Open Research and Data sharing: why and how do you do it?

Andrew.Stewart@manchester.ac.uk

@ajstewart_lang

https://github.com/ajstewartlang

Software Sustainability Institute

[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
😂🤔

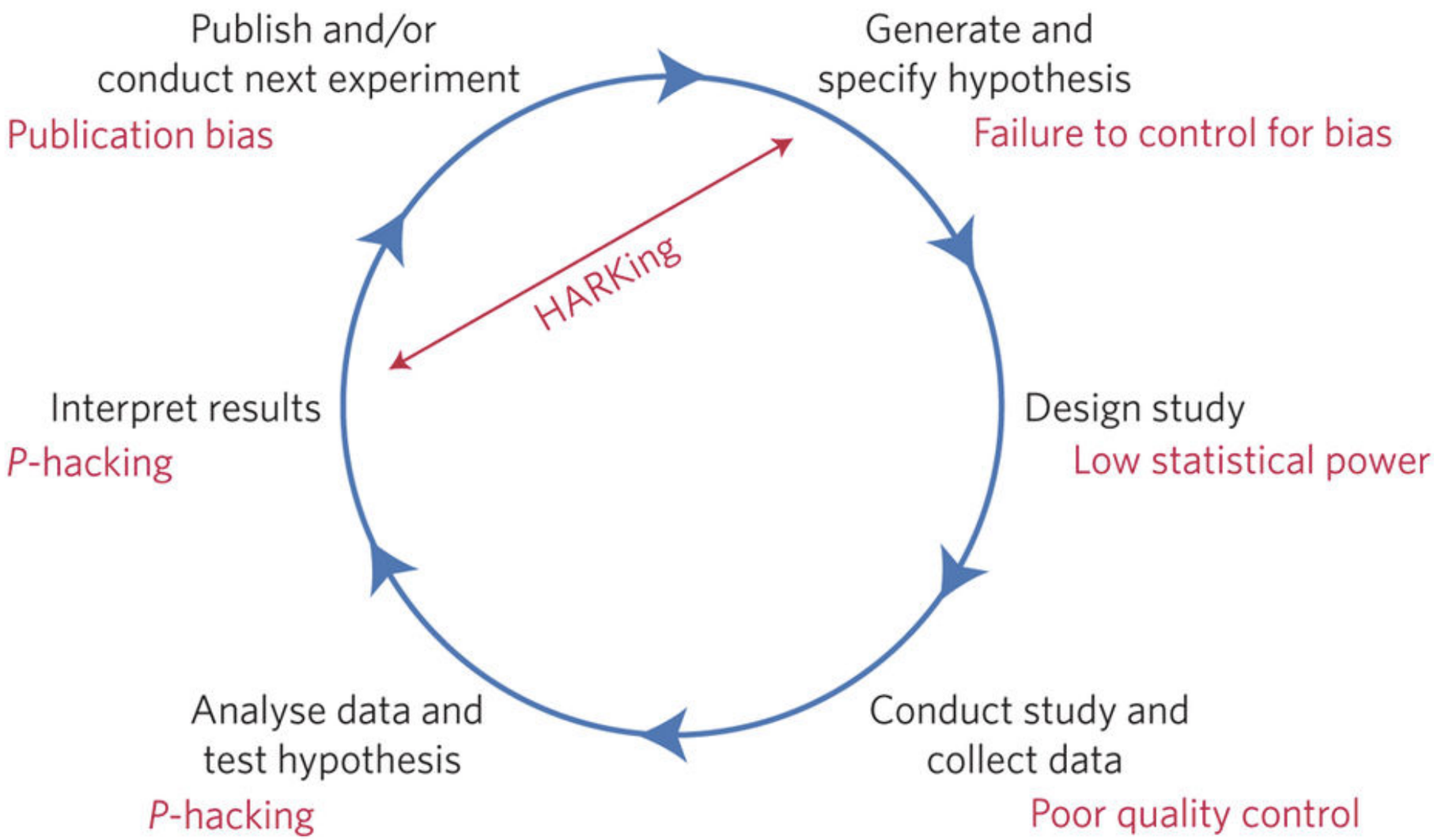
7:05 PM · Aug 7, 2019 · Twitter for iPhone

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.

In REF2021 UoA Environment…

29. The revised template will also include a **section on 'open research'**, detailing the submitting unit's open access strategy, 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. The panels will set out further guidance on this in the panel criteria.

