Danish Digitalization, Data Science and AI 3.0

OPTIMIZING REPRODUCIBILITY IN NEUROINFORMATICS: FROM DATA TO DISCOVERY

Martin Norgaard, Assistant Professor
Department of Computer Science,
University of Copenhagen

NIH National Institute of Mental Health

Stanford | Center for Reproducible Neuroscience

UNIVERSITY OF COPENHAGEN

Organizers

https://github.com/openneuropet/outreach/blob/main/D3A2025/D3A_2025_slides.pdf

Ruben P. Dörfel
PhD student
UCPH/Karolinska

Martin Nørgaard
Assistant Professor
UCPH

Llucia Coll
Postdoc
UCPH

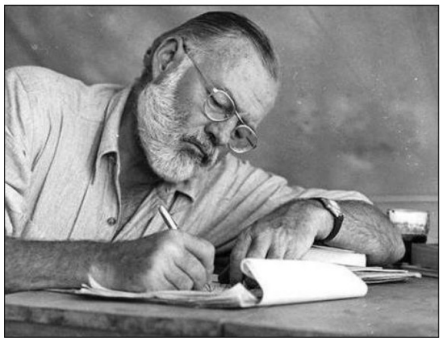
Agenda

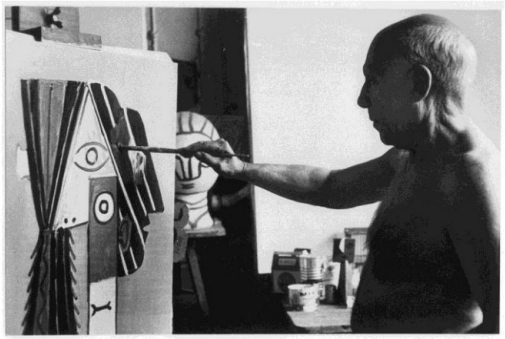
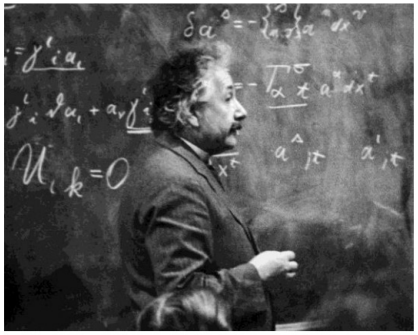
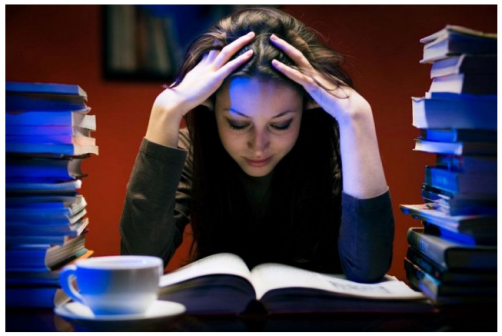
- **Introduction to Neuroinformatics and Reproducibility**
- **Case study:** Lessons learned from OpenNeuro and the NeuroImaging PREProcessing toolS (NiPreps) initiative
- **Interactive group activity:** Optimizing Neuroimaging Pipelines
- **Practical session:** Neuroinformatics Tools & Database Management
- **Discussion & wrap-up**

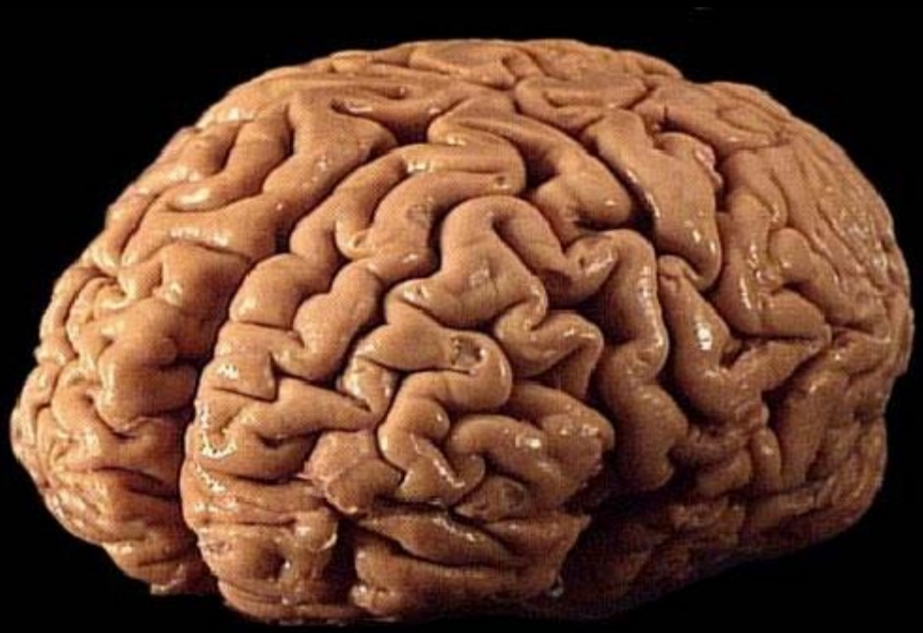
Learning goals

- Understand the scope of neuroinformatics in modern neuroscience.
- Recognize why reproducibility is a central challenge.

Introduction to Neuroinformatics and Reproducibility



unicef






Average human brain:
~1.5 kg
~1.1-1.2 l volume

What is neuroinformatics?

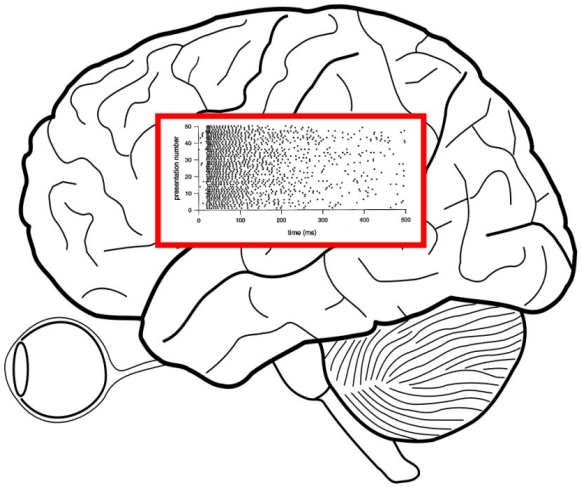
AI Overview

Neuroinformatics is an interdisciplinary field that uses computational methods to develop data and knowledge bases, software tools, and analytical techniques to organize, integrate, share, and analyze neuroscience data. It

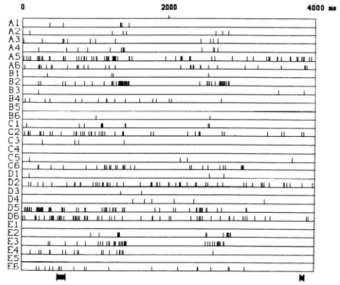
Why do we have a brain?

Stimuli in the Environment

Encoding



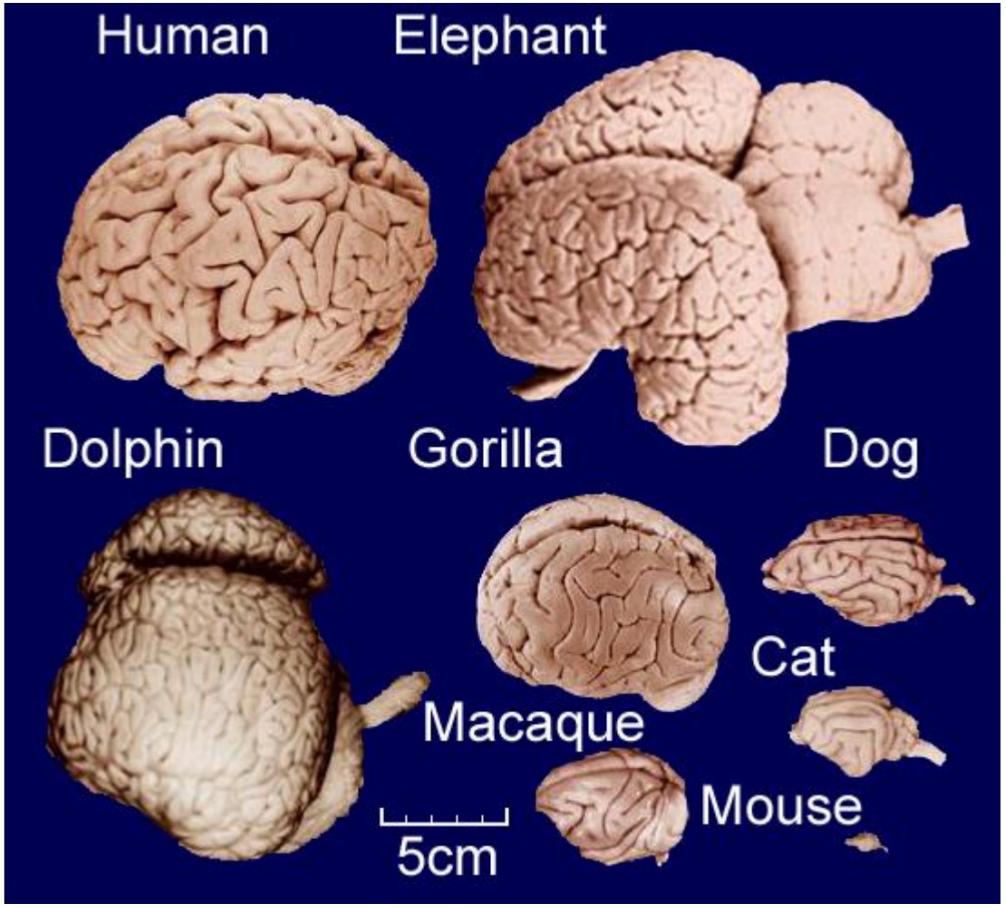
Neuronal Representation
Perception
Sensory Integration
Memory / maintenance



