Open data, open code.

Andrew.Stewart@manchester.ac.uk

@ajstewart_lang

https://github.com/ajstewartlang

Software Sustainability Institute

BETTER SOFTWARE BETTER RESEARCH

www.software.ac.uk

If you'd like to sign up to our open research working group email list go to here: https://listserv.manchester.ac.uk/cgi-bin/wa?SUBED1=open_research&A=1

Feel free to follow me in Twitter for Tweets and re-Tweets related to open research, reproducibility, and R! Twitter handle: @ajstewart_lang

On February 26th, we will have a Reproducibility half day 1400-1800 where the keynote talk will be delivered by Dorothy Bishop at 1700. Please do come along to that!

The world of research is rapidly changing...

[redacted]

#brainhackschool instructors be like: familiar with docker?
Me: 😬
Familiar with jupyter?
Me: 😬
Familiar with github?
Me: 😬
Familiar with binder?
Me: 😬
Familiar with python?
Me: 😬

It feels like I've spent these past 4 years of PhD on mars
😂🤔

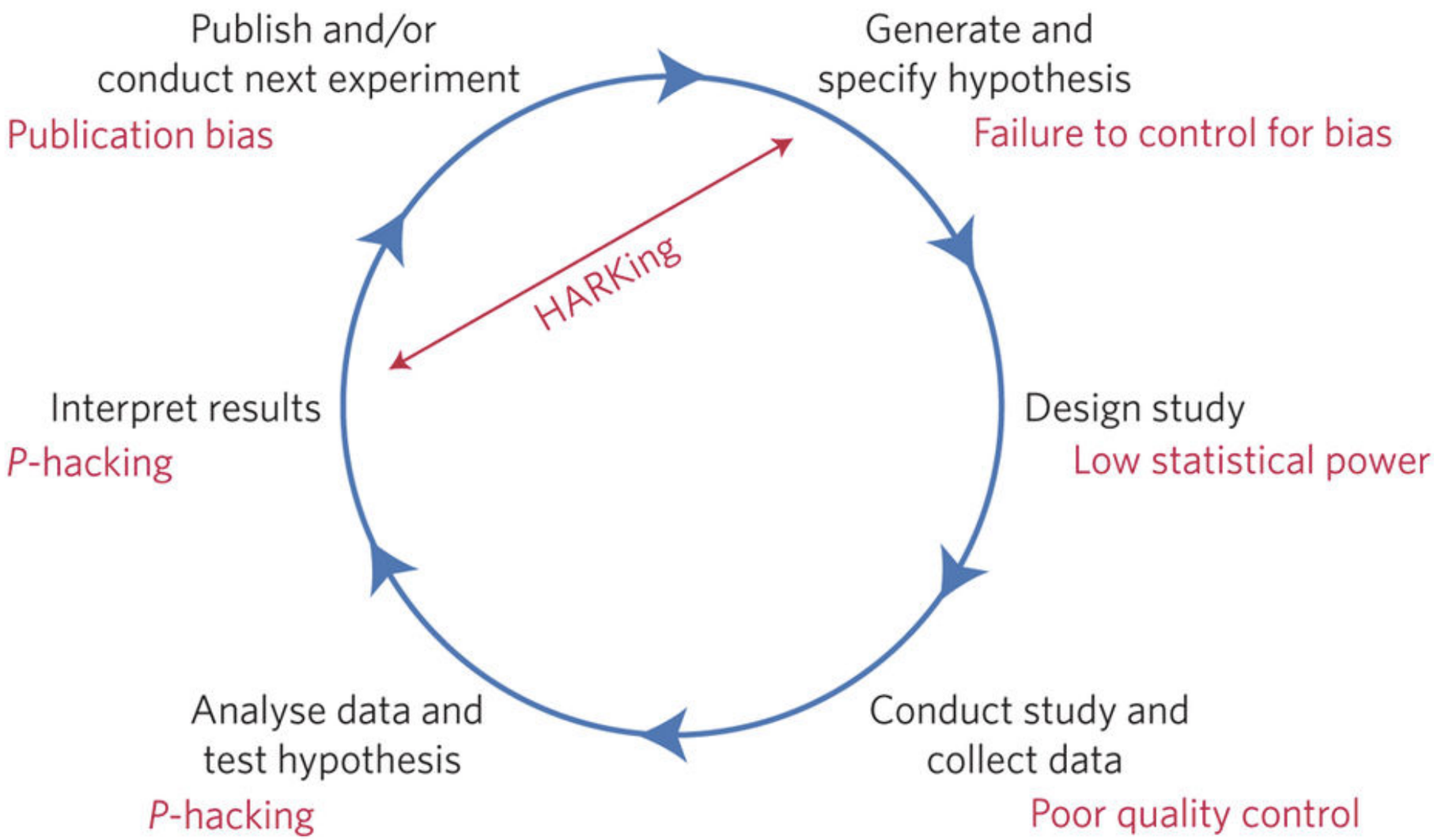
7:05 PM · Aug 7, 2019 · Twitter for iPhone

Replication and Reproducibility in Science

- Ionnidis (2005), *PLOS Medicine*, most published research findings are false.
- Prinz et al. (2011), *Nature Reviews Drug Discovery*, around 65% of cancer biology studies do not replicate.
- Button et al. (2013), *Nature Reviews Neuroscience*, small sample size undermines the reliability of neuroscience.
- MacLeod et al. (2014), *Lancet*, 85% of biomedical research resources are wasted.
- Baker (2015), *Nature*, 90% of scientists recognise a ‘reproducibility crisis’.
- Nosek & Errington (2017), *eLife*, out of first 5 replication attempts of preclinical cancer biology work, only 2 have replicated.
- Eisner (2018), *Journal of Molecular and Cellular Cardiology*. Reproducibility of science: Fraud, impact factors and carelessness.

Problems include *p*-hacking, lack of power, HARKing, failing (refusal) to share data and code, too many researcher degrees of freedom…

From: A manifesto for reproducible science



Munafo et al. (2017), *Nature Human Behaviour*

Open Science recently recognised by G7 Science Ministers…

Focus: Incentives and the researcher ecosystem

Ambition: Foster a research environment in which career advancement takes into account Open Science activities, through incentives and rewards for researchers, and valuing the skills and capabilities in the Open Science workforce.

Recommendations:

At national levels: G7 nations should each engage with research stakeholders to identify and implement enhancements to research evaluation and reward systems that take into consideration the Open Science activities carried out by researchers and research institutions. Topics that could be discussed include:

- Recognizing Open Science practices during evaluation of research funding proposals, and research outcomes;
- Recognizing and rewarding research productivity and impact that reflect open science activities by researchers during career advancement reviews;
- Including credit for service activities such as reviewing, evaluating, and curation and management of research data; and,
- Developing metrics of Open Science practices.

