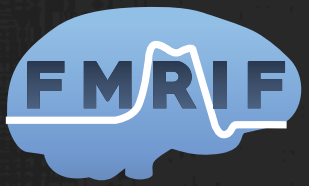
Data Sharing and Open Science in Neuroimaging

Adam Thomas
Data Science and Sharing Team, FMRIF, NIMH

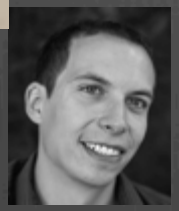
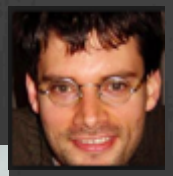
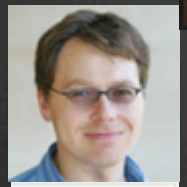
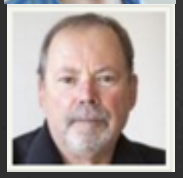


Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

Credits

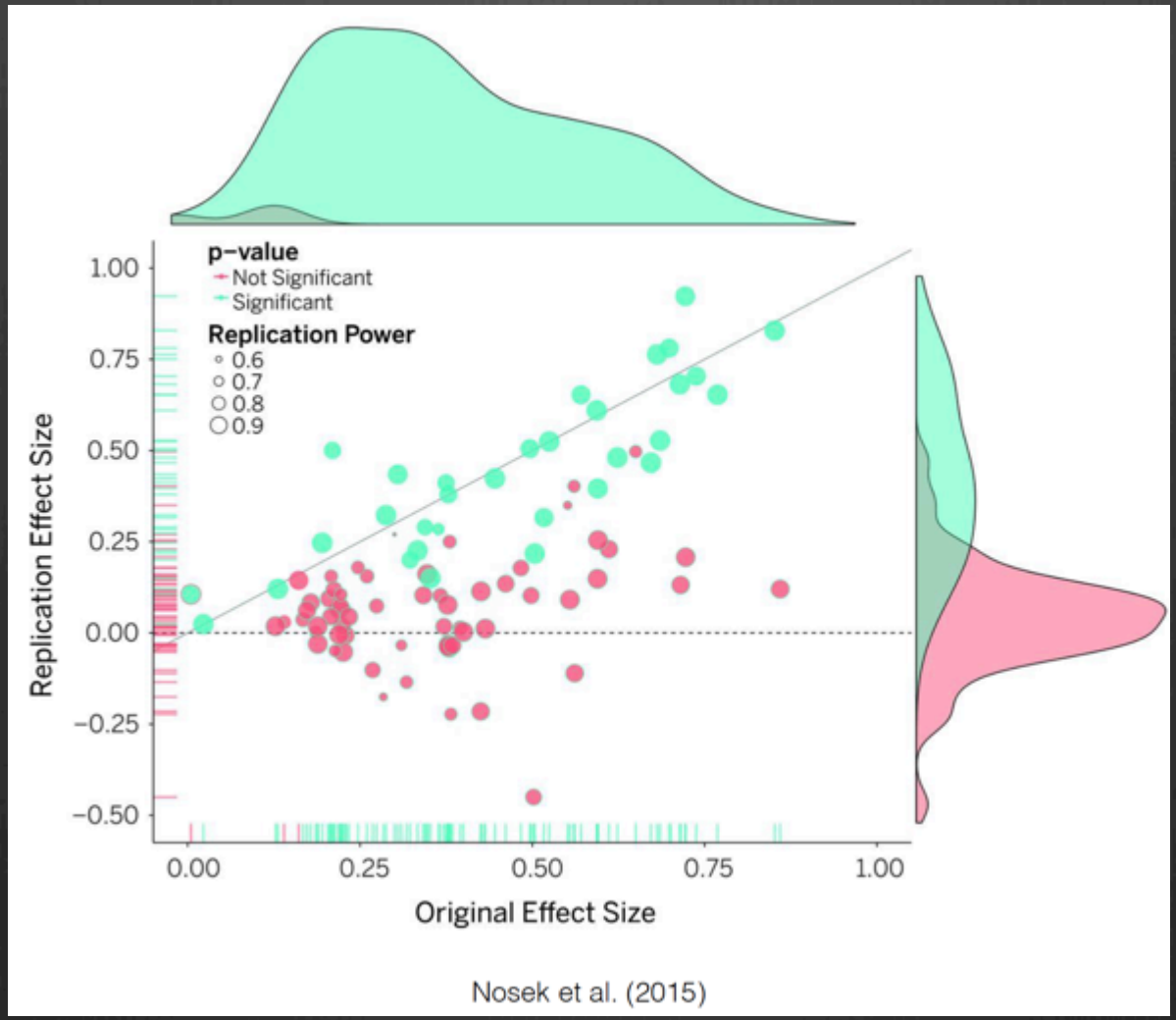
Material borrowed, adapted, and/or stolen from:

- Russ Poldrack 
- Chris Gorgolewski 
- Brian Nosek 
- Tal Yorkoni 
- Niko Kriegeskorte 
- Tom Nichols 
- Phil Bourne 

Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

Problems: Reproducibility



DOI: 10.1126/science.aac4716

Problems:
Wasted time & resources

"How much time do you spend handling, reorganizing, and managing your data as opposed to actually *doing* science?"

- Median answer is 80%

Why Open Science?
Wasted Time & resources

Unpublished Data

- File drawer problem
- Lost staff & lost metadata
- Underutilized data

Problems: The big-data revolution

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne, Jon R. Lorsch** *and* **Eric D. Green.**



Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-

recorded. All of this means that absolute numbers are hard to interpret.

