Hands-On Reproducible Neuroimaging

Are you ready?

Do you have your VirtualBox?

Do you have the course .ova?

If not, please see one of the course staff!

Please remember to complete the pre-course survey:
https://bit.ly/2MuJcP8

OHBM All Day Educational Course

Hands on Reproducible Brain Imaging

Introduction to Reproducible Neuroimaging: Motivations

ReproNim

Show of hands: Who thinks engaging in Reproducible Research is important?

Show of hands: If I pulled the neuroimaging papers just from this year (PubMed currently says that there are 3,381 of them), how many do you think would be reproducible?

- 90 - 100% ?
- Between 50 - 90% ?
- 25 - 50% ?
- Less than 25% ?
- None?

It’s not a completely fair question, as we’ve not yet really established what we mean by ‘reproducible’. Can any of the audience tell me what they mean by ‘reproducible’?

So, there are lots of different potential ways to talk about reproducibility. I’m going to set some definitions that we will use for today. There are definitely other ways to think about this, and define terms, but I think the following framework

Spectrum of Reproducibility

Original

- Data + Analysis = Result

Re-Execution

- **Exact Same Data** + **Exact Same Analysis** should yield the **Exact Same Result**

Generalization

- **Exact Same Data** + Nominally ‘*Similar*’ Analyses should yield a ‘*Similar*’ Result (i.e. FreeSurfer subcortical volumes compared to FSL FIRST)
- Nominally ‘*Similar*’ Data + **Exact Same Analysis** should yield a ‘*Similar*’ Result (i.e. my kids with autism compared to your kids with autism)
- Nominally ‘*Similar*’ Data + Nominally ‘*Similar*’ Analyses should yield a ‘*Similar*’ Result

- ‘*Similar*’ has lots of wiggle room for interpretation (both to enhance similarity and discount differences).

The premise is that a ‘true’ finding should generalize (i.e. always be true).

If a paper reaches a conclusion such as

- “The volume of the corpus callosum is reduced in children with autism”,

That statement, if generalizable, should hold for **any** *valid way of measuring corpus callosum volume*, and in **all** *children with Autism*.

Individual papers usually do not explore multiple valid ways of making a measurement (i.e. run FreeSurfer and ANTS); and, particularly for complex spectrum-style disorders, the representativeness of any finite sample and a disorder is to be questioned (it is possible that the subset of patients with autism that consent to undergo a MRI scan is not a truly representative sample of the Autism diagnosis, for example…)

The Reproducibility Problem

- A number of studies have brought the reproducibility of science into question.

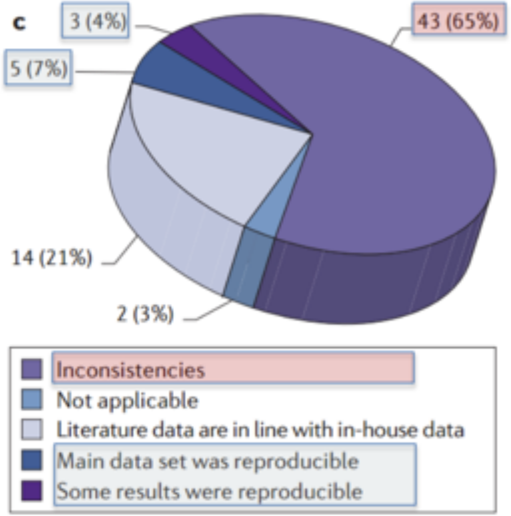


Chart from Prinz, et al. Nature Reviews Drug Discovery 10, 712 (September 2011)
Bayer Healthcare

- Definition of Reproducibility
  - **Publication-level Replication**. Take any given publication and cast it in a reproducible fashion. Problem since current publications do not actually provide complete specification.
  - **Generalizable Reproducibility** across publications. Huge problem since we dis-incentivize publication of replication studies as 'not novel'.