REF 2019/02
January 2019

Panel criteria and working methods

200. The sub-panels welcome research practice that supports reproducible science and the application of best practice. Examples include registered reports, pre-registration, publication of data sets, experimental materials, analytic code, and use of reporting checklists for publication purposes and those relating to the use of animals in research. These contribute to the evaluation of rigour for submitted outputs. Replication studies may be submitted as outputs and will be evaluated on the extent to which they contribute significant new knowledge, improved methods, or advance theory or practice[1].

346...

Within the context of the institution's strategy, how the submitting unit is progressing towards an open research environment, including where this goes above and beyond the REF open access policy requirements, and wider activity to encourage the effective sharing and management of research data, as appropriate to the discipline. Consideration of reproducibility should also be included where relevant to the discipline.

...is forming part of Universities' teaching manifestos...

Teaching with Open Science commitment:
To teach the practices and skills of open research and science in our undergraduate and postgraduate degree programmes

a. Promote open science in our teaching.
b. Design a Research Methods curriculum that teaches skills for open science and uses open science to enhance teaching (for example: teach R and use open data to practice analysis skills).
c. Learn about and adopt open educational practices in our teaching.
d. Produce and promote tools for helping student researchers adopt open practices, including training and guidance suitable to their level of study.
e. Author, share and use open educational resources to promote teaching with open science beyond our School and Institution.
f. Support our colleagues to learn the skills of teaching Open Science.

...and is required by many funders.

Wellcome signs open data concordat | Wellcome

https://www.ukri.org/files/legacy/dc

wellcome

Funding Key issues How we work About us News

All news and views Media office

News | 28 July 2016

Wellcome signs open data concordat

Wellcome has signed a concordat to ensure that research data gathered and generated by members of the UK research community is made openly available wherever possible.

HEFCE, Research Councils UK and Universities UK are also signatories.

The Concc
through cc
for sharing

- importa
- importa publica
- rights o
- expecta

Data sharing
Influencing policy
Open access

European Commission

RESEARCH & INNOVATION

Participant Portal H2020 Online Manual

Search

H2020 Online Manual
My Area - User account & roles
EU Login
Roles & access rights
Terms and Conditions of Use
Grants
Applying for funding
Find a call
Horizon 2020 structure and budget
What you need to know about Horizon 2020 calls

> H2020 Online Manual > Cross-cutting issues > Open access & Data management >

Open access | Data management

Open access & Data management

These pages guide you through

- the context and Horizon 2020 rules on **open access to scientific publications**, which is an *obligation*, and
- **open access to research data**, where opt-outs are possible, and **research data management**

Detailed guidance for both aspects is available through the buttons on top of the page and the pdf reference documents bellow.

Concordat on Open Research Data - Nine Principles

- Open access to research data is an enabler of high quality research, a facilitator of innovation and safeguards good research practice.

- There are sound reasons why the openness of research data may need to be restricted but any restrictions must be justified and justifiable.

- Open access to research data carries a significant cost, which should be respected by all parties.

- The right of the creators of research data to reasonable first use is recognised.

https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/

- Use of others' data should always conform to legal, ethical and regulatory frameworks including appropriate acknowledgement.

- Good data management is fundamental to all stages of the research process and should be established at the outset.

- Data curation is vital to make data useful for others and for long-term preservation of data.

- Data supporting publications should be accessible by the publication date and should be in a citeable form.

- Support for the development of appropriate data skills is recognised as a responsibility for all stakeholders.

https://www.ukri.org/files/legacy/documents/concordatonopenresearchdata-pdf/

TOP GUIDELINES
TRANSPARENCY AND OPENNESS PROMOTION

Transparency, open sharing, and reproducibility are core values of science, but not always part of daily practice. Journals, funders, and societies can increase research reproducibility by adopting the TOP Guidelines.

8 MODULAR STANDARDS

CITATION STANDARDS Cite shared data to incentivize their publication	**DATA TRANSPARENCY** Disclose, require, or verify shared data
ANALYTICAL METHODS TRANSPARENCY Disclose, require, or verify shared code	**RESEARCH MATERIALS TRANSPARENCY** Disclose, require, or verify shared materials
DESIGN AND ANALYSIS TRANSPARENCY Sets standards for research design disclosures	**PREREGISTRATION OF STUDIES** Specification of study details before data collection
PREREGISTRATION OF ANALYSIS PLANS Specification of analytical details before data collection	**REPLICATION** Encourages publication of replication studies

ACROSS 3 TIERS

DISCLOSURE:
The article must disclose whether or not materials are available.

REQUIREMENT:
The article must share materials when possible.

VERIFICATION:
Third party must verify that the standard is being met.

HOW TOP IS IMPLEMENTED

TOP Statements are standardized tools for disclosing research outputs such as datasets.

Open Science Badges signal transparent research.

Registered Reports protect research against biased analysis and publication.

OVER 5,000 JOURNAL SIGNATORIES

LEARN MORE AT COS.IO/TOP

The Center for Open Science is a non-profit organization with the mission of improving openness, integrity, and reproducibility in scientific research.

COS: cos.io | **OSF:** osf.io | **Email:** contact@cos.io | **CenterForOpenScience** | **@OSFramework**

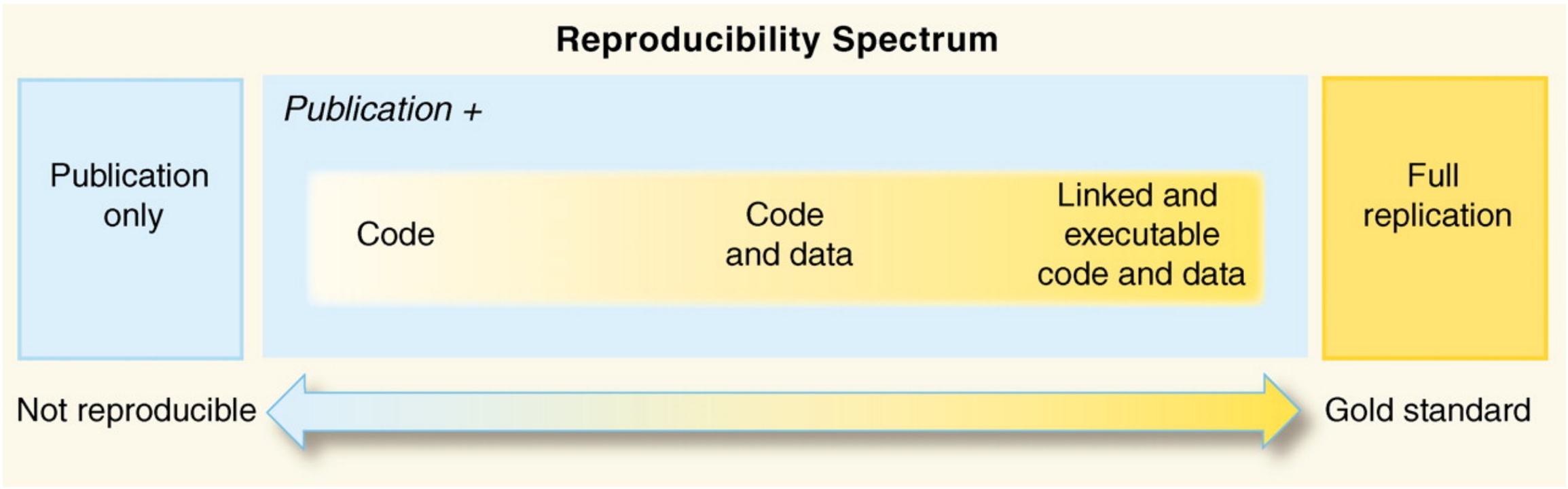
PERSPECTIVE

Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

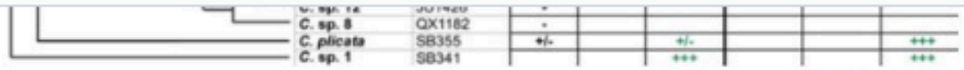
Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847



