ReproRehab

Week 2:
Introduction to Reproducible Data Analysis and Data Management

Sook-Lei Liew, PhD, OTR/L
Associate Professor & Director, Neural Plasticity and Neurorehabilitation Lab
Chair, ENIGMA Stroke Recovery Working Group
University of Southern California
sliew@usc.edu | https://chan.usc.edu/npnl/

This week's topics

- Data management for rehabilitation research (Dr. Sook-Lei Liew)
- Reproducibility in data analysis (Dr. James Finley)
- Data standardization (Dr. JB Poline)*
- *Optional (but highly recommended)*: Overview of DataLad, aka, the gloriousness that could be!

*Note that we are borrowing heavily from reproducibility lessons gleaned from the neuroimaging community – we'd love to implement many of these principles in rehabilitation research! Related, HCP = Human Connectome Project, UKBiobank = UK Biobank – examples of huge neuroimaging datasets

**Note that because there is no hands-on programming yet, the content is a little longer than it will be for most weeks

Refresher on reproducible science

1. **What is the "reproducibility crisis"?**

   Do you think that scientific reproducibility and replicability is a problem in stroke research?

2. **How can we use data science to address reproducibility?**

3. **How can we use open science to address replicability? (next week)**

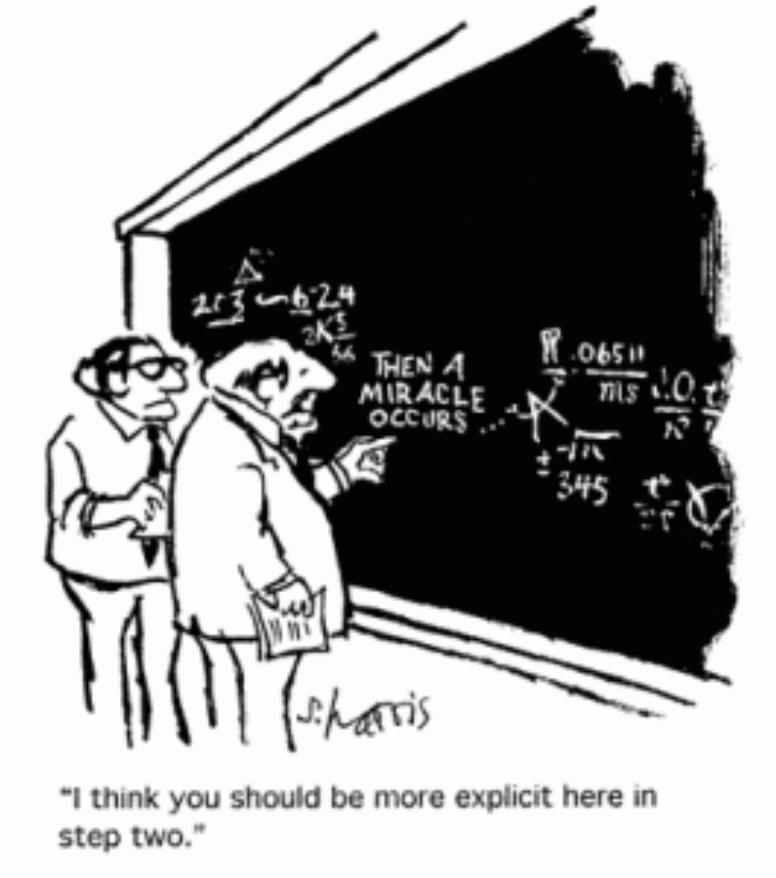
USC University of Southern California

Scientific reproducibility and replicability

Reproducibility: The ability for someone else (or yourself) to reproduce an experimental paradigm

Replicability: The ability for someone else (or yourself) to obtain consistent results, given the same experiment

1. If I read a paper, is there sufficient detail for me to implement the same experiment?
2. If I implement someone else's experiment, will I get the same results?



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

USC University of Southern California

What is the reproducibility crisis?