is beginning to appear in tenure-track job adverts…

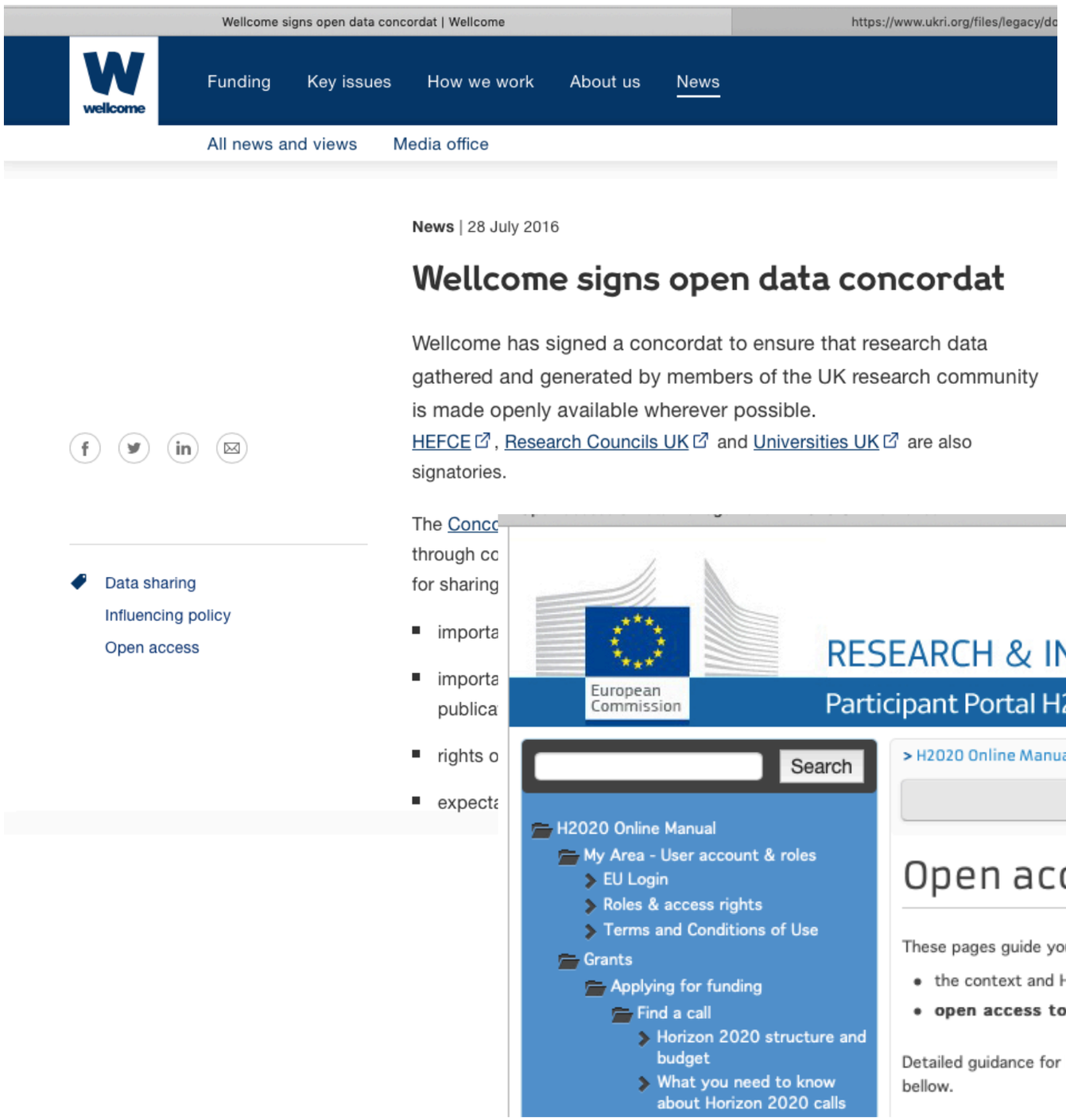
> Our Department embraces the values of open and reproducible science, and candidates are encouraged to address (in their statements and/or cover letter) how they have pursued and/or plan to pursue these goals in their work.

...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 now required by many funders.



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/

The Benefits of Open Research - Some Success Stories

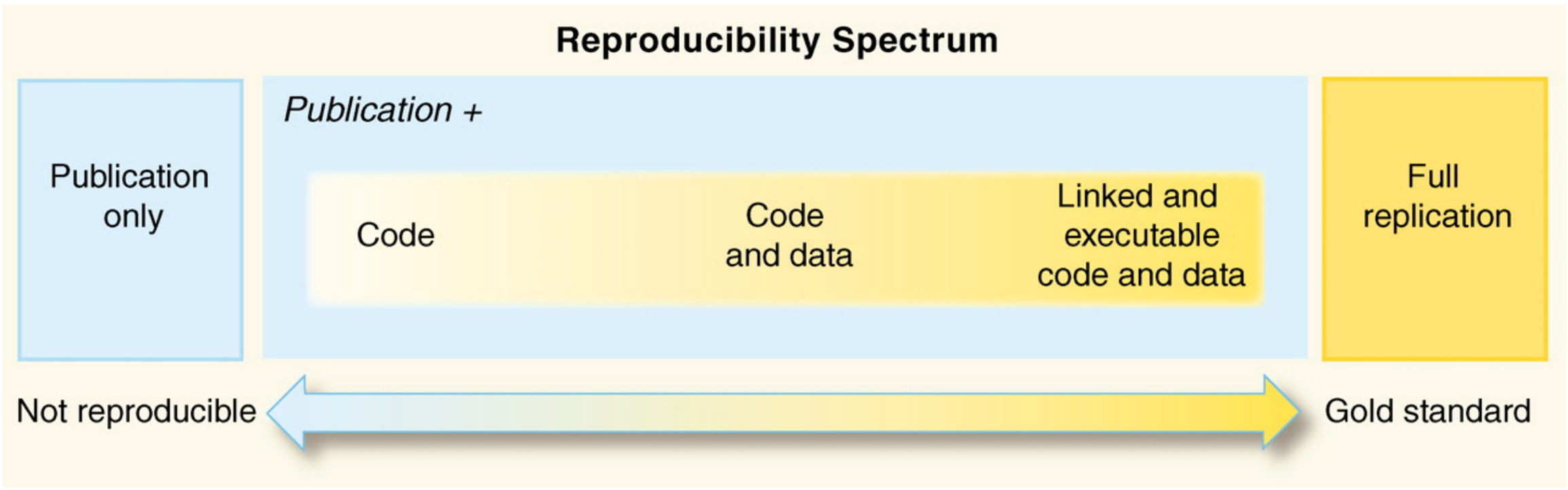
PERSPECTIVE

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 do reproducible research in a GUI*"

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?

