PERPETUAL
DECENTRALIZED MANAGEMENT OF DIGITAL OBJECTS
FOR
COLLABORATIVE OPEN-SCIENCE

Michael Hanke
Psychoinformatics lab, Institute of Psychology, Otto-von-Guericke-University, Magdeburg
Center for Behavioral Brain Sciences, Magdeburg

http://psychoinformatics.de

funded by the federal state of Sachsen-Anhalt and the European Regional Development Fund (ERDF), project: Center for Behavioral Brain Sciences (CBBS)
“The task of neural science is to explain behavior in terms of the activities of the brain.”
— Eric Kandel, Principles of Neuroscience.
INTER-INDIVIDUAL VARIABILITY? NON-COMPLIANCE? NOISE?

- three individual brains in a brain structure defined reference space (e.g., MNI)
- "diagnostic" voxels for distinguishing perception of tools and dwellings

Is brain structure alone an optimal reference for inter-individual analysis of brain function?

Mitchell et al., PLoS ONE, 2008
FUNCTIONAL HYPERALIGNMENT

Compute a transformation of a high-dimensional (representational) space based on a high-dimensional feature vector, such as the functional response to watching a movie (>1000 time points).

MORE ACCURATE PREDICTIVE MODELING (OF BRAIN ORGANIZATION)

Guntupalli, Hanke, Halchenko, Connolly, Ramadge & Haxby (2016). *A model of representational spaces in human cortex*. Cerebral Cortex, 26, 2919-2934. (suppl.)
Common pattern of involvement of brain networks in particular brain functions in real-life cognition.

Reconceptualization of inter-individual differences.

Potential to facilitate reliable clinical diagnostics.
1. Record data from lots of sensors/questionnaires
2. Determine key markers
3. Acquire normative samples
4. Describe individual sample relative to the norm

**NEO-FFI-3**

NEO Five-Factor Inventory-3

Paul T. Costa, Jr., PhD and Robert R. McCrae, PhD

**Instructions**

Write only where indicated in this Item Booklet. Carefully read all of the instructions before beginning. This questionnaire contains 60 statements. Read each statement carefully. For each statement, fill in the circle with the response that best represents your opinion. Make sure that your answer is in the correct box.

Fill in **SD** if you *strongly disagree* or the statement is definitely false.

Fill in **D** if you *disagree* or the statement is mostly false.

Fill in **N** if you are *neutral* on the statement, if you cannot decide, or if the statement is about equally true and false.

Fill in **A** if you *agree* or the statement is mostly true.

Fill in **SA** if you *strongly agree* or the statement is definitely true.

Note that the responses are numbered in rows.

**Example**

First five responses from an individual who *strongly disagrees* with items 1, 2, and 3, and agrees with items 4 and 5.

Fill in only one response for each statement. Respond to all of the statements, making sure that you fill in the correct response. **DO NOT ERASE**! If you need to change an answer, make an “X”
GUESSTIMATE MAGNITUDE OF COMPLEXITY

Too big, too risky, too expensive — for an individual lab/center

from Swaroop Guntupalli (unpublished feasibility study)
ROLE MODEL FOR COMMUNITY POTENTIAL

Concept
Give interested parties something to work on using their own resources and re-integrate their contributions for another cycle.
RESOURCES AND RESULTS TOWARDS A FUNCTIONAL BRAIN ATLAS

STUDYFORREST.ORG

- open data resource
- versatile structural imaging data
- 10+ hours of fMRI per subject, various paradigms
- simultaneous physio data, eyetracking, auxiliary datasets
- versatile movie stimulus descriptions (every spoken word (grammar, semantics); music played; emotions; body contact; eye movements, saccade targets, fixations; visible facial features; semantic conflict, space/time discontinuities)
INTERIM CONCLUSION AFTER FOUR YEARS

Was it worth being open? ABSOLUTELY!

- 16 additional, independent, published studies use these data (virtually all of them would not have been attempted by our lab)
- not a single "scoop"
- substantial boost in return-of-investment for the tax payer
- inspired similar work by others
INTERIM CONCLUSION AFTER FOUR YEARS

Was it worth being open? ABSOLUTELY!

- 16 additional, independent, published studies use these data (virtually all of them would not have been attempted by our lab)
- not a single "scoop"
- substantial boost in return-of-investment for the tax payer
- inspired similar work by others

Did we make the most out of it? ABSOLUTELY NOT!

- dozens of promises to contribute original data, none happened, yet
- starting point for users today is practically identical to 4 years ago
WHY IS THE OPEN-SCIENCE MAGIC SO WEAK?