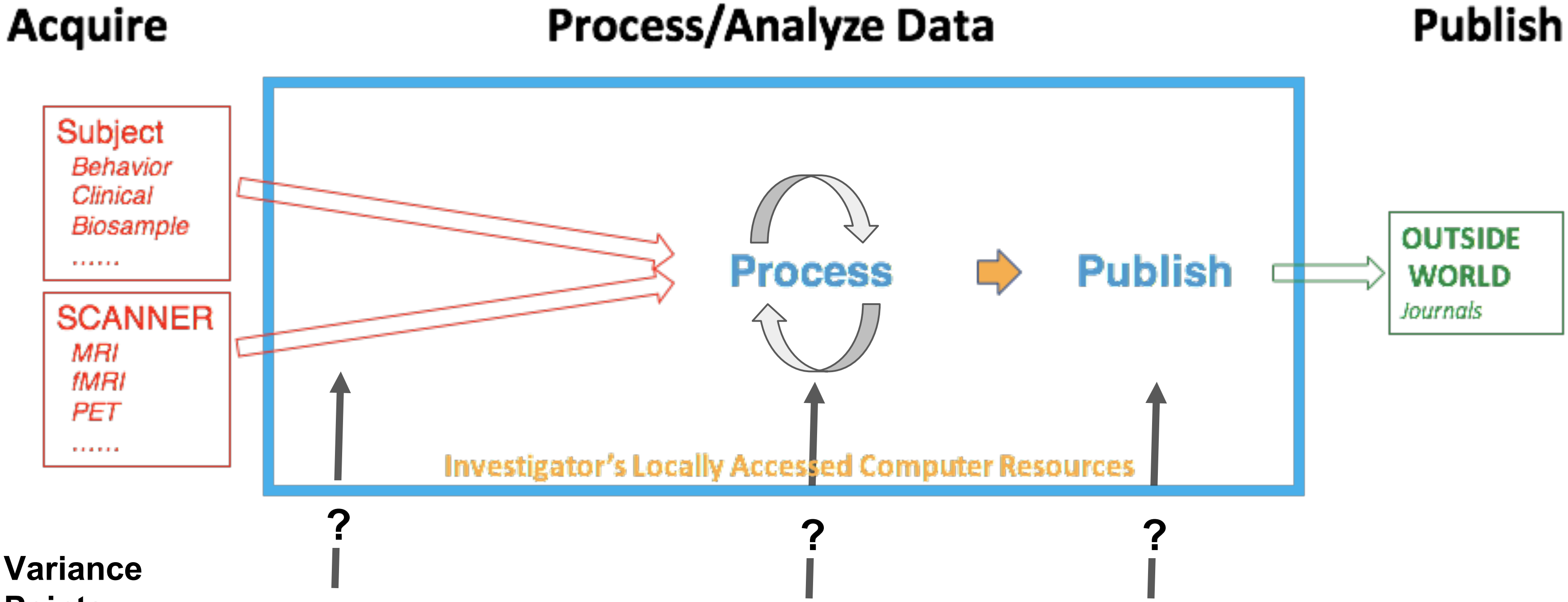
Progress in improving mental health outcomes has been slow

- Patients with mental disorders show many biological abnormalities which distinguish them from normal volunteers; however, few of these have led to tests with clinical utility. Several reasons contribute to this delay: lack of a biological ‘gold standard’ definition of psychiatric illnesses; a profusion of statistically significant, but minimally differentiating, biological findings; ‘approximate replications’ of these findings in a way that neither confirms nor refutes them; and a focus on comparing prototypical patients to healthy controls which generates differentiations with limited clinical applicability.

Causes include: Low power/Small N, Incorrect ‘Target’ (diagnosis as opposed to domain), Incomplete Methods & Results Description, Publication Bias