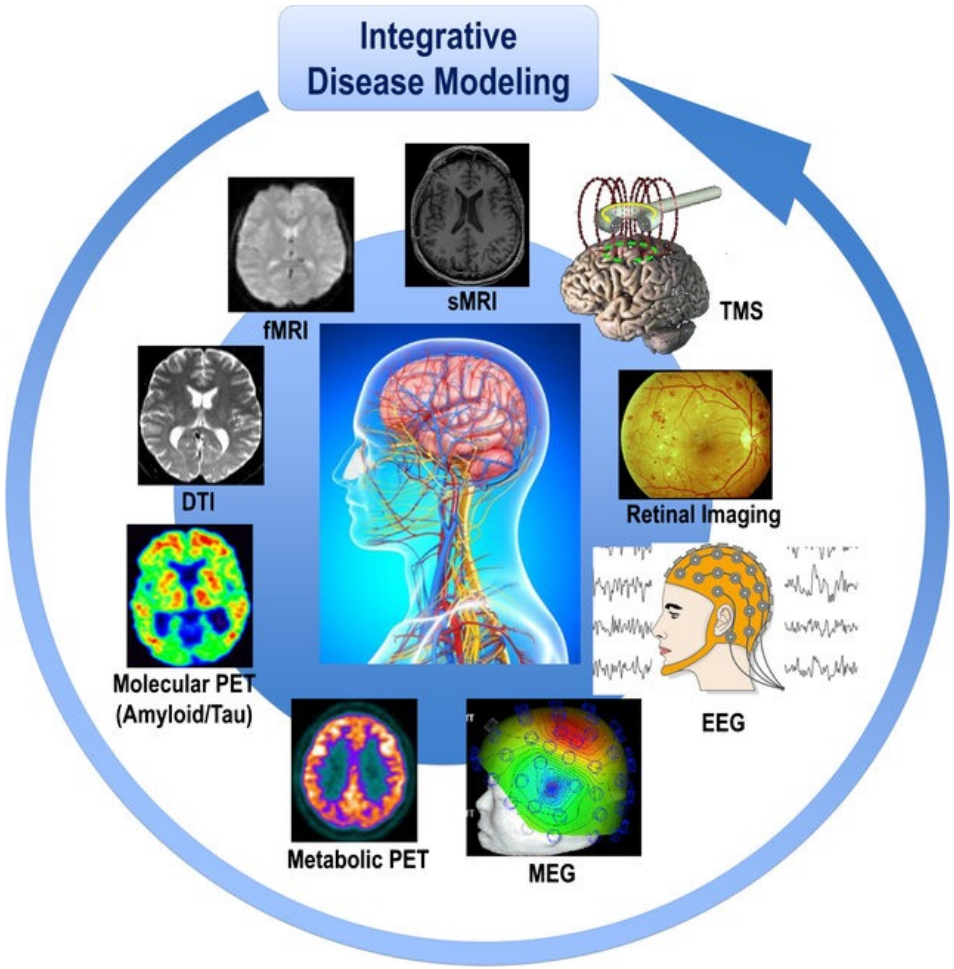
Decoding

Movement
Actions
Decisions
Behavior

Different brains can do different things

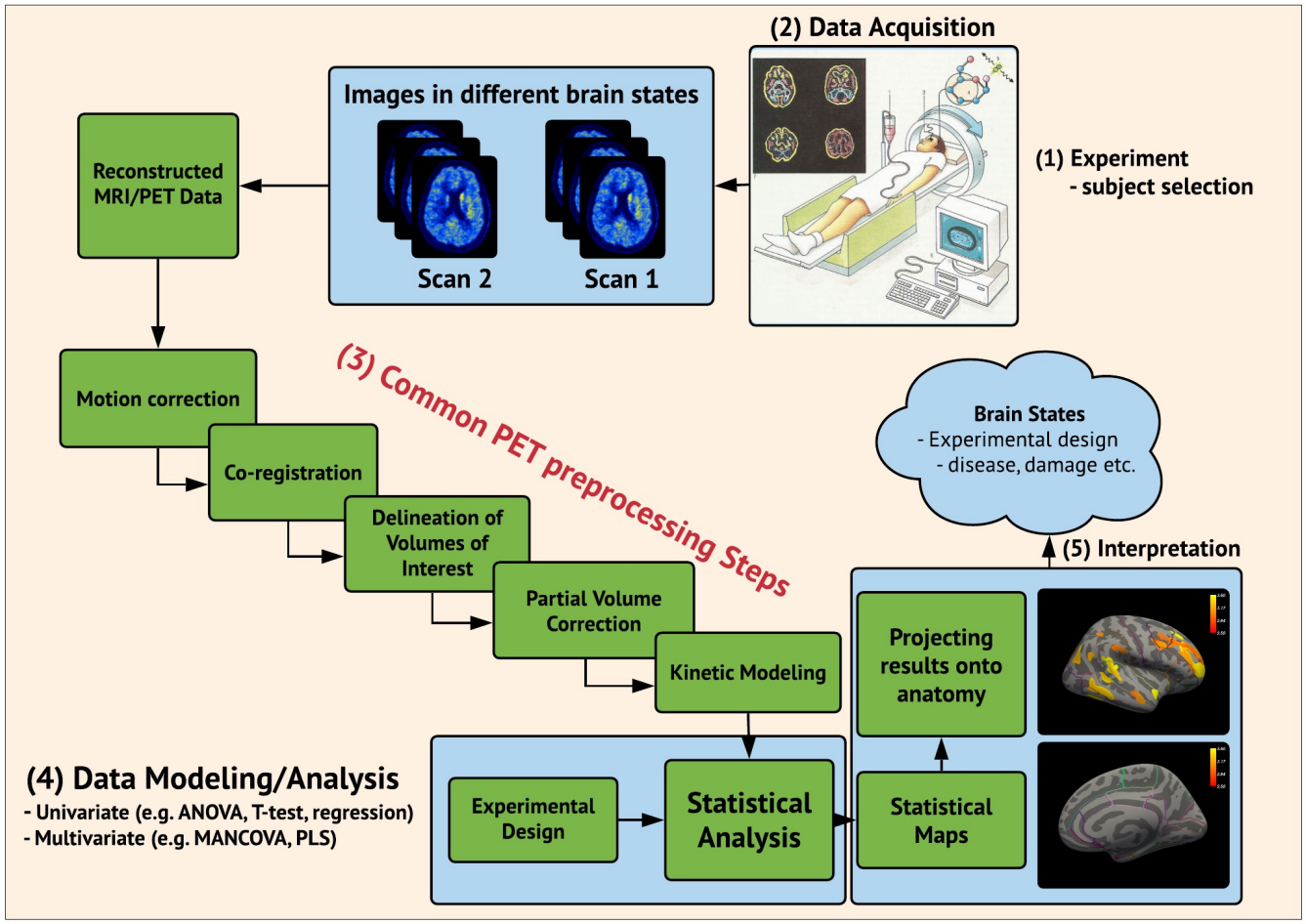


How do we measure brain data?



Common neuroimaging experiment

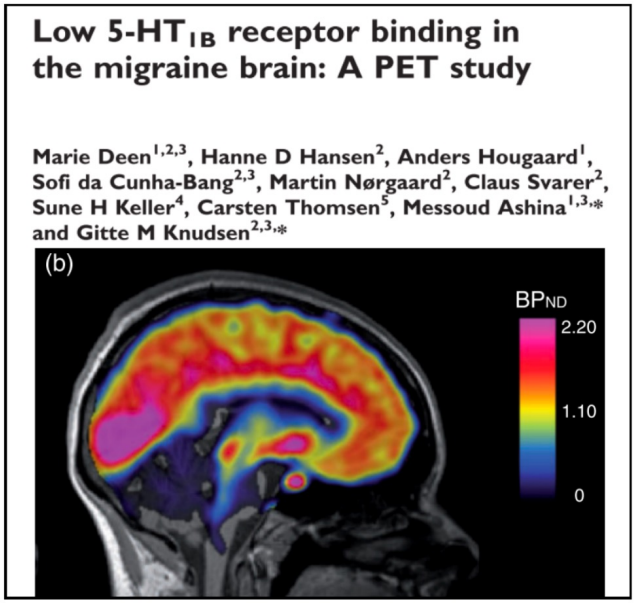
Norgaard et al. 2019, Neuroimage 199, 466-479.

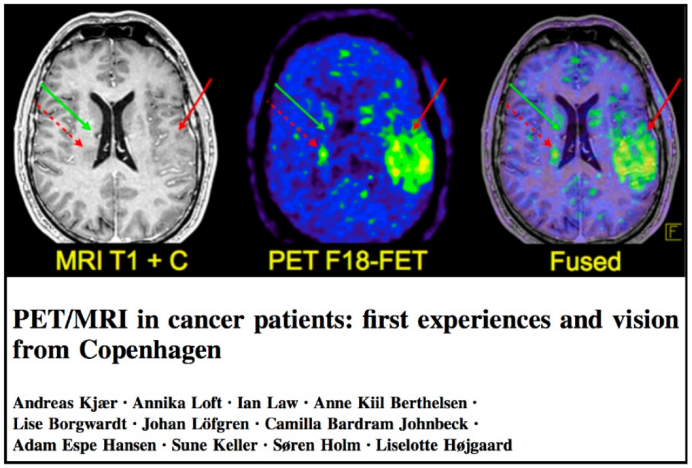


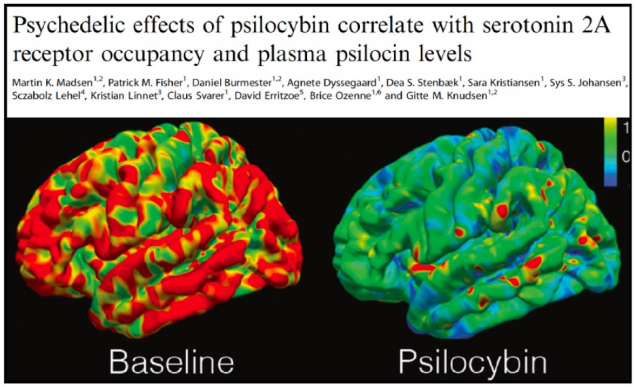
Stages

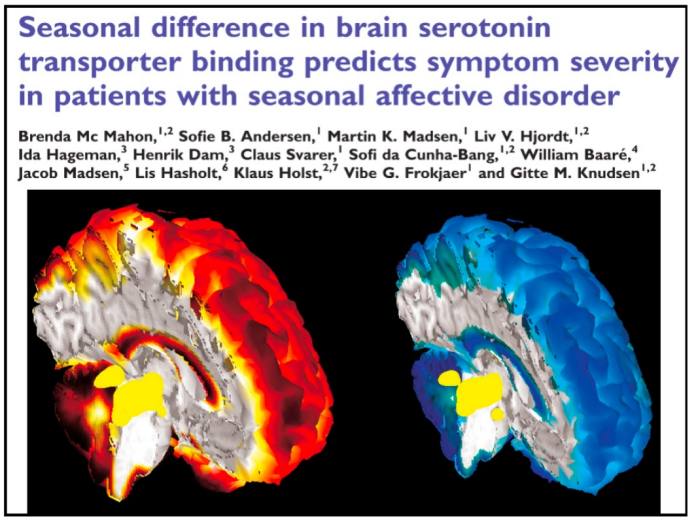
1. Subject selection
2. Data acquisition
3. Preprocessing
4. Data modeling
5. Interpretation

How are we using neuroinformatics in real life?