You can't easily do reproducible research in a GUI

Lukas Schlögl @LukasSchlogl · Sep 7

Just priceless. An estimated 20% of **genetic** research papers contain errors because **Excel** converted some gene names into calendar **dates**.



PLOS ONE PHYLOGENY/FLICKR (CC BY 2.0)

One in five genetics papers contains errors thanks to Microsoft Excel

By **Jessica Boddy** | Aug. 29, 2016 , 1:45 PM

Andrew Whitby @EconAndrew · Sep 7

This is top shelf trolling, because thanks to Excel "1 in 5" genetics papers contain errors in gene names. sciencemag.org/news/2016/08/o … twitter.com/msexcel/status…

Show this thread

186 4.6K 7.3K

Reinhart, Rogoff... and Herndon: The student who caught out the profs

By Ruth Alexander
BBC News

20 April 2013 Share

This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.

It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.

Before data and code sharing...

- You need to make sure you have a structured data management pipeline and a reproducible workflow...
- David Knight - Core Facilities - developing a data management pipeline - existing issues around (e.g.) data robustness, metadata (often too little too late). Important for data to be FAIR (more in a bit...)

Codifying your Workflow

What do you use? Python, Bash, R?

Scripts fine for small tasks but what about the multiple tasks needed in a research project?

What happens when you want to share your workflow with colleagues elsewhere or in another lab (with a different infrastructure)?

"It worked on my machine!" - not good when it doesn't work on your collaborators' machines (or on your new machine!)

Design guidelines for analysis scripts - Marijn van Vliet

1. Each analysis step is one script

2. A script either processes a single recording, or aggregates across recordings, never both

3. One master script to run the entire analysis

4. Save all intermediate results

5. Visualize all intermediate results

6. Each parameter and filename is defined only once

7. Distinguish files that are part of the official pipeline from other scripts

https://arxiv.org/pdf/1904.06163.pdf

PROTOCOLS ARTICLE

Front. Neurosci., 11 September 2018 | https://doi.org/10.3389/fnins.2018.00586

Analysis of Functional Connectivity and Oscillatory Power Using DICS: From Raw MEG Data to Group-Level Statistics in Python

 **Marijn van Vliet**[1*], **Mia Liljeström**[1,2], **Susanna Aro**[1],  **Riitta Salmelin**[1] and **Jan Kujala**[1]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland

[2]NatMEG, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

MENU SCIENTIFIC DATA

Data Descriptor | Open Access | Published: 20 January 2015

A multi-subject, multi-modal human neuroimaging dataset

Daniel G Wakeman & Richard N Henson

Scientific Data **2**, Article number: 150001 (2015) | Download Citation

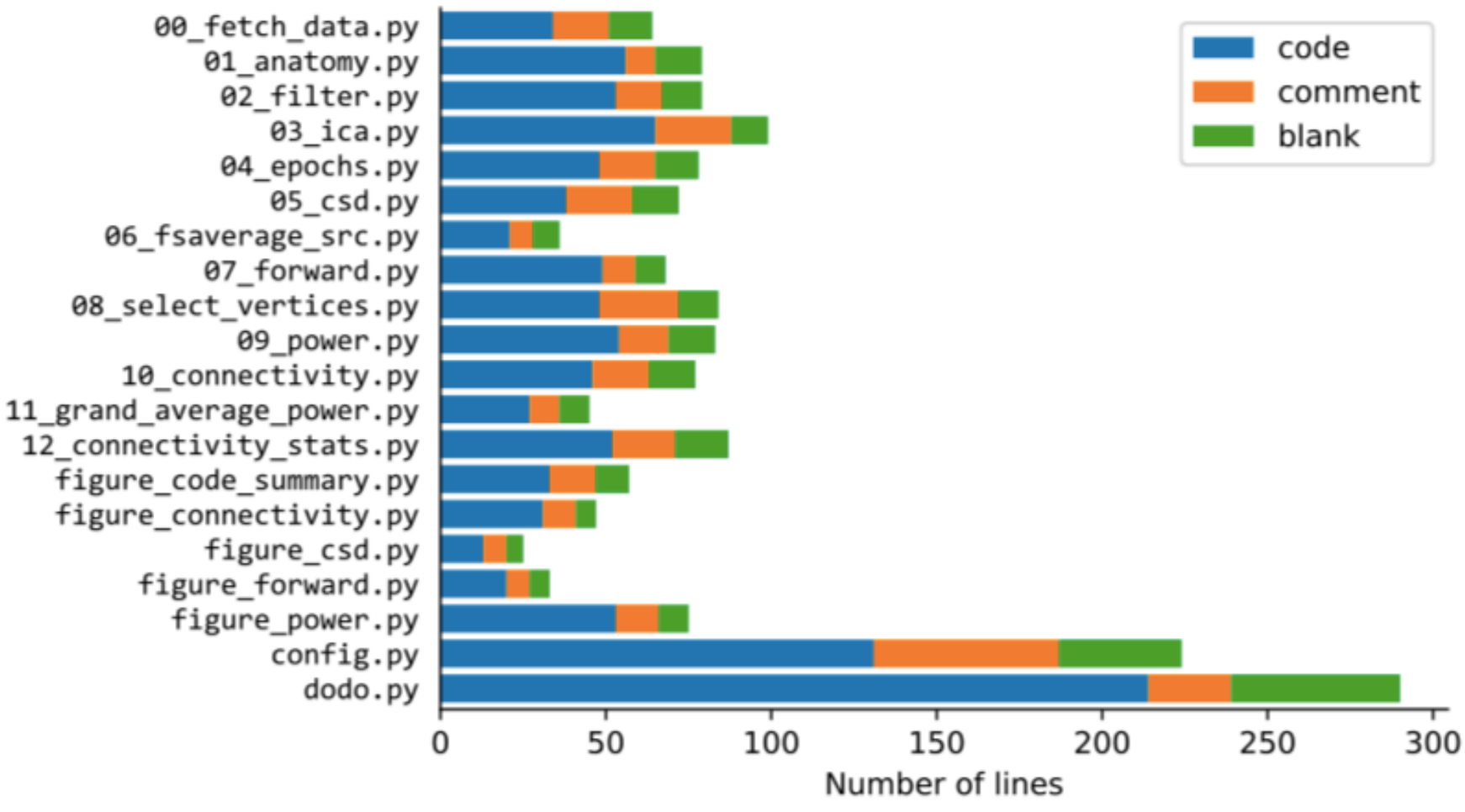


Figure 1: For each script in the analysis pipeline, the number of lines of the file, broken down into lines of programming code (code), lines of descriptive comments (comment) and blank lines (blank). The first 13 scripts perform data analysis steps, the next 5 scripts generate figures, the `config.py` script contains all configuration parameters and the `dodo.py` script is the master script that runs all analysis steps on all recordings.