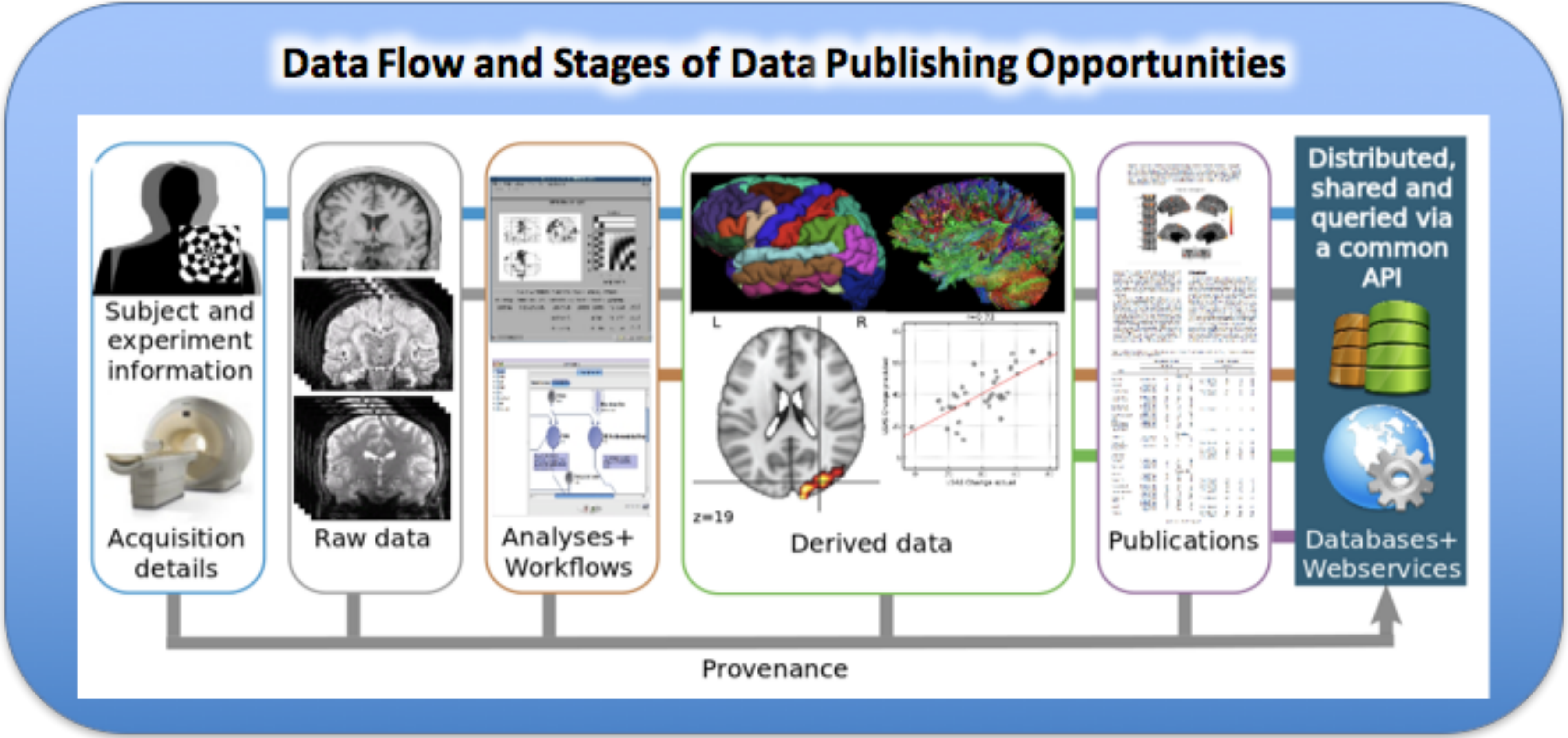
S Kapur, A G Phillips and T R Insel, *Molecular Psychiatry* (2012) 17, 1174–1179

Acquisition Lab-Centered View (Pre ReproNim)



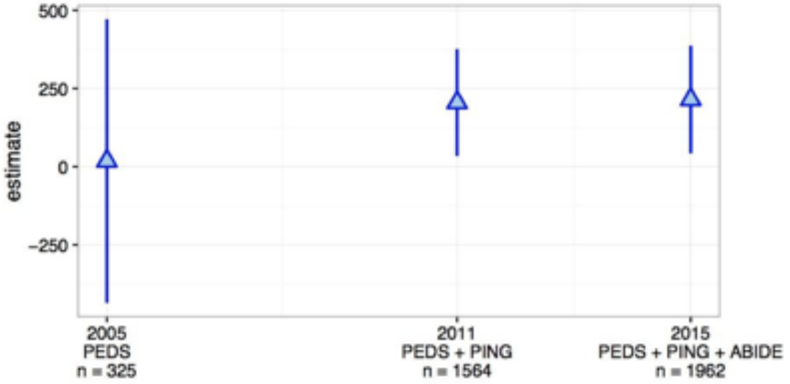
Variance Points

General Neuroimaging Workflow

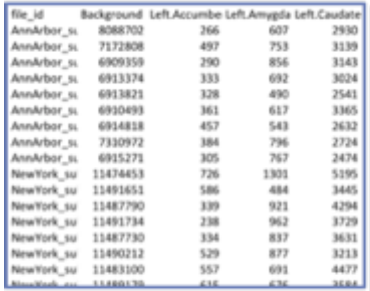

Data Flow and Stages of Data Publishing Opportunities
Subject and experiment information
Acquisition details
Raw data
Analyses+ Workflows
L
R
z=19
Derived data
Publications
Distributed, shared and queried via a common API
Databases+ Webservices
Provenance

Everything matters!