"It worked on my 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

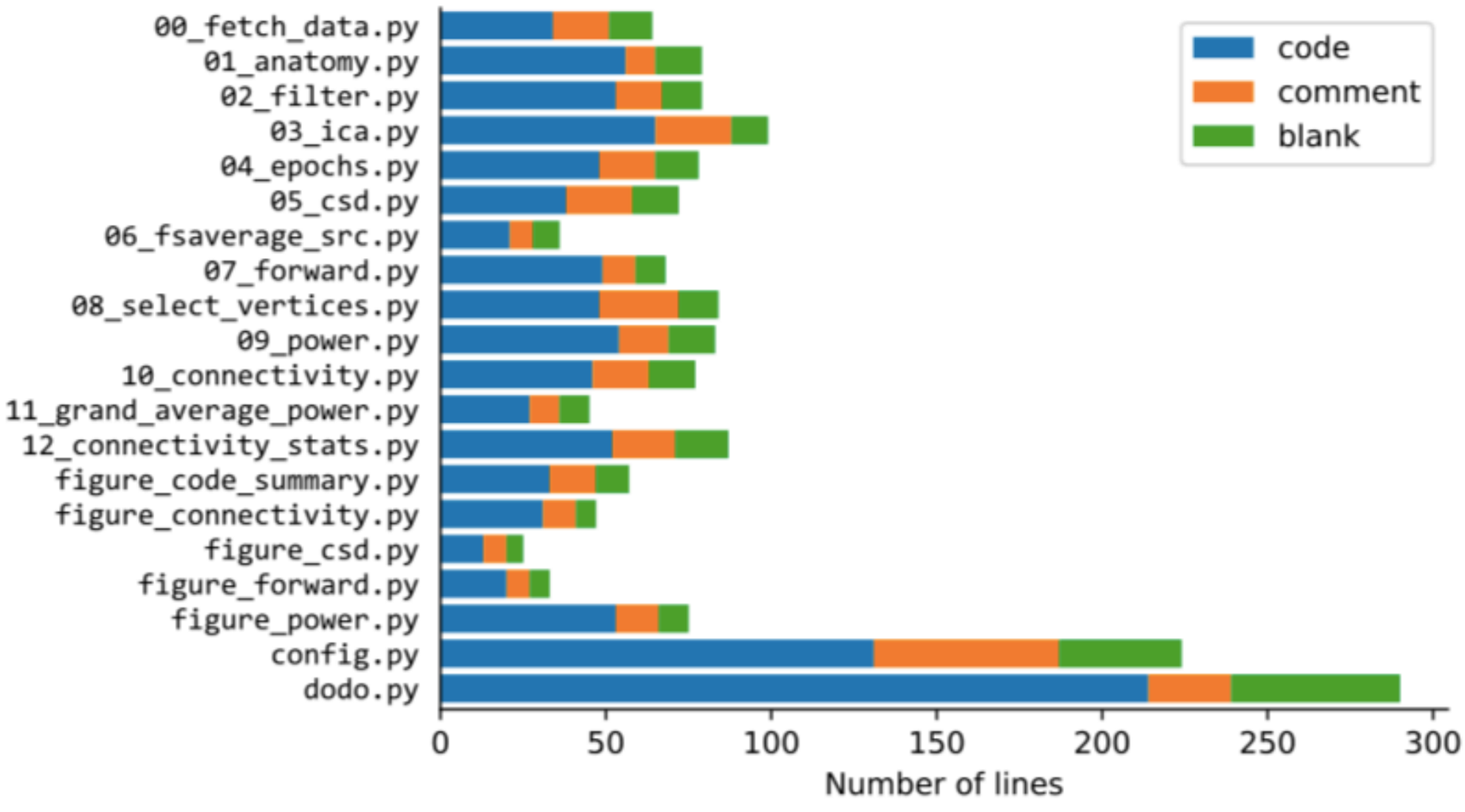


Figure 1: For each script in the analysis pipeline, the number of lines of the file, broken down into lines of programming code (code), lines of descriptive comments (comment) and blank lines (blank). The first 13 scripts perform data analysis steps, the next 5 scripts generate figures, the `config.py` 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

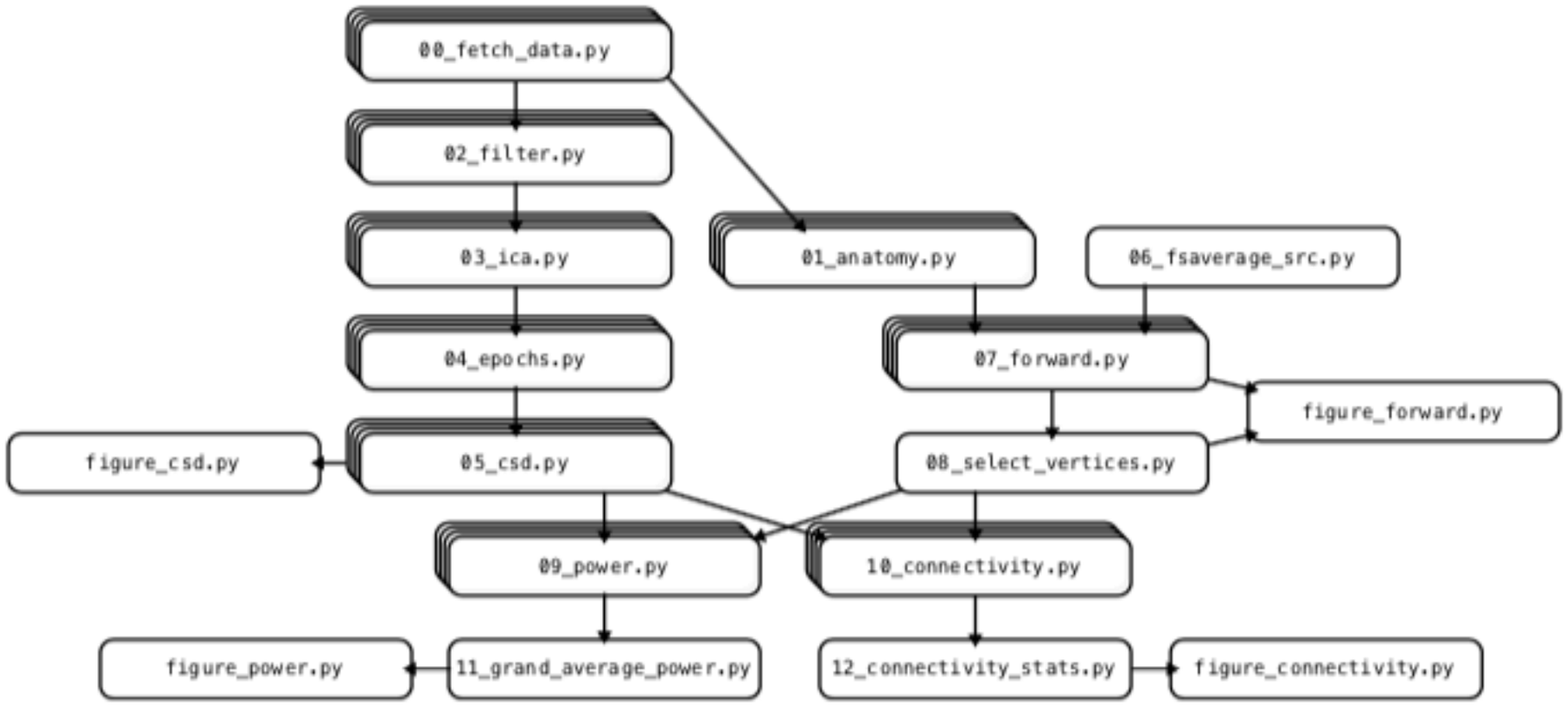
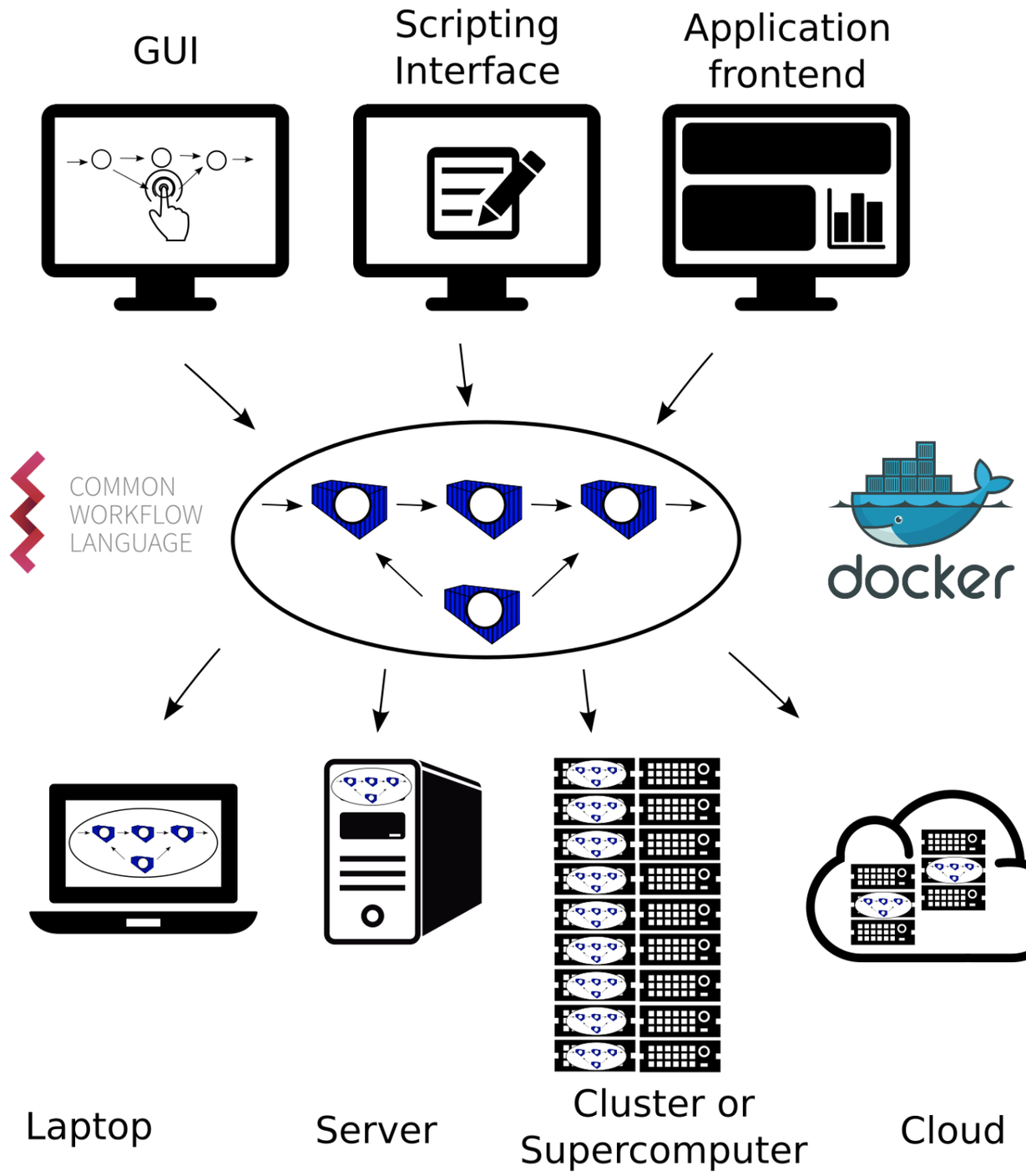