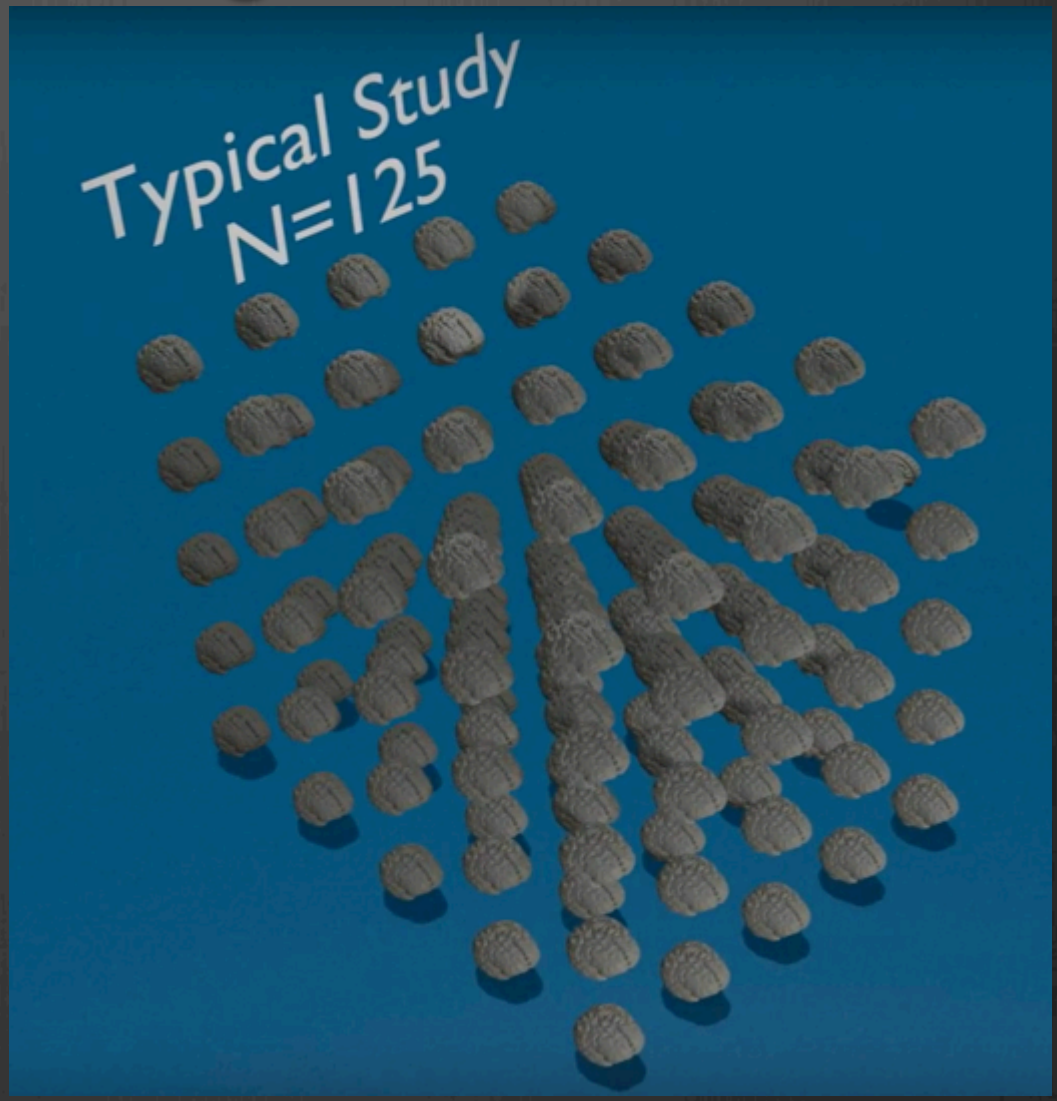
These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

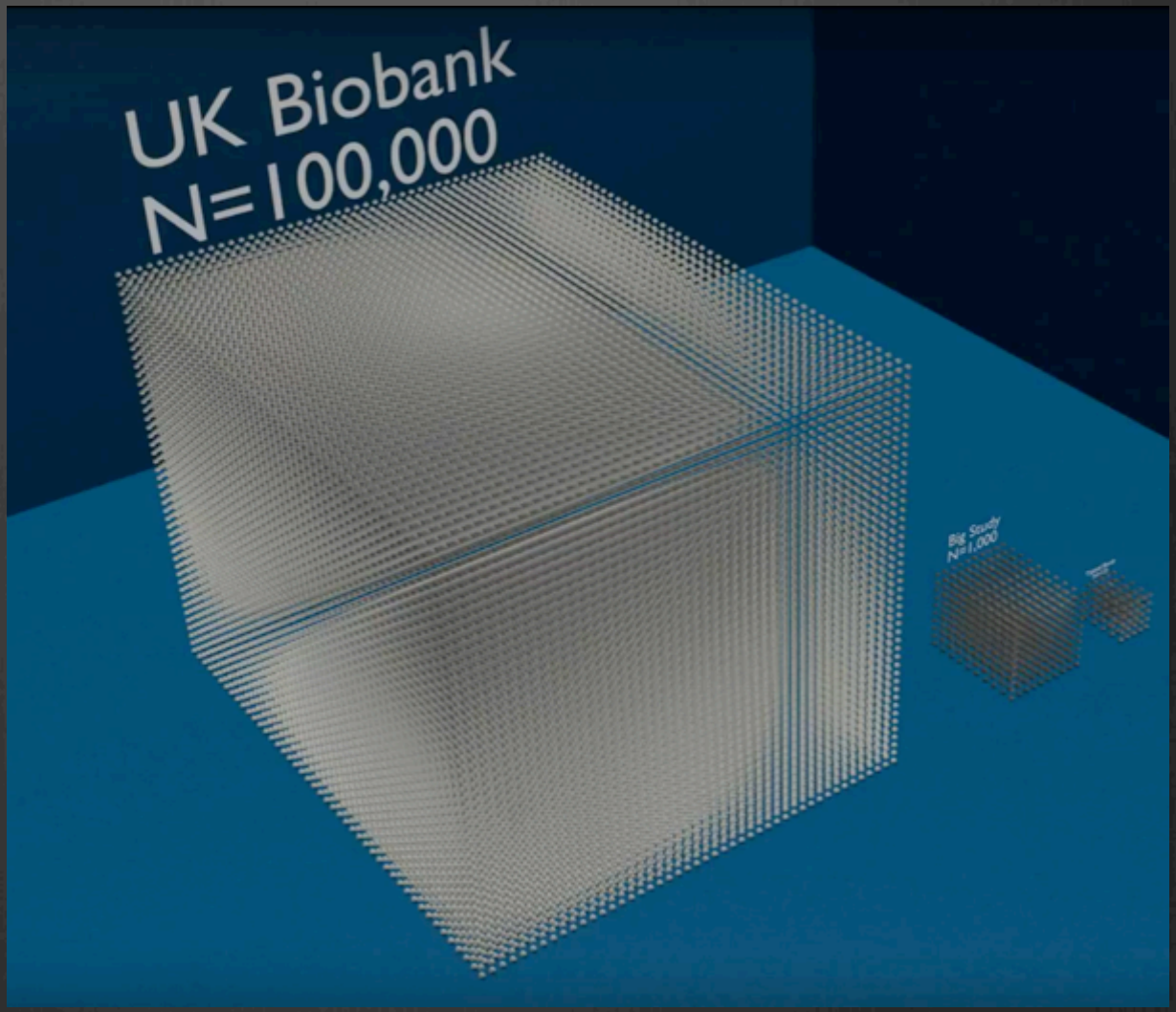
Problems: The big-data revolution

UK Biobank
Imaging
Initiative



Problems: The big-data revolution

UK Biobank Imaging Initiative



Problems: The big-data revolution

Obama's precision medicine initiative will aim to enroll a large number of people in a genetic database representing the U.S. population.

Amy West/Flickr (CC BY 2.0)

President Obama's 1-million-person health study kicks off with five recruitment centers

By Jocelyn Kaiser | Jul. 7, 2016 , 5:00 PM

Problems: The big-data revolution

Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

What is Open Science?