- Operating System Matters (i.e. Linux vs. OSX)
- Tool Selection Matters (i.e. Freesurfer vs. Ants)
- Data Matters (100 subjects versus 1000 subjects)

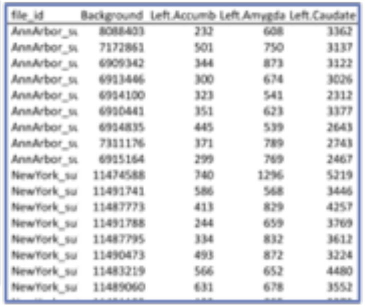


MAC OSX



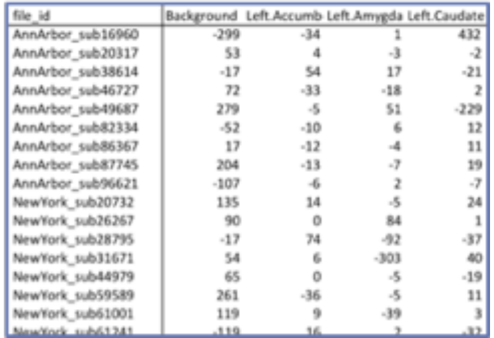
file_id	Background	Left.Accumbe	Left.Amygda	Left.Caudate
AnnArbor_su	8088702	266	607	2930
AnnArbor_su	7172808	497	753	3139
AnnArbor_su	6909359	290	856	3143
AnnArbor_su	6913374	333	692	3024
AnnArbor_su	6913821	328	490	2541
AnnArbor_su	6910493	361	617	3365
AnnArbor_su	6914818	457	543	2632
AnnArbor_su	7310972	384	796	2724
AnnArbor_su	6915271	305	767	2474
NewYork_su	11474453	726	1301	5195
NewYork_su	11491651	586	484	3445
NewYork_su	11487790	339	921	4294
NewYork_su	11491734	238	962	3729
NewYork_su	11487730	334	837	3631
NewYork_su	11490212	529	877	3213
NewYork_su	11483100	557	691	4477
NewYork_su	11489170	616	676	3584

Docker (Ubuntu 14.04)
(Reference)

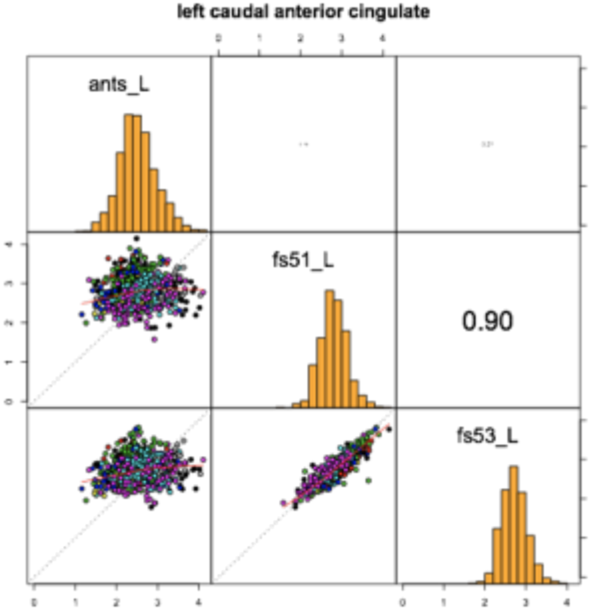


file_id	Background	Left.Accumb	Left.Amygda	Left.Caudate
AnnArbor_su	8088403	232	608	3362
AnnArbor_su	7172861	501	750	3137
AnnArbor_su	6909342	344	873	3122
AnnArbor_su	6913446	300	674	3026
AnnArbor_su	6914100	323	541	2312
AnnArbor_su	6910441	351	623	3377
AnnArbor_su	6914835	445	539	2643
AnnArbor_su	7311176	371	789	2743
AnnArbor_su	6915164	299	769	2467
NewYork_su	11474588	740	1296	5219
NewYork_su	11491741	586	568	3446
NewYork_su	11487773	413	829	4257
NewYork_su	11491788	244	659	3769
NewYork_su	11487795	334	832	3612
NewYork_su	11490473	493	872	3224
NewYork_su	11483219	566	652	4480
NewYork_su	11489060	631	678	3552

