized



Preprocessed



Less work



Derivatives

- DICOM
- Idiosyncratic organization
- Must be converted

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

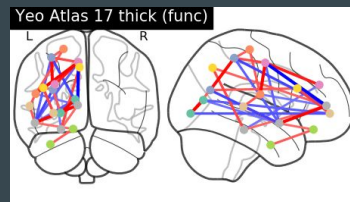
- Minimally or fully preprocessed
- Standardized organization
- Data Quality metrics
- Must be analyzed

- 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



**OpenNEURO**<sup>BETA</sup>

A free and open platform for analyzing  
and sharing neuroimaging data

<https://openneuro.org/>



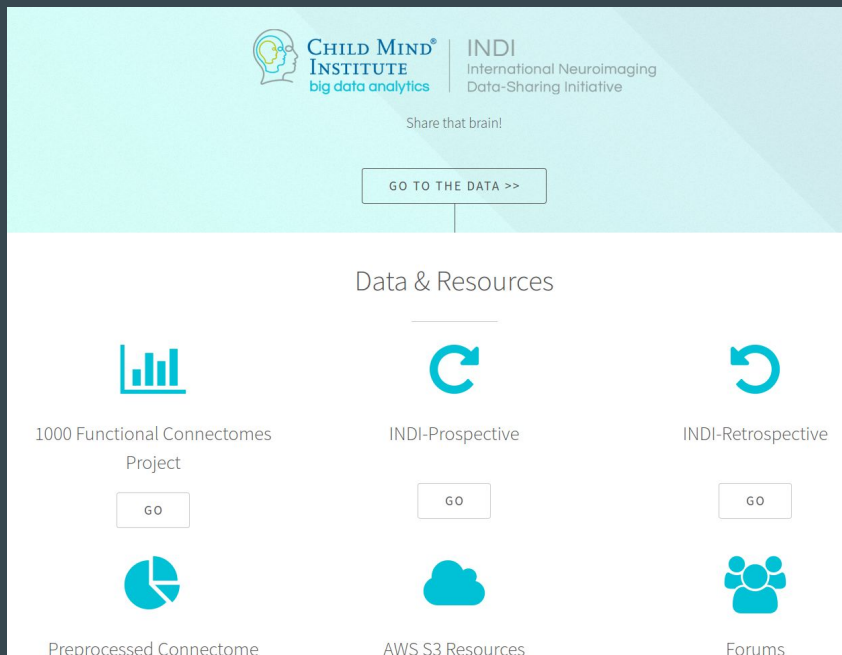
International Neuroimaging  
Data-Sharing Initiative

[http://fcon\\_1000.projects.nitrc.org/](http://fcon_1000.projects.nitrc.org/)



# FCP-INDI

[http://fcon\\_1000.projects.nitrc.org/](http://fcon_1000.projects.nitrc.org/)



The screenshot displays the FCP-INDI website. At the top, the 'CHILD MIND INSTITUTE big data analytics' logo is on the left, and the 'INDI International Neuroimaging Data-Sharing Initiative' logo is on the right. Below these logos is the text 'Share that brain!' and a button labeled 'GO TO THE DATA >>'. The main section is titled 'Data & Resources' and contains six icons arranged in a 2x3 grid. Each icon has a corresponding label and a 'GO' button below it. The icons and labels are: a bar chart for '1000 Functional Connectomes Project', a circular arrow for 'INDI-Prospective', a circular arrow for 'INDI-Retrospective', a pie chart for 'Preprocessed Connectome', a cloud for 'AWS S3 Resources', and a group of people for 'Forums'.

CHILD MIND INSTITUTE big data analytics | INDI International Neuroimaging Data-Sharing Initiative

Share that brain!