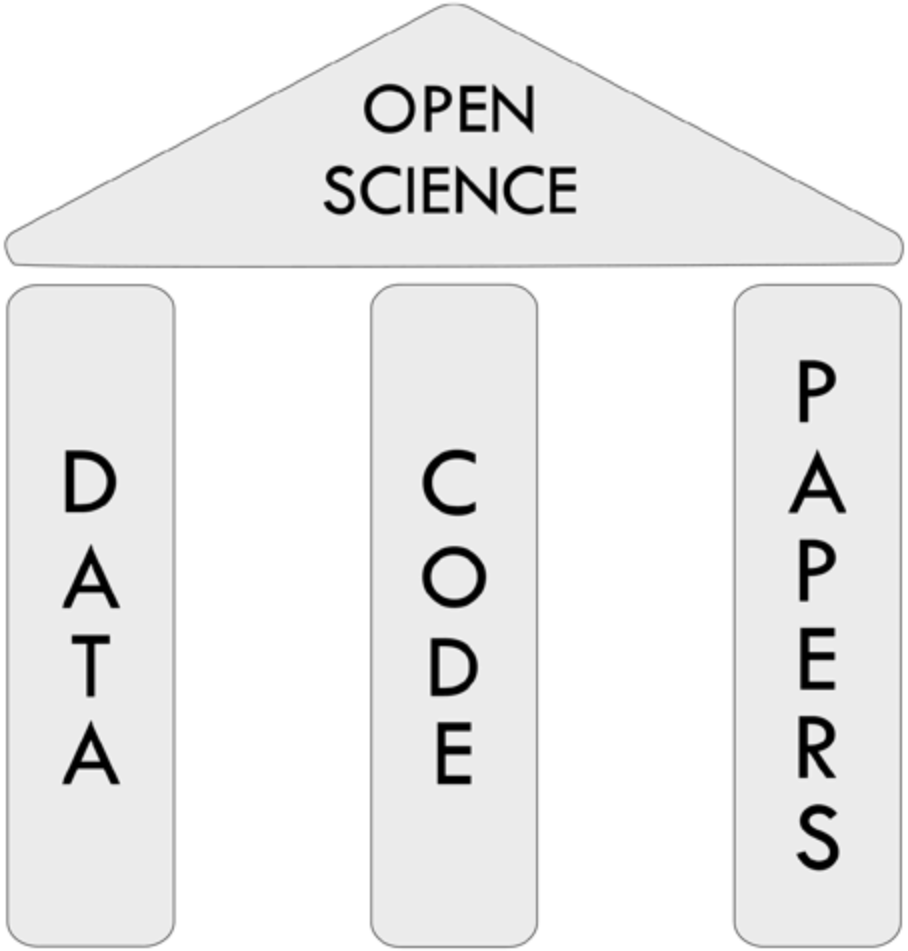


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

What is Open Data?

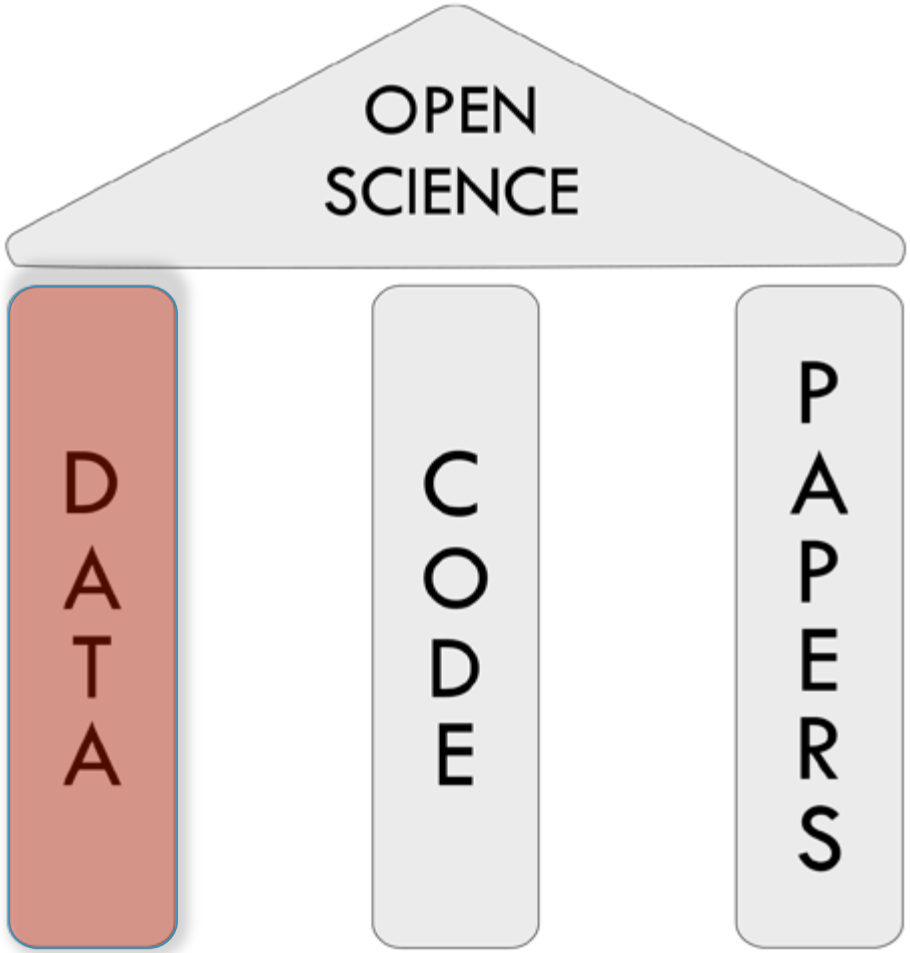


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

Data deposited in a public, community-recognized repository with a stable DOI

Follows FAIR Principle

- Findable
- Accessible
- Intra-operable
- Reusable

Should be deposited *before* publication

Open Data: Community recognized Repositories

MRI Specific Repos

- OpenfMRI
- COINS
- LORIS / OMEGA
- XNAT Central
- Neuovault*

Data Agnostic Repos

- FigShare
- Dryad
- DataVerse
- Open Science Framework
- NIMH Data Archive

* Statistical & derived data only

What is Open Code?