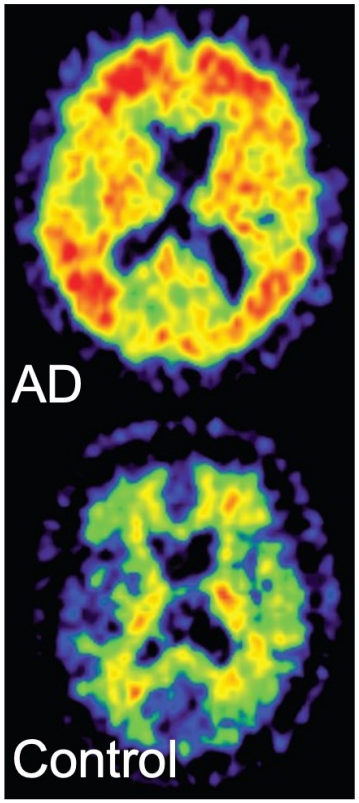


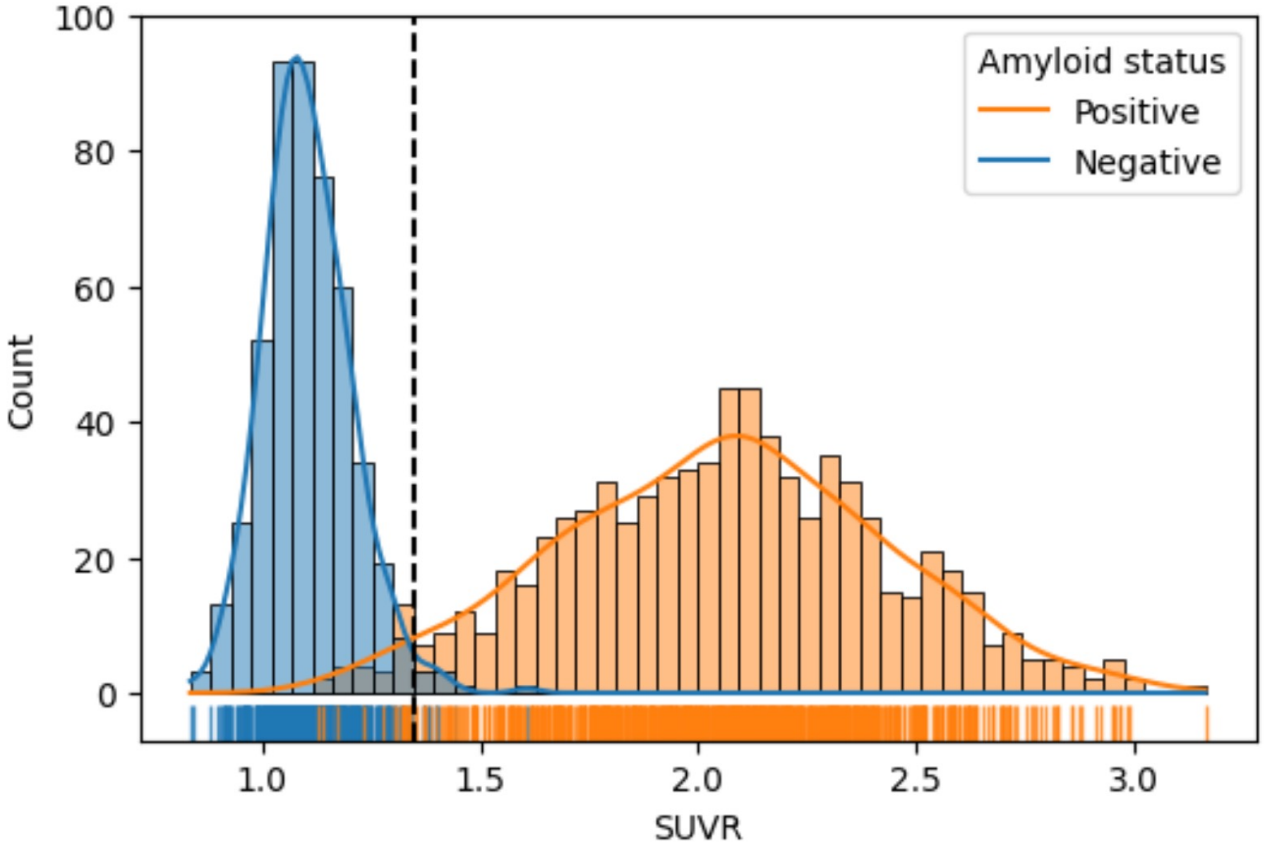






Amyloid status with [11C]PiB





Is there a problem?

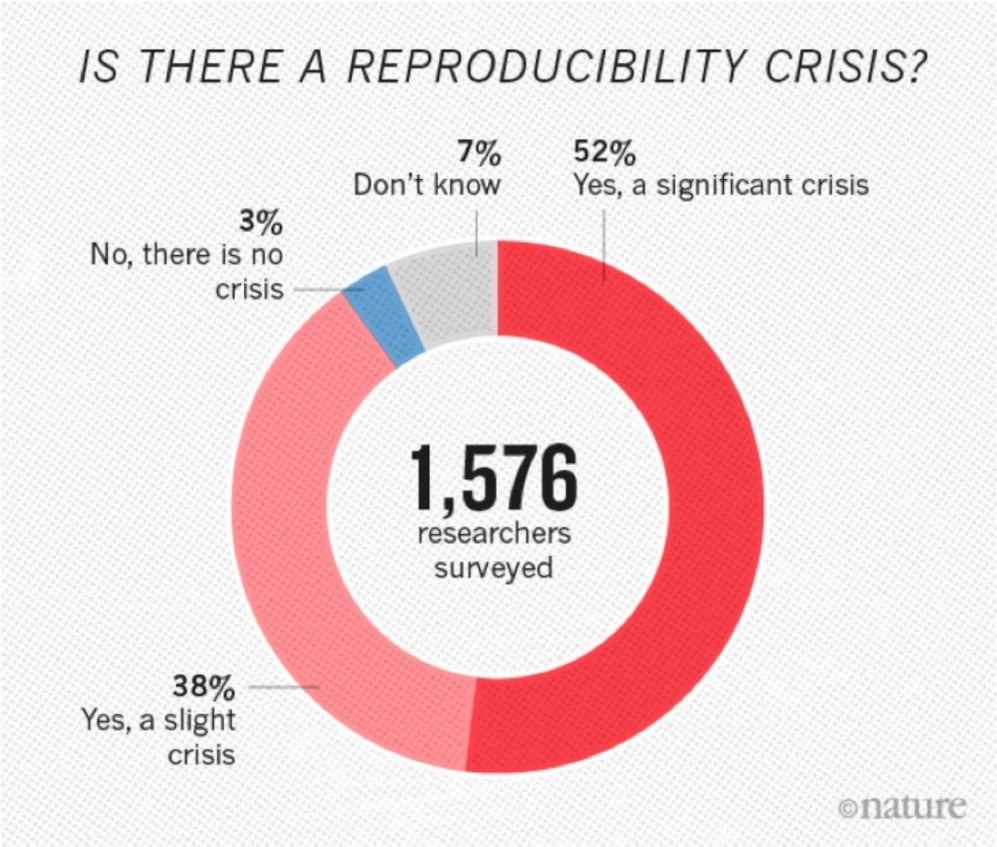
Scanning the horizon: towards transparent and reproducible neuroimaging research

Russell A. Poldrack[1], Chris I. Baker[2], Joke Durnez[1,3], Krzysztof J. Gorgolewski[1], Paul M. Matthews[4], Marcus R. Munafò[5,6], Thomas E. Nichols[7], Jean-Baptiste Poline[8], Edward Vul[9] and Tal Yarkoni[10]

Abstract | Functional neuroimaging techniques have transformed our ability to probe the neurobiological basis of behaviour and are increasingly being applied by the wider neuroscience community. However, concerns have recently been raised that the conclusions that are drawn from some human neuroimaging studies are either spurious or not generalizable. Problems such as low statistical power, flexibility in data analysis, software errors and a lack of direct replication apply to many fields, but perhaps particularly to functional MRI. Here, we discuss these problems, outline current and suggested best practices, and describe how we think the field should evolve to produce the most meaningful and reliable answers to neuroscientific questions.

Poldrack et al. 2017, Nature Reviews Neuroscience 18, 115-126.

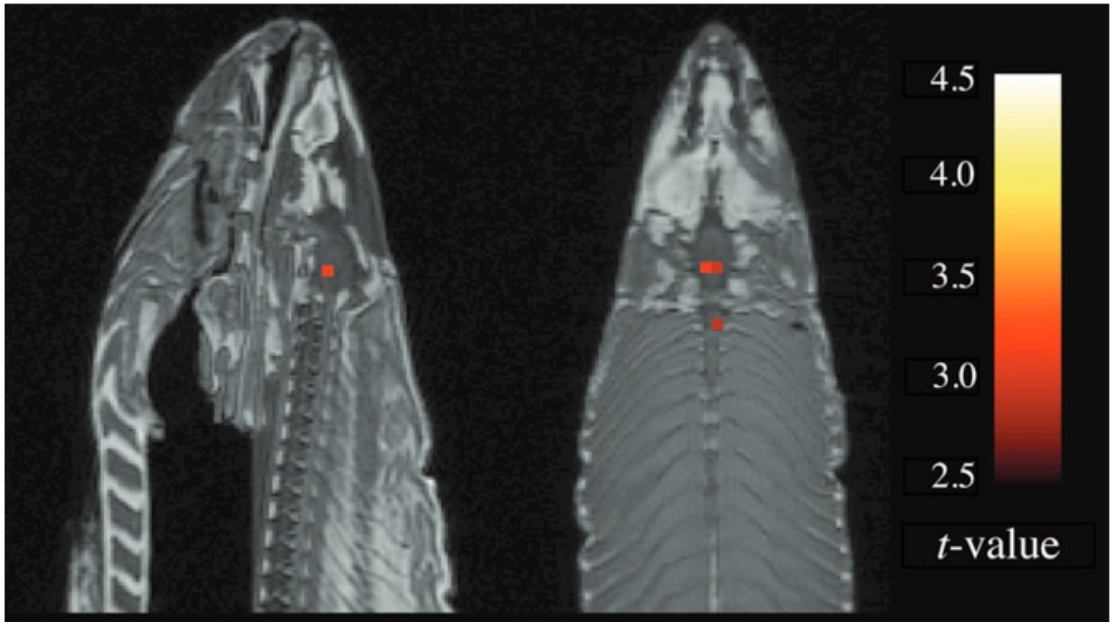
Is there a problem?



Baker M. 1,500 scientists lift the lid on reproducibility. Nature 533, 452-454, 2016.

