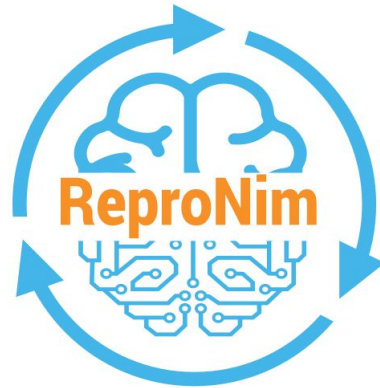# Coastal Coding

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…)
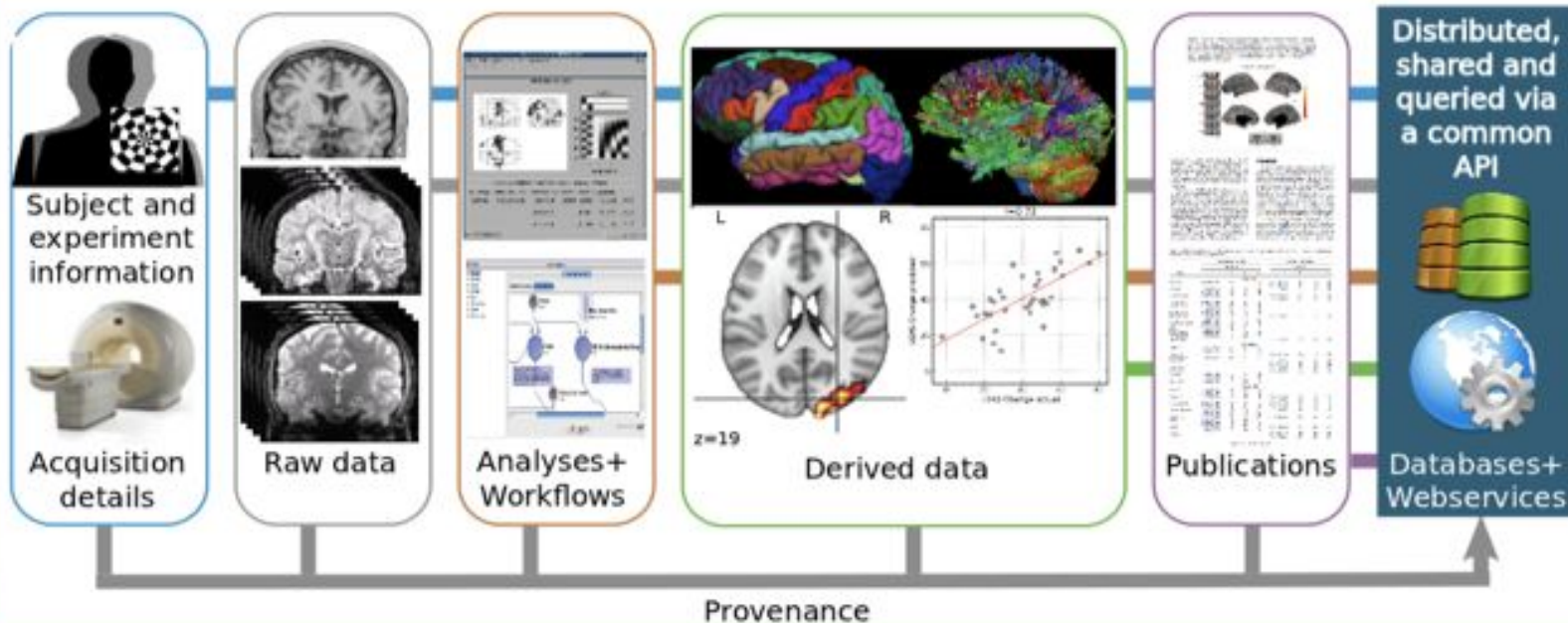
Progress in improving mental health outcomes has been slow

- 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; <u>a profusion of statistically significant, but minimally differentiating, biological findings; 'approximate replications' of these findings in a way that neither confirms nor refutes them</u>; 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
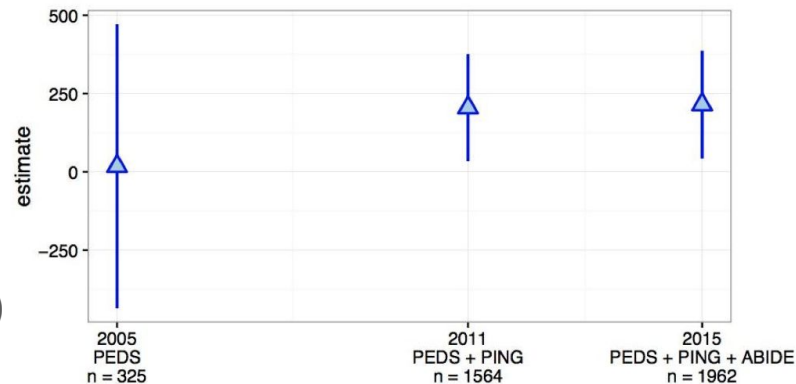
General Neuroimaging Workflow

Data Flow and Stages of Data Publishing Opportunities

Subject and experiment information

Acquisition details

Raw data

Analyses+ Workflows

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)
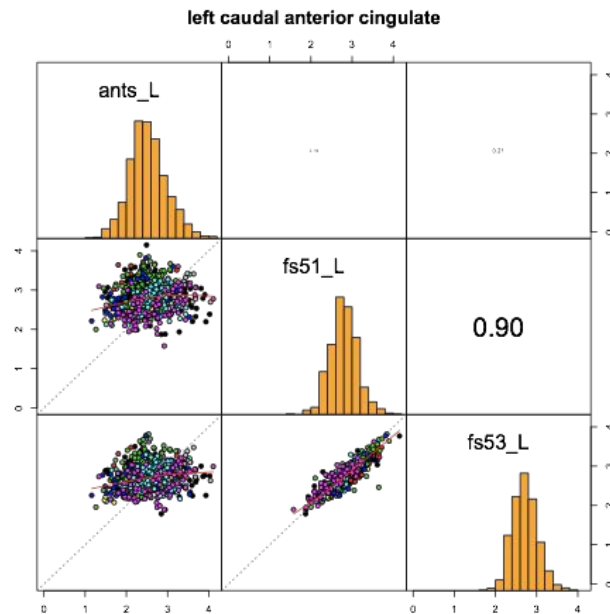
- Data Matters (100 subjects versus 1000 subjects)

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)

- 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: ...

ReproNim

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.