GO TO THE DATA >>

Data & Resources

1000 Functional Connectomes Project

GO

INDI-Prospective

GO

INDI-Retrospective

GO


Preprocessed Connectome

AWS S3 Resources

Forums

# Open-Neuro

<https://openneuro.org/>

 **OpenNEURO**

MY DASHBOARDPUBLIC DASHBOARD SUPPORTFAQUPLOAD DATASET

PUBLIC DATASETS

PUBLIC DATASETS

Search Datasets

SORT BY:

Created

Name

Uploader

Stars

Downloads

Subscriptions

UCLA Consortium for Neuropsychiatric Phenomics LA5c Study

UPLOADED BY Franklin Feingold ON 2018-03-19 - OVER 1 YEAR AGO

FILES: 49721SIZE: 5.02GBSUBJECTS: 272SESSION: 1AVAILABLE TASKS: bart, rest, scap, stopsignal, taskswitch, bht, pamenc, pamretAVAILABLE MODALITIES: T1w, dwi, bold

6688558151514

Multisubject, multimodal face processing

UPLOADED BY Richard Henson ON 2018-03-30 - OVER 1 YEAR AGO

FILES: 22244SIZE: 460.48GBSUBJECTS: 16SESSIONS: 2AVAILABLE TASKS: facerecognitionAVAILABLE MODALITIES: meg, T1w, dwi, bold, fieldmap

56117548957

Flanker task (event-related)

UPLOADED BY Chris Gorgolewski ON 2016-10-14 - ALMOST 3 YEARS AGO

FILES: 1664SIZE: 1.75GBSUBJECTS: 26SESSION: 1AVAILABLE TASKS: FlankerAVAILABLE MODALITIES: T1w, bold

456540511

Balloon Analog Risk-taking Task

UPLOADED BY Chris Gorgolewski ON 2016-10-12 - ALMOST 3 YEARS AGO

FILES: 1162SIZE: 2.25GBSUBJECTS: 16SESSION: 1AVAILABLE TASKS: balloon analog risk taskAVAILABLE MODALITIES: T1w, inplaneT2, bold

4011307523

Classification learning

UPLOADED BY Chris Gorgolewski ON 2016-10-12 - ALMOST 3 YEARS AGO

FILES: 2789SIZE: 5.4GBSUBJECTS: 17SESSION: 1AVAILABLE TASKS: deterministic classification, mixed event-related probe, probabilistic classificationAVAILABLE MODALITIES: fparticipants, T1w, inplaneT2, bold, events

3976193711

Forrest Gump

UPLOADED BY Joseph Wexler ON 2018-08-12 - 12 MONTHS AGO

FILES: 45879SIZE: 18.72GBSUBJECTS: 37SESSIONS: 8AVAILABLE TASKS: Forrest Gump, objectcategories, movielocalizer, retmapcow, retmapcon, retmapexp, movie, retmapclw, coverage, orientation, auditory perceptionAVAILABLE MODALITIES: bold, T1w, T2w, angio, dwi, fieldmap

347279799712

Neural Processing of Emotional Musical and Nonmusical Stimuli in Depression

UPLOADED BY William James ON 2017-08-18 - ALMOST 2 YEARS AGO

FILES: 1791SIZE: 5.71GBSUBJECTS: 39SESSION: 1AVAILABLE TASKS: Music, Non-MusicAVAILABLE MODALITIES: T1w, bold

2694258754

A test-retest fMRI dataset for motor, language and spatial attention functions.