https://arxiv.org/pdf/1904.06163.pdf

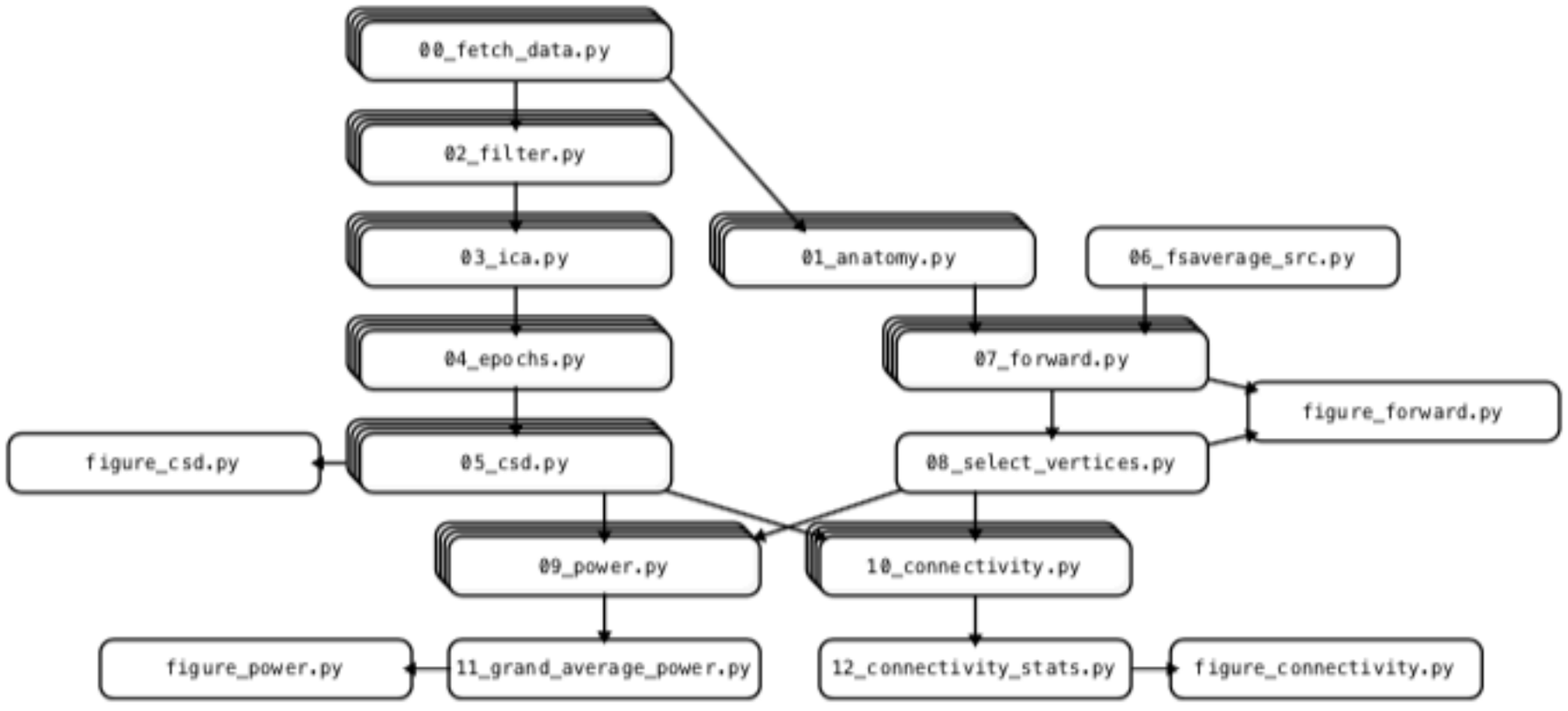


Figure 2: Dependency graph showing how the output of one script is used by another. Stacked boxes indicate scripts that are run for each participant.