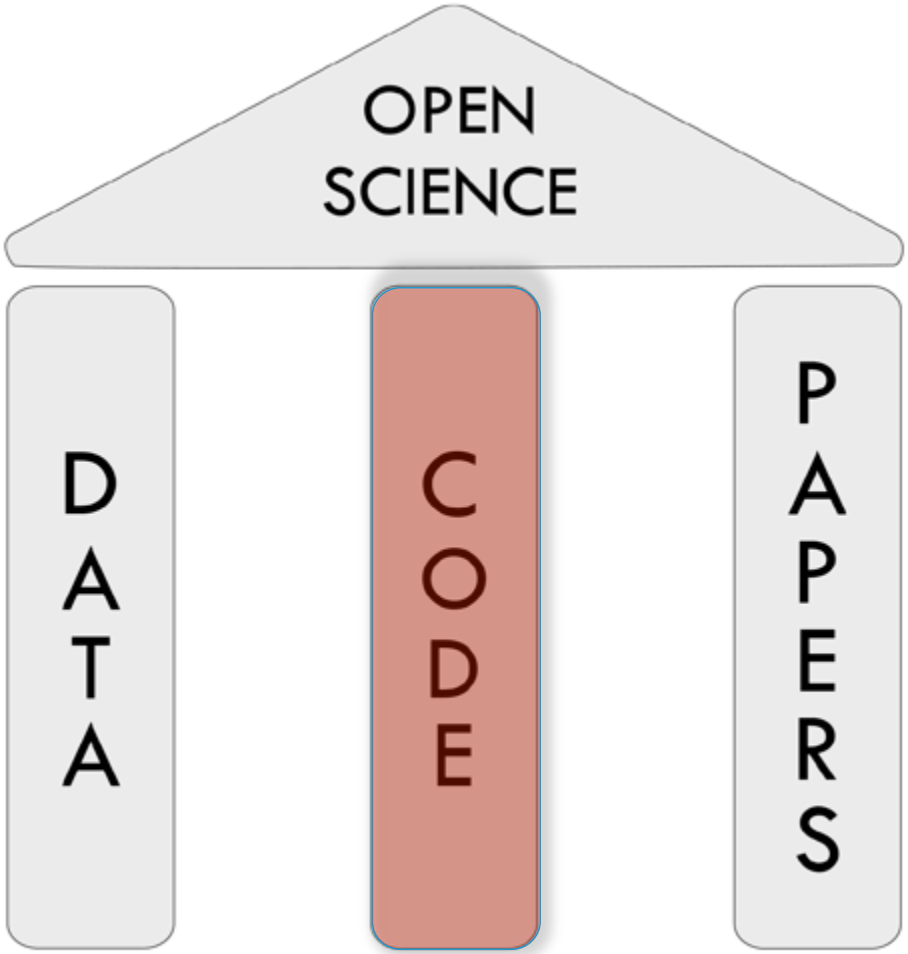
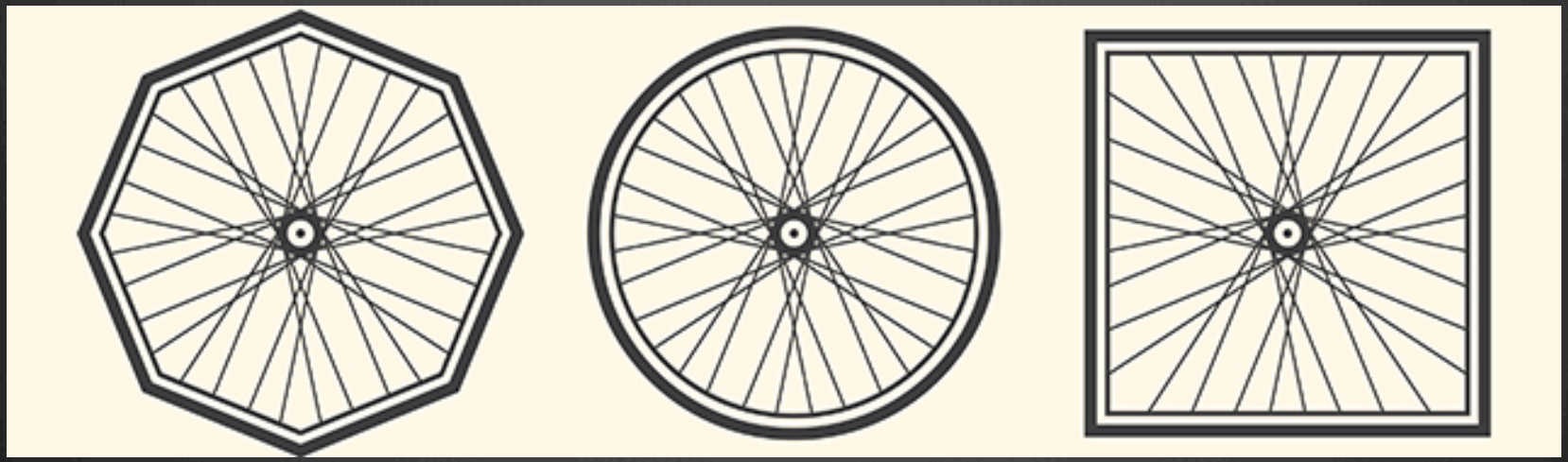


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

Open code enables greater reproducibility (includes non-code methods)

Open Code – Don't Reinvent



Reuse and improve

Open Code - Version Control

Version control systems allows you to:

- Store all of your analysis in a central repository
- Keep a history of "snapshots" of your evolving analysis
- Quickly switch between different versions of your analysis
- Adopt and modify code from other scientists
- Collaborate

GitHub

Atlassian Bitbucket

What are Open Papers?

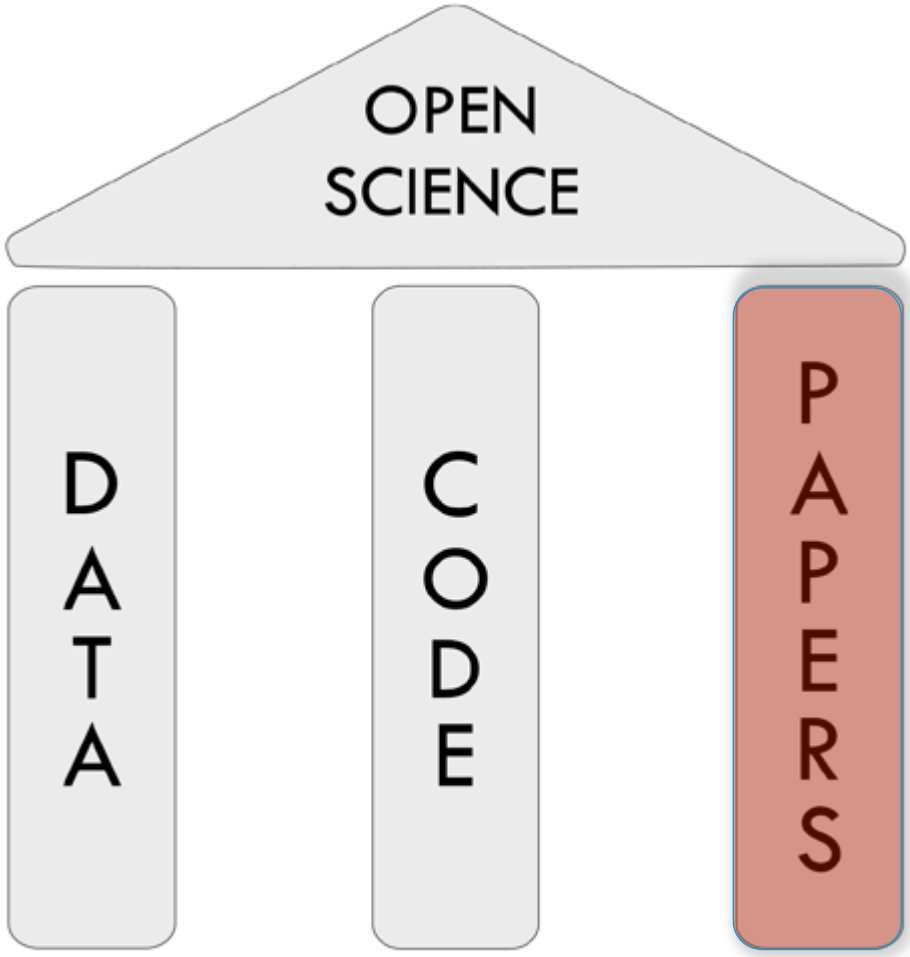


Fig 1. Three pillars of Open Science: data, code, and papers.