Figure 2: Dependency graph showing how the output of one script is used by another. Stacked boxes indicate scripts that are run for each participant.

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



https://blog.esciencecenter.nl/reproducible-science-the-common-workflow-language-6437889b33b4

Data Sharing - FAIR data

Findable

- The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.

- F1. (Meta)data are assigned a globally unique and persistent identifier

- F2. Data are described with rich metadata (defined by R1 below)

- F3. Metadata clearly and explicitly include the identifier of the data they describe

- F4. (Meta)data are registered or indexed in a searchable resource

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

Accessible

- Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

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

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

Interoperable

- The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- I2. (Meta)data use vocabularies that follow FAIR principles

- I3. (Meta)data include qualified references to other (meta)data

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

Reusable

- The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

- 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

The FAIR Guiding Principles for scientific data management and stewardship MD Wilkinson, M Dumontier, IJJ Aalbersberg… - Scientific data, 2016, https://rdcu.be/bQbKF

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

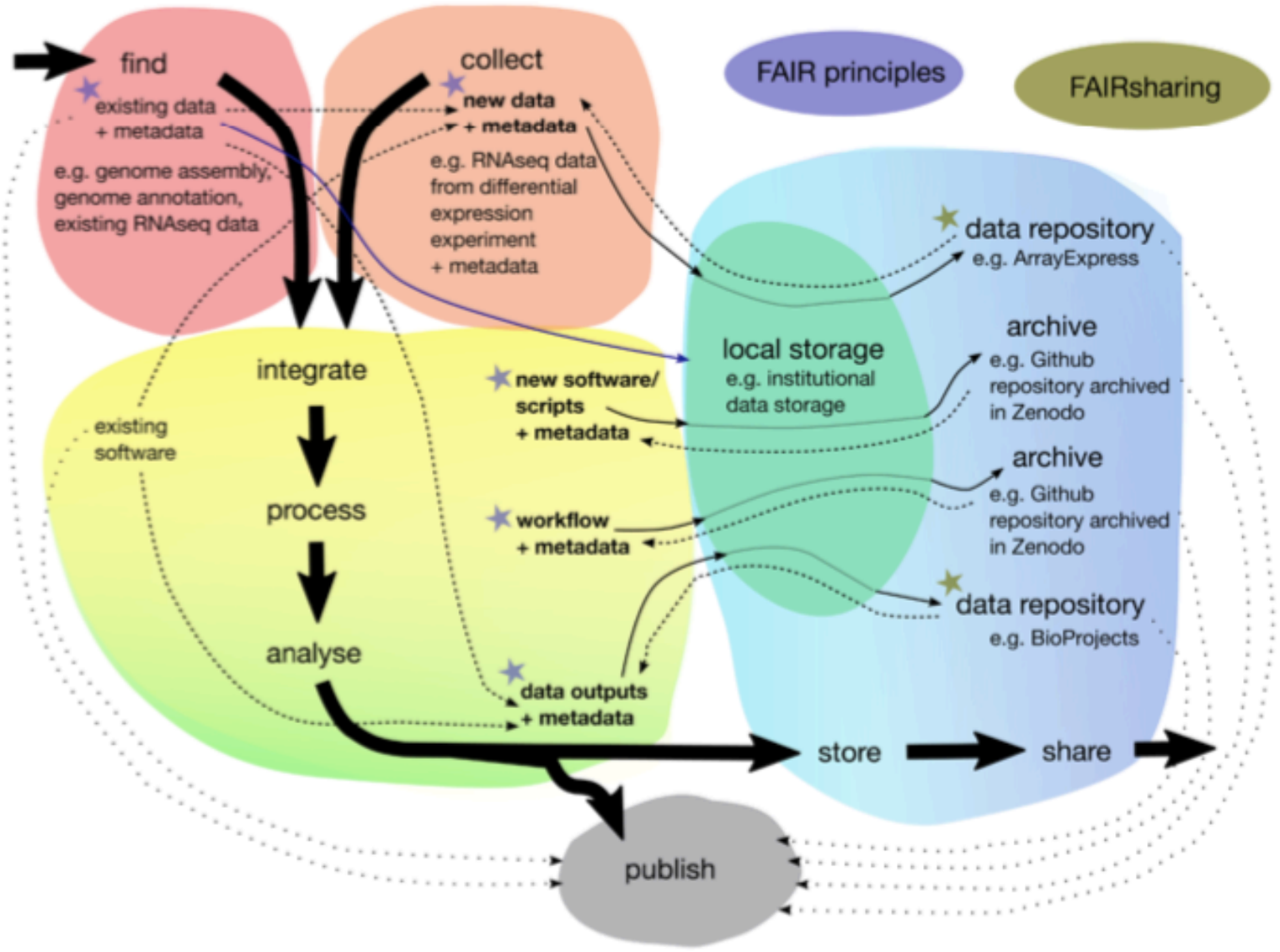


Figure 2. Flowchart of the data life cycle stages applied to an example research project. Bold text indicates new data, software or workflow objects created during the project. Solid thin arrows indicate movement of objects from creation to storage and sharing. Dashed thin arrows indicate where downstream 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

