USC University of Southern California

Reliability and Reproducibility in Neurorehabilitation Research

Sook-Lei Liew, PhD, OTR/L

University of Southern California

James Finley, PhD

University of Southern California

Keith Lohse, PhD, PStat

University of Utah

ASNR

October 16, 2019

Outline

Reproducible Science

- What's the problem?
- What can be done?
    - Data Management
    - Data Analysis
    - Data Visualization

- (Bonus topic: Open science!)

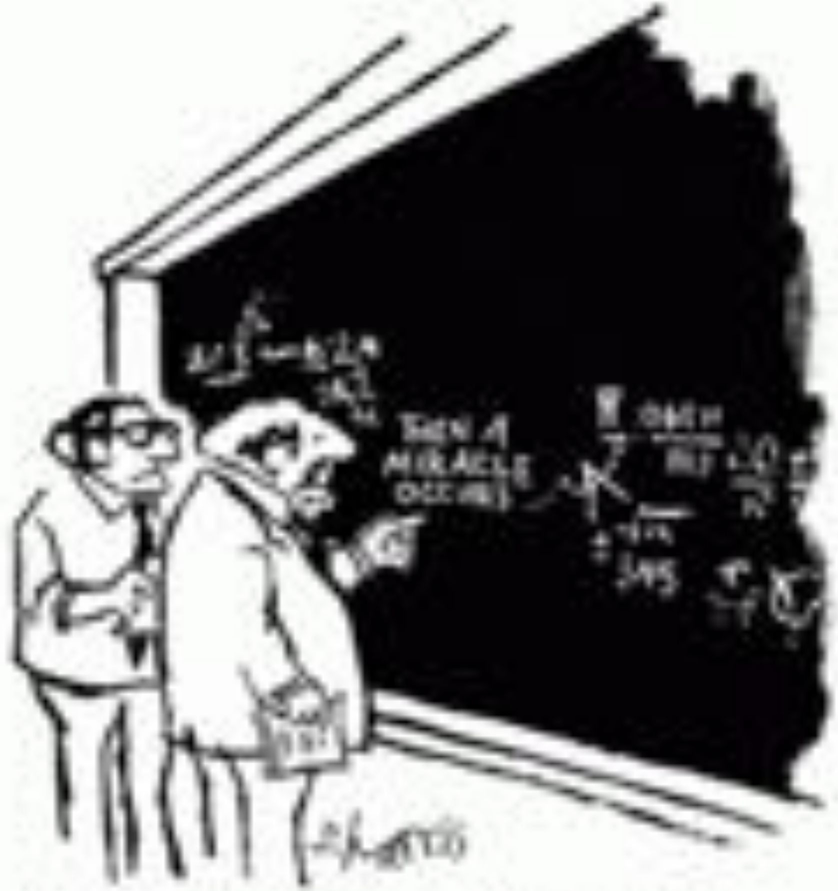
Reproducible Science

What is reproducibility?

The ability for someone else (or yourself) to reproduce an entire experiment and results

What is reliability?

The degree to which the result of a measurement, *calculation*, or specification can be depended on to be accurate.



"I think you should be more explicit here in step two."

The Reproducibility Crisis

- What is the reproducibility crisis?

- Psychology – only 39 of 100 replication attempts were successful
  - https://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248

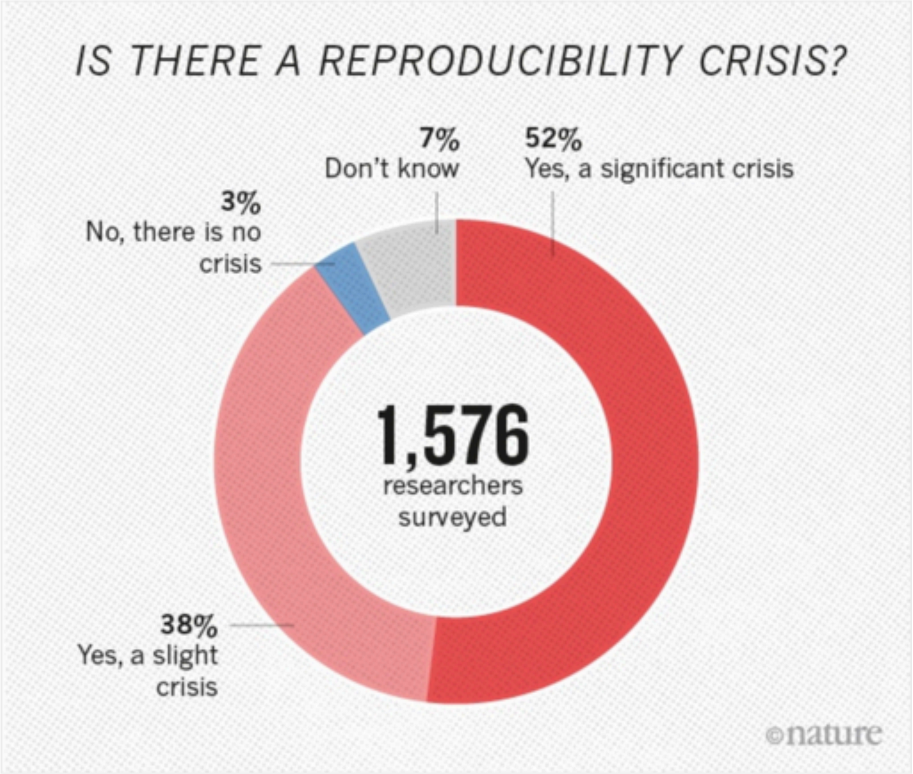
- More than 70% of scientists have tried and failed to reproduce another scientist’s experiments:
  - https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Contributing Factors

1. Positive publication bias
2. Logistical limitations – limited money, time leading to underpowered samples
3. Heterogeneity and random variance in the sample that provides a positive effect (e.g., a few very special participants)
4. Normal human errors

Clinical Implications of Irreproducibility

- Small studies with promising results → Large (expensive, time-consuming) studies with null results
- Wasted time/hope for patients
- Setting back science…



What Can Be Done?

- Collecting really large heterogeneous datasets, either prospectively at one site:
    - UK Biobank
    - All of Us
    - Human Connectome Project

- Or retrospectively pooling data across sites:
    - Consortiums or multi-site experiments where experiments can be replicated across sites
    - ADNI (Alzheimer’s Disease Neuroimaging Initiative)
    - ABIDE
    - 1000 Functional Connectomes
    - ENIGMA

What Can Be Done?

An experiment can only be reproduced if all the details of the experiment are properly recorded and shared. It's everyone's job!