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1*], Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]



The platform dependence of analysis results

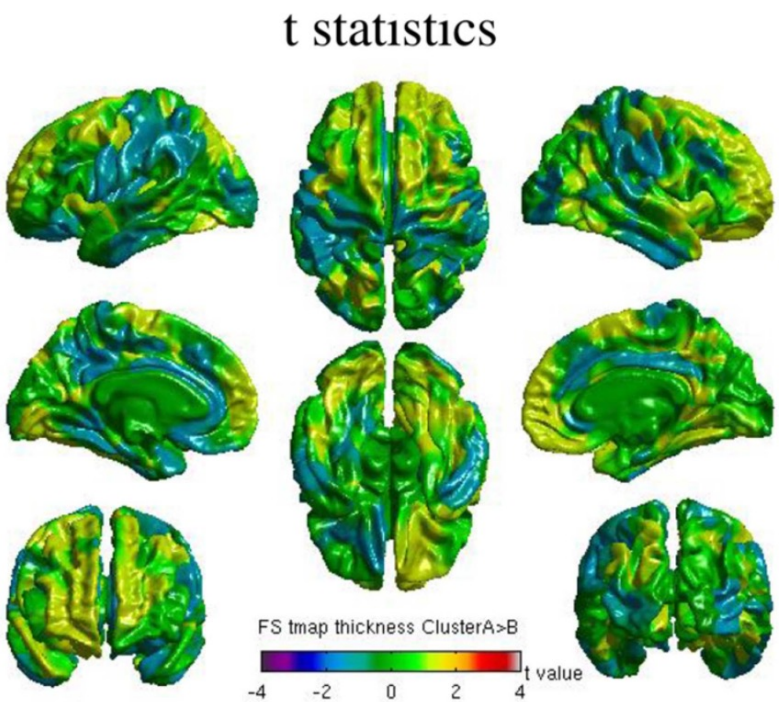
frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 24 April 2015
doi: 10.3389/fninf.2015.00012

Reproducibility of neuroimaging analyses across operating systems

*Tristan Glatard[1,2], Lindsay B. Lewis[1], Rafael Ferreira da Silva[3], Reza Adalat[1], Natacha Beck[1], Claude Lepage[1], Pierre Rioux[1], Marc-Etienne Rousseau[1], Tarek Sherif[1], Ewa Deelman[3], Najmeh Khalili-Mahani[1] and Alan C. Evans[1]**

Freesurfer thickness differences b/w linux versions
(same software build)



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124

The Earth Is Round ($p < .05$)

Jacob Cohen

December 1994 • American Psychologist 997

Copyright 1994 by the American Psychological Association, Inc. 0003-066X/94/$2.00

Vol. 49, No. 12, 997–1003

ANALYSIS

Science 28 August 2015:
Vol. 349 no. 6251
DOI: 10.1126/science.aac4716

‹ Prev | Table of Contents | Next ›
Read Full Text to Comment (0)

RESEARCH ARTICLE

Estimating the reproducibility of psychological science

Open Science Collaboration*,†

± Author Affiliations

↵†Corresponding author. E-mail: nosek@virginia.edu

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

NATURE REVIEWS | NEUROSCIENCE VOLUME 14 | MAY 2013 | 365

© 2013 Macmillan Publishers Limited. All rights reserved

Courtesy of Stephen Strother

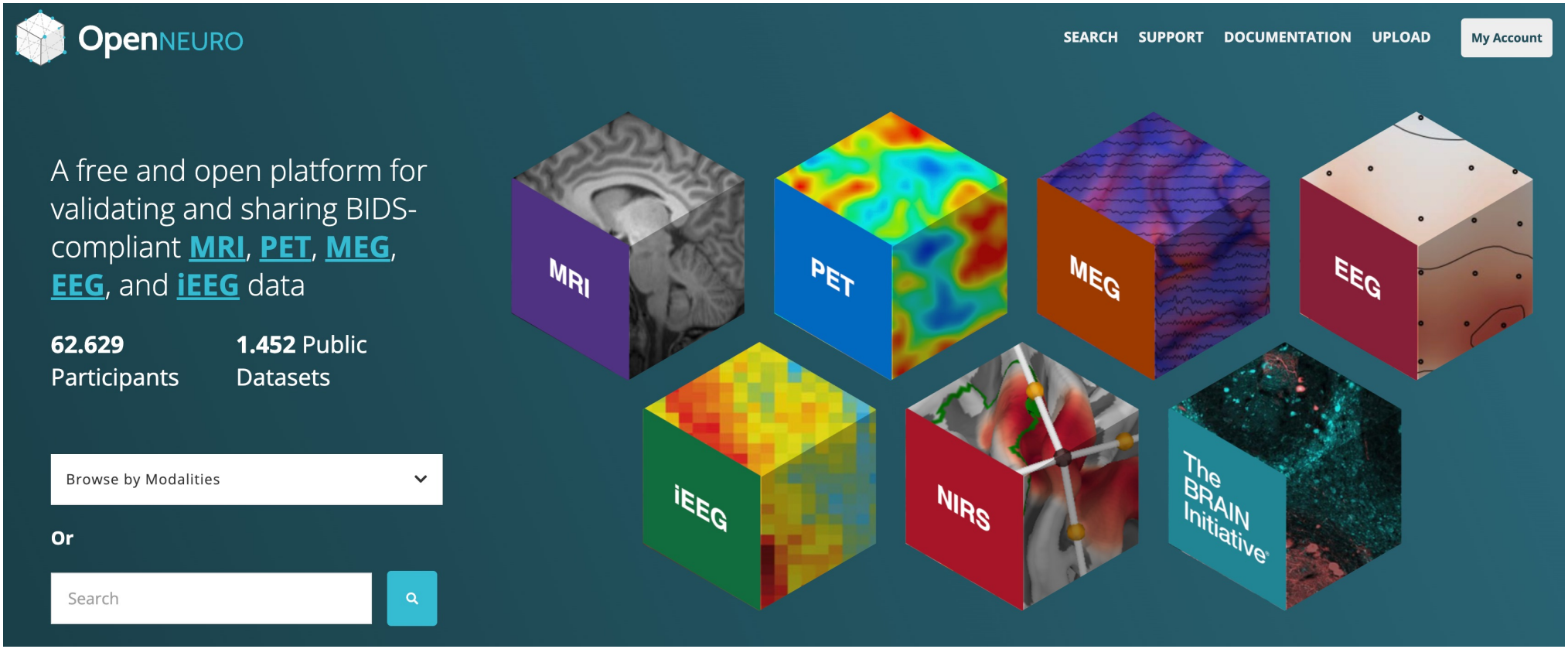
Summary

- Neuroinformatics is the field of organizing, integrating, sharing and analyzing neuroscience data

- All stages of a neuroinformatics workflow may impact the reproducibility, ultimately affecting the neuroscientific conclusions that can be drawn from the data

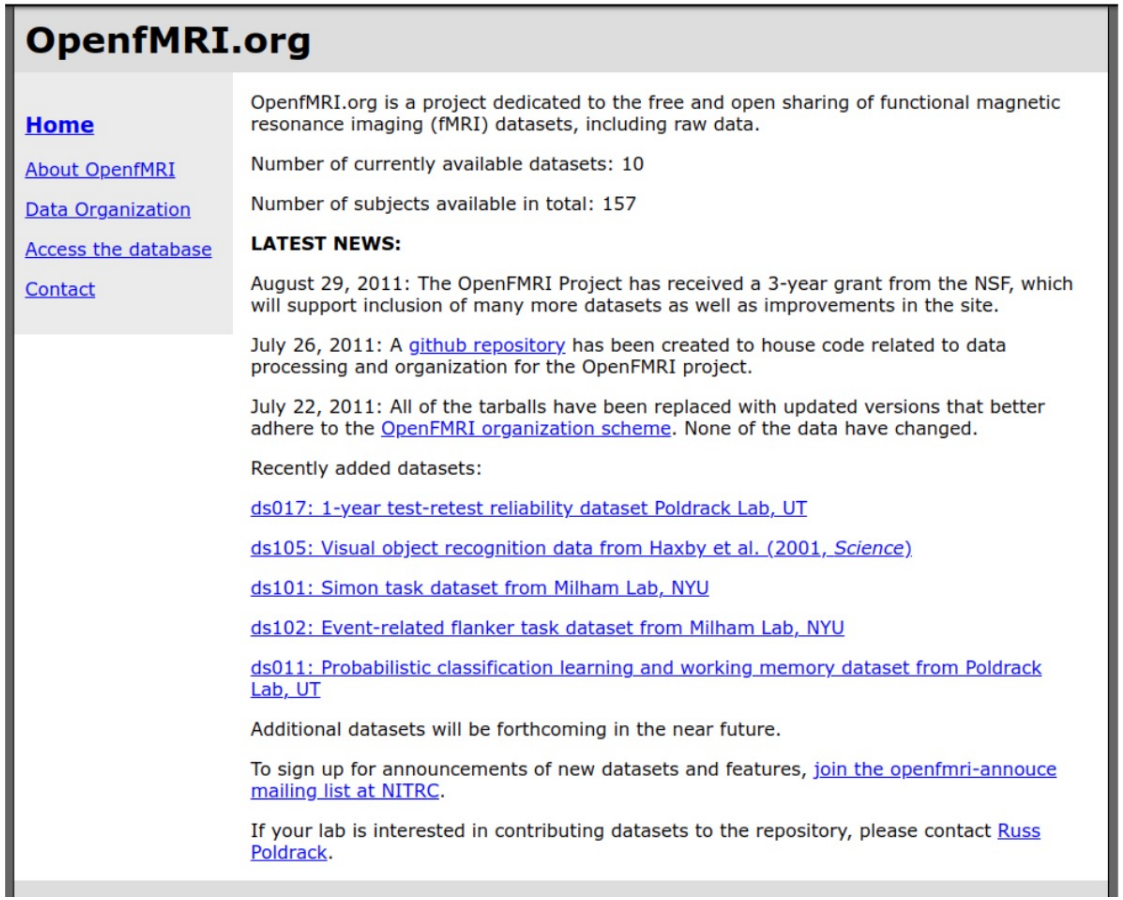
Case study: Lessons learned from OpenNeuro and the NeuroImaging PREProcessing toolS (NiPreps) initiative

OpenNeuro (www.openneuro.org)



Background of OpenNeuro: OpenfMRI

- Launched ~2011
- Accrued almost 100 datasets from 2010-2017



OpenfMRI.org

Home

About OpenfMRI

Data Organization

Access the database

Contact

OpenfMRI.org is a project dedicated to the free and open sharing of functional magnetic resonance imaging (fMRI) datasets, including raw data.

Number of currently available datasets: 10

Number of subjects available in total: 157

LATEST NEWS:

August 29, 2011: The OpenFMRI Project has received a 3-year grant from the NSF, which will support inclusion of many more datasets as well as improvements in the site.

July 26, 2011: A github repository has been created to house code related to data processing and organization for the OpenFMRI project.

July 22, 2011: All of the tarballs have been replaced with updated versions that better adhere to the OpenFMRI organization scheme. None of the data have changed.

Recently added datasets:

ds017: 1-year test-retest reliability dataset Poldrack Lab, UT

ds105: Visual object recognition data from Haxby et al. (2001, *Science*)

ds101: Simon task dataset from Milham Lab, NYU

ds102: Event-related flanker task dataset from Milham Lab, NYU

ds011: Probabilistic classification learning and working memory dataset from Poldrack Lab, UT

Additional datasets will be forthcoming in the near future.

To sign up for announcements of new datasets and features, join the openfmri-annouce mailing list at NITRC.

If your lab is interested in contributing datasets to the repository, please contact Russ Poldrack.

From: NeuroImage <ni@elsevier.com>
Subject: NIMG-11-2734: Decision
Date: January 17, 2012 9:17:41 PM CST
To: poldrack@mail.utexas.edu

Reviewer #1: This is an article by Poldrack, Barch, Mitchell, Wager, Wagner, Cumba, and Milham in which, once again, we are told how great the science will be when we all happily share our data. The community is told that we need to contribute our data to the latest, greatest database effort for the betterment of brain research, and while that goal certainly is a laudable one, the "real" motivations which govern people's interests for so doing are never discussed and certainly not done so in this article. Moreover, the authors appear to be promising that "this time it will be different!" One wonders haven't we heard this tune before? Regrettably, this article does little to be convincing that this effort is any different from previous data sharing enterprises. …

With the OpenfMRI project simply the latest in a series of flavor of the month data sharing enterprises, the authors have a very long way to go before the project and this paper are mature enough to be a solid demonstration of the benefits of sharing for the neuroimaging field. It is a year away and at least a dozen more data sets away from a proper initial publishable report. At present, as viewed through this manuscript, it's just the most recent neuroimaging databasing effort playing the same old, tired, unconvincing song.