Diff



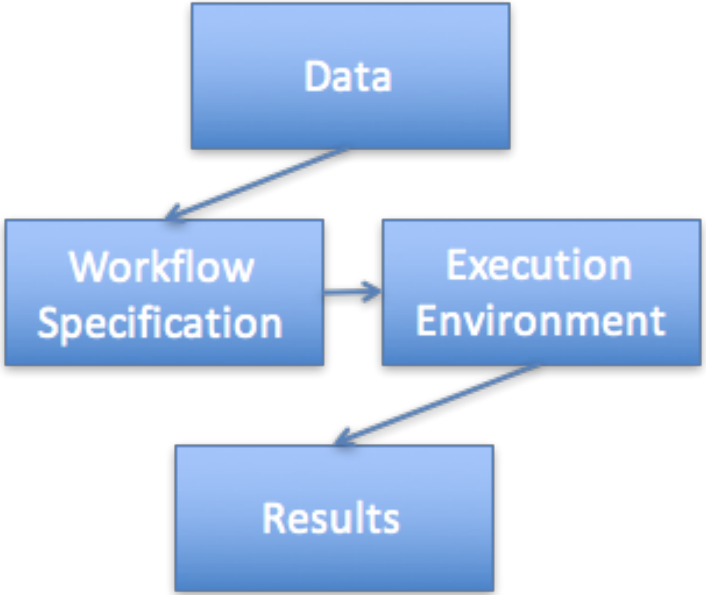
file_id	Background	Left.Accumb	Left.Amygda	Left.Caudate
AnnArbor_sub16960	-299	-34	1	432
AnnArbor_sub20317	53	4	-3	-2
AnnArbor_sub38614	-17	54	17	-21
AnnArbor_sub46727	72	-33	-18	2
AnnArbor_sub49687	279	-5	51	-229
AnnArbor_sub82334	-52	-10	6	12
AnnArbor_sub86367	17	-12	-4	11
AnnArbor_sub87745	204	-13	-7	19
AnnArbor_sub96621	-107	-6	2	-7
NewYork_sub20732	135	14	-5	24
NewYork_sub26267	90	0	84	1
NewYork_sub28795	-17	74	-92	-37
NewYork_sub31671	54	6	-303	40
NewYork_sub44979	65	0	-5	-19
NewYork_sub59589	261	-36	-5	11
NewYork_sub61001	119	9	-39	3
NewYork_sub61241	-119	16	2	-32



A Vision for a Next-Generation Publication

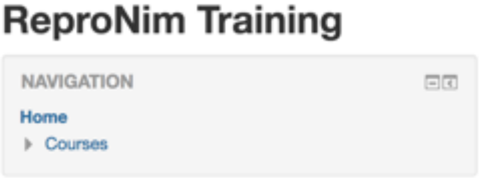
The Reproducible Publication of the future includes:

- Words, as usual, PLUS the following supplemental information:
  - Data
  - Workflow Specification
  - Execution Environment Specification
  - Complete Results



In other words, given the data, workflow specification and execution environment specification, a third party can generate (and validate) the exact results independently.

Ghosh SS, Poline JB, Keator DB *et al.* A very simple, re-executable neuroimaging publication. *F1000Research* 2017, 6:124 (doi: 10.12688/f1000research.10783.2)