1. **Data Management**
   - Managing your own data
   - Managing and compiling data from others

2. **Data Analysis**
   - Simulating, checking, processing, and analyzing data

3. **Data Visualization**
   - From quality assurance to full papers, showing individual and group results

4. **Reproducible Papers** (Jupyter Notebook, R Markdown, Matlab Markup)

4. Reproducible Papers

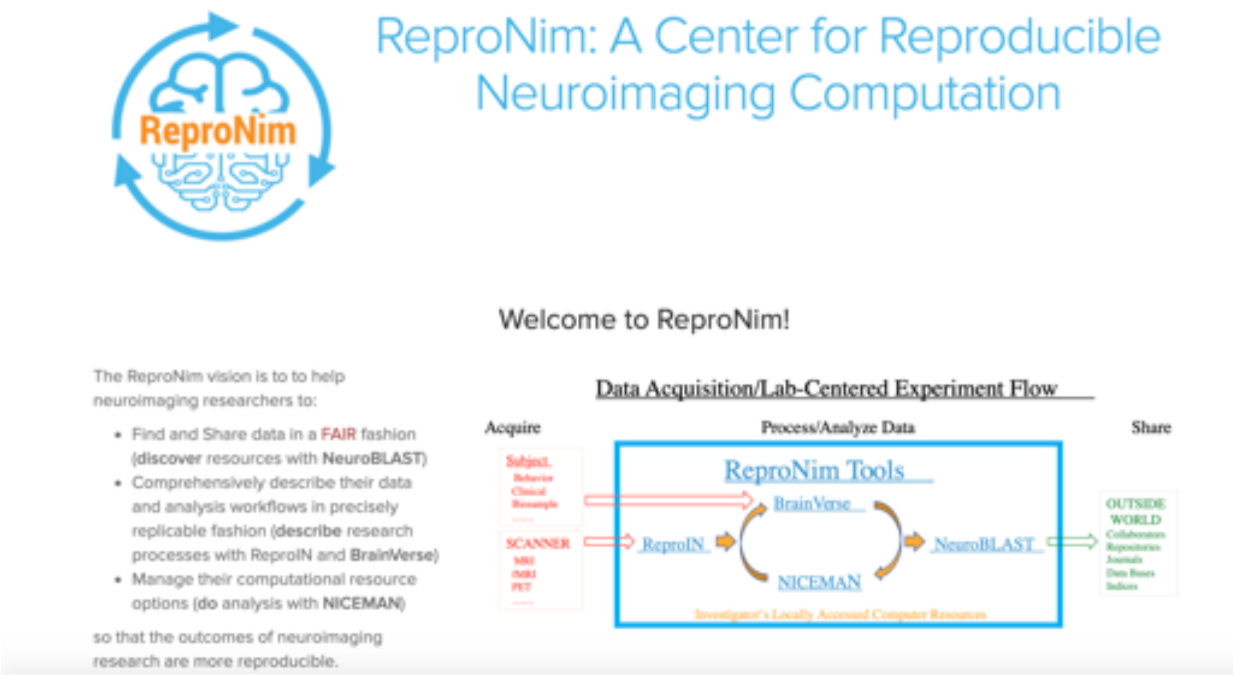
An experiment can only be reproduced if all the details of the experiment are properly recorded and shared

- At the very least, report:
  - Actual values (means +/- SD)
  - Effect sizes (cohen’s d, f, hedge’s g for smaller samples)

- But better yet is sharing the raw data so people can analyze it for themselves.
- Doing so allows others to replicate your results and power their studies based on your results

4. Reproducible Papers

- Using a tool such as Jupyter Notebook or R Markdown, you can provide the code for analyses, as well as interactive results
- Reproducible Neuroimaging: http://www.reproducibleimaging.org/



ReproNim: A Center for Reproducible Neuroimaging Computation

Welcome to ReproNim!

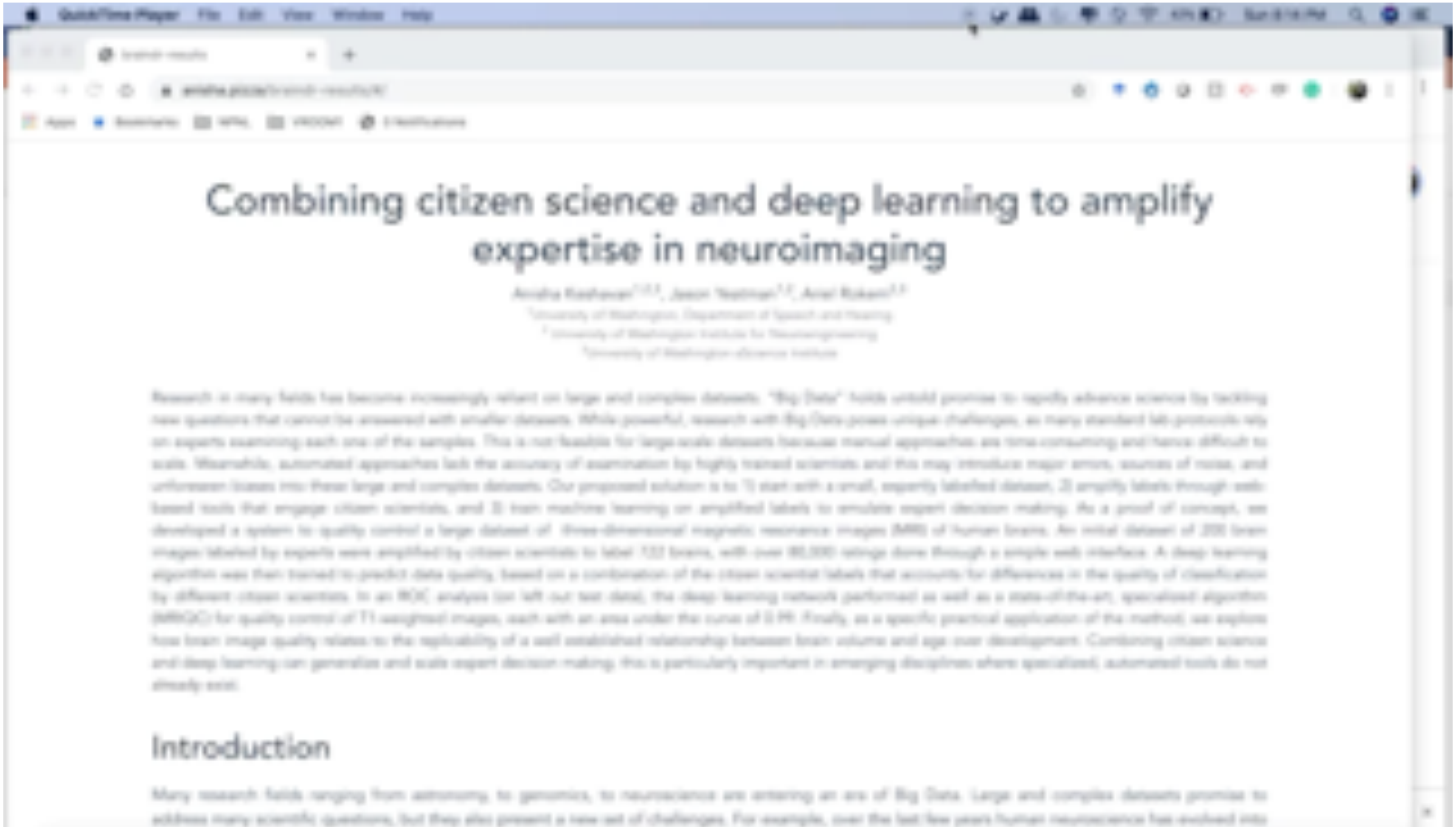
The ReproNim vision is to to help neuroimaging researchers to:

- 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.

4. Reproducible Papers

Example: Keshavan et al., 2019: braindr.results.us


QuickTime Player File Edit View Window Help
braindr results
anisha.pizza/braindr-results/#/
Apps
Bookmarks
3 Notifications

Combining citizen science and deep learning to amplify expertise in neuroimaging

Anisha Keshavan[1,2,3], Jason Yeatman[1,2], Ariel Rokem[2,3]

[1]University of Washington, Department of Speech and Hearing

[2] University of Washington Institute for Neuroengineering

[3]University of Washington eScience Institute

Research in many fields has become increasingly reliant on large and complex datasets. "Big Data" holds untold promise to rapidly advance science by tackling new questions that cannot be answered with smaller datasets. While powerful, research with Big Data poses unique challenges, as many standard lab protocols rely on experts examining each one of the samples. This is not feasible for large-scale datasets because manual approaches are time-consuming and hence difficult to scale. Meanwhile, automated approaches lack the accuracy of examination by highly trained scientists and this may introduce major errors, sources of noise, and unforeseen biases into these large and complex datasets. Our proposed solution is to 1) start with a small, expertly labelled dataset, 2) amplify labels through web-based tools that engage citizen scientists, and 3) train machine learning on amplified labels to emulate expert decision making. As a proof of concept, we developed a system to quality control a large dataset of three-dimensional magnetic resonance images (MRI) of human brains. An initial dataset of 200 brain images labeled by experts were amplified by citizen scientists to label 722 brains, with over 80,000 ratings done through a simple web interface. A deep learning algorithm was then trained to predict data quality, based on a combination of the citizen scientist labels that accounts for differences in the quality of classification by different citizen scientists. In an ROC analysis (on left out test data), the deep learning network performed as well as a state-of-the-art, specialized algorithm (MRIQC) for quality control of T1-weighted images, each with an area under the curve of 0.99. Finally, as a specific practical application of the method, we explore how brain image quality relates to the replicability of a well established relationship between brain volume and age over development. Combining citizen science and deep learning can generalize and scale expert decision making; this is particularly important in emerging disciplines where specialized, automated tools do not already exist.

Introduction

Many research fields ranging from astronomy, to genomics, to neuroscience are entering an era of Big Data. Large and complex datasets promise to address many scientific questions, but they also present a new set of challenges. For example, over the last few years human neuroscience has evolved into

A few notes on reproducibility…

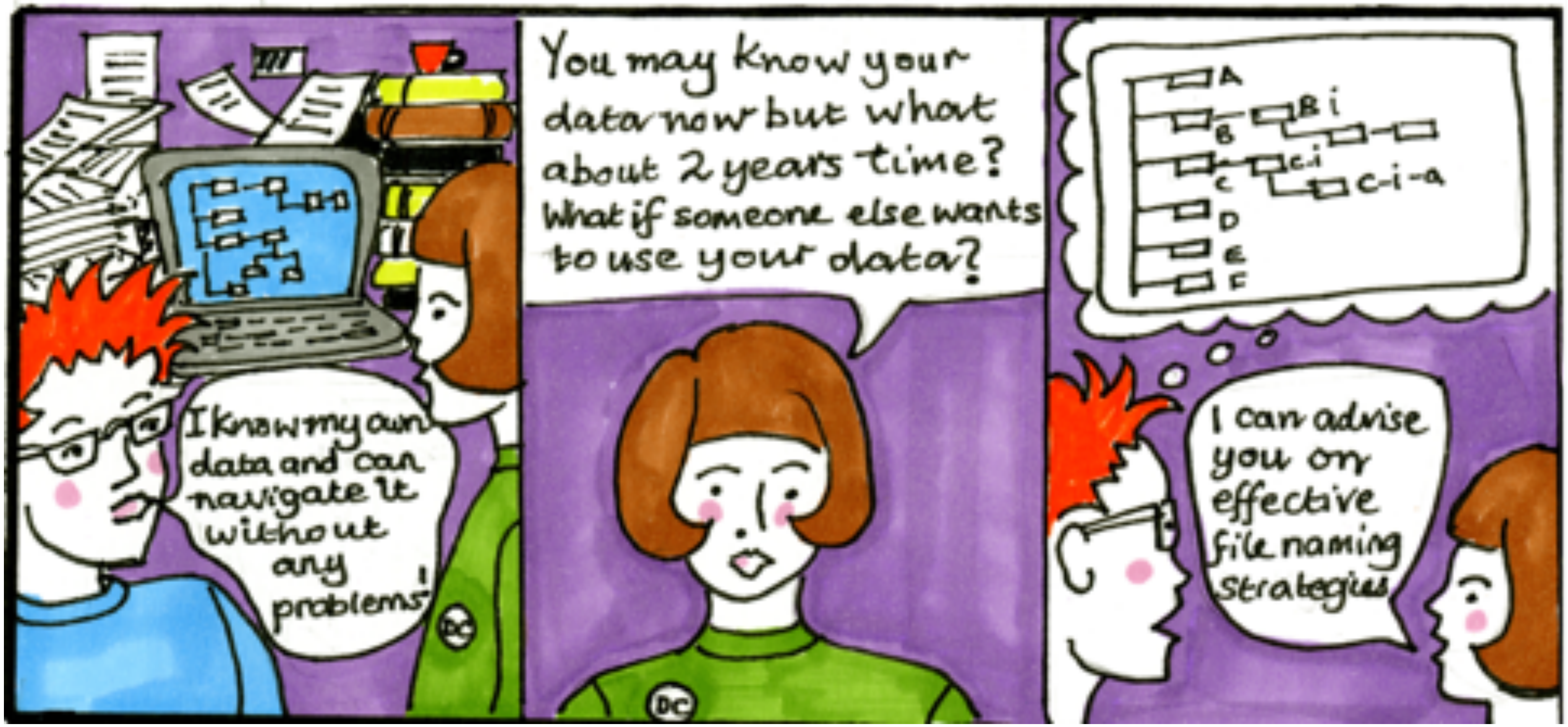
- Emphasis on reproducibility is growing!
- Reproducibility awards and competitions
- Emphasis on sharing data for reproducible papers (will discuss more in open science next)
- Efforts for pre-registered reports: https://www.journals.elsevier.com/cortex/news/registered-reports-a-new-article-format-from-cortex
- And growing acceptance of null results as just as valid as positive results!