Keep the faith! The first real contributions are happening right now.
LESSONS FROM OPEN-SCIENCE
Don't be special whenever possible, or risk being too expensive to work with.

Reporting standards
http://www.humanbrainmapping.org/cobidas

Standard data structures
The Brain Imaging Data Structure: a protocol for standardizing and describing outputs of neuroimaging experiments. Scientific Data.
http://bids.neuroimaging.io

Code review/release necessity
MAKE YOUR SCIENTIFIC OUTPUT...

F indable
A ccessible
I nteroperable
R eusable

https://www.go-fair.org/fair-principles
FAIR PRINCIPLES

F1  (Meta)data are assigned a globally unique and persistent identifier
F2  Data are described with rich metadata
F3  Metadata clearly and explicitly include the identifier of the data they describe
F4  (Meta)data are registered or indexed in a searchable resource

A1  (Meta)data are retrievable by their identifier using a standardised ... protocol
A1.1 The protocol is open, free, and universally implementable
A1.2 The protocol allows for an authentication and authorisation procedure
A2  Metadata are accessible, even when the data are no longer available

I1  (Meta)data use a formal, accessible ... language for knowledge representation.
I2  (Meta)data use vocabularies that follow FAIR principles
I3  (Meta)data include qualified references to other (meta)data

R1  Meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1 (Meta)data are released with a clear and accessible data usage license
R1.2 (Meta)data are associated with detailed provenance
R1.3 (Meta)data meet domain-relevant community standards

https://www.go-fair.org/fair-principles
AN OPEN-SCIENCE PROJECT IS NEVER REALLY FINISHED

The utility of your contribution declines in the absence of continued investment.
FAIR today is not FAIR forever.

- what worked yesterday will eventually need updating to remain useful (especially analysis code)
- data can be "broken" too!
- sticking to "old" standards will ultimately make you special, and too expensive to work with
DATALAD

A software suite that aids managing the evolution of digital objects (incl. code and data)...

...and *also* yields FAIR resources that can be shared with anyone.
DATALAD PRINCIPLES
• There are only two things in the world: datasets and files.

• A dataset is a Git repository.

• 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.

• Maximize use of existing 3rd-party infrastructure.
INSTALL AN EXISTING DATASET

request via standard URL,
(each dataset has a UUID, and each dataset location another UUID)

$ datalad install http://example.com/ds1
OBTAI N DATASE T CONTENT

request via user-friendly local file path, not internal ID, regardless of remote actual storage solution properties

ds1/ $ datalad get file2
TRACKING "REMOTE" DATA EVOLUTION

ability to track any number of dataset "siblings", in Git or non-Git data stores

```
ds1/ $ datalad update
```
apply changes from default or selected sibling while maintaining local data availability status

ds1/ $ datalad update --merge --reobtain-data
DATASET LINKAGE

$ datalad install --dataset . --source http://example.com/ds inputs/rawdata

$ git diff HEAD~1
diff --git a/.gitmodules b/.gitmodules
new file mode 100644
index 0000000..c3370ba
--- /dev/null
+++ b/.gitmodules
@@ -0,0 +1,3 @@
+ [submodule "inputs/rawdata"]
+ path = inputs/rawdata
+ url = http://example.com/importantds
diff --git a/inputs/rawdata b/inputs/rawdata
new file mode 160000
index 0000000..fabf852
--- /dev/null
+++ b/inputs/rawdata
@@ -0,0 +1 @@
+Subproject commit fabf8521130a13986bd6493cb33a70e580ce8572
"actionable" links to subdatasets/files, seamless handling of dataset trees, each dataset can be individually management by a different curator
"COMPLETE" PROVENANCE CAPTURE

- for any local command

```
$ datalad run -m "Perform eye movement event detection"
   --input 'inputs/raw_eyegaze/sub-*/beh/sub-*/...tsv.gz'
   --output 'sub-*/'
   bash code/compute_all.sh
```

- for any containerized app (can be tracked in the dataset too)

```
$ datalad containers-run -n nilearn
   --input 'inputs/mri_aligned/sub-*/in_bold3Tp2/sub-*_task-avmovie_run-*_bold*'
   --output 'sub-*/LC_timeseries_run-*_run-*.csv'
   "bash -c 'for sub in sub-*/; do for run in run-1 ... run-8; do python3 code/extract_lc_timeseries.py \$sub \$run; done; done'"
```

Complete capture of any input data, computational environment, code, parameters, and outputs possible — without sacrificing modularity