UPLOADED BY Chris Gorgolewski ON 2017-06-19 - ABOUT 2 YEARS AGO

FILES: 1791SIZE: 5.71GBSUBJECTS: 39SESSION: 1AVAILABLE TASKS: Music, Non-MusicAVAILABLE MODALITIES: T1w, bold

254657532

# 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/](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]

## 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:**

<code>fetch_atlas_craddock_2012([data_dir, url, ...])</code>	Download and return file names for the Craddock 2012 parcellation
<code>fetch_atlas_destrieux_2009([lateralized, ...])</code>	Download and load the Destrieux cortical atlas (dated 2009)
<code>fetch_atlas_harvard_oxford(atlas_name[, ...])</code>	Load Harvard-Oxford parcellations from FSL.
<code>fetch_atlas_msdl([data_dir, url, resume, ...])</code>	Download and load the MSDl brain atlas.
<code>fetch_coords_power_2011()</code>	Download and load the Power et al.
<code>fetch_atlas_smith_2009([data_dir, mirror, ...])</code>	Download and load the Smith ICA and BrainMap atlas (dated 2009)
<code>fetch_atlas_yeo_2011([data_dir, url, ...])</code>	Download and return file names for the Yeo 2011 parcellation.
<code>fetch_atlas_aal([version, data_dir, url, ...])</code>	Downloads and returns the AAL template for SPM 12.
<code>fetch_atlas_basc_multiscale_2015([version, ...])</code>	Downloads and loads multiscale functional brain parcellations
<code>fetch_atlas_allen_2011([data_dir, url, ...])</code>	Download and return file names for the Allen and MIALAB ICA atlas (dated 2011).
<code>fetch_atlas_pauli_2017([version, data_dir, ...])</code>	Download the Pauli et al.
<code>fetch_coords_dosenbach_2010([ordered_regions])</code>	Load the Dosenbach et al.
<code>fetch_abide_pcp([data_dir, n_subjects, ...])</code>	Fetch ABIDE dataset
<code>fetch_adhd([n_subjects, data_dir, url, ...])</code>	Download and load the ADHD resting-state dataset.
<code>fetch_haxby([data_dir, n_subjects, ...])</code>	Download and loads complete haxby dataset
<code>fetch_icbm152_2009([data_dir, url, resume, ...])</code>	Download and load the ICBM152 template (dated 2009)
<code>fetch_icbm152_brain_gm_mask([data_dir, ...])</code>	Downloads ICBM152 template first, then loads 'gm' mask image.
<code>fetch_localizer_button_task([data_dir, url, ...])</code>	Fetch left vs right button press contrast maps from the localizer.
<code>fetch_localizer_contrasts(contrasts[, ...])</code>	Download and load Brainomics localizer dataset (24 subjects)