Data Management

Data Management

1. You should plan to keep your data in a format that can be (easily) shared with anyone at any time.

2. This means thinking about your data and analyses BEFORE you collect it. (And evaluating DURING).


I know my own data and can navigate it without any problems!
You may know your data now but what about 2 years time? What if someone else wants to use your data?
A
B
Bi
C
ci
c-i-a
D
E
F
I can advise you on effective file naming strategies

By: Clare Trowell

Data Management

Today we'll cover two types of data management:

1) Data that is yours:

1. Recording your experimental protocol
2. File and naming conventions
3. Recording meta-data

2) Combining data from others:

1. Database tools
2. Data processing pipelines
3. Tools for (reproducibly) checking and manipulating data

Data Management

Today we’ll cover two types of data management:

1) Data that is yours:

   1. Recording your experimental protocol
   2. File and naming conventions
   3. Recording meta-data

2) Combining data from others:

   1. *Database tools*
   2. *Data processing pipelines*
   3. *Tools for (reproducibly) checking and manipulating data*

I, too, was young once.

And I wish I had done better with data management!

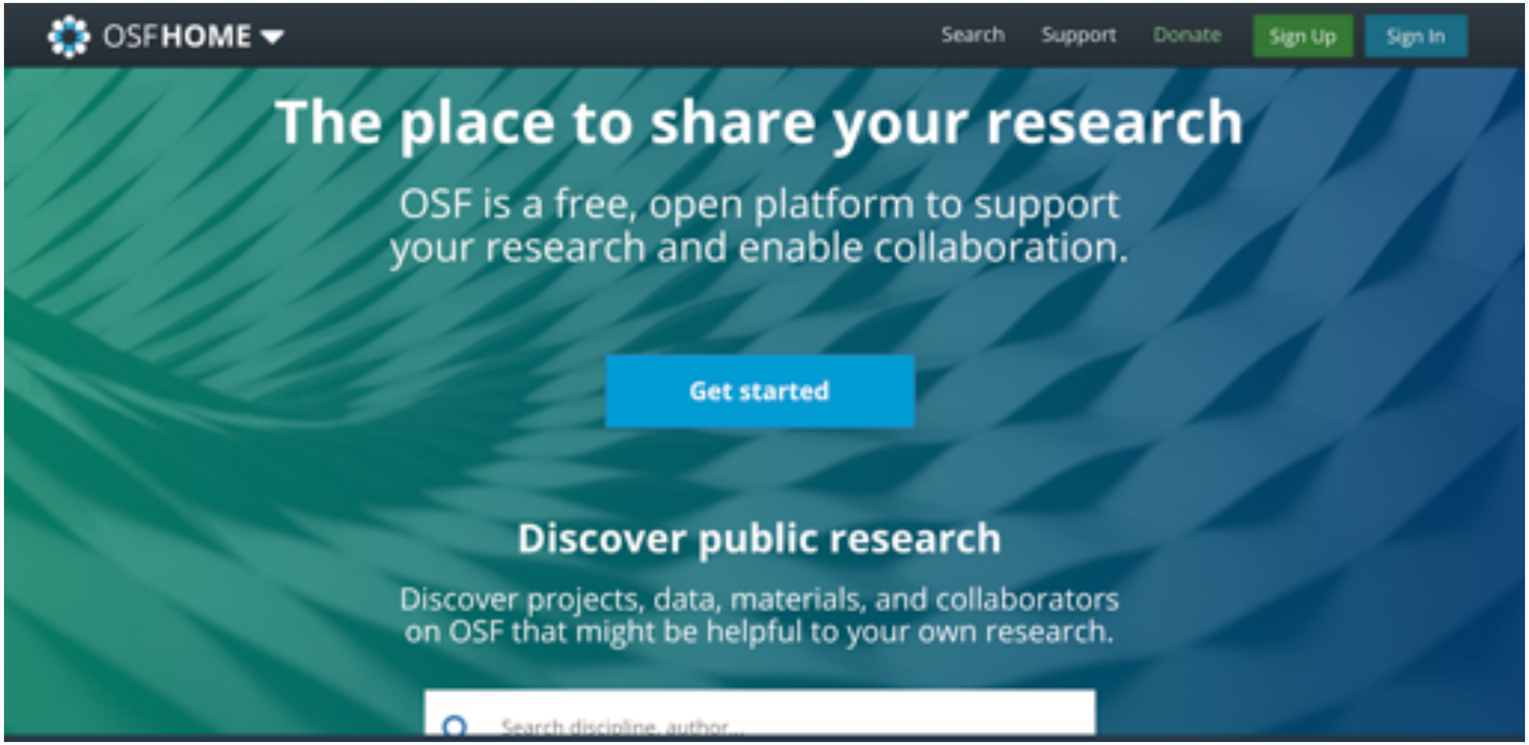
1. Experimental Protocols

How do you currently keep track of your experiments?

- Record a “brief” protocol with an overview of all steps, including IRB used, consent, general experimental steps, subject payment amount, etc.
- Record a “detailed” protocol with every step, including instructions to participants
- Have a second person review the detailed protocol with you to ensure they can replicate it as detailed
- Keep this somewhere accessible (e.g., lab google drive)

1. Experimental Protocols

Consider an organizational framework, such as a lab google drive with all the necessary components for each project and a similar format per project, or Open Science Framework (https://osf.io/dashboard)



The place to share your research

OSF is a free, open platform to support your research and enable collaboration.

Discover public research

Discover projects, data, materials, and collaborators on OSF that might be helpful to your own research.