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

Factors contributing to the problem

Methods (Reproducibility)

- Underutilized reproducible methods:
  - Human error in manual processes (data entry, analysis)
  - Inconsistent keeping record across different team members

Results (Replicability)

- Positive publication bias
- Logistical limitations:
  - Limited money, time, and participant availability can lead to biased and underpowered samples

Potential solutions

Methods (Reproducibility) → **Data Science**

- Underutilized reproducible methods:
  - Human error in manual processes (data entry, analysis)
  - Inconsistent keeping record across different team members

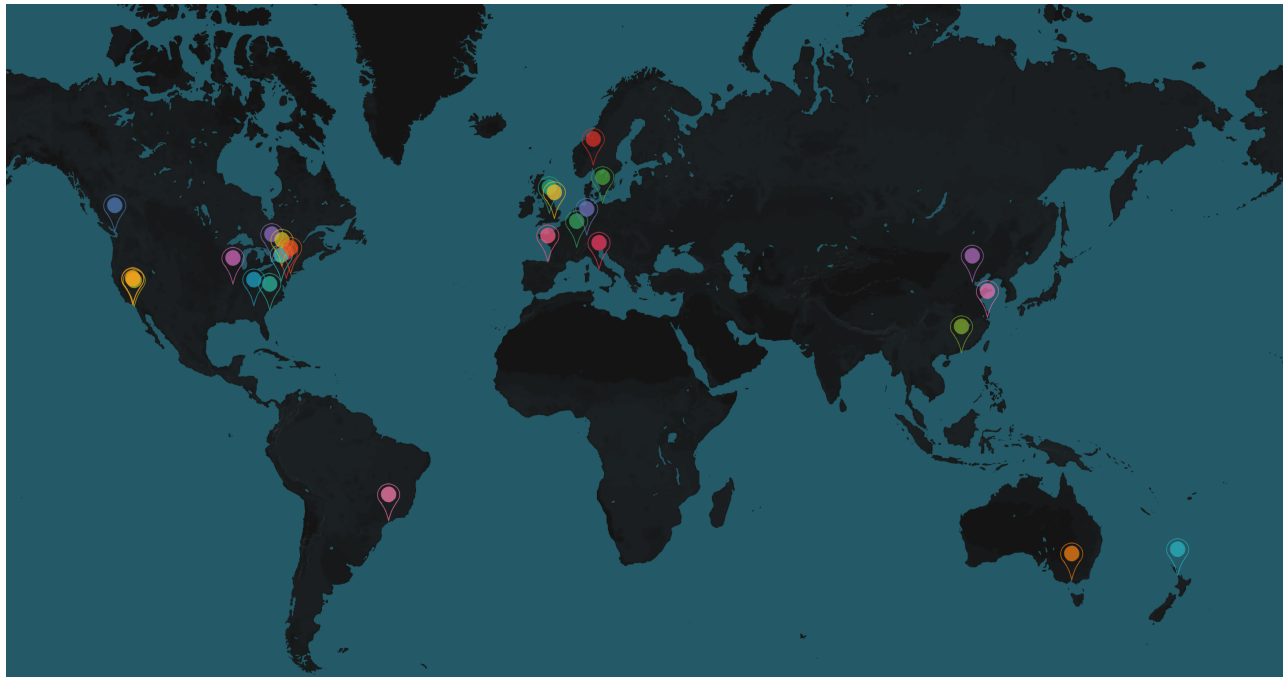
Results (Replicability) → **Big Data / Open Science**

- Positive publication bias
- Logistical limitations:
  - Limited money, time, and participant availability can lead to biased and underpowered samples

ENIGMA Stroke Recovery Working Group

100+ researchers from 45+ research cohorts worldwide

2000+ high-resolution stroke MRIs + behavior, and growing



New Data
- Receive new MRI and behavioral site data

Data Scrubbing
- Ensure MRI and behavioral data quality
- Clarify and rename any columns to match database
- Convert data to BIDS format

FreeSurfer
- Run FreeSurfer 5.3 and FreeSurfer 6.0 on all data
- Extract subcortical volumes, cortical surface area, cortical thickness measures
- Perform manual quality control on all FreeSurfer regions of interest

Lesion Analysis
- Perform semi-automated lesion segmentation, manual quality control, and analysis of lesion volume and overlap with different brain structures

SQLite
- Enter all data (FreeSurfer measures, lesion overlaps, behavior) into SQLite database

Analyses
- Use custom scripts to generate datasets from SQLite database based on analysis (e.g., left versus right hemisphere stroke)
- Perform analyses using scripts in Python and R

Liew et al., 2020, *Human Brain Mapping*
Liew et al., 2021, *Brain Communications*

USC University of Southern California

ENIGMA Stroke Recovery Working Group

- Inside look at how different researchers organize and manage their stroke data
- Over 100 different behavioral measures and MR scan types
- Turned to data science tools to organize, scrub, and harmonize these complex stroke datasets
- **Takeaway:** Education on data science and programming principles early on can help researchers manage data better from the start so it can be more useful, harmonizable, and "AI-able" for the future

USC University of Southern California

What can be done?

