Data science and open science: Impacts on reproducibility in stroke rehabilitation research

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/

Outline

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

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

   - Yes – results of stroke research studies are often difficult to reproduce/replicate.
   - No - results of stroke research studies are often easy to reproduce/replicate.
   - Not sure

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

3. **How can we use open science to address replicability?**

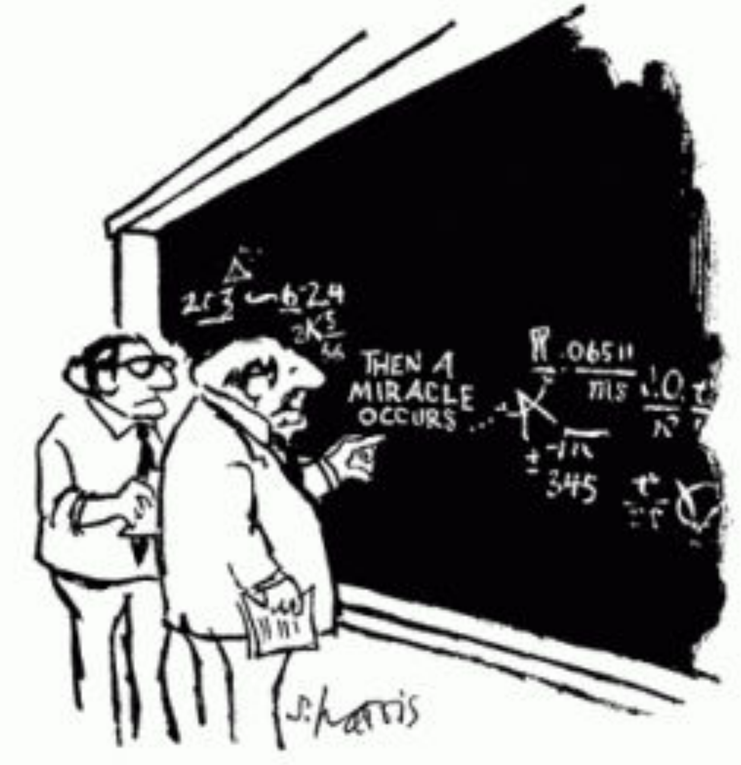
USC University of Southern California

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."

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

USC University of Southern California

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**

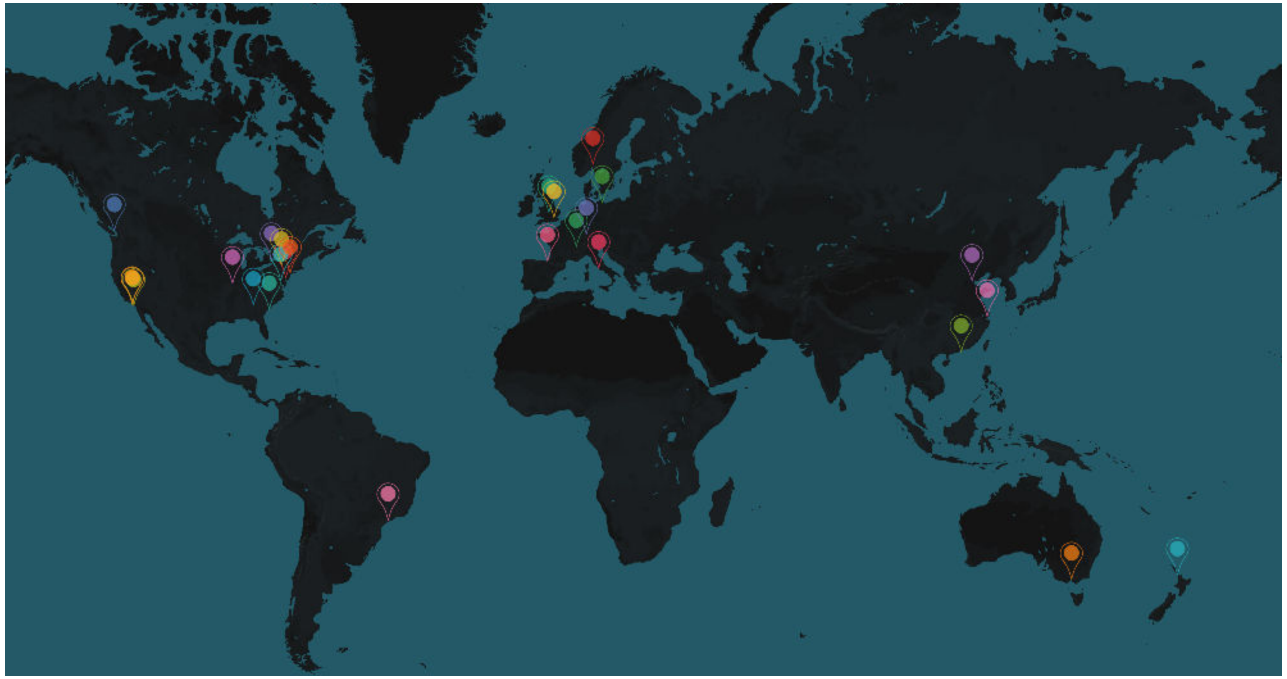
- Positive publication bias
- Logistical limitations:
  - Limited money, time, and participant availability can lead to biased and underpowered samples