https://arxiv.org/pdf/1904.06163.pdf

Data Sharing - FAIR data

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf

Data Sharing

F1000Research 2018, 6:1618 Last updated: 25 JUL 2019

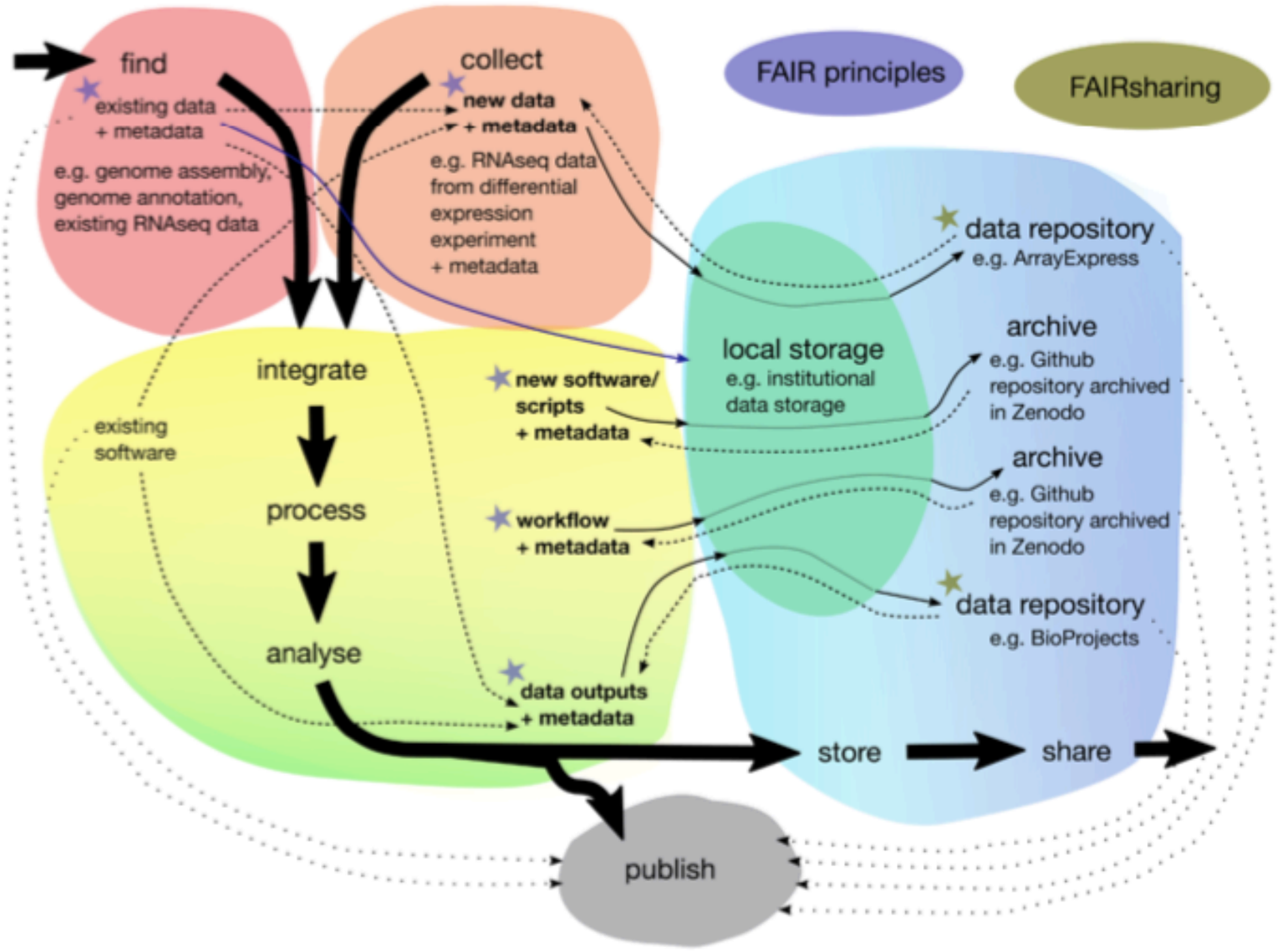


Figure 2. Flowchart of the data life cycle stages applied to an example research project. Bold text indicates new data, software or workflow objects created during the project. Solid thin arrows indicate movement of objects from creation to storage and sharing. Dashed thin arrows indicate where downstream entities should influence decisions made at a given step. (For example, the choice of format, granularity, metadata content and structure of new data collected may be influenced by existing software requirements, existing data characteristics and requirements of the archive where the data will be deposited.) Purple stars indicate objects for which the FAIR principles[9] can provide further guidance. Dotted thin arrows indicate citation of an object using its unique persistent identifier. Brown stars indicate where FAIRsharing can help identify appropriate archives for storing and sharing.

https://f1000research.com/articles/6-1618

Table 1. Overview of some representative databases, registries and other tools to find life science data. A more complete list can be found at FAIRsharing.

Database/ registry	Name	Description	Datatypes	URL
Database	Gene Ontology	Repository of functional roles of gene products, including: proteins, ncRNAs, and complexes.	Functional roles as determined experimentally or through inference. Includes evidence for these roles and links to literature	http://geneontology.org/
Database	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Repository for pathway relationships of molecules, genes and cells, especially molecular networks	Protein, gene, cell, and genome pathway membership data	http://www.genome.jp/kegg/
Database	OrthoDB	Repository for gene ortholog information	Protein sequences and orthologous group annotations for evolutionarily related species groups	http://www.orthodb.org/
Database with analysis layer	eggNOG	Repository for gene ortholog information with functional annotation prediction tool	Protein sequences, orthologous group annotations and phylogenetic trees for evolutionarily related species groups	http://eggnogdb.embl.de/
Database	European Nucleotide Archive (ENA)	Repository for nucleotide sequence information	Raw next-generation sequencing data, genome assembly and annotation data	http://www.ebi.ac.uk/ena
Database	Sequence Read Archive (SRA)	Repository for nucleotide sequence information	Raw high-throughput DNA sequencing and alignment data	https://www.ncbi.nlm.nih.gov/sra/
Database	GenBank	Repository for nucleotide sequence information	Annotated DNA sequences	https://www.ncbi.nlm.nih.gov/genbank/
Database	ArrayExpress	Repository for genomic expression data	RNA-seq, microarray, CHIP-seq, Bisulfite-seq and more (see https://www.ebi.ac.uk/arrayexpress/help/experiment_types.html for full list)	https://www.ebi.ac.uk/arrayexpress/
Database	Gene Expression Omnibus (GEO)	Repository for genetic/genomic expression data	RNA-seq, microarray, real-time PCR data on gene expression	https://www.ncbi.nlm.nih.gov/geo/
Database	PRIDE	Repository for proteomics data	Protein and peptide identifications, post-translational modifications and supporting spectral evidence	https://www.ebi.ac.uk/pride/archive/
Database	Protein Data Bank (PDB)	Repository for protein structure information	3D structures of proteins, nucleic acids and complexes	https://www.wwpdb.org/
Database	MetaboLights	Repository for metabolomics experiments and derived information	Metabolite structures, reference spectra and biological characteristics; raw and processed metabolite profiles	http://www.ebi.ac.uk/metabolights/
Ontology/ database	ChEBI	Ontology and repository for chemical entities	Small molecule structures and chemical properties	https://www.ebi.ac.uk/chebi/
Database	Taxonomy	Repository of taxonomic classification information	Taxonomic classification and nomenclature data for organisms in public NCBI databases	https://www.ncbi.nlm.nih.gov/taxonomy
Database	BioStudies	Repository for descriptions of biological studies, with links to data in other databases and publications	Study descriptions and supplementary files	https://www.ebi.ac.uk/biostudies/

F1000Research 2018, 6:1618 Last updated: 25 JUL 2019

Platforms for hosting data...

PUBLIC DATASETS

Search Datasets

SORT BY: Created | Name | Uploader | Stars | **Downloads** | Subscriptions

UCLA Consortium for Neuropsychiatric Phenomics LA5c Study

695 1061100 16 16

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

FILES: 49721 | SIZE: 5.02GB | SUBJECTS: 272 | SESSION: 1 | AVAILABLE TASKS : bart, rest, scap, stopsignal, taskswitch, bht, pamenc, pamret | AVAILABLE MODALITIES : T1w, dwi, bold

Multisubject, multimodal face processing

595 180485 5 7

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

FILES: 22244 | SIZE: 460.48GB | SUBJECTS: 16 | SESSIONS: 2 | AVAILABLE TASKS : facerecognition | AVAILABLE MODALITIES : meg, T1w, dwi, bold, fieldmap

Flanker task (event-related)

481 5702 1 1

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

FILES: 1664 | SIZE: 1.75GB | SUBJECTS: 26 | SESSION: 1 | AVAILABLE TASKS : Flanker | AVAILABLE MODALITIES : T1w, bold

Classification learning

455 64847 2 2

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

FILES: 2789 | SIZE: 5.4GB | SUBJECTS: 17 | SESSION: 1 | AVAILABLE TASKS : deterministic classification, mixed event-related probe, probabilistic classification | AVAILABLE MODALITIES : /participants, T1w,

Balloon Analog Risk-taking Task

434 13800 2 3

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

FILES: 1162 | SIZE: 2.25GB | SUBJECTS: 16 | SESSION: 1 | AVAILABLE TASKS : balloon analog risk task | AVAILABLE MODALITIES : T1w, inplaneT2, bold