2010: The year data sharing broke in neuroimaging

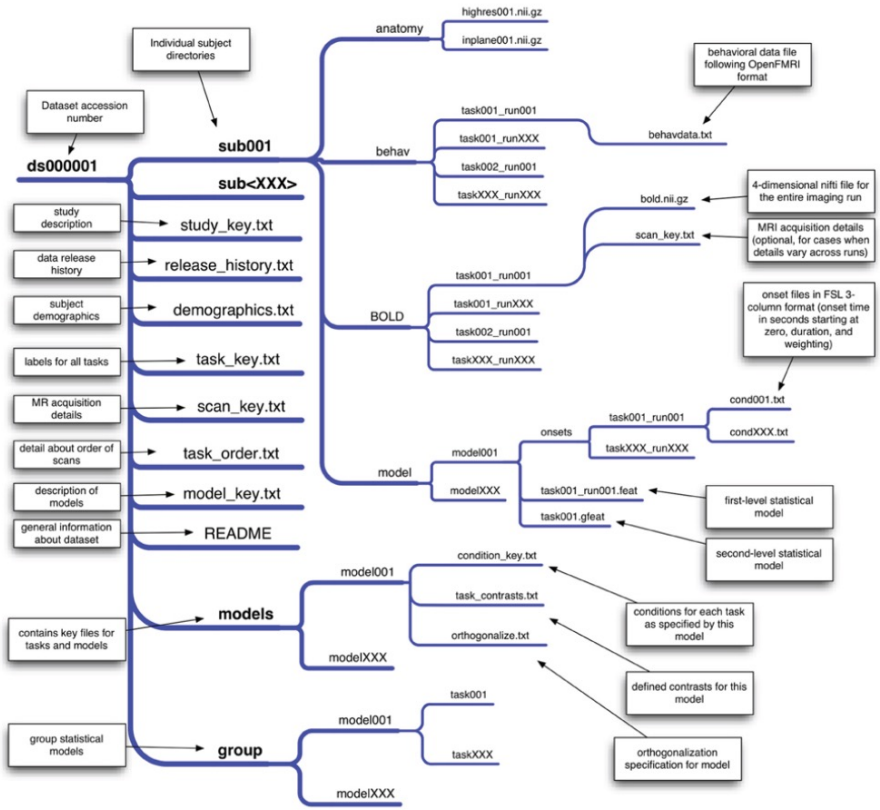
Toward discovery science of human brain function

Bharat B. Biswal[a], Maarten Mennes[b], Xi-Nian Zuo[b], Suril Gohel[a], Clare Kelly[b], Steve M. Smith[c], Christian F. Beckmann[c], Jonathan S. Adelstein[b], Randy L. Buckner[d], Stan Colcombe[e], Anne-Marie Dogonowski[f], Monique Ernst[g], Damien Fair[h], Michelle Hampson[i], Matthew J. Hoptman[j], James S. Hyde[k], Vesa J. Kiviniemi[l], Rolf Kötter[m], Shi-Jiang Li[n], Ching-Po Lin[o], Mark J. Lowe[p], Clare Mackay[c], David J. Madden[q], Kristoffer H. Madsen[f], Daniel S. Margulies[r], Helen S. Mayberg[s], Katie McMahon[t], Christopher S. Monk[u], Stewart H. Mostofsky[v], Bonnie J. Nagel[w], James J. Pekar[x], Scott J. Peltier[y], Steven E. Petersen[z], Valentin Riedl[aa], Serge A. R. B. Rombouts[bb], Bart Rypma[cc], Bradley L. Schlaggar[dd], Sein Schmidt[ee], Rachael D. Seidler[ff,u], Greg J. Siegle[gg], Christian Sorg[hh], Gao-Jun Teng[ii], Juha Veijola[jj], Arno Villringer[ee,kk], Martin Walter[ll], Lihong Wang[q], Xu-Chu Weng[mm], Susan Whitfield-Gabrieli[nn], Peter Williamson[oo], Christian Windischberger[pp], Yu-Feng Zang[qq], Hong-Ying Zhang[ii], F. Xavier Castellanos[b,j], and Michael P. Milham[b,1]

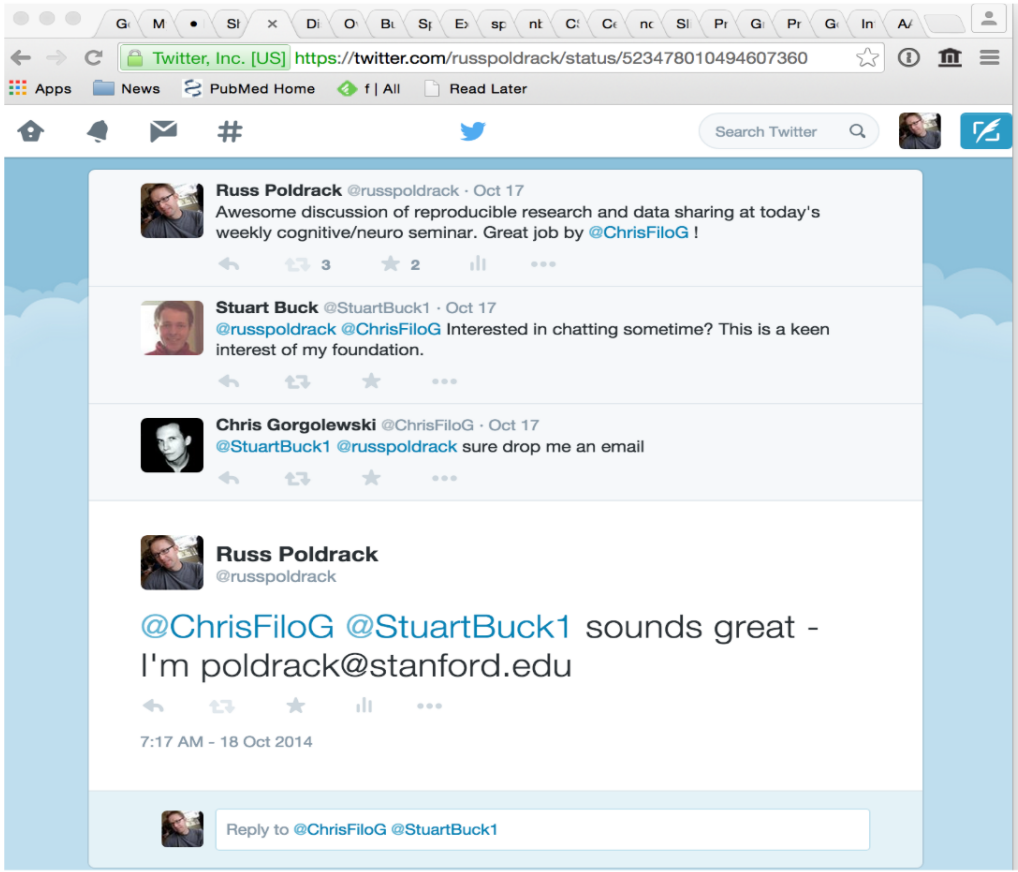
- Comprehensive mapping of the functional connectome, and its subsequent exploitation to discern genetic influences and brain–behavior relationships, will require multicenter collaborative datasets. Here we initiate this endeavor by gathering R-fMRI data from 1,414 volunteers collected independently at 35 international centers. We demonstrate a universal architecture of positive and negative functional connections, as well as consistent loci of inter-individual variability. …

Challenges of OpenfMRI

- Data were manually curated
  - Resulted in very slow ingestion of datasets
- The data organization scheme was ad hoc
- There was no way to validate a dataset
  - Other than running it through the automated analysis workflow and seeing if it crashed!

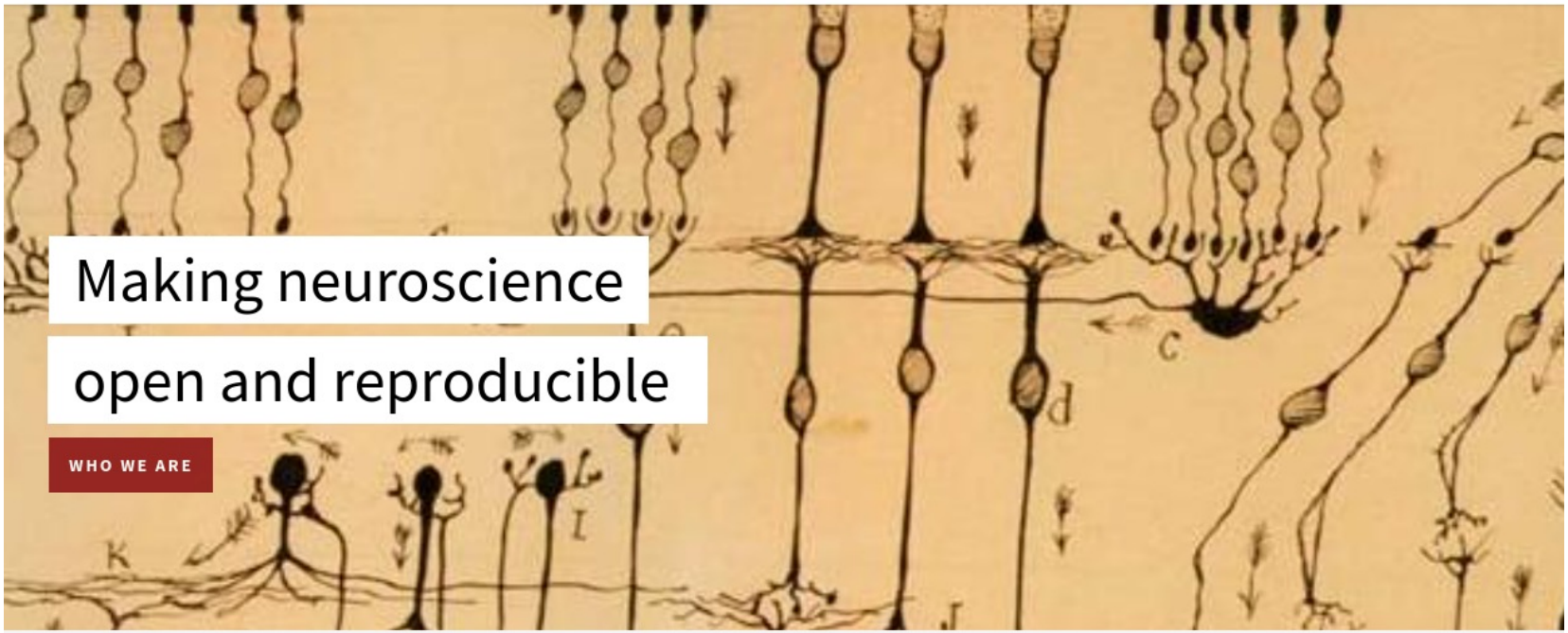


The birth of OpenNeuro



Stanford | Center for Reproducible Neuroscience

HOME ABOUT THE CENTER WHO WE ARE RESOURCES BLOG THE OHBM REPLICATION AWARD



Reproducibility matters

Neuroscience research is the basis for critical decisions about health and society. Our first goal as researchers is to ensure that the results of our research will stand the test of time.