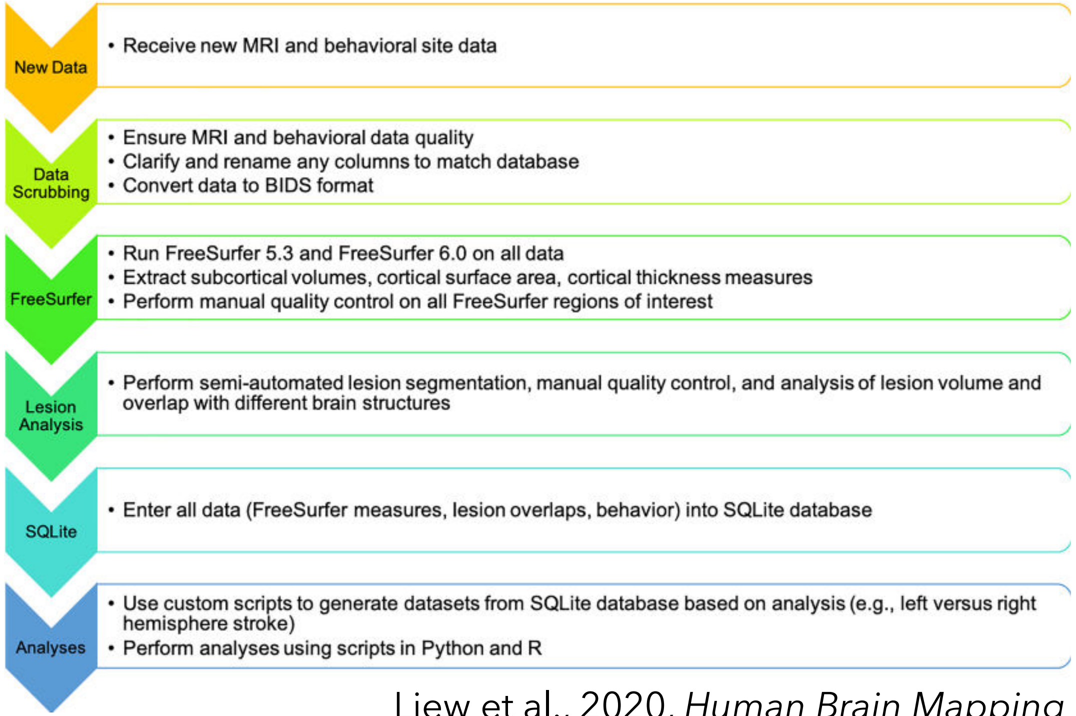
USC University of Southern California

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 and "AI-able" for the future

What can be done?