# How do I get the data



Directly from the Amazon S3 bucket



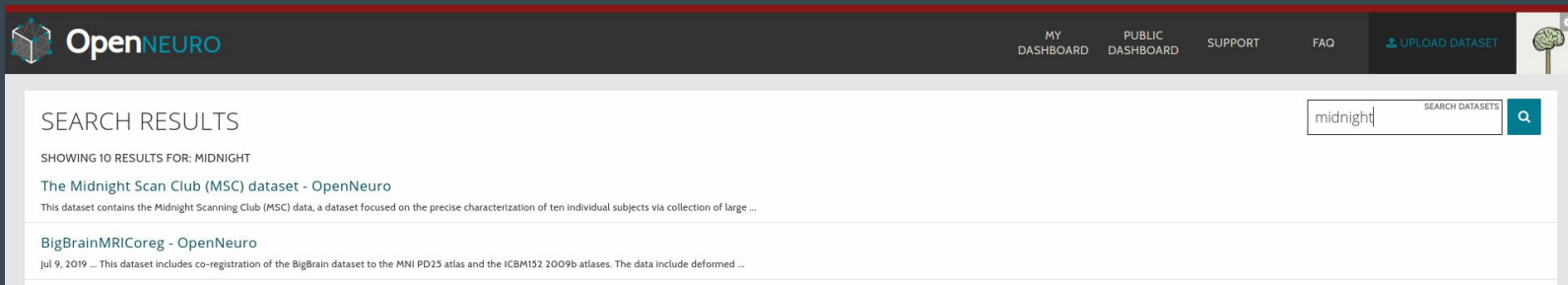
Through the nilearn data grabber

# Amazon S3 example

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

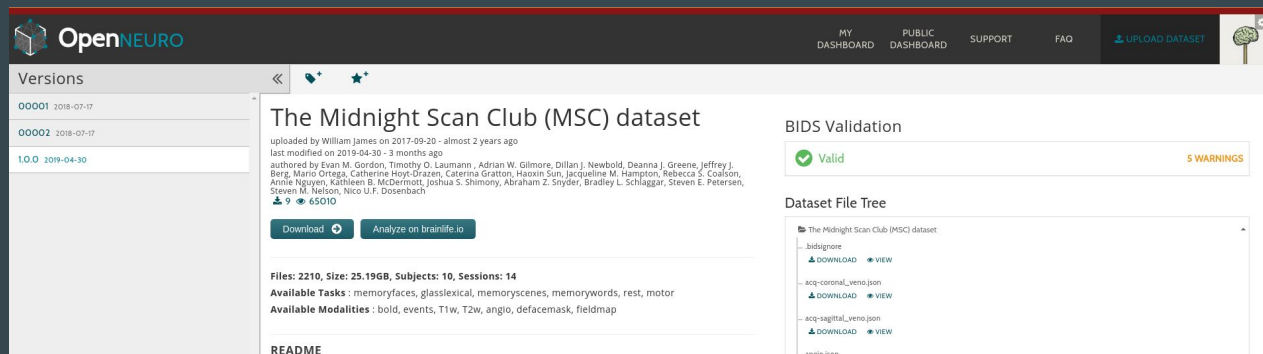
# Amazon S3 - pick a dataset

- Go to [openneuro.org](https://openneuro.org), search for 'midnight'



The screenshot shows the OpenNeuro website's search results page. The header includes the OpenNeuro logo and navigation links: MY DASHBOARD, PUBLIC DASHBOARD, SUPPORT, FAQ, and an UPLOAD DATASET button. A search bar on the right contains the text 'midnight'. Below the search bar, the page displays 'SEARCH RESULTS' and 'SHOWING 10 RESULTS FOR: MIDNIGHT'. The first result is 'The Midnight Scan Club (MSC) dataset - OpenNeuro', with a description: 'This dataset contains the Midnight Scanning Club (MSC) data, a dataset focused on the precise characterization of ten individual subjects via collection of large ...'. The second result is 'BigBrainMRICoreg - OpenNeuro', with a description: 'Jul 9, 2019 ... This dataset includes co-registration of the BigBrain dataset to the MNI PD25 atlas and the ICBM152 2009b atlases. The data include deformed ...'.

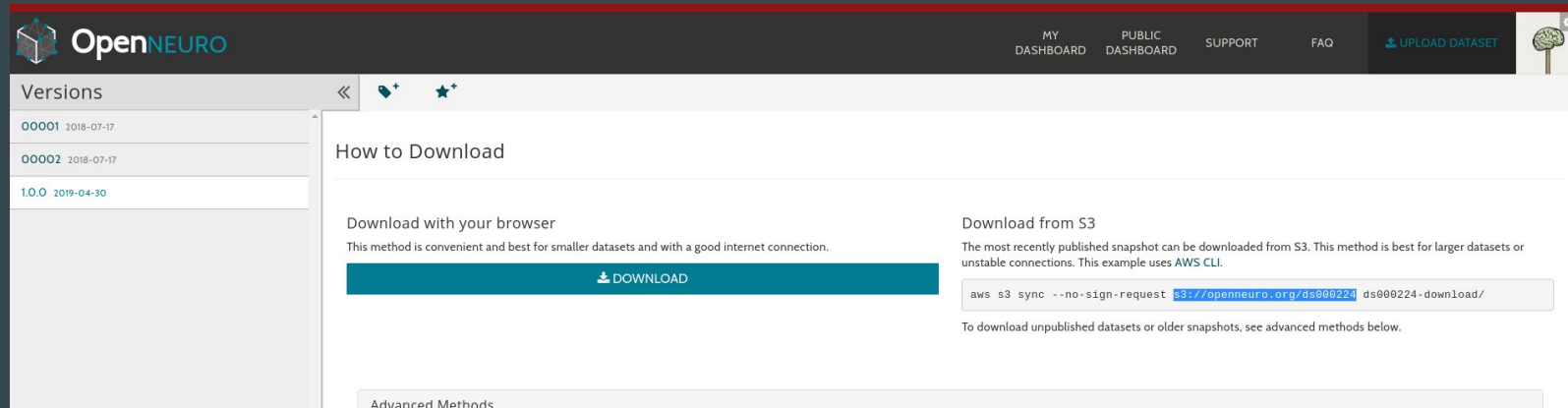
- Click 'download'