2. Naming and File Structures

An experiment can only be reproduced if all the details of the experiment are properly recorded and shared

Basics

- Keep all file names machine readable!
  - Alphanumeric (no spaces, slashes or symbols) – no initials
    - Consistent capitalization
    - Subj01, subj01, subj_01, Subj_01, SUBJ01…. – Pick one and stick with it
  - Consistent number of digits - that means anticipating how many subjects, groups, etc. (e.g., padded 0s; subj01, subj02..)

- How to write dates so it becomes logically organized: 20191016

- If you have multiple timepoints on the same day, add time in HHMMSS if needed (military format): 20191016_133402

2. Naming and File Structures

An experiment can only be reproduced if all the details of the experiment are properly recorded and shared

Basics

- Create templates of subject folders to copy and populate

- After the first subject's data is collected, analyze all data and refine file structures

- Use a consistent planned naming convention with no spaces and few symbols so it can be easily machine-sorted if needed:
  - sub-c01g01s001-t01

    - This way you can easily extract just: *c01*, *g01*, *t01*

2. Naming and File Structures

Basics

- Consistent file structure for each subject:
  - studyFolder
    - subjFolder
      - subjData
      - subjAnalyses
      - subjResults
      - old
    - groupFolder
      - groupData
      - groupAnalyses
      - groupResults
      - old
    - writeups
      - 20180925_draft
  - strokeStudy
    - subj01
      - data
      - analyses
      - results
      - old
    - group
      - data
      - analyses
      - results
      - old
    - writeups
      - 20180925_draft
- For neuroimaging - see BIDS format (https://bids.neuroimaging.io/)

3. Meta-Data

- Useful to also have a “meta-data” file that describes what the various analyses are, if subgroup analyses were performed, what they were and why, etc.
  - Sex: 1=female, 2=male
  - FMUE is out of 60, not 66 points (left out reflexes)
  - 9 hole peg is normalized to other hand

- Think of this as your detailed methods section so you can revisit it 5 years later and make sense of what you did and why

- Keep track of this with the experimental protocol and in the data file – will also allow you to easily combine data with other projects or other collaborators

Data Management

Today we'll cover two types of data management:

1) Data that is yours:

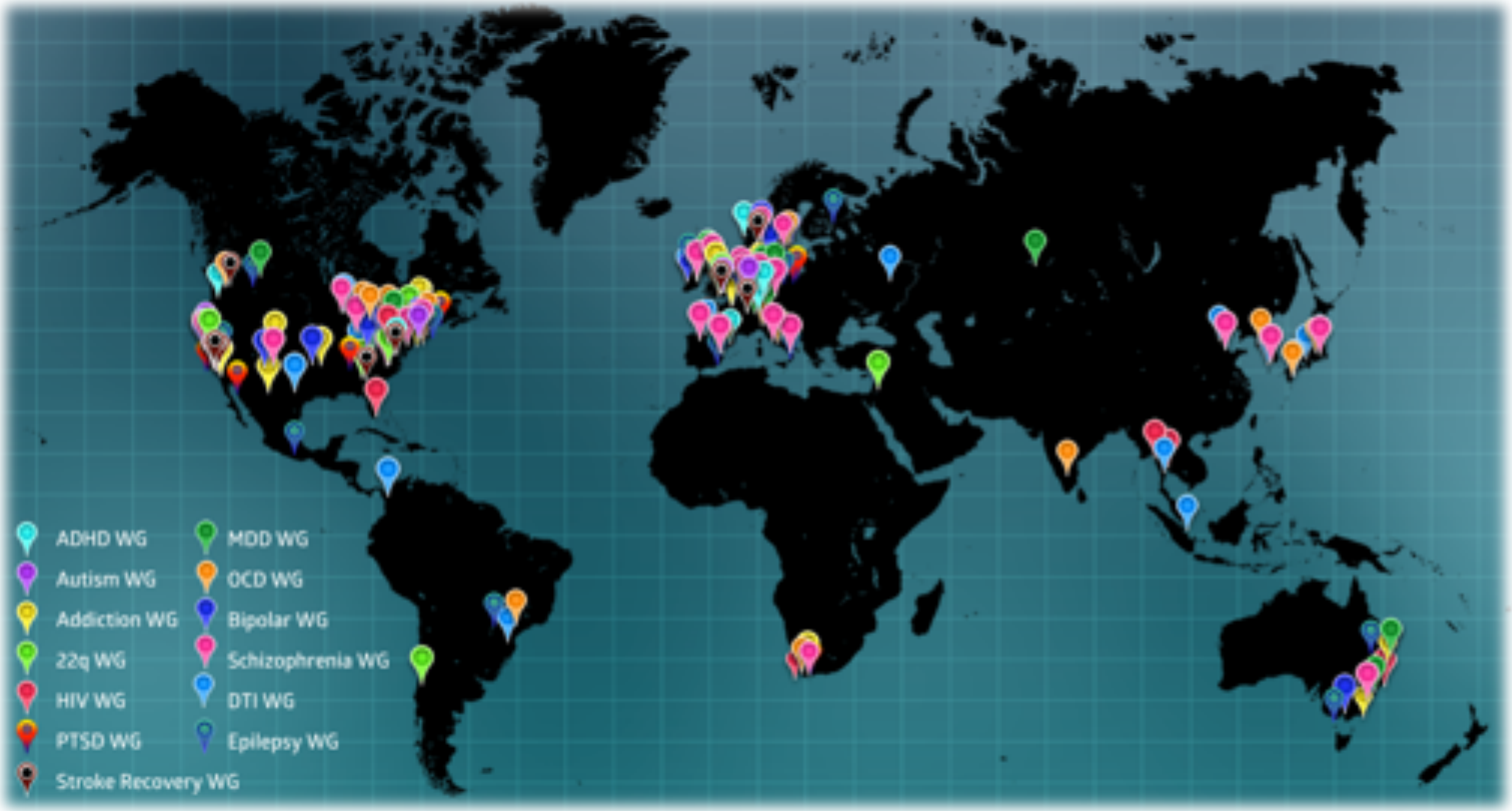
1. *Recording your experimental protocol*
2. *File and naming conventions*
3. *Recording meta-data*

2) Combining data from others:

1. Database tools
2. Data processing pipelines
3. Tools for (reproducibly) checking and manipulating data

ENIGMA Consortium

Largest-Ever Worldwide Analysis of Brain Scans and Genetic Data
53,000+ people from over 35 countries studying 18 brain diseases