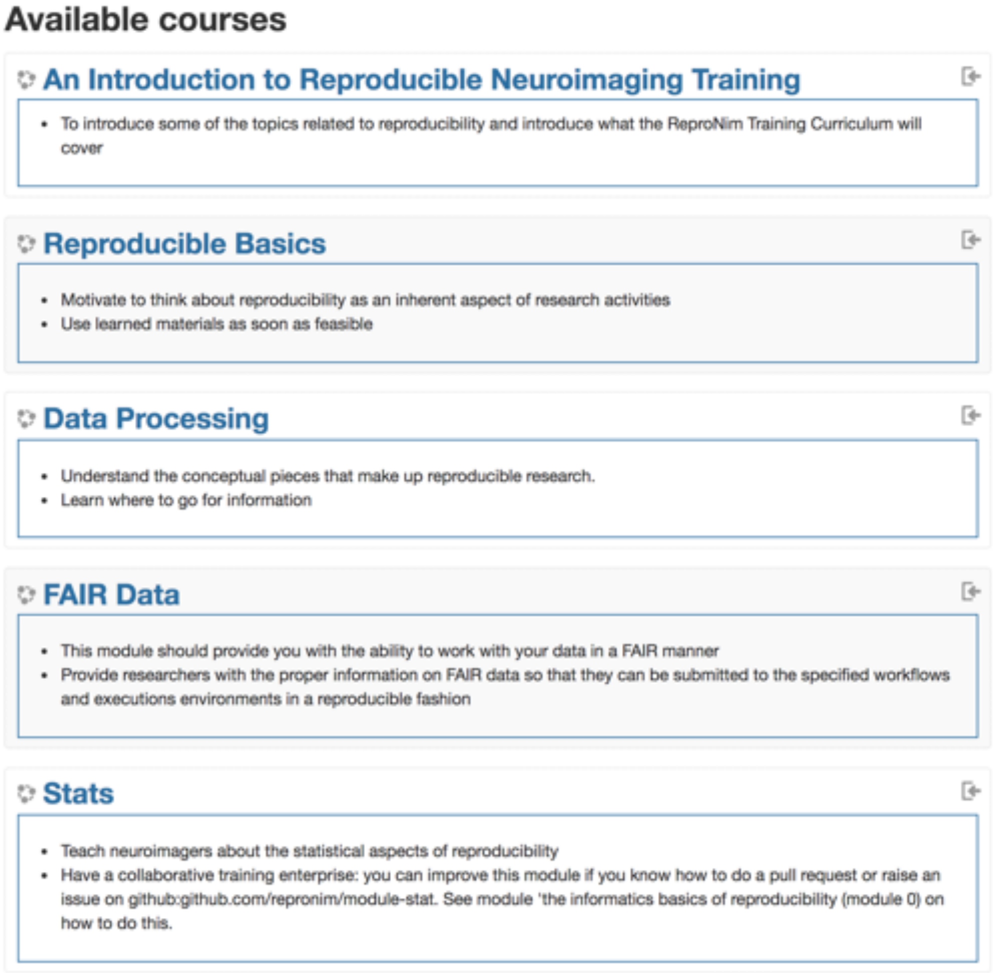
ReproNim Training

The materials we are using today are part of a larger curriculum

training.repronim.org

It borrows from the concepts of Software and Data Carpentry, and ultimately has a “Train the Trainer” design.

As this full curriculum is designed to potentially spans many months, spending 1 day on this will potentially feel **“Appropriately Overwhelming”**. That’s ok (and even the intent). We all will still have a lot of learning and practice to do after today's session.



Available courses

An Introduction to Reproducible Neuroimaging Training

- To introduce some of the topics related to reproducibility and introduce what the ReproNim Training Curriculum will cover

Reproducible Basics

- Motivate to think about reproducibility as an inherent aspect of research activities
- Use learned materials as soon as feasible

Data Processing

- Understand the conceptual pieces that make up reproducible research.
- Learn where to go for information

FAIR Data

- This module should provide you with the ability to work with your data in a FAIR manner
- Provide researchers with the proper information on FAIR data so that they can be submitted to the specified workflows and executions environments in a reproducible fashion