Enabling better research

The OpenNeuro (previously OpenfMRI) project is a free and open platform that enables the dissemination of neuroimaging data.

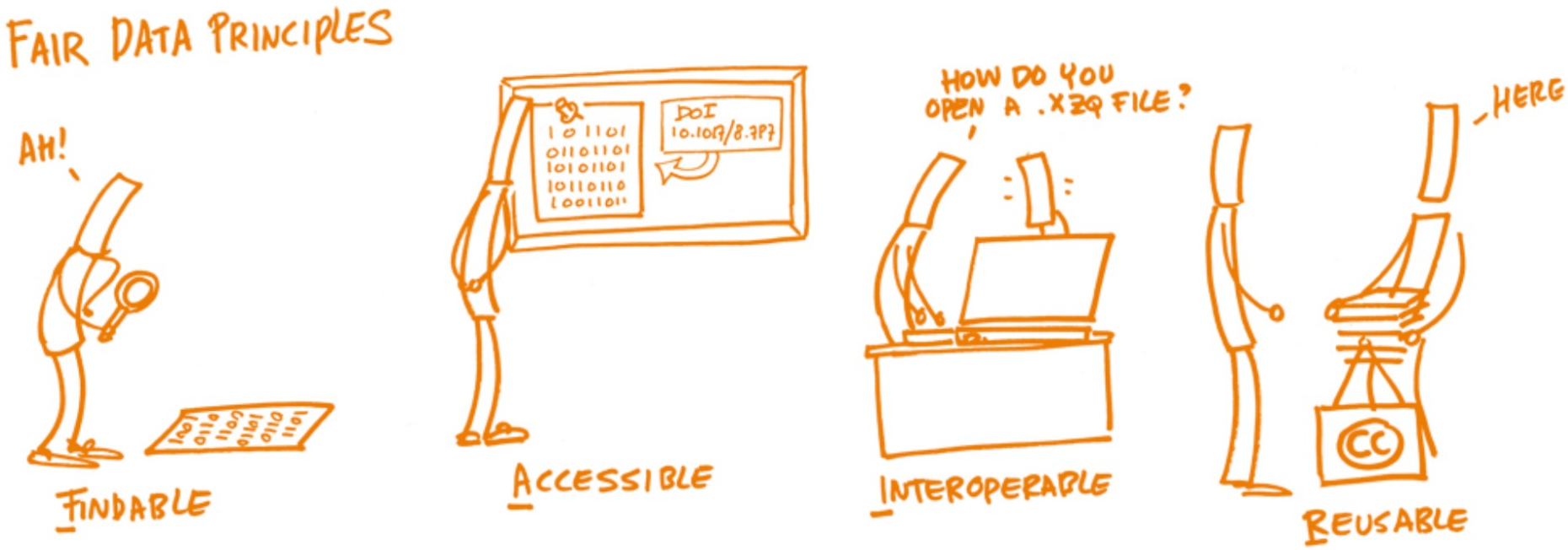
From data to discovery

We will provide neuroimaging researchers with leading-edge tools to analyze and share large datasets, with a focus on quantifying the reproducibility of the results.

Simply sharing data is not sufficient

It must be shared in a way that makes it useful!

The FAIR principles for open data



The birth of BIDS

2015-01-27 Bay area 2015 - NIDM-Experiment meeting: Agenda

File Edit View Insert Format Tools Extensions Help Last edit was made on June 30, 2015 by D...

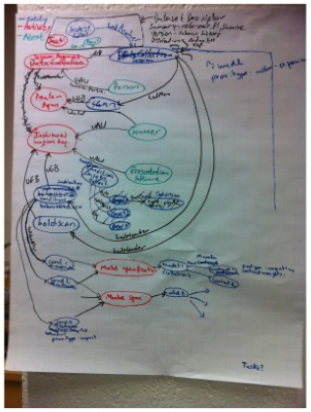
Share

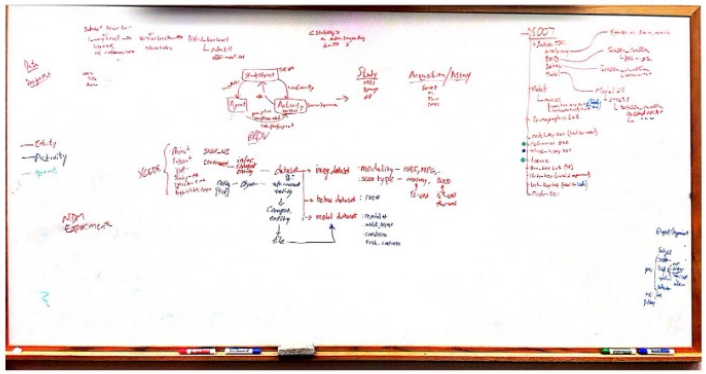
NIDM-experiment Bay Area January 27th-30th, 2015

Participants

JB Poline (JB), Nolan Nichols (NN), Camille Maumet (CM), Guillaume Flandin (GF), Karl Helmer (KH), Chris Filo Gorgolewski (CG), David Keator (DKea), Russ Poldrack (RP), Mathew Abrams (MA), Ariel Rokem (AR), Vanessa Sochat (VS), Michel Dumontier (MD)

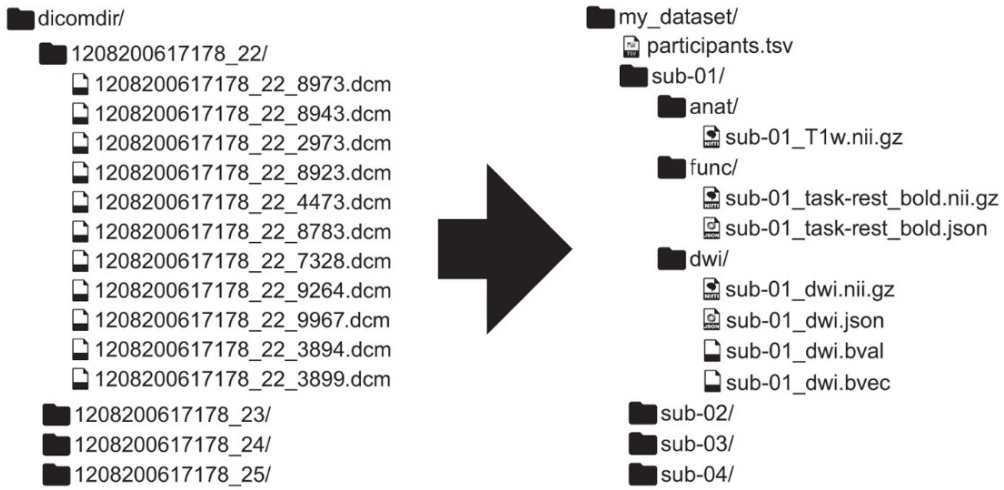
Subgroup1: OBIDS (Open Brain Imaging Data Structure) format proposal (derived from the openfmri format)



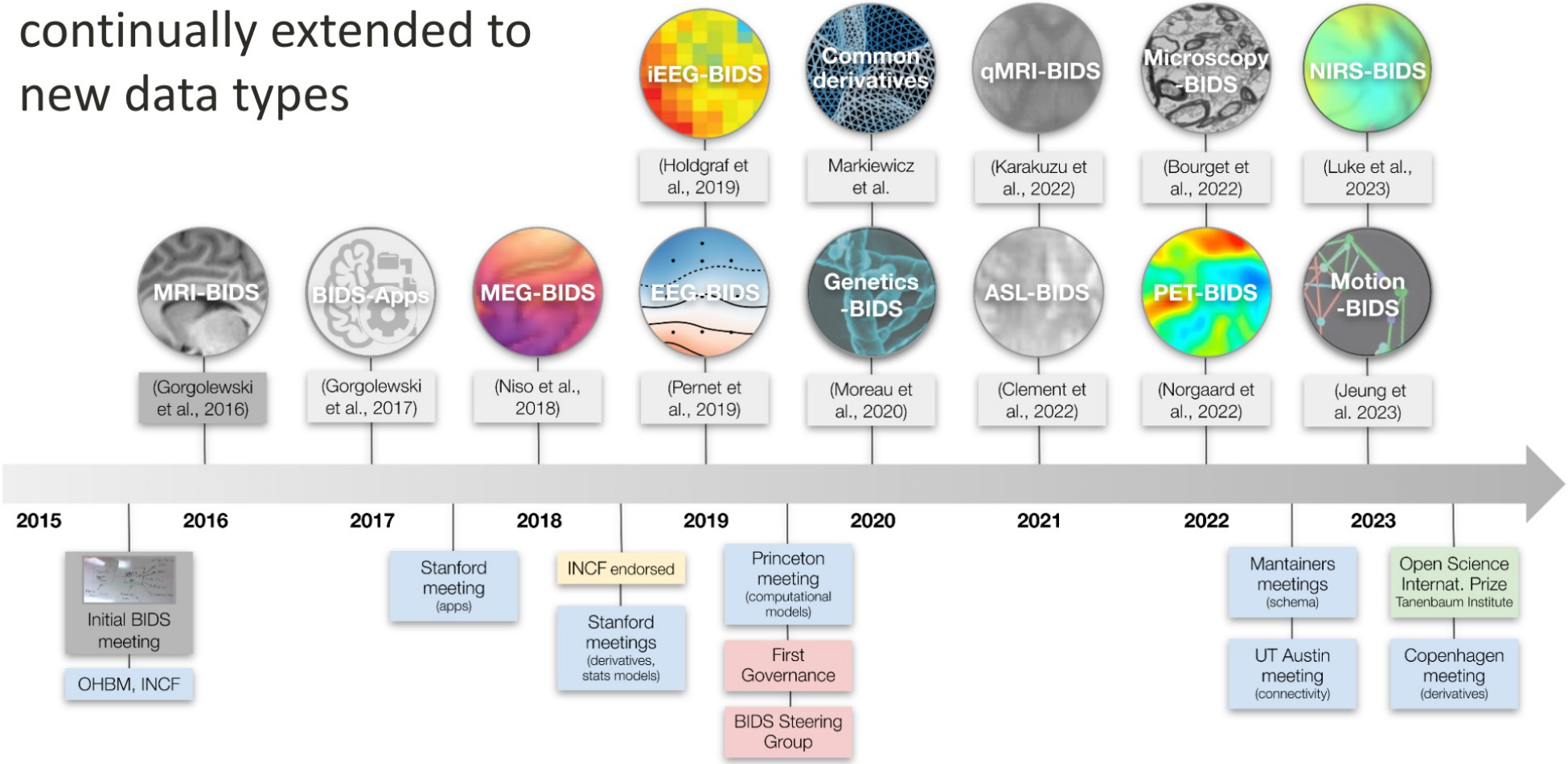


The Brain Imaging Data Structure (BIDS)

- A community-based open standard for neuroimaging data
  - A file organization standard
  - A metadata standard