doi:10.1371/journal.pbio.1002506.g001

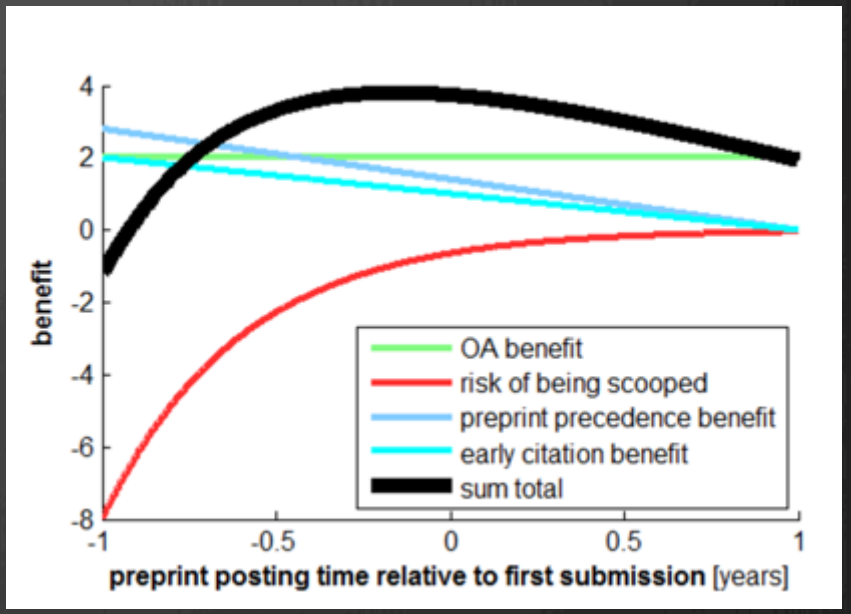
- Preprint posting
- Open access
- Open review

Open Papers: Preprint posting

arXiv.org

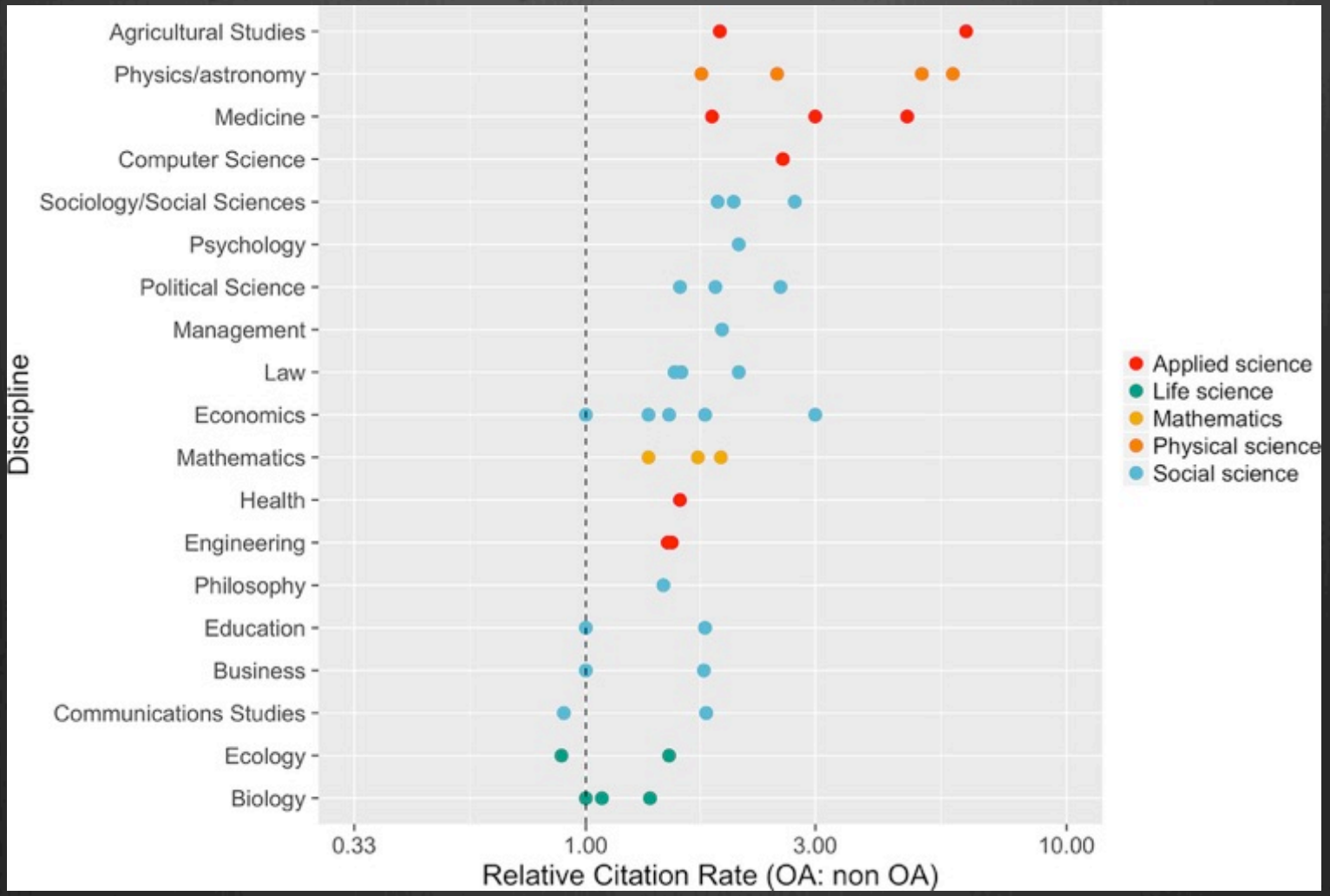
bioRxiv beta
THE PREPRINT SERVER FOR BIOLOGY

- Benefits:
  - Open access
  - Catch errors
  - Earlier citation
  - Earlier precedence, prevent scooping
  - Speed and improve final submission



Kriegeskorte (2016) DOI:10.15200/winn.145838.88372

Open Access
Open access publication are cited more



https://elifesciences.org/content/5/e16800%20

mean citation rate of OA articles divided by mean citation rate of non-OA articles

Open Review

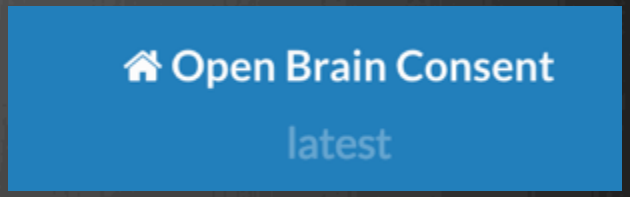
- Public discussion of pros and cons of submission
- Optional anonymity
- Prevent low-quality and or biased review

Outline

- Why do we need Open Science?
- What is Open Science?
- How do I do Open Science?