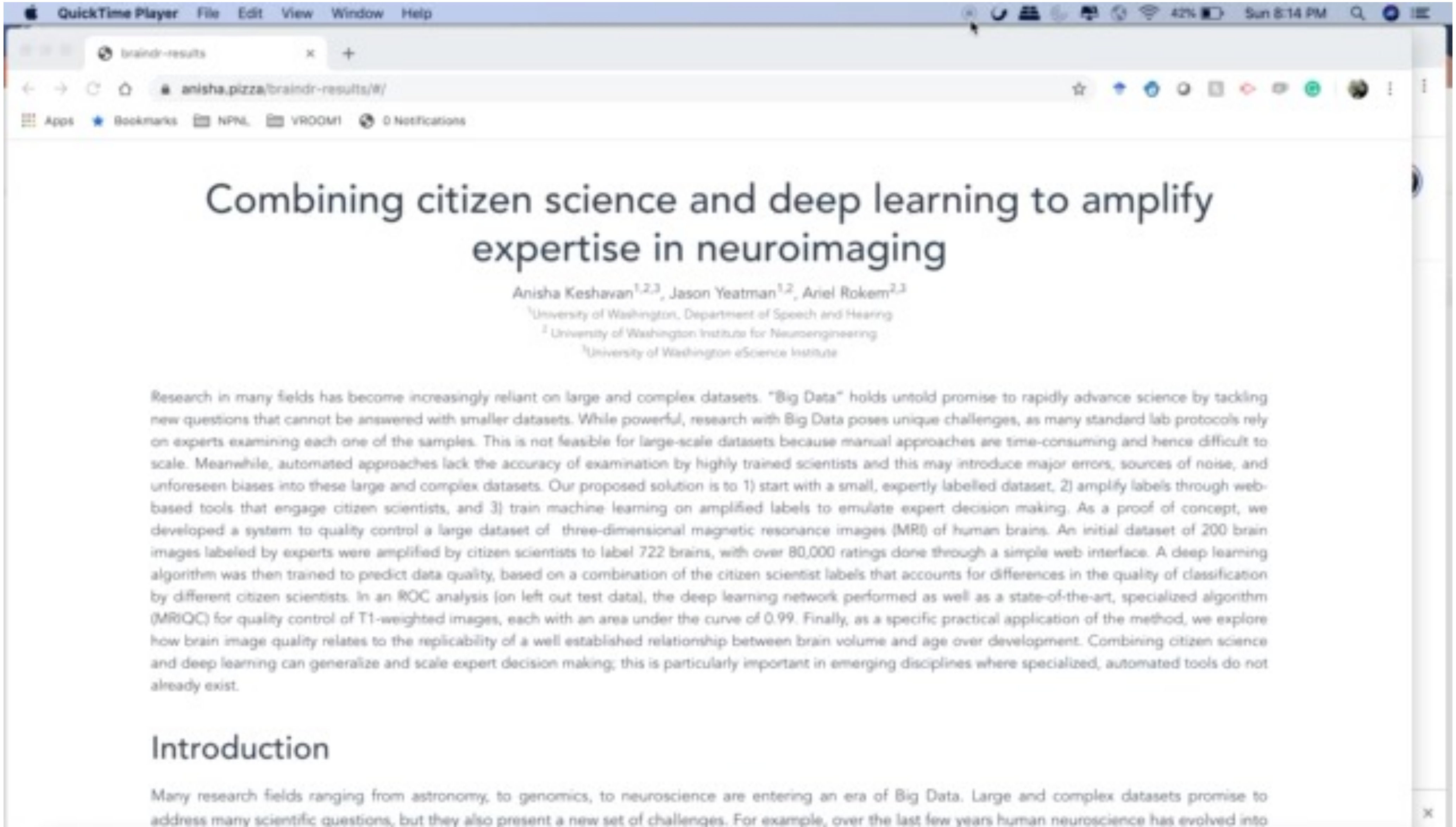
Methods (Reproducibility) → **Data Science**

- 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/

Reproducible paper example (Keshavan et al., 2019)

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



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

Resources for data science in rehab research

- Mobilize Center: http://mobilize.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
- **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 or email us at reprorehab@gmail.com.