BIDS has been continually extended to new data types

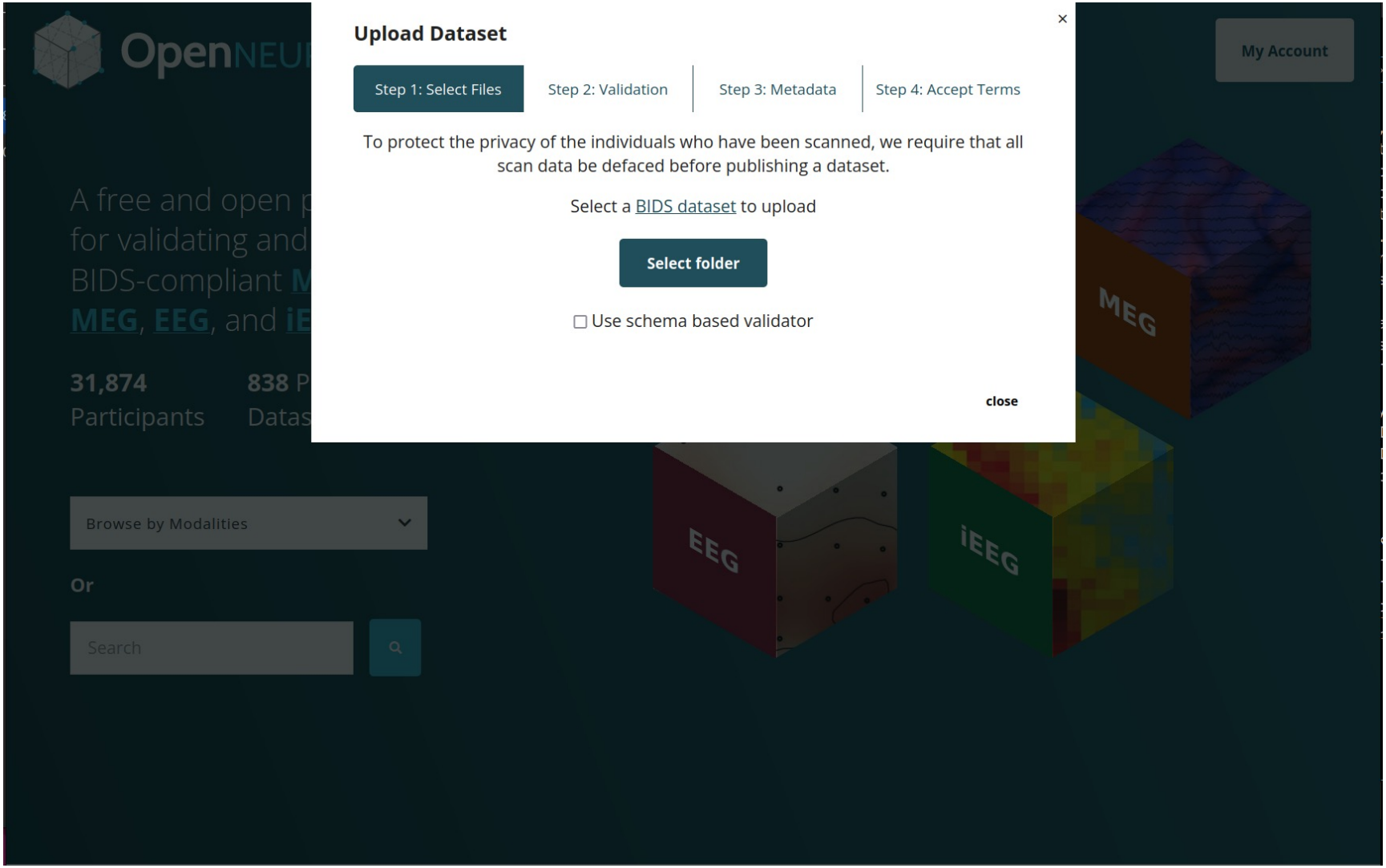


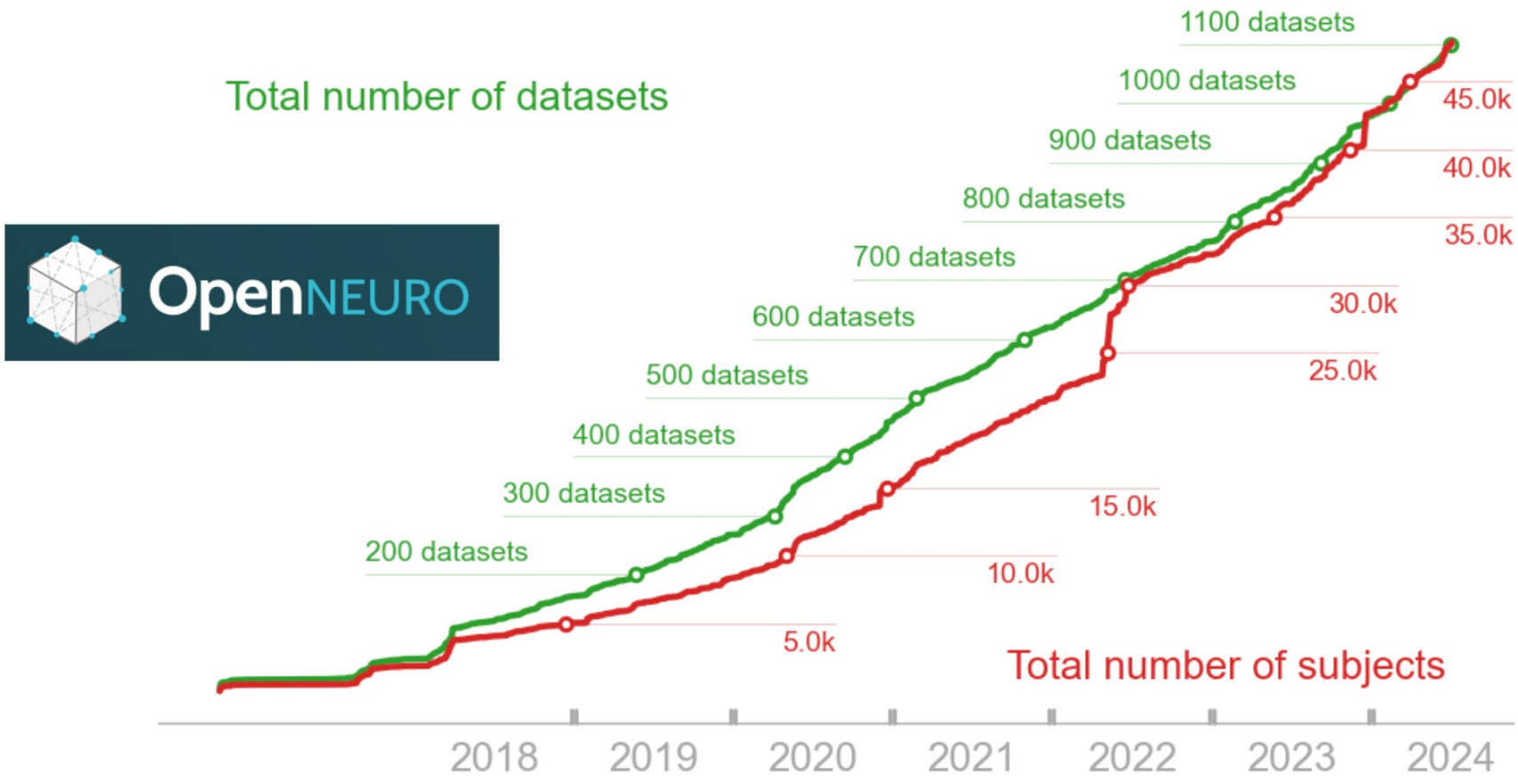
Imaging Neuroscience
an open access journal

Review

The past, present, and future of the brain imaging data structure (BIDS)

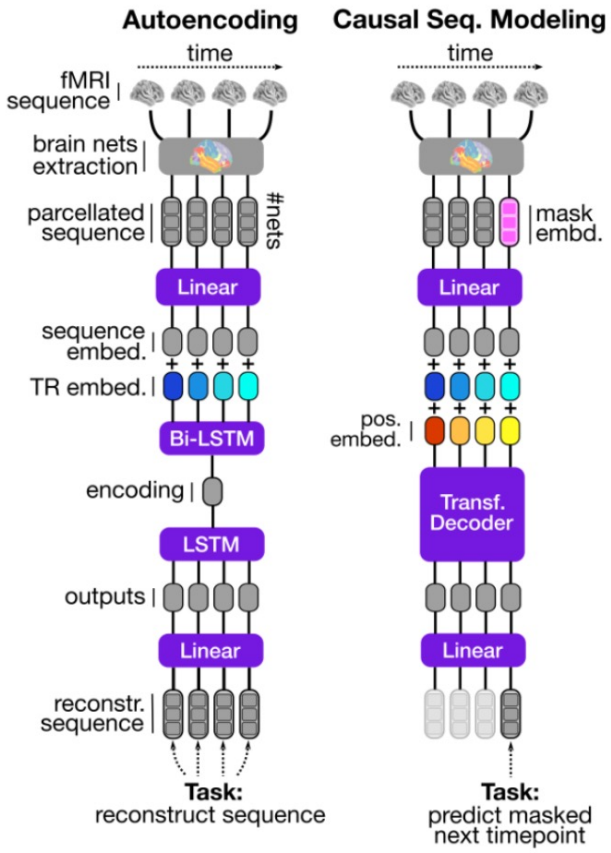
Russell A. Poldrack[1], Christopher J. Markiewicz[1], Stefan Appelhoff[2], Yoni K. Ashar[3], Tibor Auer[4,5], Sylvain Baillet[6], Shashank Bansal[7], Leandro Beltrachini[8], Christian G. Benar[9], Giacomo Bertazzoli[10,11,12,13,14], Suyash Bhogawar[15], Ross W. Blair[1], Marta Bortoletto[10], Mathieu Boudreau[16], Teon L. Brooks[1], Vince D. Calhoun[17], Filippo Maria Castelli[18,19], Patricia Clement[20,21], Alexander L. Cohen[22], Julien Cohen-Adad[16], Sasha D'Ambrosio[23,24], Gilles de Hollander[25], María de la Iglesia-Vayá[26], Alejandro de la Vega[27], Arnaud Delorme[28], Orrin Devinsky[29], Dejan Draschkow[30], Eugene Paul Duff[31], Elizabeth DuPre[1], Eric Earl[32], Oscar Esteban[33], Franklin W. Feingold[1], Guillaume Flandin[34], Anthony Galassi[32], Giuseppe Gallitto[35,36], Melanie Ganz[37,38], Rémi Gau[39], James Gholam[40], Satrajit S. Ghosh[41], Alessio Giacomel[42], Ashley G. Gillman[43], Padraig Gleeson[44], Alexandre Gramfort[45], Samuel Guay[46], Giacomo Guidali[47], Yaroslav O. Halchenko[48], Daniel A. Handwerker[32], Nell Hardcastle[1], Peer Herholz[49], Dora Hermes[50], Christopher J. Honey[51], Robert B. Innis[32], Horea-Ioan Ioanas[48], Andrew Jahn[52], Agah Karakuzu[16], David B. Keator[53,54,55], Gregory Kiar[56], Balint Kincses[35,36], Angela R. Laird[57], Jonathan C. Lau[58], Alberto Lazari[59], Jon Haitz Legarreta[60], Adam Li[61], Xiangrui Li[62], Bradley C. Love[63], Hanzhang Lu[64], Eleonora Marcantoni[65], Camille Maumet[66], Giacomo Mazzamuto[67], Steven L. Meisler[68], Mark Mikkelsen[69], Henk Mutsaerts[70,71], Thomas E. Nichols[72], Aki Nikolaidis[73], Gustav Nilsonne[74,75], Guiomar Niso[76], Martin Norgaard[32,37], Thomas W. Okell[59], Robert Oostenveld[77,78], Eduard Ort[79], Patrick J. Park[80], Mateusz Pawlik[81], Cyril R. Pernet[38], Franco Pestilli[27], Jan Petr[82], Christophe Phillips[83], Jean-Baptiste Poline[84], Luca Pollonini[85,86], Pradeep Reddy Raamana[87], Petra Ritter[88,89,90,91,92], Gaia Rizzo[93,94], Kay A. Robbins[95], Alexander P. Rockhill[96], Christine Rogers[97], Ariel Rokem[98], Chris Rorden[99], Alexandre Routier[100], Jose Manuel Saborit-Torres[26], Taylor Salo[101], Michael Schirner[88,89,90,91,92], Robert E. Smith[102,103], Tamas Spisak[35,104], Julia Sprenger[105], Nicole C. Swann[106], Martin Szinte[105], Sylvain Takerkart[105], Bertrand Thirion[45], Adam G. Thomas[32], Sajjad Torabian[107], Gael Varoquaux[108], Bradley Voytek[109], Julius Welzel[110], Martin Wilson[111], Tal Yarkoni[112], Krzysztof J. Gorgolewski[1]