How – Plan Ahead

- Get data sharing in your protocol:
  - https://open-brain-consent.readthedocs.io



- When designing, collecting, and analyzing consult with standards documents:
  - Enhancing Quality and Transparency of Health Research (EQUATOR) http://www.equator-network.org
  - Best Practices in Data Analysis and Sharing in Neuroimaging using MRI (COBIDAS) http://dx.doi.org/10.1101/054262
  - Good practices for conducting and reporting MEG research http://dx.doi.org/10.1016/j.neuroimage.2012.10.001

Standards – EQUATOR & COBIDAS

- EQUATOR: Different standards for different designs
  - RCT, crossover, observational, etc.
- COBIDAS Sections
  1. Experimental Design
  2. Image Acquisition
  3. Preprocessing
  4. Statistical Modeling
  5. Results
  6. Data Sharing
  7. Reproducibility
- Both EQUATOR and COBIDAS focus on reporting
- Reviewing them in advance will help you plan and design your study
- Also useful reference when reviewing papers

Standards – EQUATOR & COBIDAS

Checklists

CONSORT 2010 checklist of information to include when reporting a randomised trial*

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	
	2b	Specific objectives or hypotheses	
Methods			
Trial design	3a	Description of trial design (such as p	
	3b	Important changes to methods after	
Participants	4a	Eligibility criteria for participants	
	4b	Settings and locations where the dat	
Interventions	5	The interventions for each group wit actually administered	
Outcomes	6a	Completely defined pre-specified pri were assessed	
	6b	Any changes to trial outcomes after	
Sample size	7a	How sample size was determined	
	7b	When applicable, explanation of any	
Randomisation:			

Table D.1. Experimental Design Reporting

Aspect	Notes	Mandatory
Number of subjects	*Elaborate each by group if have more than one group.*	
Subjects approached		N
Subjects consented		N
Subjects refused to participate	Provide reasons.	N
Subjects excluded	Subjects excluded after consenting but before data acquisition; provide reasons.	N
Subjects participated and analyzed	Provide the number of subjects scanned, number excluded after acquisition, and the number included in the data analysis. If they differ, note the number of subjects in each particular analysis.	Y
Inclusion criteria and descriptive statistics	*Elaborate each by group if have more than one group.*	
Age	Mean, standard deviation and range.	Y
Sex	Absolute counts or relative frequencies.	Y
Race & ethnicity	Per guidelines of NIH or other relevant agency.	N

COBIDAS – 7. Reproducibility

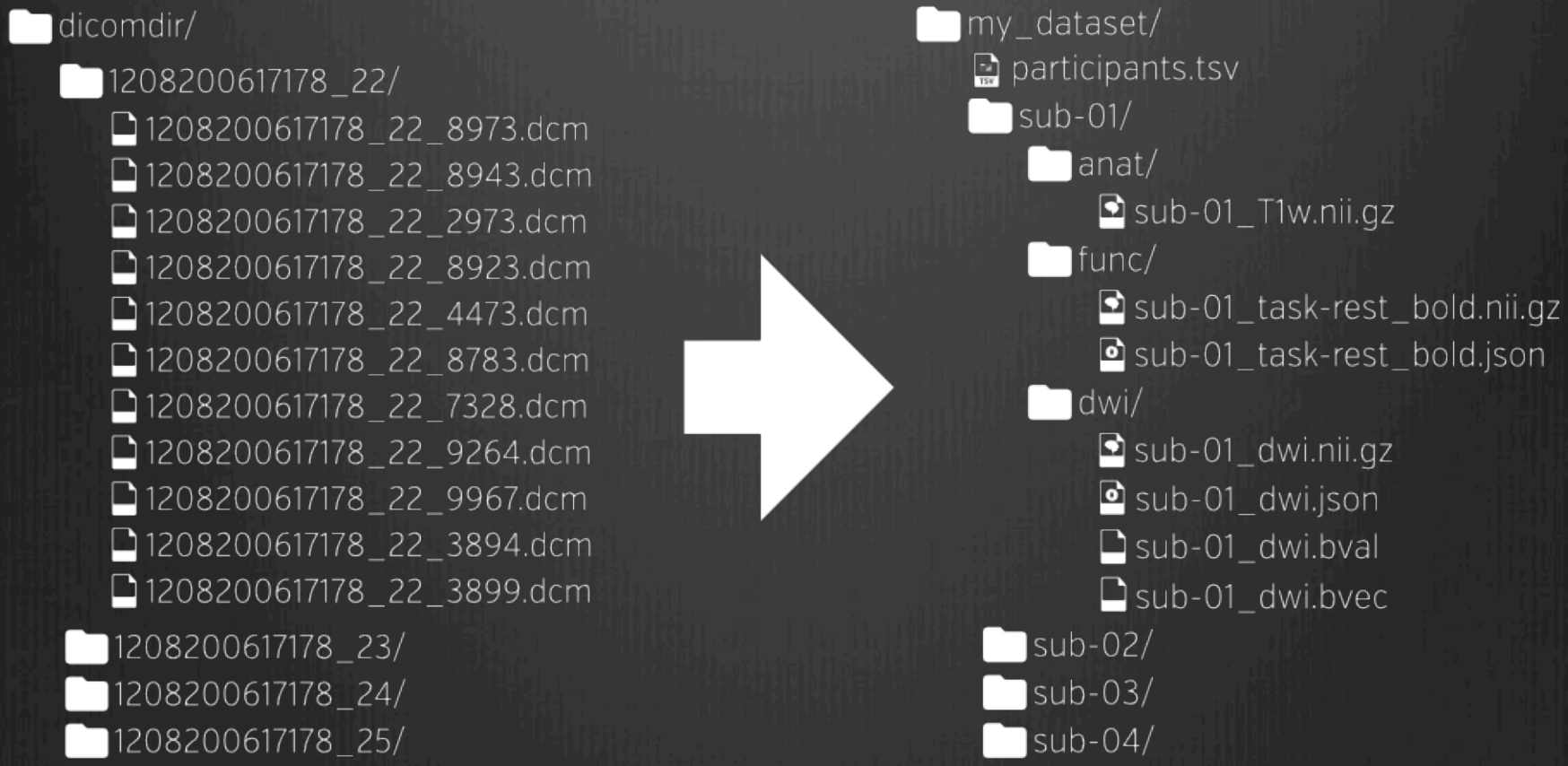
Archiving: Think long term

- Open-source software is more likely to be available long term
- URLs "decay" over time. Use Digital Object Identifiers (DOI) instead
- Deposit data in community recognized repositories that are committed to long-term availability