The screenshot shows the OpenNeuro website's dataset page for 'The Midnight Scan Club (MSC) dataset'. The header is the same as the previous screenshot. On the left, there is a 'Versions' sidebar showing two versions: '00001 2018-07-17' and '00002 2018-07-17', with the current version being '1.0.0 2019-04-30'. The main content area displays the dataset title 'The Midnight Scan Club (MSC) dataset', followed by a description: 'uploaded by William James on 2017-09-20 - almost 2 years ago last modified on 2019-04-30 - 3 months ago'. Below this, it lists the authors: 'authored by Evan M. Gordon, Timothy O. Laumann, Adrian W. Gilmore, Dillan J. Newbold, Deanna J. Greene, Jeffrey J. Berg, Mario Ortega, Catherine Hoyt-Orazen, Caterina Gratton, Haomin Sun, Jacqueline M. Hampton, Rebecca S. Coalson, Annie Nguyen, Kathleen B. McDermott, Joshua S. Shimony, Abraham Z. Snyder, Bradley L. Schlaggar, Steven E. Petersen, Steven M. Nelson, Nico U.F. Dosenbach'. There are two buttons: 'Download' and 'Analyze on brainlife.io'. Below the buttons, it lists the dataset statistics: 'Files: 2210, Size: 25.19GB, Subjects: 10, Sessions: 14'. It also lists the 'Available Tasks' and 'Available Modalities'. At the bottom, there is a 'README' section. On the right, there is a 'BIDS Validation' section showing a green checkmark and 'Valid', and a 'Dataset File Tree' section showing the file structure: 'The Midnight Scan Club (MSC) dataset' containing 'bidsignore', 'acq-coronal\_venio.json', 'acq-sagittal\_venio.json', and 'angio.json'.

# Amazon S3 - pick a dataset

- Copy the AWS S3 path



The screenshot shows the OpenNEURO website interface. The top navigation bar includes links for 'MY DASHBOARD', 'PUBLIC DASHBOARD', 'SUPPORT', 'FAQ', and 'UPLOAD DATASET'. The main content area is titled 'How to Download' and is divided into two columns. The left column, 'Download with your browser', includes a 'DOWNLOAD' button. The right column, 'Download from S3', provides instructions for downloading datasets from Amazon S3, including a code snippet for the AWS CLI command. A sidebar on the left lists dataset versions: '00001' (2018-07-17), '00002' (2018-07-17), and '1.0.0' (2019-04-30).

OpenNEURO

MY DASHBOARD PUBLIC DASHBOARD SUPPORT FAQ UPLOAD DATASET

Versions

00001 2018-07-17

00002 2018-07-17

1.0.0 2019-04-30

### How to Download

#### Download with your browser

This method is convenient and best for smaller datasets and with a good internet connection.

DOWNLOAD

#### Download from S3

The most recently published snapshot can be downloaded from S3. This method is best for larger datasets or unstable connections. This example uses AWS CLI.

```
aws s3 sync --no-sign-request s3://openneuro.org/ds000224 ds000224-download/
```

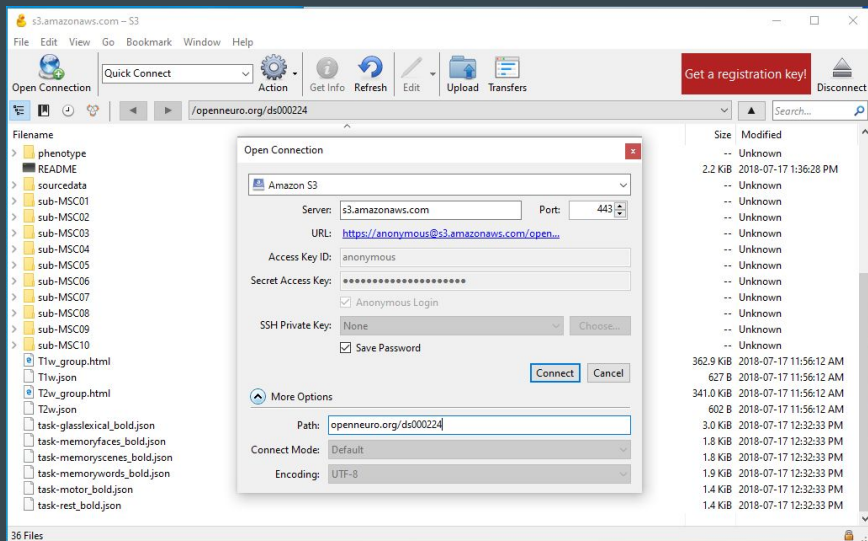
To download unpublished datasets or older snapshots, see advanced methods below.