USC University of Southern California

What can be done?

Results (Reliability) → **Open Science**

- Overcoming positive publication bias and logistical limitations by testing samples from:
  - Retrospective datasets that have been archived
  - Pooled samples across retrospective/prospective datasets from diverse research sites (e.g., ENIGMA)
  - Large prospective datasets (e.g., UK Biobank)

  - All of these would benefit from data science for accurate data management, analysis across sites

Open Science: What is it?

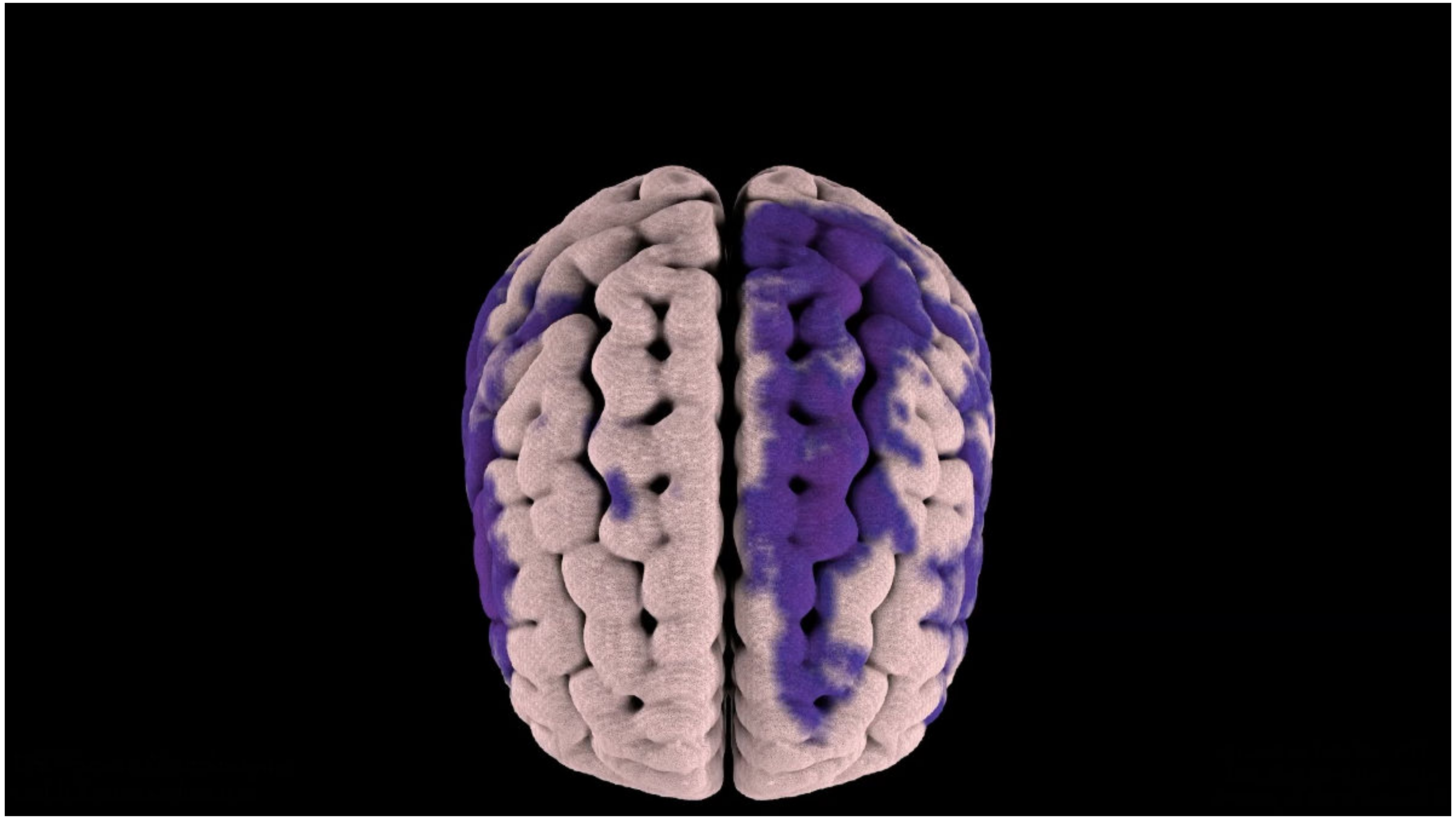
- Open science movement: Sharing (published & unpublished) data, code, protocols, resources