Enables enigma-style computing — analyze data that you don't have!
DataLad can serve as a transport layer for arbitrary metadata
Metadata plurality: no need to decide on a single standard
JSON-LD format (for true semantic graphs, or simple dumps)
Concept:
  - Metadata are automatically (and repeatedly) extracted from source
  - Dataset authors/curators decide on extractor selection
  - Metadata can be aggregated into super-datasets
  - (Super)datasets can be queried for all available metadata of any content, regardless of that content being locally available or not
• Easily extensible with additional metadata standard support
• Build metadata-driven apps, e.g. bids2scidata for metadata submission to Scientific Data

http://docs.datalad.org/en/latest/metadata.html#internal-metadata-representation
METADATA-BASED SEARCH FOR INDIVIDUAL FILES
across datasets, without a DB (server)

```bash
$ datalad \
  -c datalad.search.index-egrep-documenttype=files \
  -f json_pp \
  search \
    bids.subject.sex:female \n    bids.type:t1 \n    bids.subject.age:24
{
  "dsid": "4842e188-7df5-11e6-8e6b-002590f97d84",
  "metadata": {
    "@context": {...},
    "bids": {...},
    "datalad_core": {
      "url": [
        "http://openneuro.s3.amazonaws.com/ds000008/ds000008_R1.1.0/...MZ92g",
        "http://openneuro.s3.amazonaws.com/ds000008/ds000008_R2.0.0/..._flBz"
      ],
    },
    "nifti1": {...},
    "parentds": "/tmp/mega/openfmri/ds000008",
    "path": "/tmp/mega/openfmri/ds000008/sub-15/anat/sub-15_T1w.nii.gz",
    "query_matched": {
      "bids.subject.age(years)": "24",
      "bids.subject.sex": "female",
      "bids.type": "T1"
    },
    "refcommit": "b18692ef1beefd88055bc0578b7567a8f4f6f8f9",
    "type": "file"
  }
...```

alternative output formats: JSON stream, custom, ...
Supports a variety of **consumer storage solutions** (SSH-servers, GIN, DropBox, Box.com, Google, WEBDAV, bittorrent, IPFS, ...) via Git-annex

- Built-in support for strong **data encryption**
- Per-target configuration of accepted content, with **configurable permissions and authorization** mechanisms
- Export of dataset to FigShare and similar storage solutions
- Multiple redundant synchronized publication targets are supported (seemingly "publish 2TB on GitHub")

**Datasets**
- are lightweight (typically <<10MB, even when tracking TBs)
- can be attached to a traditional paper to enable direct access to original data, analysis code, computational environments and results
- have machine-readable metadata attached
- support redundant storage
- insure utility against failure of career, institutions, publishers
EXTEND DATALAD

- Separate Python packages, anyone can develop their own
  https://github.com/datalad/datalad-extension-template

- Means for tailored solutions with narrower scope or specific audiences

- Extensions can provide **additional commands, procedures, metadata extractors, webapps**

- Available extensions
  - **containers**: support for containerized computational environments
  - **crawler**: track web resources in automated data distributions
  - **neuroimaging**: neuroimaging research data and workflow
  - **hirni**: imaging raw data management/entry, automatic BIDS-conversion
  - **htcondor**: cluster/cloud/grid-based remote code execution
  - **webapp**: REST API for querying/manipulating datasets

http://docs.datalad.org/en/latest/customization.html#extension-packages
MODULAR DECENTRALIZED MANAGEMENT OF RESEARCH COLLABORATIONS

Consume, create, curate, analyze, publish, and query data with full provenance capture and "universal" metadata support.

Early adopters: Canadian Open Neuroscience Platform (McGill), OpenNeuro (Stanford)

DataLad is free and open source (MIT-licensed).
OMG! I AM TOO OLD FOR THIS...

All this is possible, but not necessary!

• You just need to use two commands:

```bash
# 1. start something
myproject/ % datalad rev-create
# 2. do something
# ...
# 3. save state
myproject/ % datalad rev-save
# go to (2)
```

• The result is a dataset that
  ▪ captures the full history of a project (all data, all code, all changes ever done to them)
  ▪ is compatible with everything that was shown previously in this talk
ACKNOWLEDGEMENTS

- Yaroslav Halchenko
- Joey Hess (git-annex)
- Benjamin Poldrack
- Kyle Meyer
- 20+ additional contributors

Website + Demos: http://datalad.org
Development: http://github.com/datalad
Chat: https://matrix.to/#/#datalad:matrix.org
Open data: http://datasets.datalad.org