Advanced Methods



# 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/
```



# 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:**

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

[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

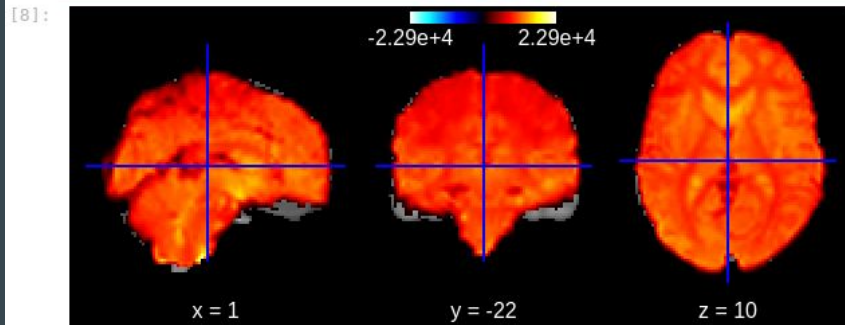
```
[6]: # Look at the location of the data  
data['func']
```

```
[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_' scalars to be interpreted as an index  
    return _nd_image.find_objects(input, max_label)
```



# 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

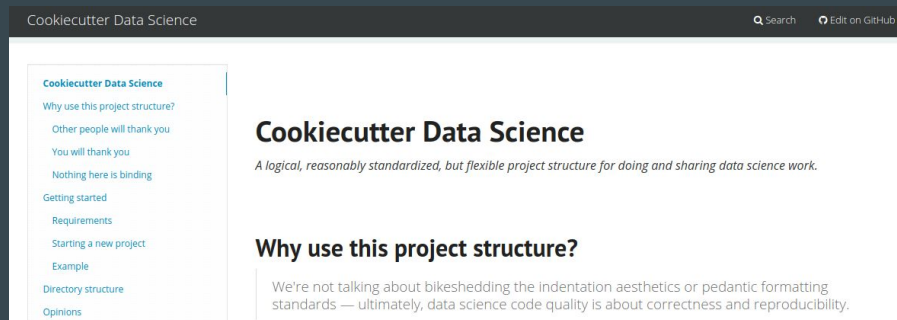
# How do I work with the data

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



<https://zenodo.org/>

Repository for data associated  
with publication

Digital object identifier

Any license

Hosted by CERN (cool)



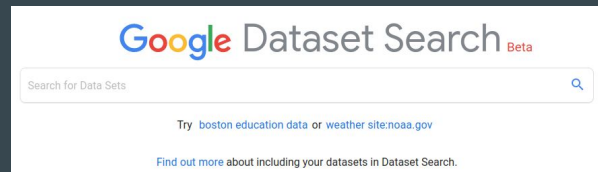
<https://figshare.com/>

Repository for data

Digital object identifier

Creative Commons license

Has commercial side



<https://toolbox.google.com/datasetsearch>

Dataset search engine

Doesn't store anything

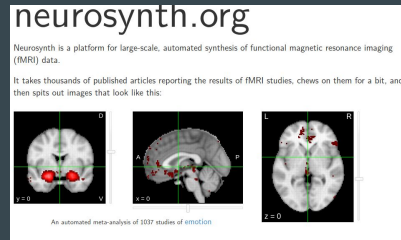
Let's you search other databases

# Some additional resources



<https://neurovault.org/>

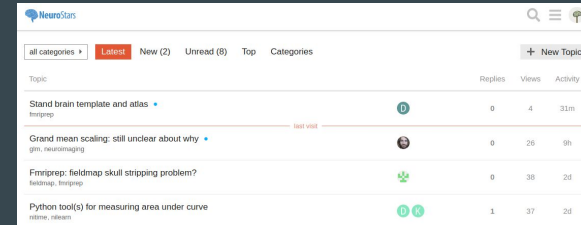
Repository of statistical maps of completed studies



<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