- Why do it? To improve scientific reproducibility and replicability and build the capacity of the scientific community (especially trainees)

- What's involved? Usually free to download, with some agreement you won't abuse/sell the data. That's it!

Open-source data sharing to advance research

Anatomical Tracings of Lesions After Stroke (ATLAS) v2.0
N=955 stroke T1-w high resolution MRIs and lesion masks
Liew et al., 2018, *Scientific Data;* Liew et al., 2021, *medRxiv*



http://fcon_1000.projects.nitrc.org/indi/retro/atlas.html

USC University of Southern California

Open Data: What types of data are there?

- Data types: Surveys, behavioral measures, demographics, kinematic data, videos, physiological data (e.g., brain imaging)

- Prospective data collections (protocol is set prior to data collection)
- Retrospective data archives (usually study-specific data)
- Health services / medical records

Open Data: What types of data are there?

- **Rehabilitation-Related Data Archives (NCMRR-funded)**
  - CLDR: https://www.utmb.edu/cldr/
    - Center for Large Data Research and Data Sharing in Rehabilitation
    - Many types including health services research (e.g., medical records) and retrospective study-specific rehabilitation data
  - ICPSR/ADDEP: https://www.icpsr.umich.edu/web/pages/ADDEP/index.html
    - Archive of Data on Disability to Enable Policy and research
    - Retrospective study-specific rehabilitation data
  - OpenSim: https://opensim.stanford.edu/
    - Free motion simulation toolbox and trained models for different populations: http://simtk.org/

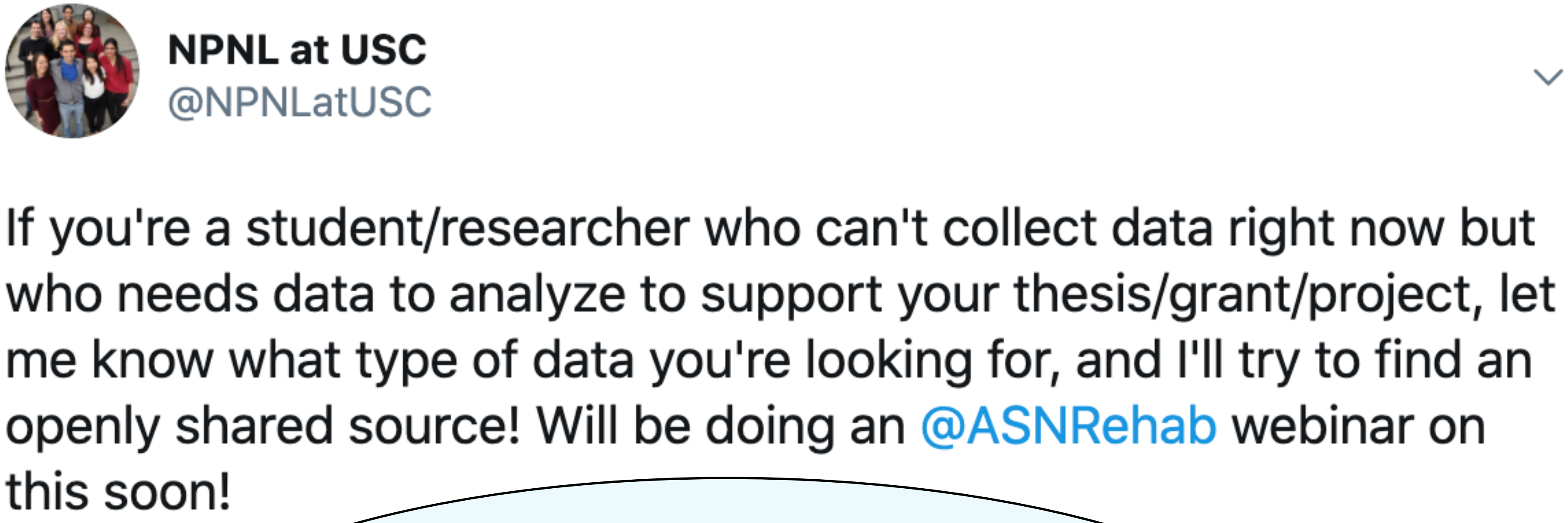
Open Data: What types of data are there?