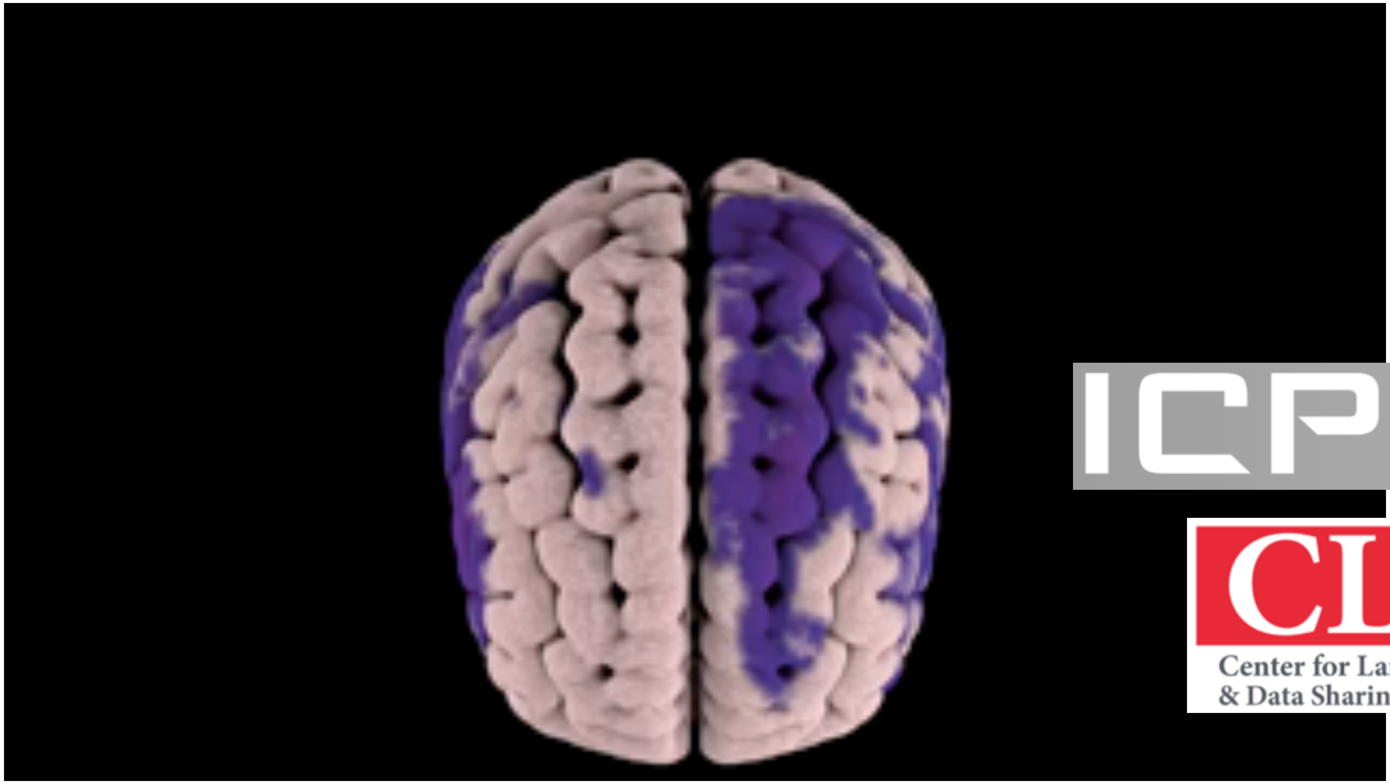


http://enigma.ini.usc.edu

Shared Dataset of Stroke MRIs and Lesion Masks

ATLAS: Anatomical Tracings of Lesions After Stroke

Open-source dataset of 304 stroke T1-weighted MRIs with manually hand-traced lesion masks, along with neuroradiology evaluations and meta-data (Liew et al., 2018, Nature *Scientific Data, https://www.nature.com/articles/sdata201811*)



ICPSR

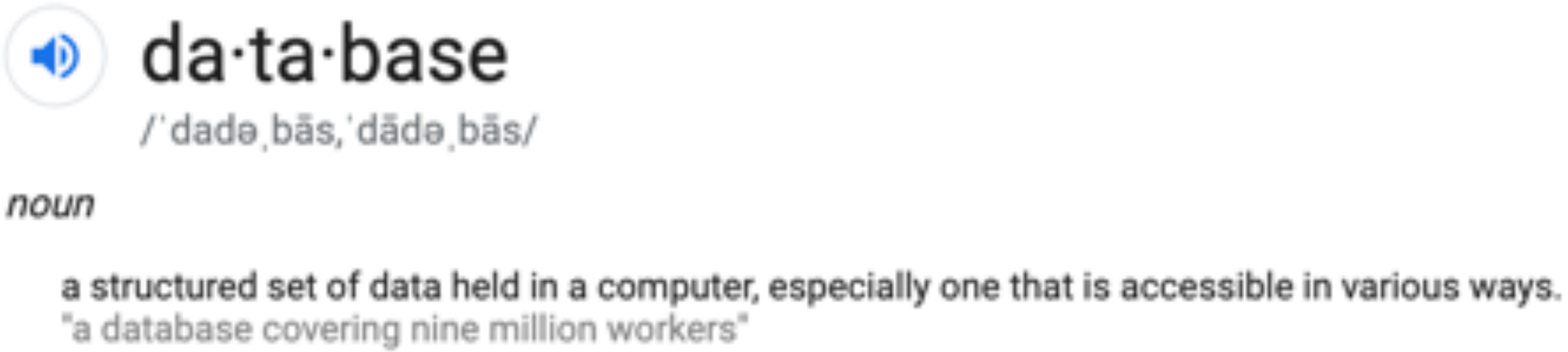
CLDR

Center for Large Data Research
& Data Sharing in Rehabilitation

https://www.youtube.com/watch?v=Ag5CUsRNY9Q

Data Management

Data versus a database



da·ta·base

/ˈdadəˌbās,ˈdādəˌbās/

noun

a structured set of data held in a computer, especially one that is accessible in various ways.

"a database covering nine million workers"

Most data is (hopefully) stored in a database

Database Management

1. CSV file
   - Can be read into Matlab, R, Python, etc.
   - Matlab: Data structure
   - R/Python: Data frame

2. If higher dimensional data (e.g., you have multiple CSV files (or excel sheet tabs)), consider a relational database, such as SQL (or SQLite)
   - Tables for subject demographics, behavioral measures, brain imaging measure, etc.
   - Fast queries, non-redundant fields
   - For many columns, it’s a little more human readable

Database Management

Common data elements: What are we going to record, and how are we going to record it?