Methods (Reproducibility) → **Data Science**

- Reproducible methods from data science:
  - Data management with consistent formatting
  - Data analysis using executable scripts (Matlab, R, Python)
  - Version control across across different team members, analyses
  - End goal: **Reproducible papers**

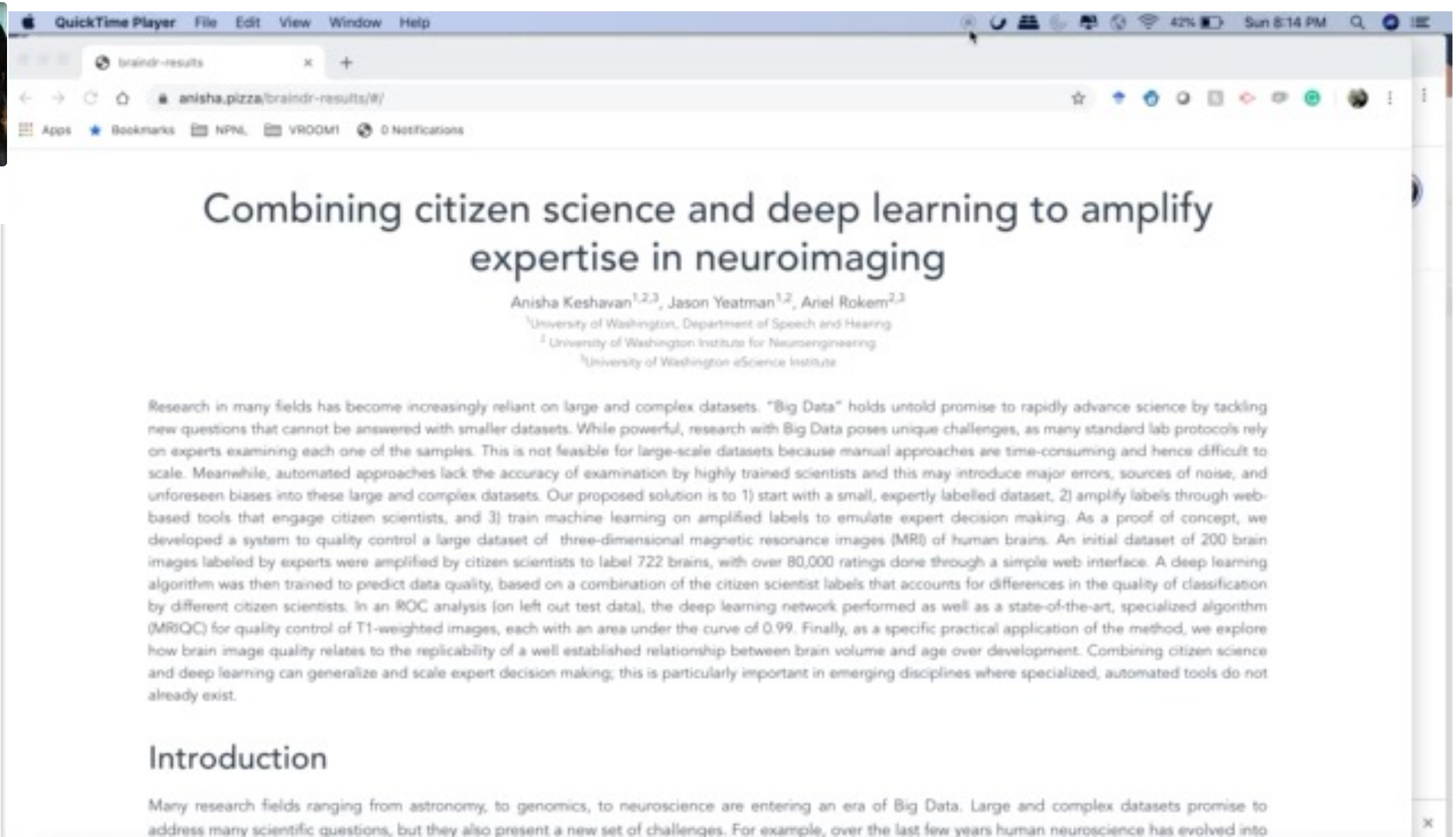
  - See Center for Reproducible Neuroimaging (ReproNim) as an example: https://www.repronim.org/