Versions

00001 2018-07-18

00002 2018-07-18

00016 2018-07-18

UCLA Consortium for Neuropsychiatric Phenomics LA5c Study

uploaded by Franklin Feingold on 2018-03-19 - over 1 year ago

last modified on 2018-07-18 - about 1 year ago

authored by Bilder, R, Poldrack, R, Cannon, T, London, E, Freimer, N, Congdon, E, Karlsgodt, K, Sabb, F

517 810652

Download

Files: 49721, Size: 5.02GB, Subjects: 272, Session: 1

Available Tasks : bart, rest, scap, stopsignal, taskswitch, bht, pamenc, pamret

Available Modalities : T1w, dwi, bold

README

UCLA CONSORTIUM FOR NEUROPSYCHIATRIC PHENOMICS LA5C STUDY

Preprocessed data described in

Gorgolewski KJ, Durnez J and Poldrack RA. Preprocessed Consortium for Neuropsychiatric Phenomics dataset. F1000Research 2017, 6:1262

BIDS Validation

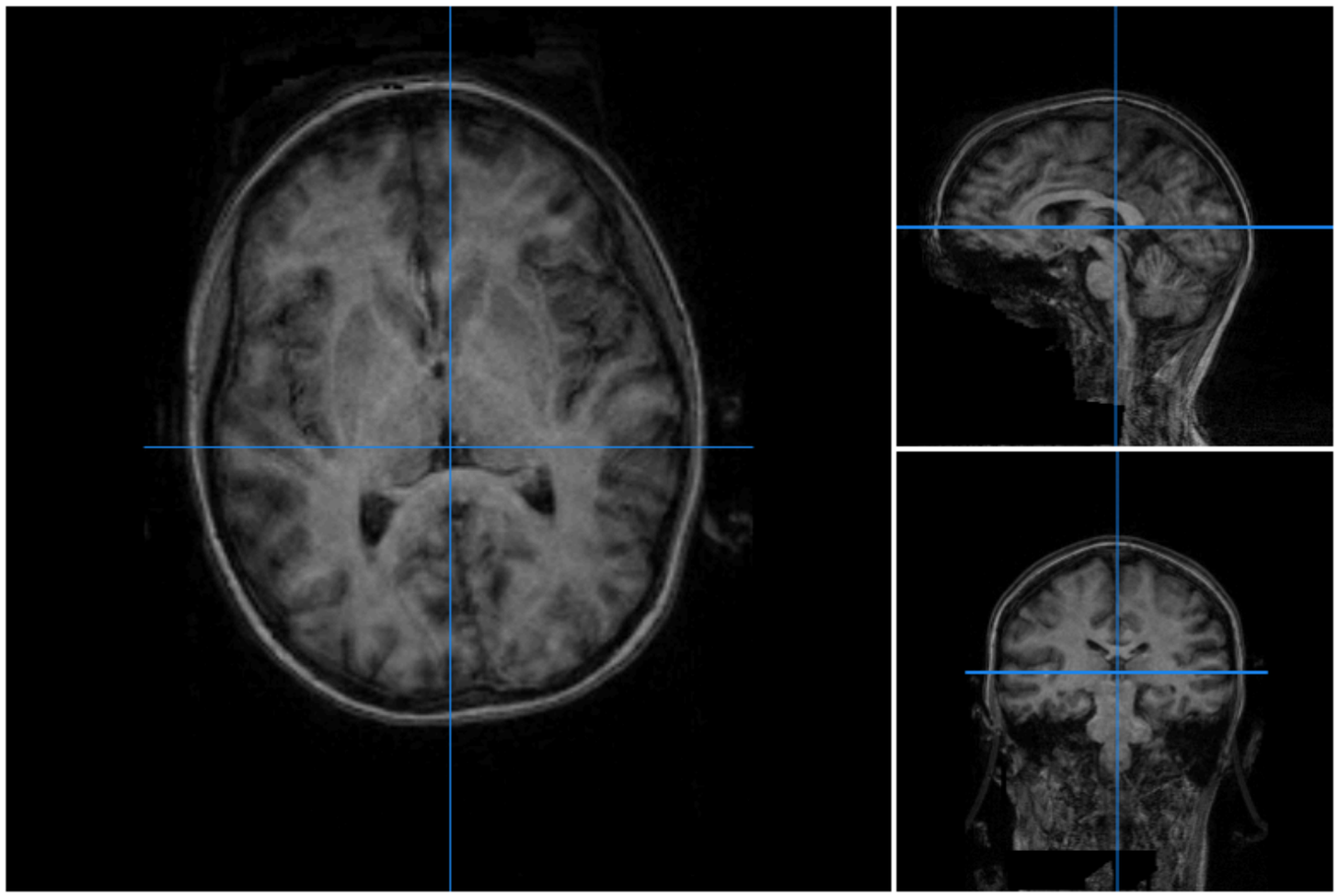
Valid 2 WARNINGS

Dataset File Tree

UCLA Consortium for Neuropsychiatric Phenomics LA5c Study

- CHANGES — DOWNLOAD VIEW
- dataset_description.json — DOWNLOAD VIEW
- participants.tsv — DOWNLOAD VIEW
- README — DOWNLOAD VIEW
- task-bart_bold.json — DOWNLOAD VIEW
- task-bht_bold.json — DOWNLOAD VIEW
- task-pamenc_bold.json — DOWNLOAD VIEW
- task-pamret_bold.json