- What are we going to record? Requires consensus
  - Stroke Recovery and Rehabilitation Roundtable (SRRR)

- How are we going to record it? Requires forethought
  - Time since stroke (days, months, years…)
  - Units (metric, not-metric)
  - Default to the highest resolution where possible

Database Steps

Key aspects of database management (when receiving data from others)

1. Data schema (organization)

2. Data scrubbing

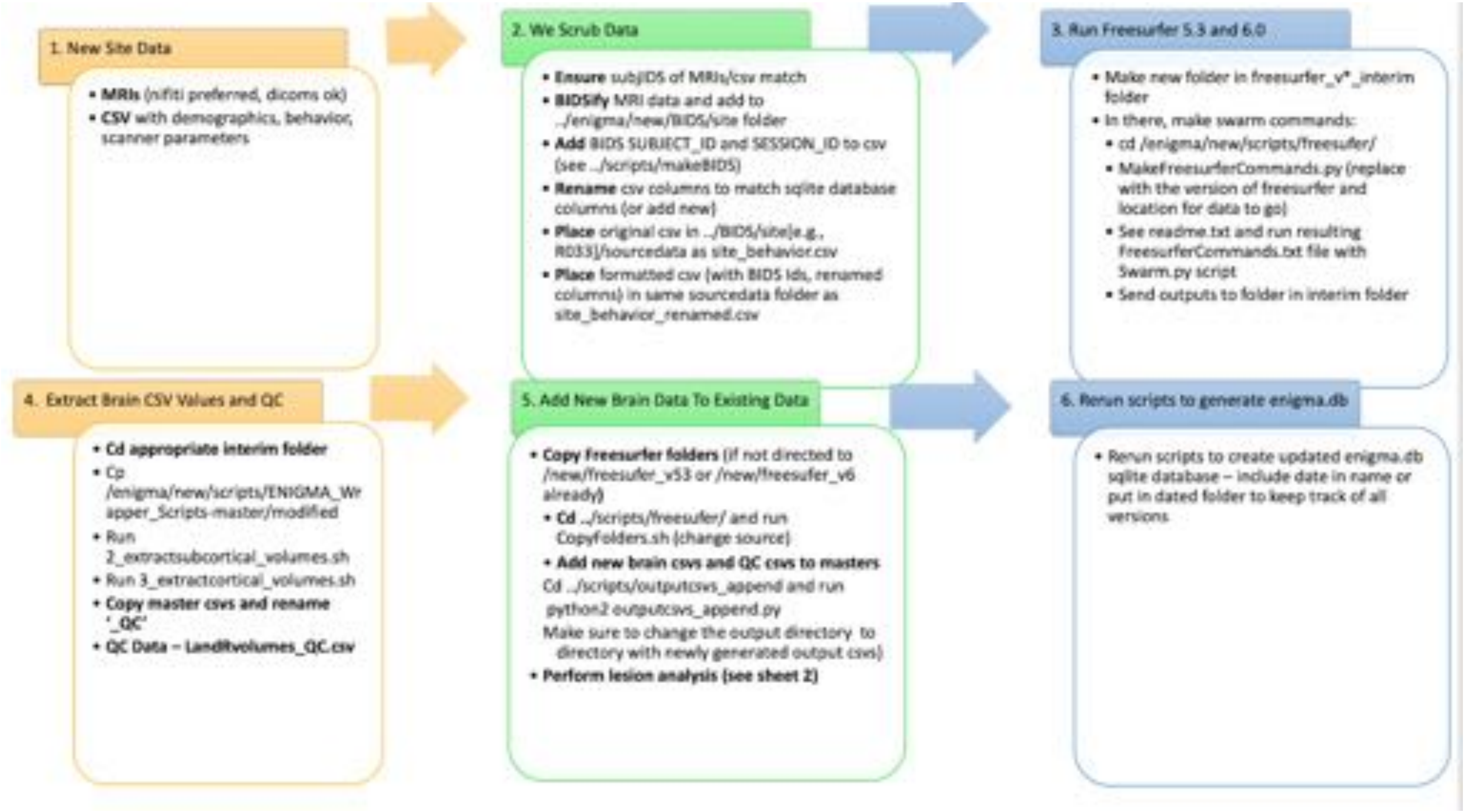
3. Data scrubbing

4. Data re-organization and data scrubbing

…

99. Data analysis

Database Management



1. New Site Data

- **MRIs** (nifiti preferred, dicoms ok)
- **CSV** with demographics, behavior, scanner parameters

2. We Scrub Data

- **Ensure** subjIDS of MRIs/csv match
- **BIDSify** MRI data and add to ../enigma/new/BIDS/site folder
- **Add** BIDS SUBJECT_ID and SESSION_ID to csv (see ../scripts/makeBIDS)
- **Rename** csv columns to match sqlite database columns (or add new)
- **Place** original csv in ../BIDS/site[e.g., R033]/sourcedata as site_behavior.csv
- **Place** formatted csv (with BIDS Ids, renamed columns) in same sourcedata folder as site_behavior_renamed.csv

3. Run Freesurfer 5.3 and 6.0

- Make new folder in freesurfer_v*_interim folder
- In there, make swarm commands:
  - cd /enigma/new/scripts/freesufer/
  - MakeFreesurferCommands.py (replace with the version of freesurfer and location for data to go)
  - See readme.txt and run resulting FreesurferCommands.txt file with Swarm.py script
  - Send outputs to folder in interim folder

4. Extract Brain CSV Values and QC

- **Cd appropriate interim folder**
- Cp /enigma/new/scripts/ENIGMA_Wrapper_Scripts-master/modified
- Run 2_extractsubcortical_volumes.sh
- Run 3_extractcortical_volumes.sh
- **Copy master csvs and rename '_QC'**
- **QC Data – LandRvolumes_QC.csv**

5. Add New Brain Data To Existing Data

- **Copy Freesurfer folders** (if not directed to /new/freesufer_v53 or /new/freesufer_v6 already)
  - **Cd** ../scripts/freesufer/ and run CopyFolders.sh (change source)
  - **Add new brain csvs and QC csvs to masters**

  Cd ../scripts/outputcsvs_append and run python2 outputcsvs_append.py

  Make sure to change the output directory to directory with newly generated output csvs)
- **Perform lesion analysis (see sheet 2)**

6. Rerun scripts to generate enigma.db

- Rerun scripts to create updated enigma.db sqlite database – include date in name or put in dated folder to keep track of all versions