OpenNEURO

A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data

Sign in with Google
Sign in with ORCID

Search Datasets

Browse All Public Datasets

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

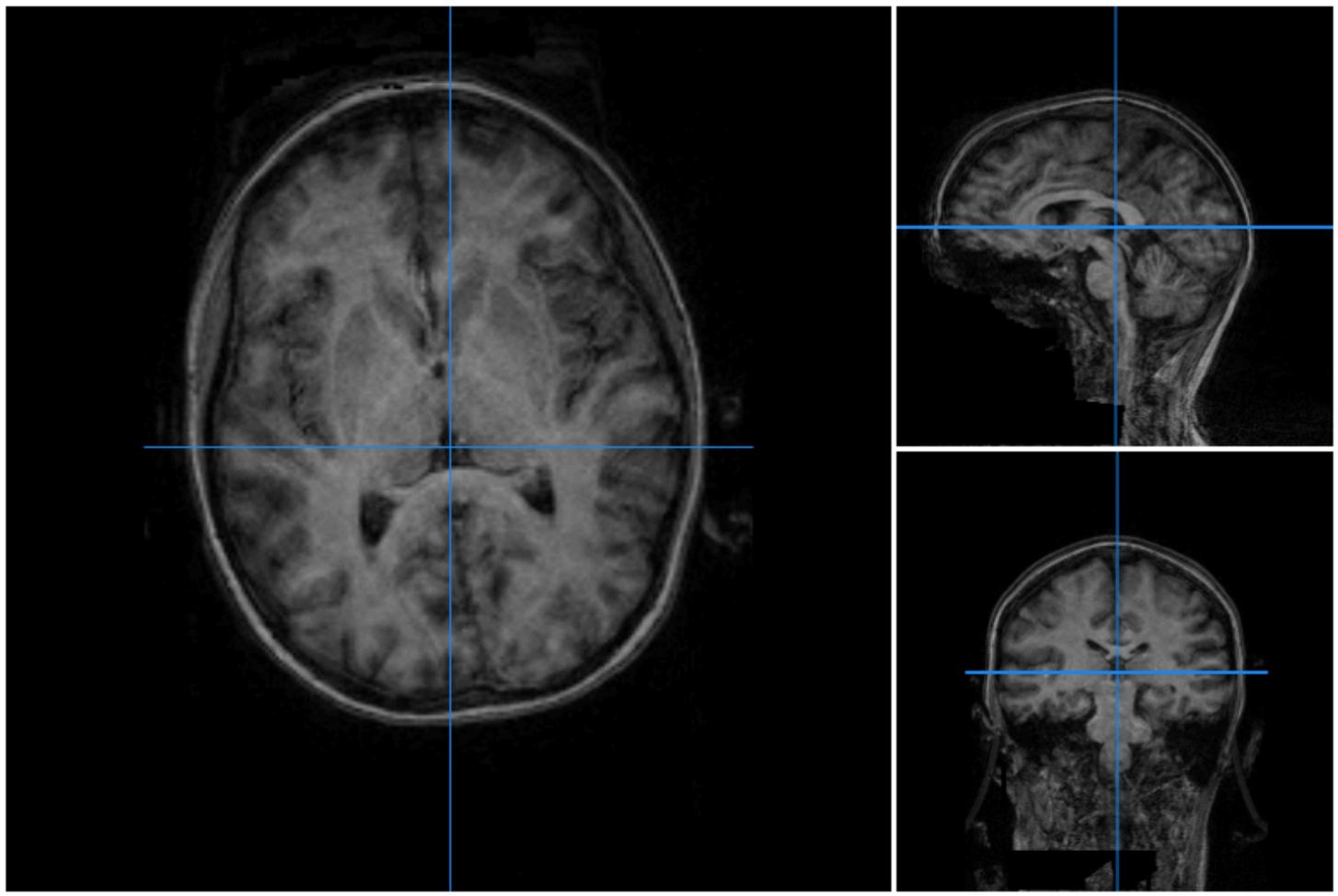
Valid 2 WARNINGS

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

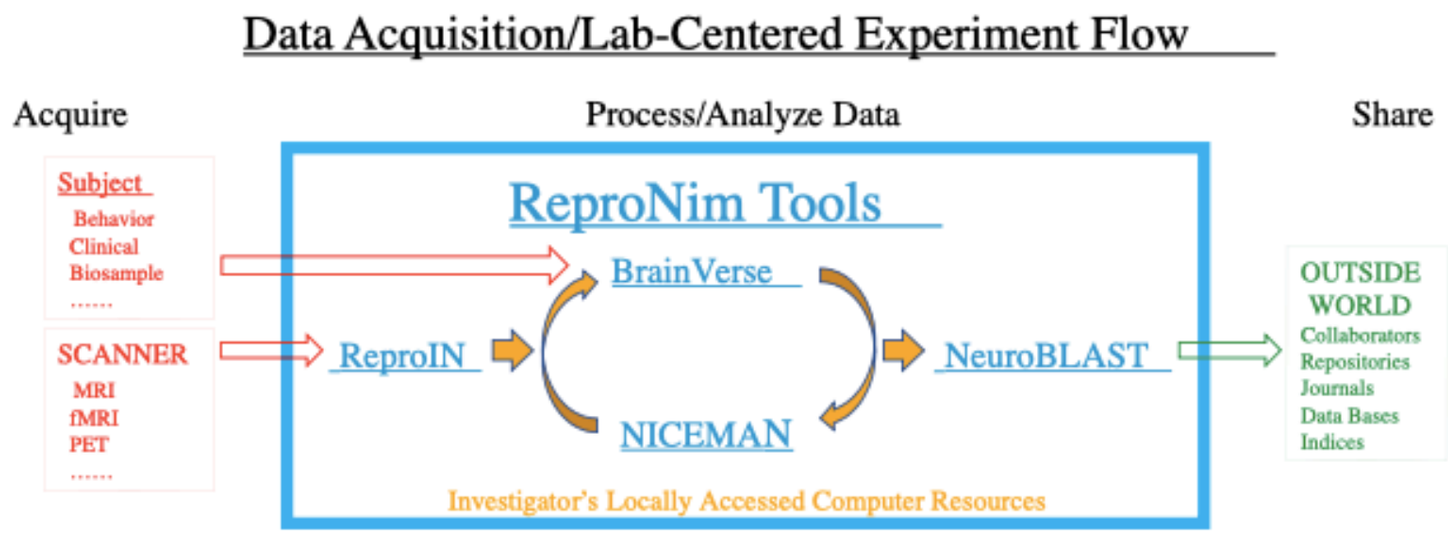
ReproNim: A Center for Reproducible Neuroimaging Computation

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.

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Log in or Register

A curated, informative and educational resource on data and metadata *standards*, inter-related to *databases* and data *policies*.

HOW CAN WE HELP?

We guide consumers to discover, select and use these resources with confidence, and producers to make their resource more discoverable, more widely adopted and cited.

Researchers in academia, industry and government

Identify and cite the standards, databases or repositories that exist for your discipline when creating a data management plan, releasing data or submitting a manuscript to a journal...
[read more]

https://fairsharing.org

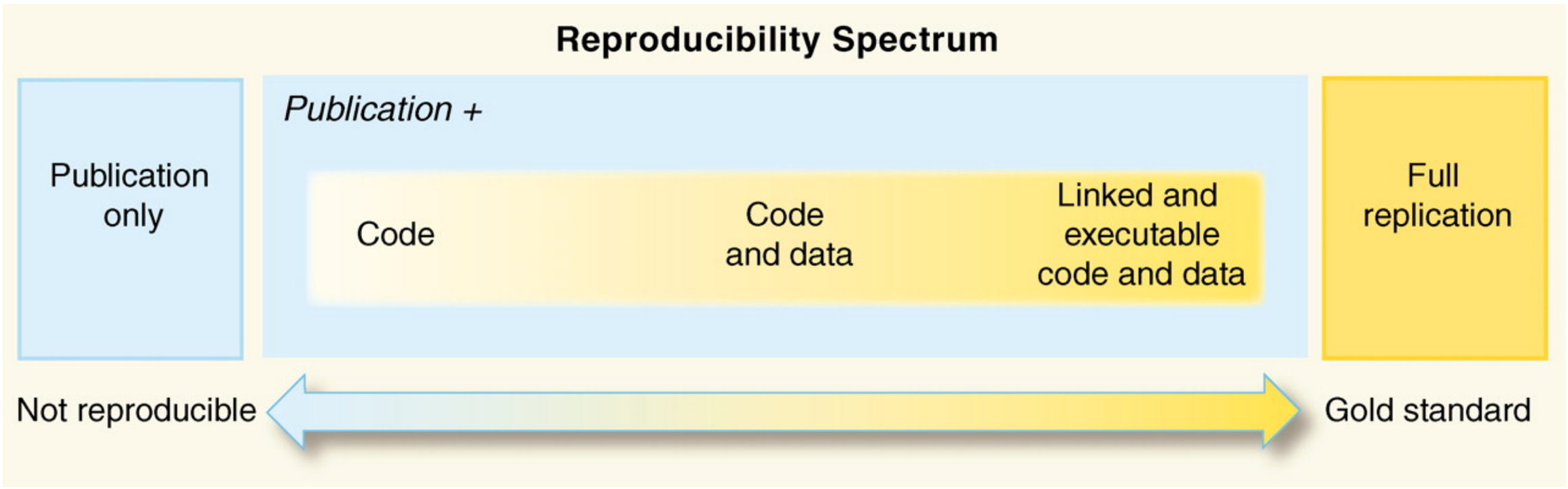
PERSPECTIVE

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



Why do we need to reproduce the computational environment?

- Quite often analysis code ‘breaks’ - often in one of two ways:

- Code that worked previously now doesn’t - maybe a function in an R package was updated (e.g., `lsmeans` became `emmeans` so old code using `lsmeans` wouldn’t now run).

- Code that worked previously still works - but produces a slightly different result or now throws a warning where it didn’t previousy (e.g., convergence/singular fit warnings in `lme4` version 1.1-19 vs. version 1.1-20).