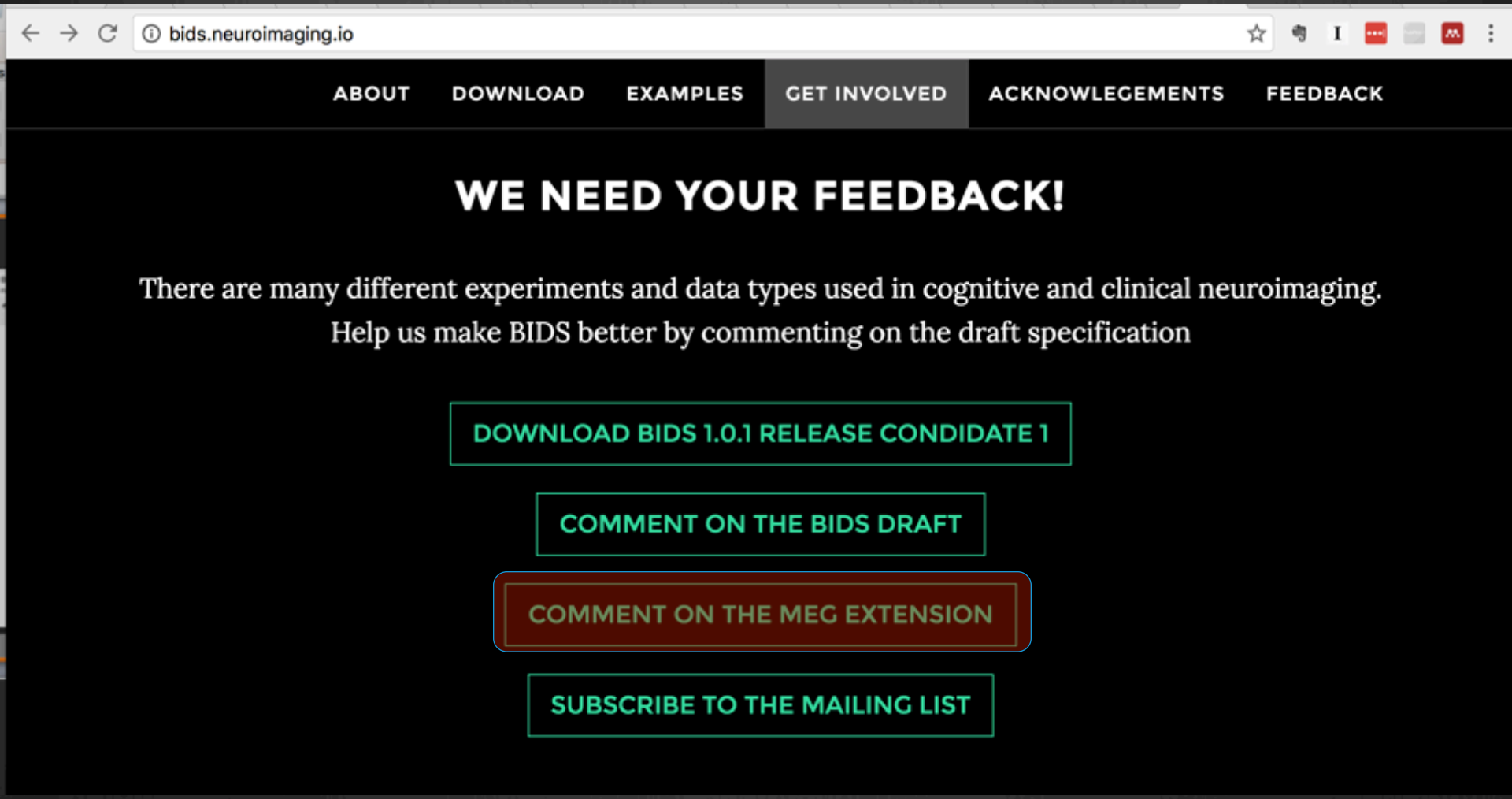
Organizing your data - BIDS

A simple and intuitive way to organize and describe your neuroimaging and behavioral data.

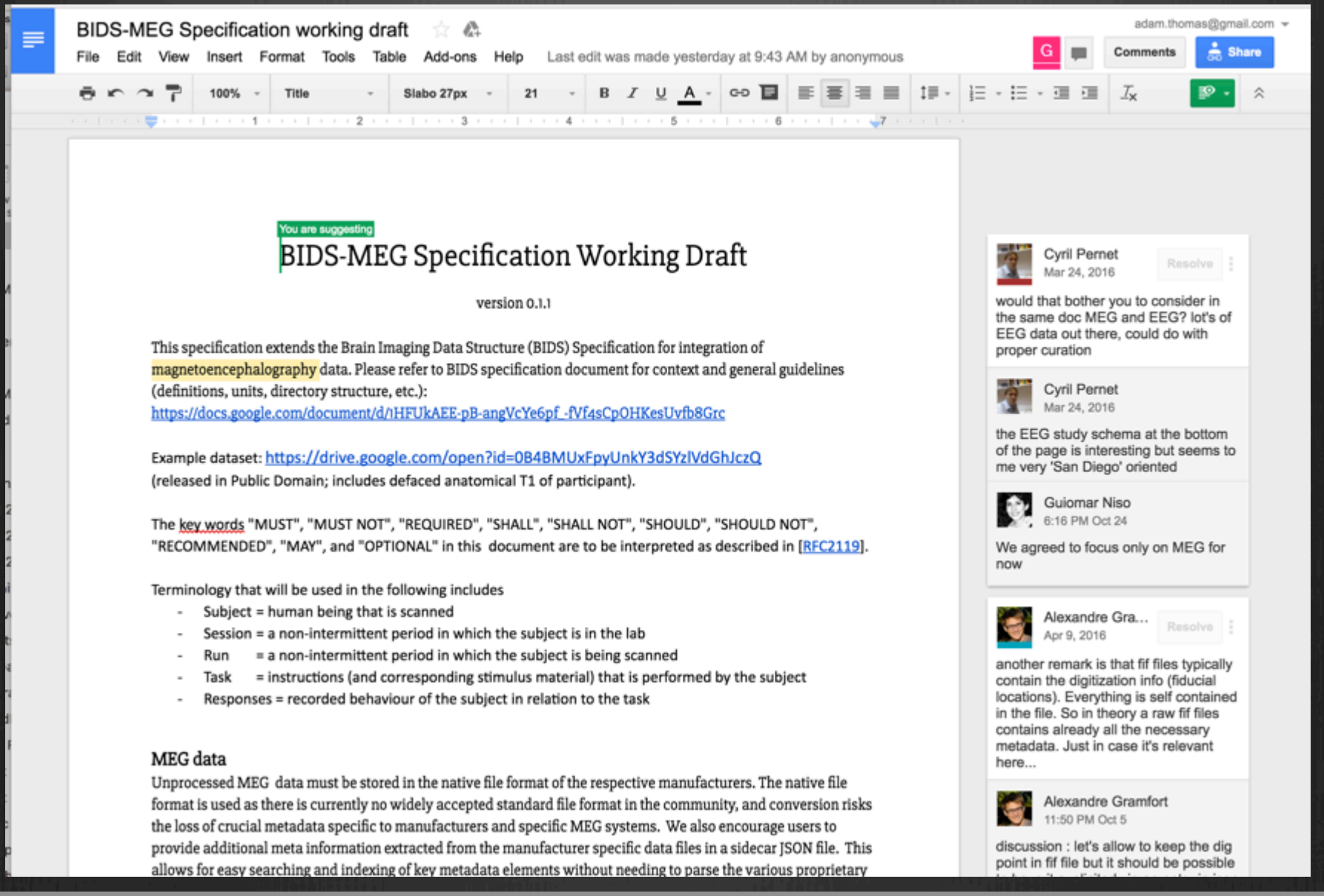
http://bids.neuroimaging.io



Organizing your data - BIDS



Organizing your data - BIDS



How to be Open – Choose your battles

Be open where you can, as you can

SCIENTIFIC STANDARDS

Promoting an open research culture

Author guidelines for journals could help to promote transparency, openness, and reproducibility