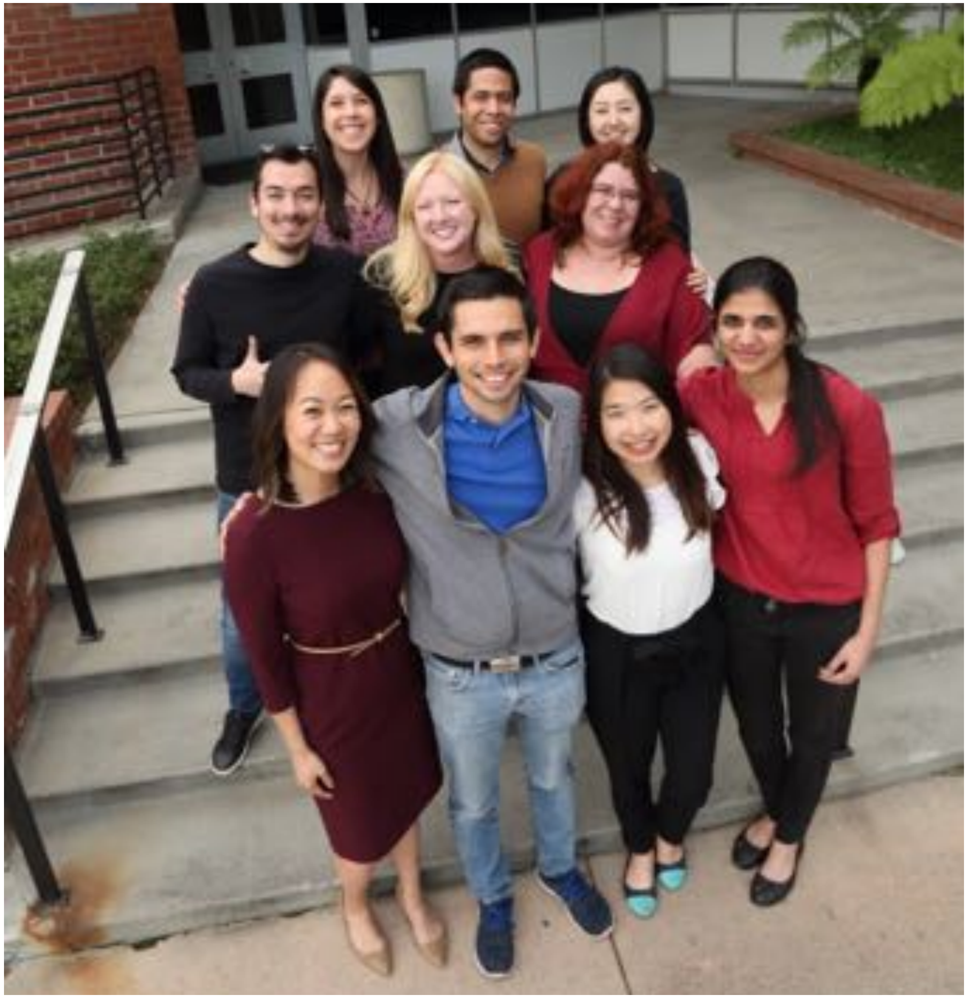
Hands-On Database Tools

- SQLite (walk-through demo)
- Python with PANDAS to manipulate csv files in jupyter notebook (hands on)

Thank you!

The Neural Plasticity and Neurorehabilitation Laboratory

http://npnl.usc.edu



Email me:

sliew@usc.edu

Twitter us:

@NPNLatUSC

DISCOVER INNOVATE TRANSITION ARL

NIH

American Heart Association

CLDR

Center for Large Data Research & Data Sharing in Rehabilitation

Bonus Topic!

- Open Science
  - What is it?
  - Why share?
  - Resources

What is Open Science?

- **Open science** is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional.

- Encompasses:
  - Open data – shared data
  - Open source code – shared code and analyses
  - Open methods – shared experimental paradigms
  - Open peer review – comments and responses posted
  - Open access – access to journal articles
  - Open educational resources – online tutorials, etc.

Why Open Science?

- **Why not share:**
  - Someone else could steal my ideas and publish them before me!
  - I worked hard to collect this data!
  - I worked hard to create this code!

But… open science doesn’t mean you give everything up and get nothing in return!

Why Open Science?

- **Why share:**
  - You can publish a paper on your data or code and get cited for creating these valuable resources
  - Someone else could find new and exciting findings from the data you collected, or build on your ideas, or make use of the code you worked so hard on
  - You can also find new collaborations to keep building on your current work
  - You are contributing not only to your own scientific career but to science in general!

What Can I Share?

Open data – shared data

- Nature's Scientific Data, GigaScience Data Reports, and others are highly cited

- Many journals now request data shared with your paper submission (hence the repro steps before)

- Many grants now mandate that you share your data on a site, such as github, figshare, etc.

- There are many repositories for archiving data – e.g., ICPSR, FCP/INDI
- See: https://www.nature.com/sdata/policies/repositories

What Can I Share?

Open source code – shared code and analyses

- Github!
  - This allows others to report bugs and recommend enhancements
  - Also allows you to track users of your data
  - Also allows others to contribute to your code and analyses – to make it better
    - The open science community is usually friendly and might recommend a pull to help you out!
    - You can even hire someone via bitcoin bids to help you with your code :D
- https://guides.github.com/activities/hello-world/
- https://product.hubspot.com/blog/git-and-github-tutorial-for-beginners

What Can I Share?

Open methods – shared experimental paradigms

- If code, also can be done via Github!
- Provide your exact paradigm so others can try to replicate it in different contexts
- Provide your tools, substrates, etc. so others can try it at their sites (→ usually involves some data or material transfers via your local tech transfer office)

What Can I Share?

Open peer review – comments and responses posted

- Preprint archives such as bioRxiv: https://www.biorxiv.org

- Allows for comments from the community, twitter and views metrics (for cover letters), and uploads of revised versions; will linked to published version later

- Possible to get open peer review and point editors to this (but still beta)

  - https://www.biorxiv.org/content/10.1101/441451v1?versioned=true

What Can I Share?

Open access and open educational resources

- Consider publishing in open access journals
- Consider making your tutorials videotaped and shared on youtube
- I just learned that Quicktime lets you record your screen plus audio easily :D https://support.apple.com/guide/quicktime-player/record-your-screen-qtp97b08e666/mac
- Watch out for this as it will grow in the future!

A few other notes…

- Some people may still try to misuse your data or be entitled, so it’s important to have adequate protections over it (e.g., an education license for any code shared on github)

- You don’t actually own your work (your university does) so you do need to disclose any work you create

- Open science is growing and always changing

    - Exciting to get into it now because it’s likely to become mandated soon :D

    - It’s a great community full of supportive people ☺

Some Resources

- http://www.reproducibleimaging.org/module-reproducible-basics/
- https://software-carpentry.org/
- https://datacarpentry.org/