- **Prospective/Coordinated Brain Imaging, Clinical/Behavior**
  - Human Connectome Project: https://www.humanconnectome.org/
    - Lifespan, young adult, clinical populations, with harmonized behavior
  - UK Biobank: https://www.ukbiobank.ac.uk/
    - UK health records data including brain imaging, genetics, clinical variables
    - Working up to 100,000 individuals
  - All of Us: https://allofus.nih.gov/
    - On beta release; will be US health records data including brain imaging, genetics, clinical variables and questionnaires
    - Working up to 1 million individuals

Open Data: What types of data are there?

- **Community (Study-Specific) Brain Imaging**
  - Open Neuro: https://openneuro.org/
    - 372 MRI, MEG, EEG, ECoG datasets
  - INDI: http://fcon_1000.projects.nitrc.org/
    - International Neuroimaging Data-Sharing Initiative: Prospective and retrospective data
    - Resting state fMRI, structural MRI, diffusion MRI with behavioral measures
  - NITRC: https://www.nitrc.org/
    - Neuroimaging Tools and Resources Collaboratory: Atlases, data, and tons of software/tools

Open Data: But I want something specific?



NPNL at USC
@NPNLatUSC

If you're a student/researcher who can't collect data right now but who needs data to analyze to support your thesis/grant/project, let me know what type of data you're looking for, and I'll try to find an openly shared source! Will be doing an @ASNRehab webinar on this soon!

- Myelin water fraction MRI with behavior
- EEG data during FES-evoked movements
- Walking data with EMG, kinematics
- Resting state EEG with motor learning

Open Data: Collaborative data sharing

- If you have a specific need, you may consider reaching out to someone who has published a dataset that you'd like to utilize

- General guidelines:
  - Collaborate on the data (including authorship)
  - Receive useful insight on the data wrt how you use it
  - No one's data is perfect!
  - Maybe help organize their data into a data archive that you both can also publish (see journals like *Scientific Data*, *GigaScience*) or cite

USC University of Southern California

Open Data: I want to share data

- Everyone should think now about data sharing
  - Include consent/IRB language for sharing de-identified data
  - Learn about good data management

- Learn more about FAIR principles and reproducible methods for open science: https://www.repronim.org/index.html

- Happy to discuss best place to archive or other questions: sliew@usc.edu

Thank You!

The Neural Plasticity and Neurorehabilitation Laboratory

http://npnl.usc.edu
sliew@usc.edu

Twitter: @NPNLatUSC

USC Chan Division of Occupational Science and Occupational Therapy

USC Division of Biokinesiology and Physical Therapy

Keck Medicine of USC

USC Viterbi
Department of Biomedical Engineering

USC Stevens Neuroimaging and Informatics Institute

DISCOVER INNOVATE TRANSITION ARL

NIH

American Heart Association®

CLDR
Center for Large Data Research & Data Sharing in Rehabilitation

USC University of Southern California