USC University of Southern California

Reproducible paper example (Keshavan et al., 2019)

https://anisha.pizza/braindr-results/#/

Anisha Keshavan
akeshavan



Resources for data science in rehab research

- **ReproRehab!** A new NIH R25 education research program aimed at teaching data science skills to rehabilitation researchers
  - https://www.reprorehab.usc.edu/
  - follow us @ReproRehab on twitter or email us at reprorehab@gmail.com
- Mobilize Center: http://mobilize.stanford.edu
- Restore Center: https://restore.stanford.edu/
- Center for Large Data Research and Data Sharing in Rehabilitation: https://www.utmb.edu/cldr
- ReproNim (https://www.repronim.org/), NeuroHackademy (https://neurohackademy.org/neurohack_year/2020/)
- 2019 ASNR Symposium: Reliability and Reproducibility in Neurorehabilitation Research
  - Hands-on tutorials and slides on Github: https://github.com/npnl/ASNR_2019

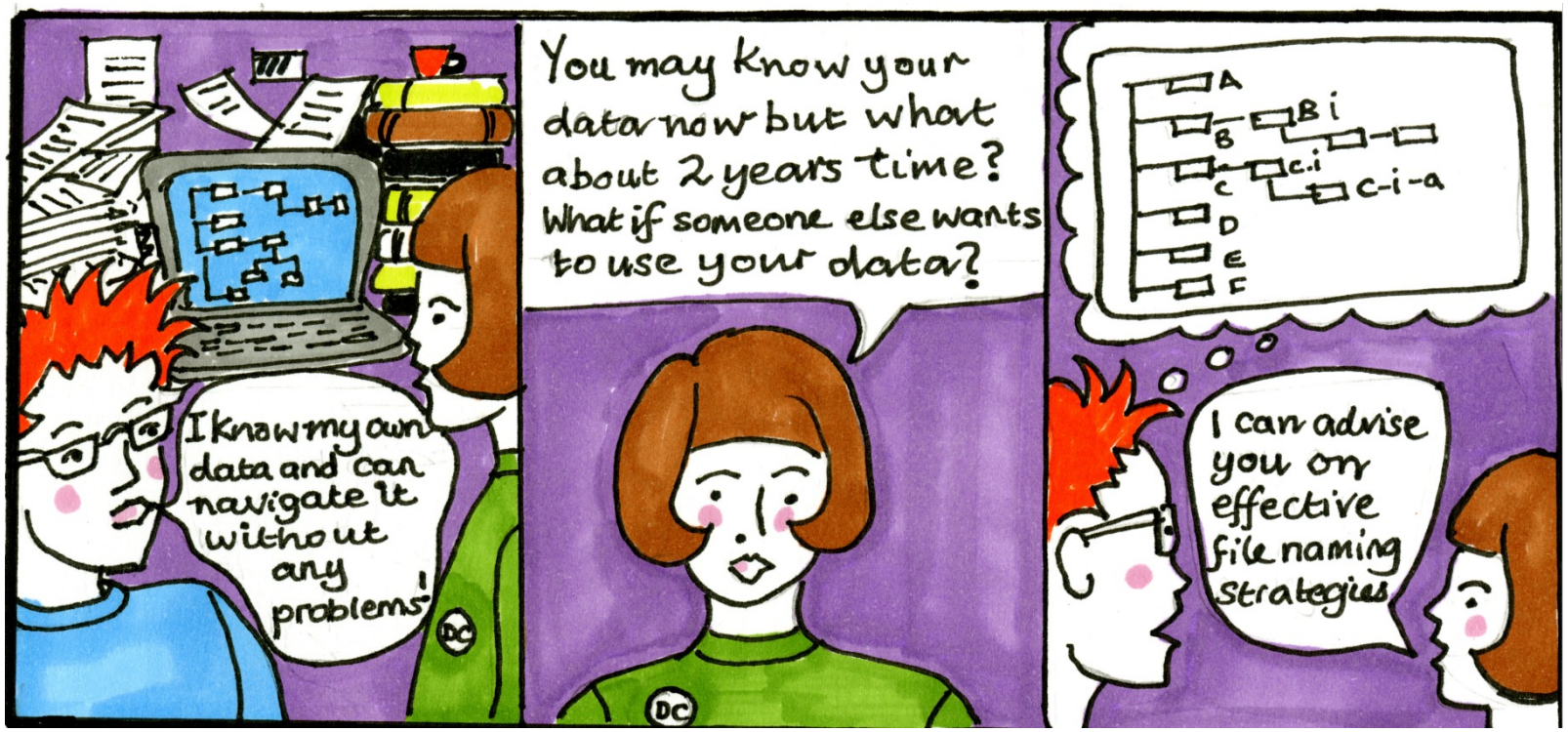
USC University of Southern California

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).



Data Management

Managing 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!