Capturing your local computational environment

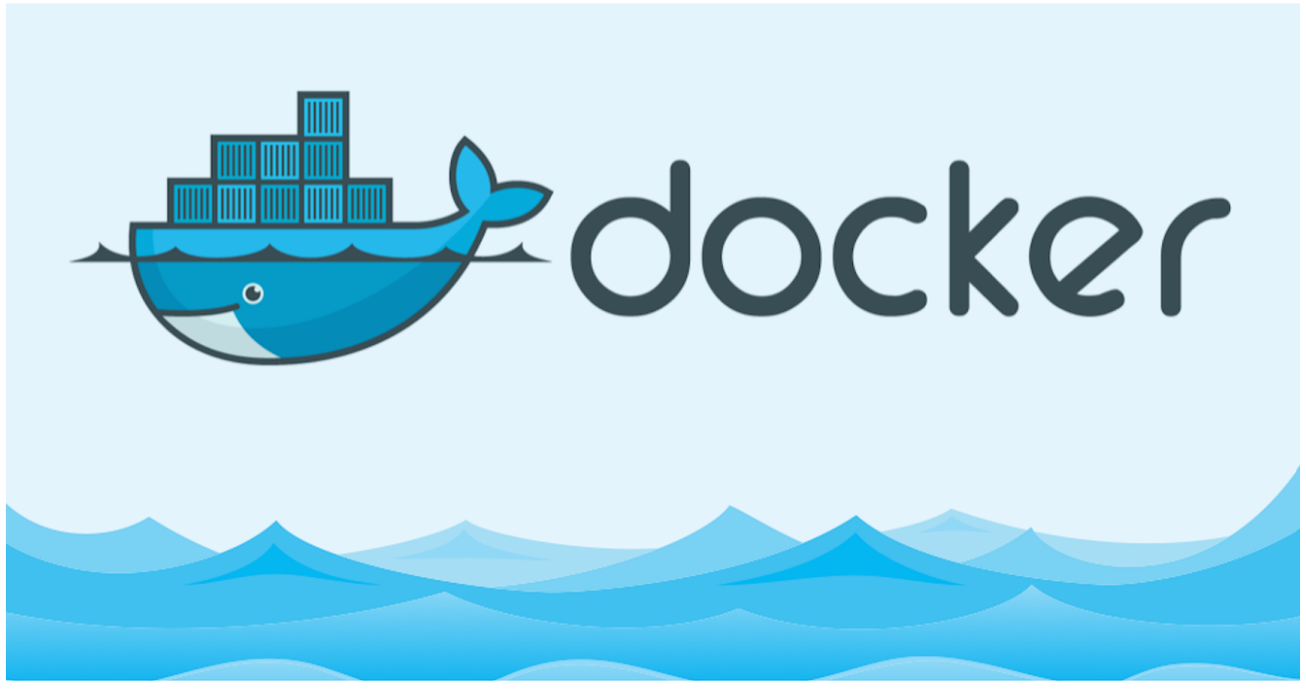
- You need to capture the versions of the different your software packages (plus their dependencies incl. system-level ones).

- May sound trivial but trying running some old analysis code and be amazed at how many things now don’t work as they once did!

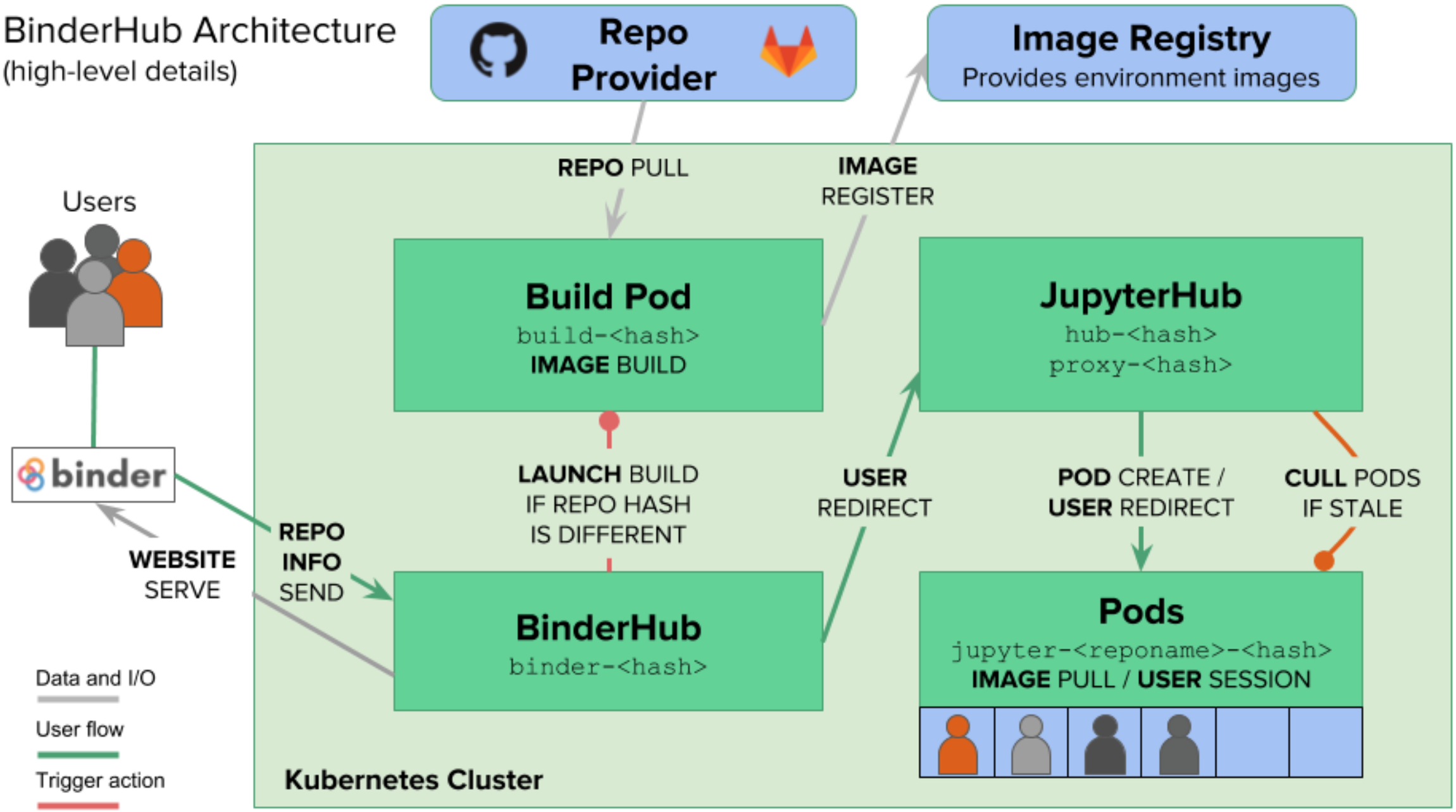
Docker for beginners

Docker packages your data, code and all its dependencies in the form called a docker container to ensure that your application works seamlessly in any environment.