ds000016 2018-07-18



https://www.repronim.org

ReproNim

DISCOVER DESCRIBE DO TEACH COMMUNITY

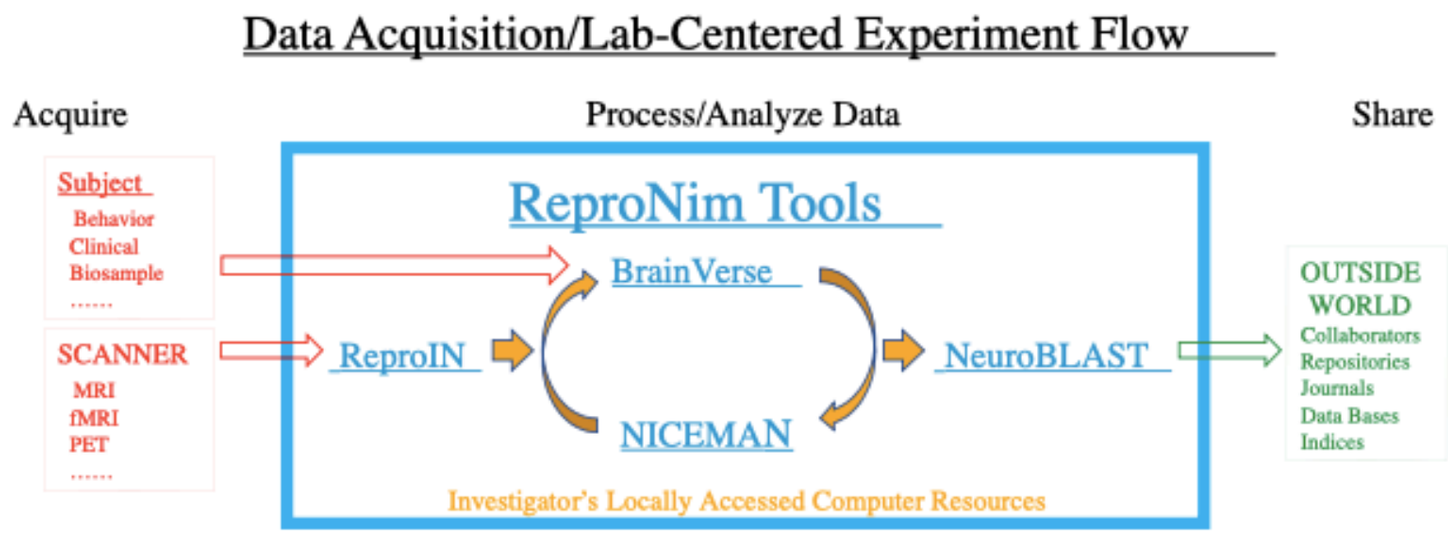
ReproNim: A Center for Reproducible Neuroimaging Computation

Welcome to ReproNim!

The ReproNim vision is to to help neuroimaging researchers to:

- Find and Share data in a FAIR fashion (**discover** resources with **NeuroBLAST**)
- Comprehensively describe their data and analysis workflows in precisely replicable fashion (**describe** research processes with ReproIN and **BrainVerse**)
- Manage their computational resource options (**do** analysis with **NICEMAN**)

so that the outcomes of neuroimaging research are more reproducible.



Including training resources...

ReproNim Introduction

Why do we care about reproducibility? Can we do anything to improve the reproducibility of our neuroimaging work? Let's get motivated to change the world!

Goto module.

Reproducibility Basics

Shells, version control, package managers, and other tools to embrace "Reproducibility By Design"!

Goto module.

FAIR Data

FAIR is a collection of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable. We look at ways to ensure that a researcher's data is properly managed and published in support of reproducible research.

Goto module.

Data Processing

What do we need to know to conduct reproducible analysis? Learn to: Annotate, harmonize, clean, and version data; and Create and maintain reproducible computational environments.

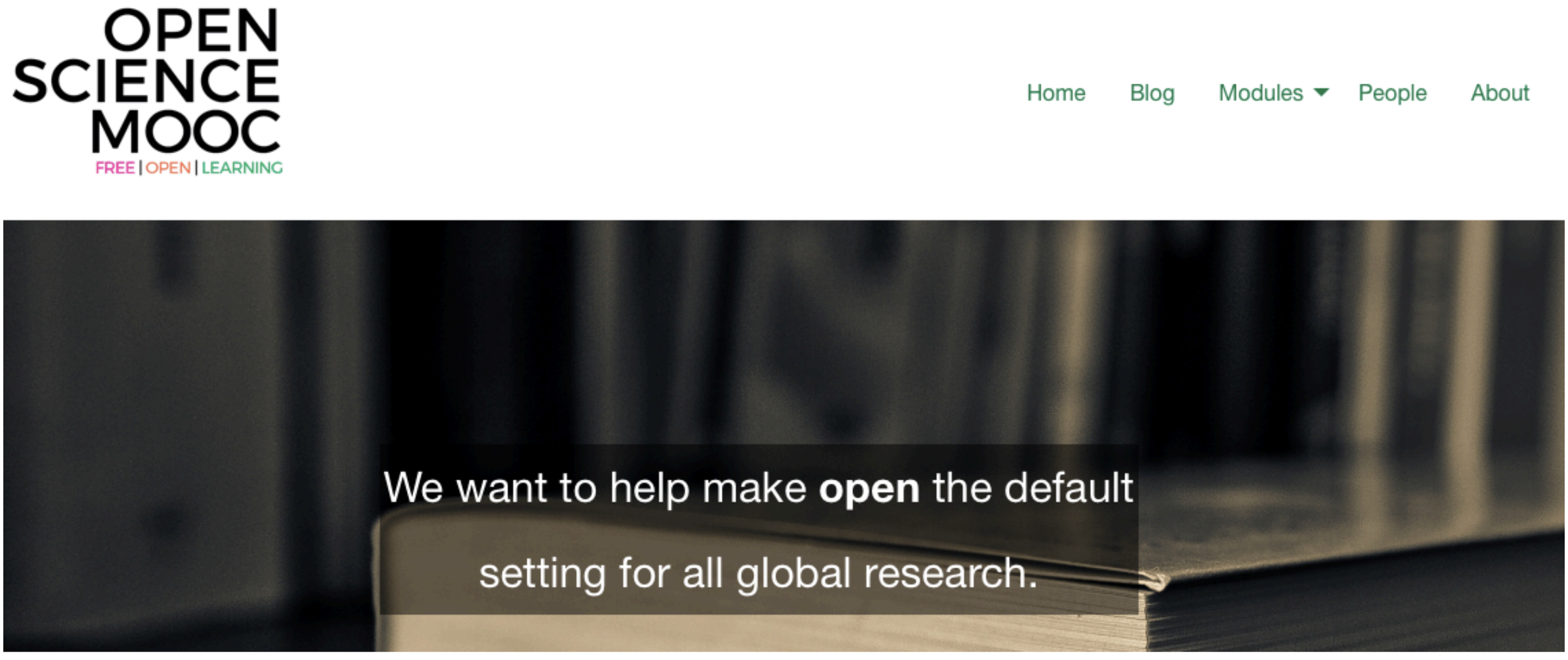
Goto module.

Statistics

Here we describe some key statistical concepts, and how to use them to make your research more reproducible. Everything you ever wanted to know about power, effect size, P-values, sampling and everything else.