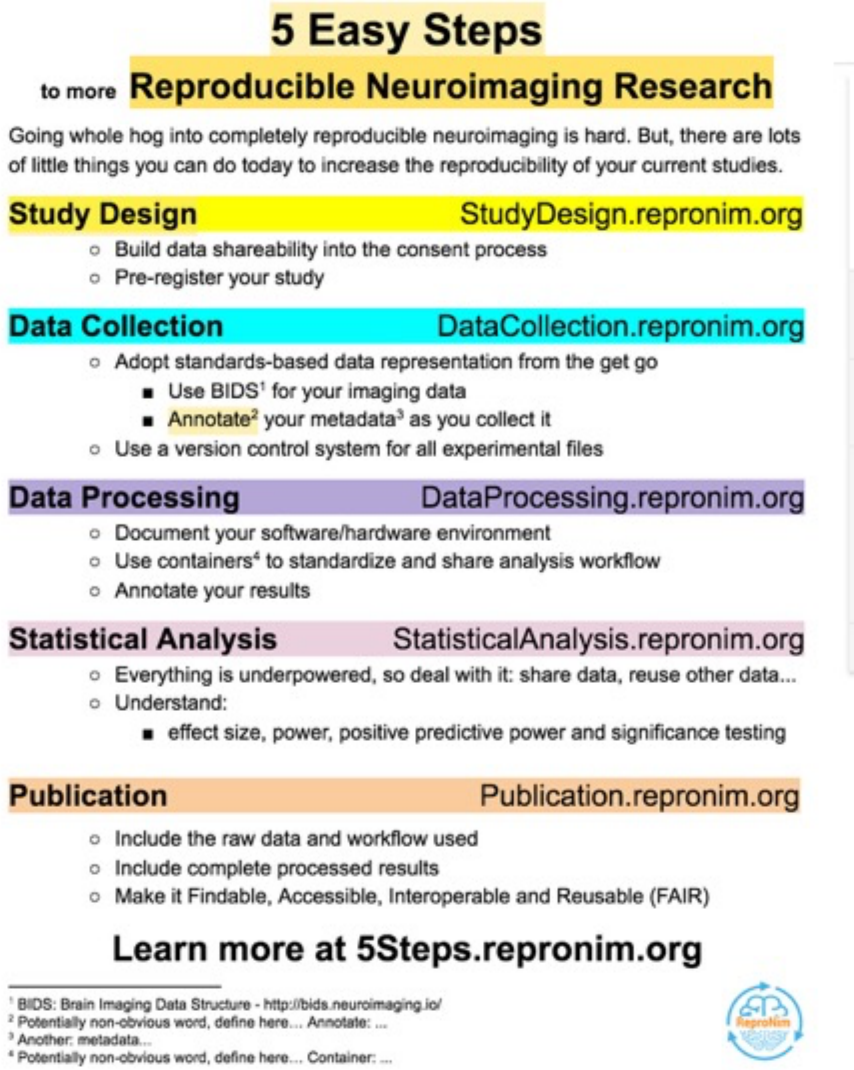
Stats

- Teach neuroimagers about the statistical aspects of reproducibility
- Have a collaborative training enterprise: you can improve this module if you know how to do a pull request or raise an issue on github:github.com/repronim/module-stat. See module 'the informatics basics of reproducibility (module 0) on how to do this.

- Identify simple things researchers WANT to do.
- Give them instructions for things that they will understand that they CAN do.

- Support/embrace incremental improvements!

- Think Locally; Act Globally
- Think reproducibly; Act re-executably



5 Easy Steps

to more Reproducible Neuroimaging Research

Going whole hog into completely reproducible neuroimaging is hard. But, there are lots of little things you can do today to increase the reproducibility of your current studies.

Study Design StudyDesign.repronim.org

- Build data shareability into the consent process
- Pre-register your study

Data Collection DataCollection.repronim.org

- Adopt standards-based data representation from the get go
  - Use BIDS[1] for your imaging data
  - Annotate[2] your metadata[3] as you collect it
- Use a version control system for all experimental files

Data Processing DataProcessing.repronim.org

- Document your software/hardware environment
- Use containers[4] to standardize and share analysis workflow
- Annotate your results

Statistical Analysis StatisticalAnalysis.repronim.org

- Everything is underpowered, so deal with it: share data, reuse other data...
- Understand:
  - effect size, power, positive predictive power and significance testing

Publication Publication.repronim.org

- Include the raw data and workflow used
- Include complete processed results
- Make it Findable, Accessible, Interoperable and Reusable (FAIR)

Learn more at 5Steps.repronim.org

[1] BIDS: Brain Imaging Data Structure - http://bids.neuroimaging.io/
[2] Potentially non-obvious word, define here... Annotate: ...
[3] Another: metadata...
[4] Potentially non-obvious word, define here... Container: ...

Schedule

8:30-10:00

FAIR Data - BIDS datasets - Jeffrey Grethe, UCSD

10:15-11:45

Computational basis - Yaroslav Halchenko, Dartmouth College and Michael Hanke, Magdeburg

13:00-14:30

Neuroimaging Workflows - Dorota Jarecka and Satrajit Ghosh, MIT

14:45-16:00

Statistics for reproducibility - Celia Greenwood and Jean-Baptiste Poline, McGill

YAHOO!