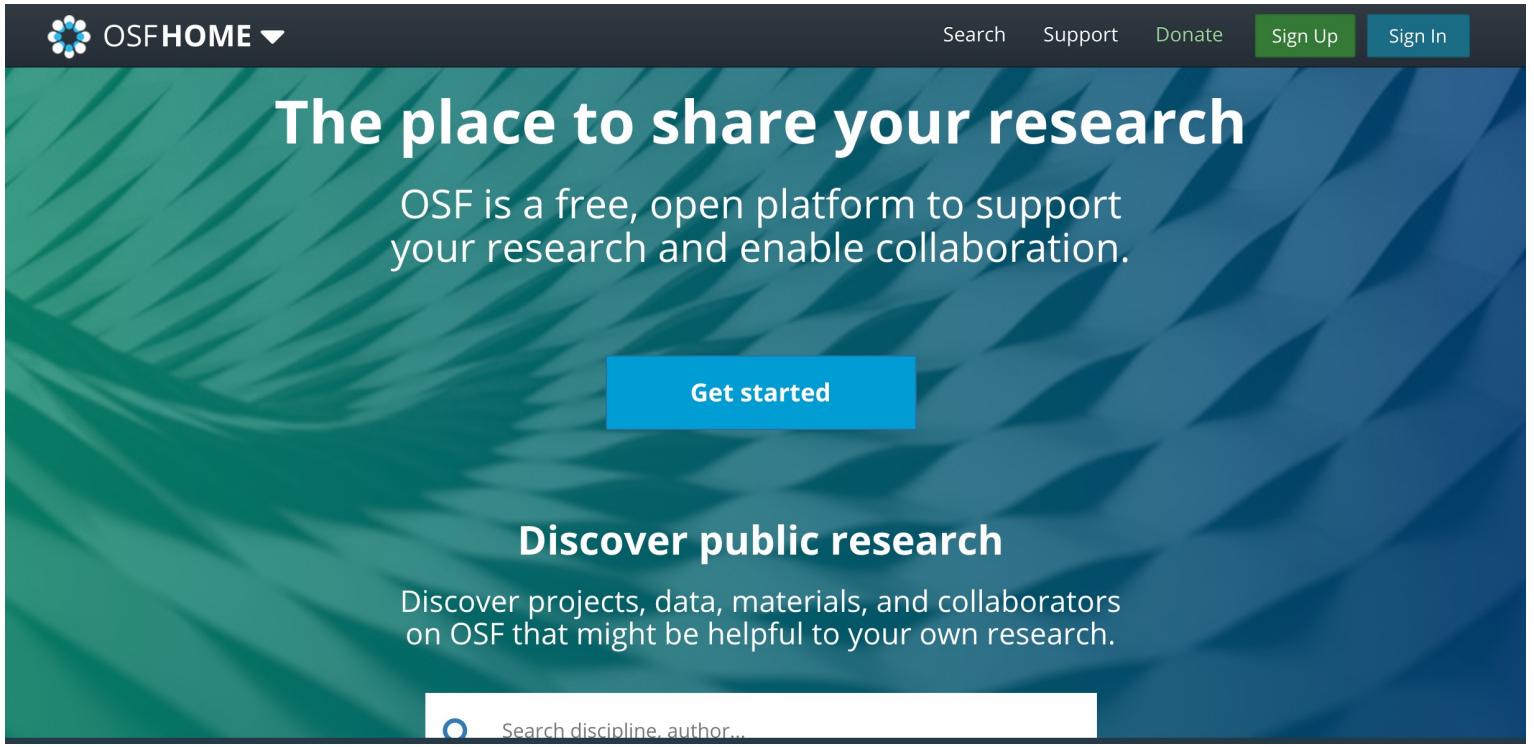
1. Experimental Protocols

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)
- (Shared our getting started with a study, and sub-study form in case it’s useful for anyone who is getting started – on our github: www.github.com/reprorehab/reprorehab2022)

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)



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/) – also see Dr. Poline's talk this week!

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

- This is often recorded in a **json** file (stores simple data structures and objects in JavaScript Object Notation) – lightweight, text-based, human-readable, can be edited using a text editor such as Atom (https://atom.io/)

Thank you!

Visit us at http://npnl.usc.edu

Contact Us:
Me: sliew@usc.edu
Lab: npnl@usc.edu
Twitter: @NPNLatUSC

Special thanks to:

- ReproRehab Team
- Coralie Phanord
- Grace Song
- ENIGMA Stroke Recovery Team

- NIH R25
- NIH R01 NS115845
- NIH K01 091283

USC University of Southern California

NIH

CLDR
Center for Large Data Research & Data Sharing in Rehabilitation

enigma.ini.usc.edu