By **B. A. Nosek,* G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni**

DOI: 10.1126/science.aab2374

Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

How – Getting help

- Training

- Asking for help

Summary and Take Homes

- Science is changing (for the better) in both scope (big) and culture (open) to address future challenges
- Open science strives to maximize reproducibility and transparency of data, code, and papers which is essential for complex techniques for MEG
- Adopting open science practices yields benefits in productivity, impact, reproducibility, and the ability to conduct large scale analyses
- You don't have to do it all at once, and you don't have to do it alone

Thanks!

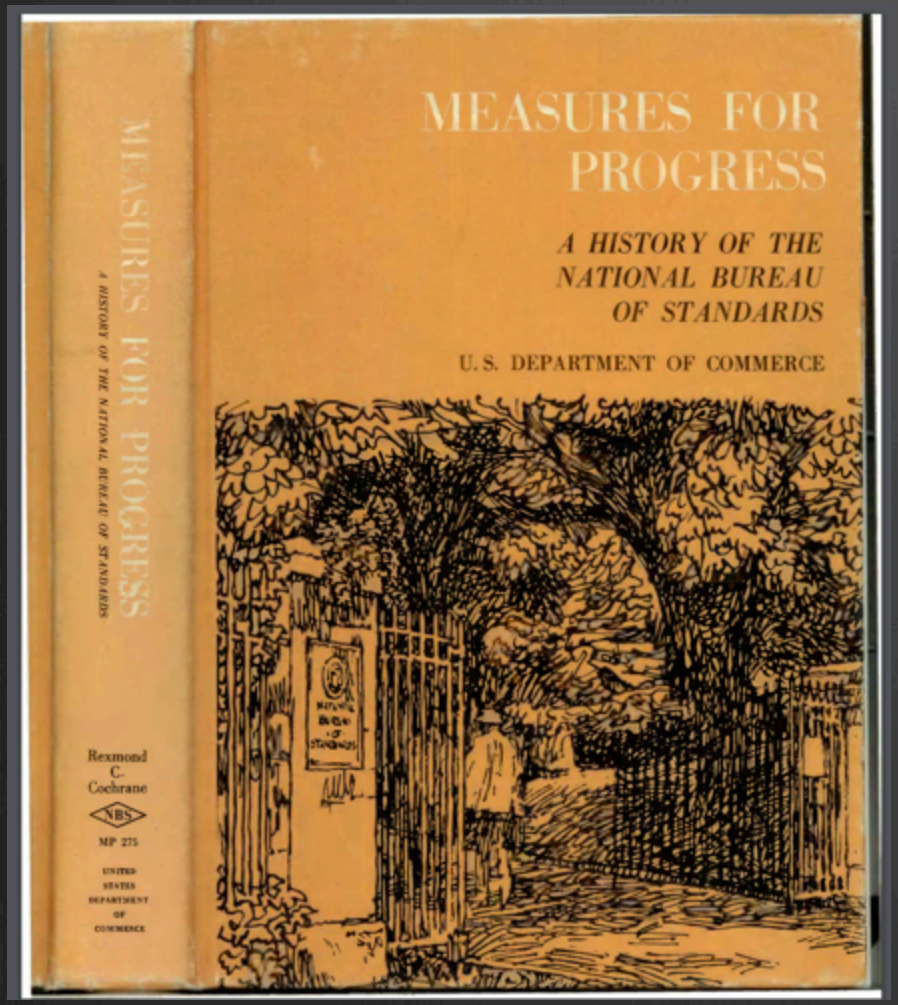
Slides: https://github.com/agt24

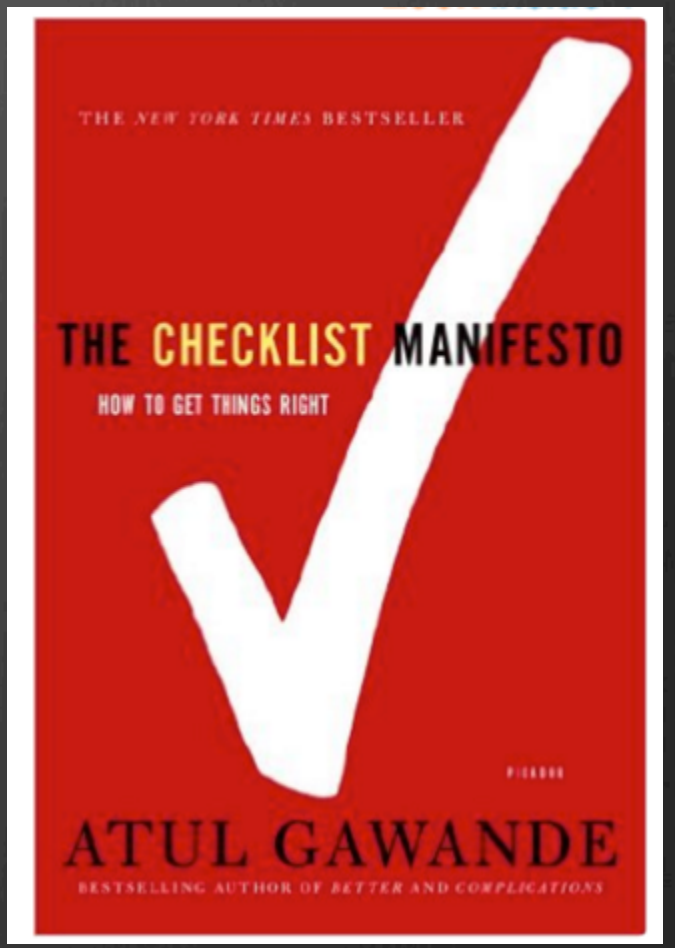
adamt@nih.gov

Questions?

The Problem

- Science vs. Art: The importance of standardization





COBIDAS – Highlights

- Report scan parameters by exporting exam cards
- Preprocessing include *all* steps applied to the data before and must be reported
- For maximal transparency, report all regions of interest (ROIs) and/or experimental conditions examined as part of the research, so that the reader can gauge the degree of any HARKing
  - Hypothesizing After The Results are Known
  - It's OK to explore your data, just be clear that that is what you're doing