Achieving Reproducibility: Perspectives and New Tools

Professor Brian Wandell

Director, Stanford Center for Cognitive and Neurobiological Imaging (CNI)

Deputy Director, Wu Tsai Neurosciences Institute

Thanks to the Simons Foundation and the Arnold Foundation for their support.
Commercial disclosure

I am an active participant in Flywheel Exchange LLC, a commercial venture related to the work described here.
Early days of MRI - hardware oriented, little ability to do much with data management

1992

- MRI acquisition methods have become more complex
- MRI computations (reconstruction, data analysis) have become more complex
- Networking and computer technology have advanced
Implementing our MRI center made us think about data management
Motivation

Help people check and share their work to advance scientific understanding together.
**Replication**

Obtaining the data again, usually by independent investigators using similar methods, equipment and protocols

**Computational reproducibility**

Confirm the calculations (e.g., statistics, images, numerical relationships) calculated from existing data

**We can’t help people do this**

**We can help people do this**
Computational reproducibility ... must be designed into a project from the beginning. **One does need to develop a whole set of programming and research disciplines** with the end result in mind and stick with them.
Computational reproducibility: The MRI Center perspective

**User needs**
- Reproducible data and computation
- Tools, not rules

**What Flywheel built**
- Acquisition, metadata, visualization
- User-rights
- Search and data reuse
- Embedded Gears
- Collaborating
Data management is a pre-requisite to computational management are both important

- Most MRI Centers provide one of these data retrieval options from the MRI scanner
  - Copy data to CD, DVDs, USB hard drive
  - Copy to a server and remote login
- Data and metadata are transferred to a system controlled by a student or post-doc
- **Limitations** – reuse and sharing become burdensome; metadata and pre-processing information are frequently lost
Data management is a pre-requisite to computational management are both important

• Archiving MRI Center data eliminates the need for users to gather the data again for publication

• The data should be available through a platform-independent web browser to simplify access

• Basic tools, such as search, visualization, and pre-processing can be available through the browser

• The data are ready for sharing and reuse; metadata can be stored; pre-processing methods shared can be shared
Data Capture Architecture

- Data capture must classify and interpret many different MR data types
- Users ask to add legacy data, or to include PACS (medical) data
- Adding behavioral and physiological data is valuable for reproducibility
- Permitting custom MRI reconstruction methods has been important at several MRI Centers
Tools are better than rules

- Users have different preferences about how to prepare their data (pre-process) for detailed analysis.
- Project members change over time, so user rights must be easy to manage.
- Many users seek help with pre-processing, managing project notes, and data visualization.
- The Center and users have a shared interest in Quality Assurance (QA).
- Tools to help upload to an archival site (e.g., NDA, Journal) simplifies grant and publication compliance.
• Information about each project is organized for sharing and editing

• Users can be added or removed from the project with different permission levels

• Notes, papers, rules for processing the data can be specified for each project
Providing users with tools

- Many types of calculations can be installed and shared between users.

- Here is a list of the ‘Gears’ on my lab’s site that I share with Grill-Spector:

- There are dozens, including, FreeSurfer, FSL, HCP, tractography, file format (DCM), directory format (BIDS), and ...
Computational record keeping - Gears and their parameters

- The Gears you used and their parameters are stored in the Provenance tab.
- Here is a FreeSurfer run.
Taking advantage of cloud-scale computing

- **Cloud-scale** means you can select and run many Gears on different data sets at the same time.
- The cloud providers set up the machines upon request; they have a lot.
• **Cloud-scale** means you can select and run many Gears on different data sets at the same time.

• The cloud providers set up the machines upon request; they have a lot.
Computation: Exploration and methods development

- Language bindings (SDK)
  - Python
  - Matlab
  - R
- REST database endpoints
- Computational sharing
Computation: Exploration and methods development

- Language bindings (SDK)
  - Python
  - Matlab
  - R
- REST database endpoints
- Computational sharing
The SDK can be used with Jupyter Notebooks or Live Script to make.

This shows the Live Script interface to search, read data, perform analyses, store the analyses in the system.
Databases are designed for different purposes

- At time of publication, curated data may be placed on a shared site
- These sites add value by checking for quality, providing some pipelines
- They differ in many ways from our MRI Center goal – mainly they are not designed to manage ongoing projects, support a broad range of cloud computations
- These are valuable tools, but with another purpose

LORIS, COINS, XNAT, OTHERS ...
Statistical reproducibility differs from computational reproducibility

- Null hypothesis statistical testing and group comparisons are deeply ingrained in the neuroscience publication culture, particularly cognitive and clinical neuroscience.

- Our tools are for many types of computations, not the final statistical hypothesis testing.
Flywheel is an implementation of the FAIR principles for reproducible research in MRI

The FAIR principles are a set of community-developed guidelines to ensure that data or any digital object are

- Findable
- Accessible
- Interoperable
- Reusable

We are trying to support computational reproducibility following the FAIR principles