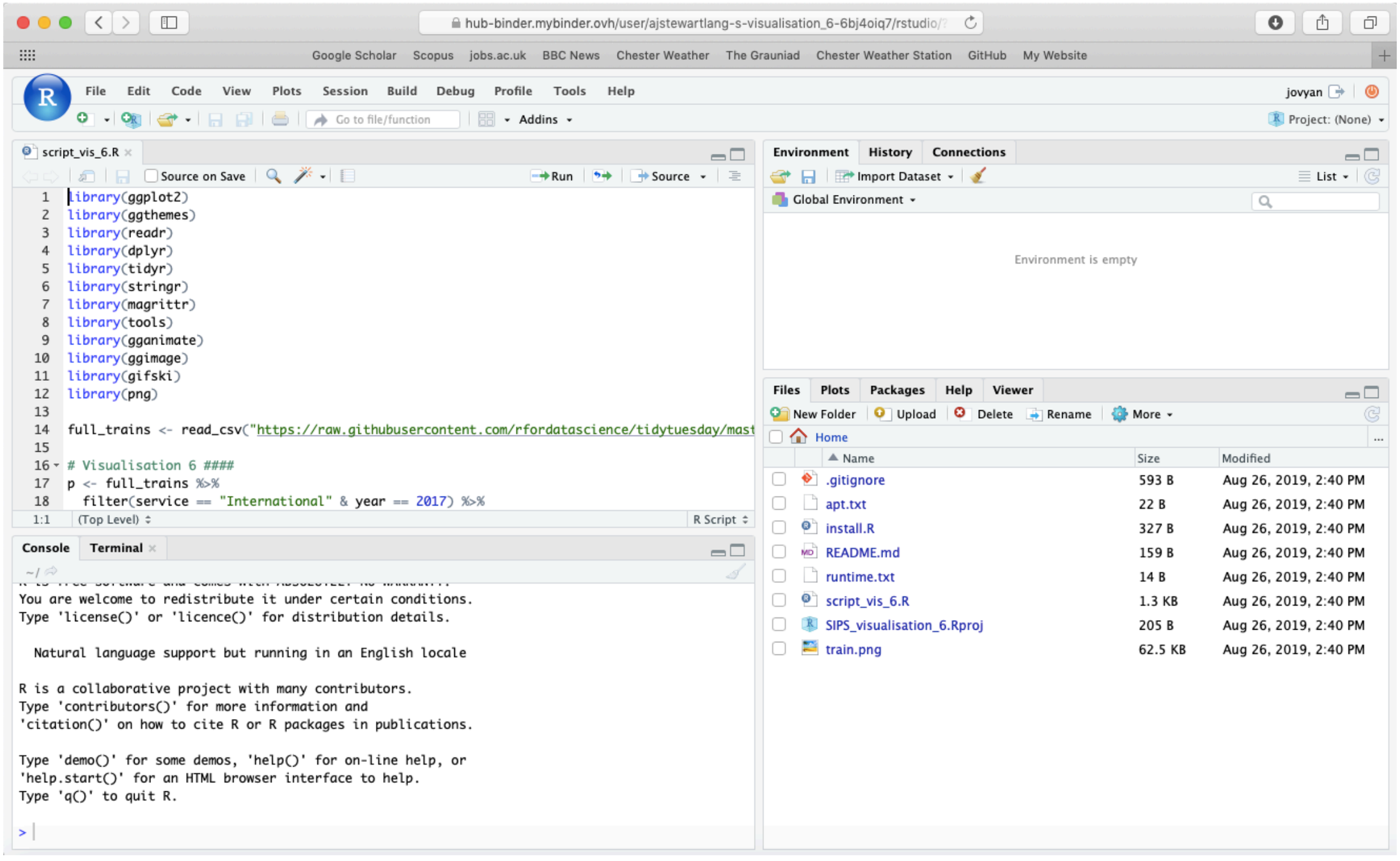
When you run a docker container it's like running your analysis on a computer that has the same configuration as our own one at the point in time when you ran the analysis.



https://medium.com/the-andela-way/docker-for-beginners-61e8e0ce6a19



https://binderhub.readthedocs.io/en/latest/index.html


hub-binder.mybinder.ovh/user/ajstewartlang-s-visualisation_6-6bj4oiq7/rstudio/?
Google Scholar
Scopus
jobs.ac.uk
BBC News
Chester Weather
The Grauniad
Chester Weather Station
GitHub
My Website
File Edit Code View Plots Session Build Debug Profile Tools Help
jovyan
Go to file/function
Addins
Project: (None)
script_vis_6.R
Source on Save
Run
Source
1 library(ggplot2)
2 library(ggthemes)
3 library(readr)
4 library(dplyr)
5 library(tidyr)
6 library(stringr)
7 library(magrittr)
8 library(tools)
9 library(gganimate)
10 library(ggimage)
11 library(gifski)
12 library(png)
13
14 full_trains <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/mast
15
16 # Visualisation 6 ####
17 p <- full_trains %>%
18 filter(service == "International" & year == 2017) %>%
1:1
(Top Level)
R Script
Console
Terminal
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Home
Name
Size
Modified
.gitignore 593 B Aug 26, 2019, 2:40 PM
apt.txt 22 B Aug 26, 2019, 2:40 PM
install.R 327 B Aug 26, 2019, 2:40 PM
README.md 159 B Aug 26, 2019, 2:40 PM
runtime.txt 14 B Aug 26, 2019, 2:40 PM
script_vis_6.R 1.3 KB Aug 26, 2019, 2:40 PM
SIPS_visualisation_6.Rproj 205 B Aug 26, 2019, 2:40 PM
train.png 62.5 KB Aug 26, 2019, 2:40 PM

[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