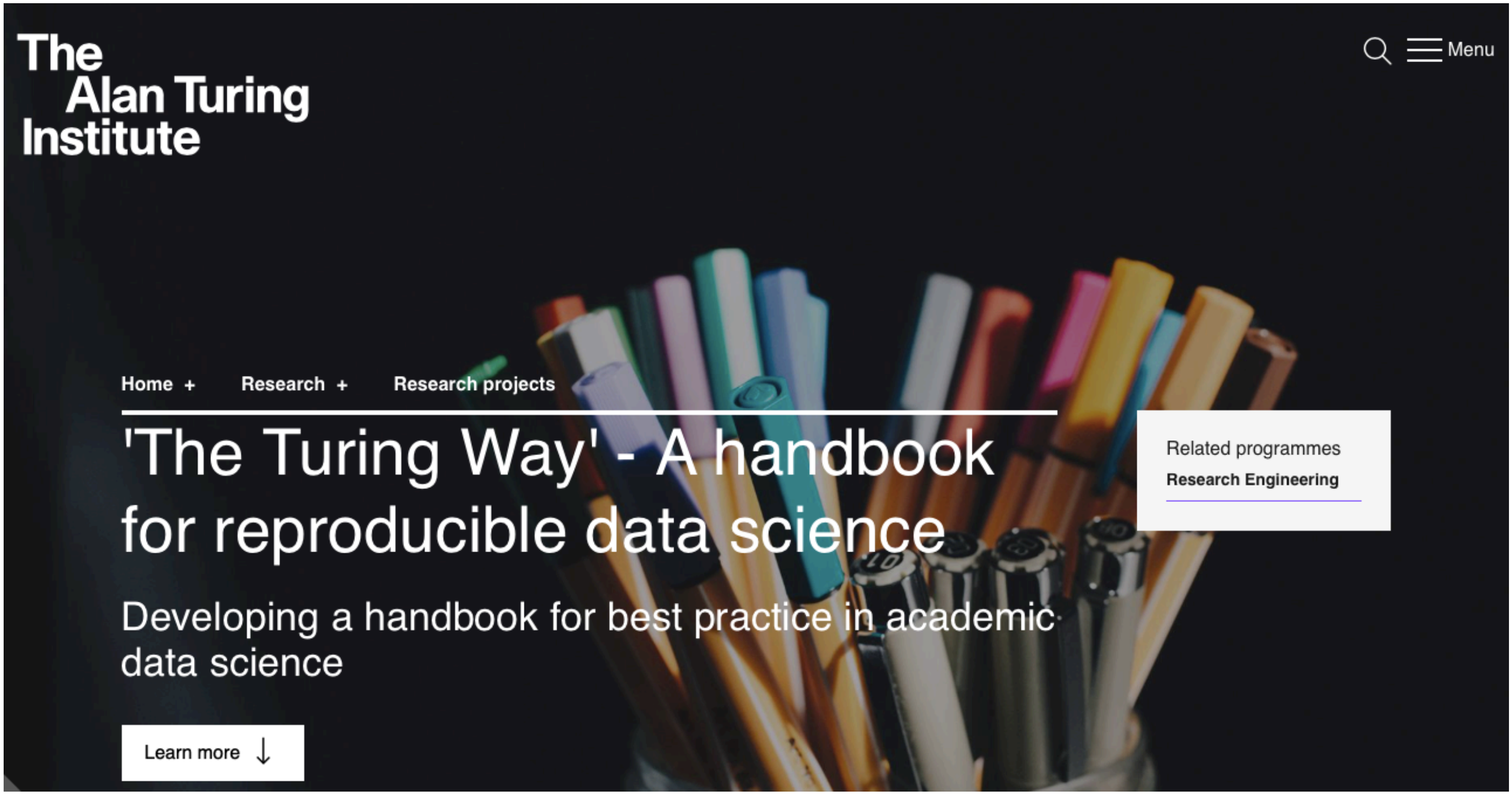
Goto module.

More training resources...



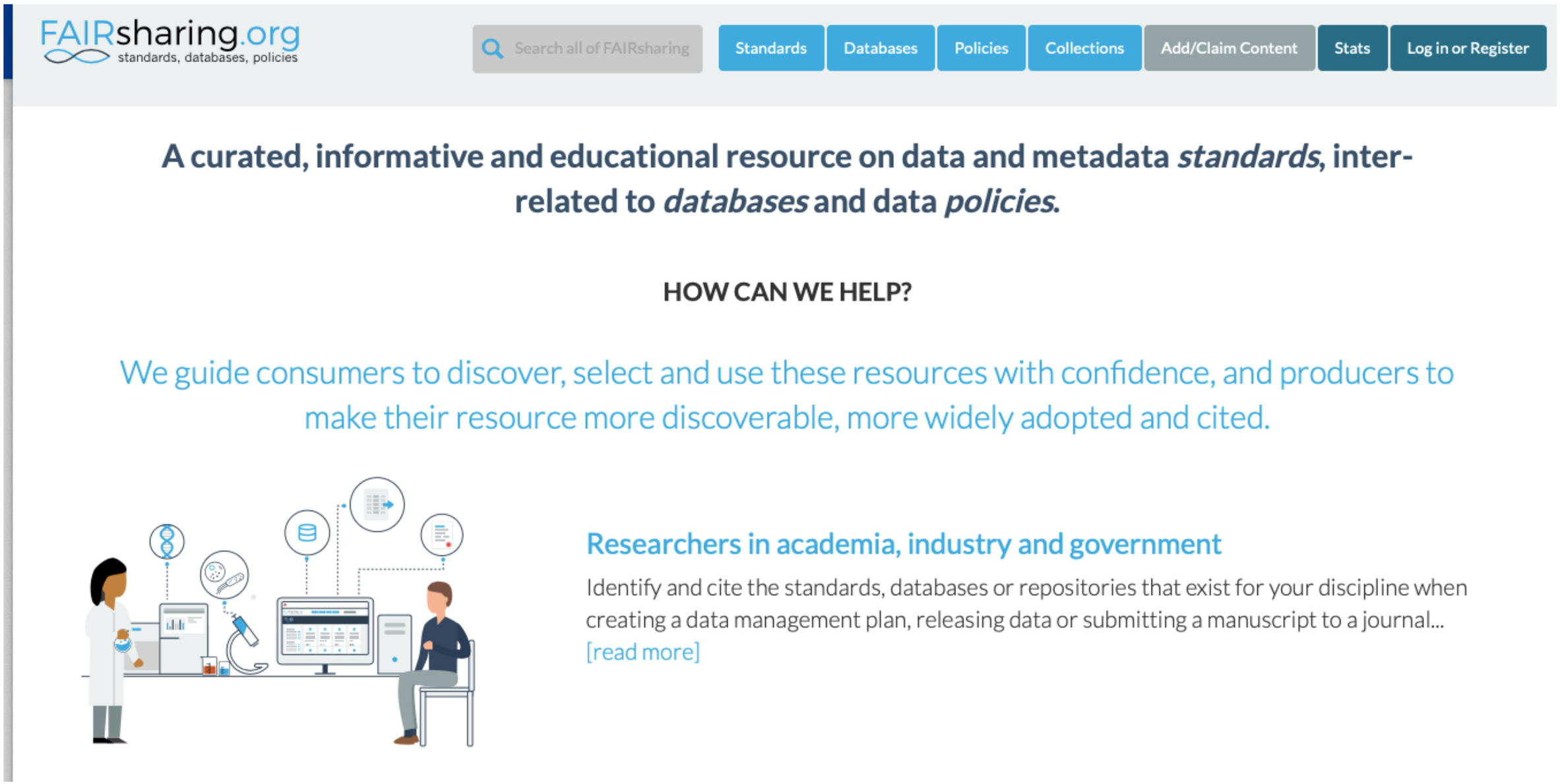
https://opensciencemooc.eu

More training resources...



https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science

More training resources...



https://fairsharing.org