Upload Dataset
Step 1: Select Files
Step 2: Validation
Step 3: Metadata
Step 4: Accept Terms
To protect the privacy of the individuals who have been scanned, we require that all scan data be defaced before publishing a dataset.
Select a BIDS dataset to upload
Select folder
Use schema based validator
close
My Account
MEG
EEG
iEEG
31,874
Participants
Browse by Modalities
Or
Search

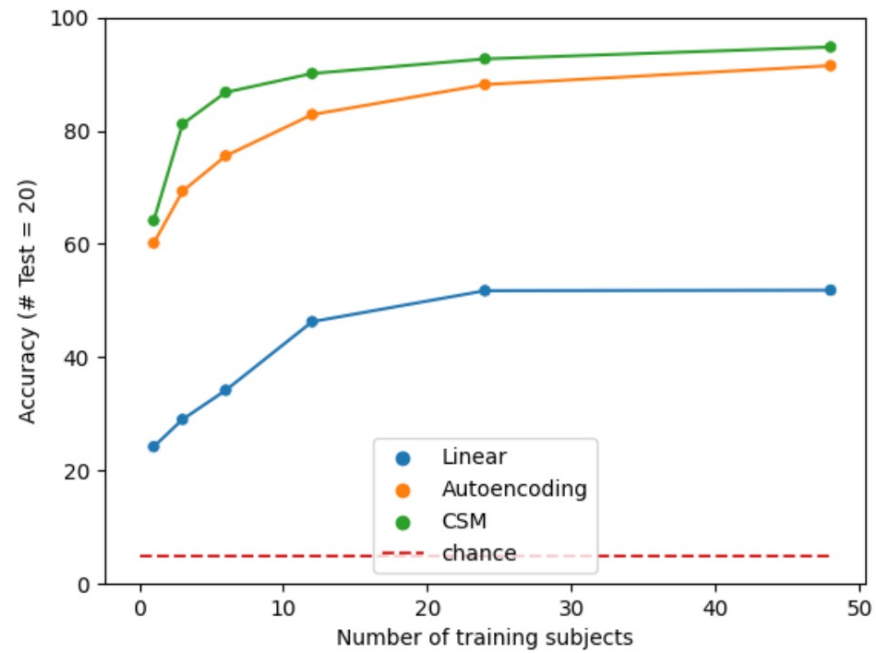


updated from Markiewicz et al, 2021, *eLife*

Example of OpenNeuro reuse: A foundation model for fMRI



Models trained on 11,980 fMRI runs from 1,726 individuals across 34 OpenNeuro datasets



Thomas et al. 2022, NeurIPS

An open source software ecosystem for neuroimaging

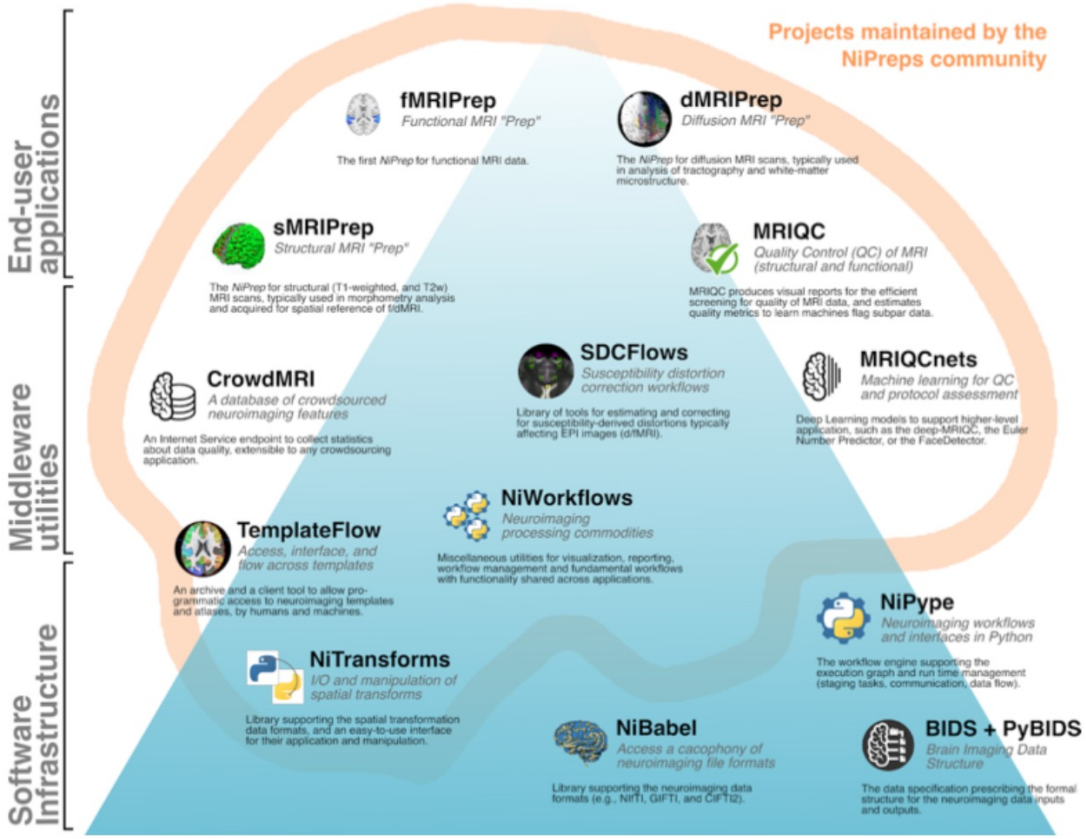
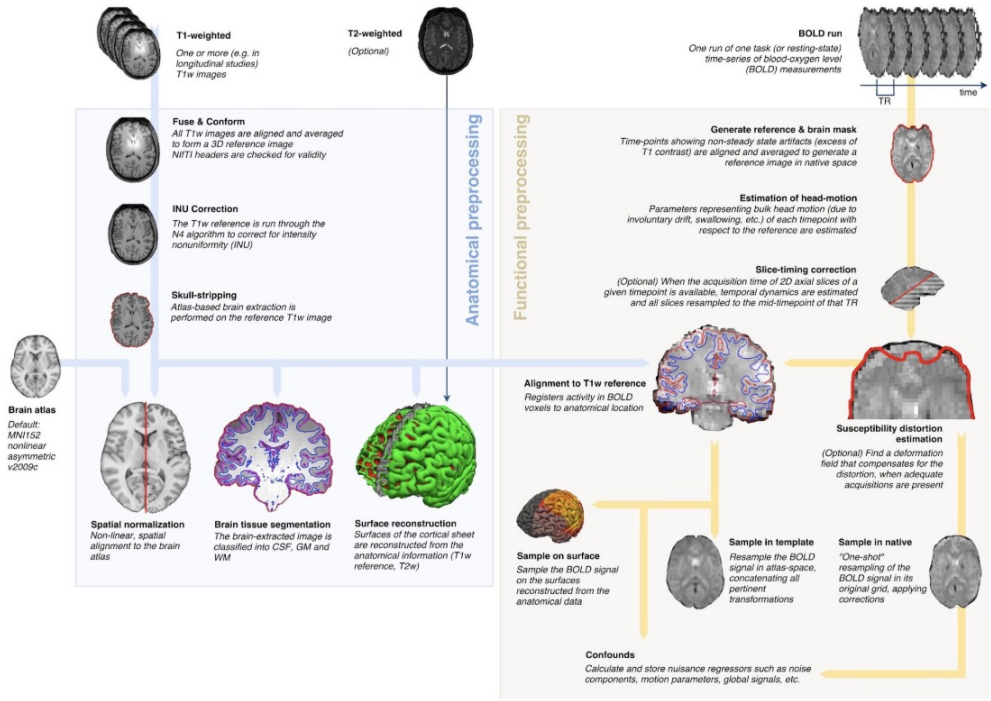


Figure 1. The NIPreps framework. *NIPreps involve some existing neuroimaging tools and their extensions to integrate new modalities and support the preprocessing of populations of interest and species other than neurotypical humans.*

fMRIPrep: a new design pattern for fMRI preprocessing

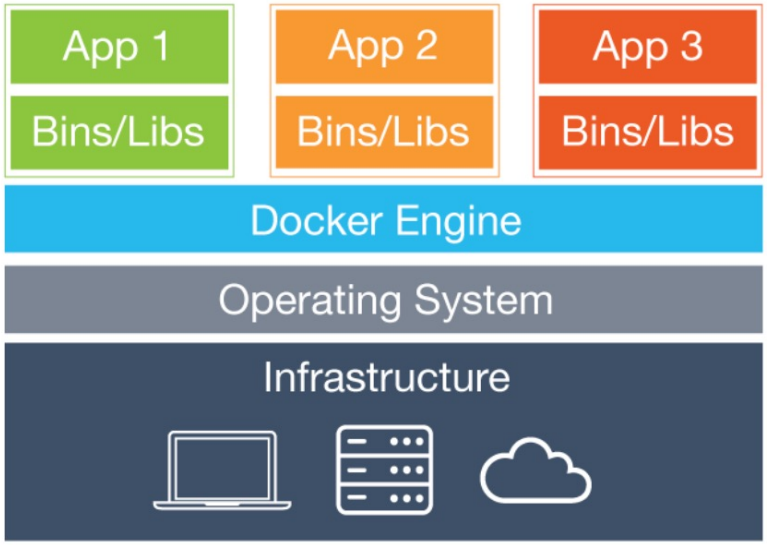
- **A robust workflow for preprocessing fMRI data**
  - Adapt to each dataset and processes it in the best way possible
  - Outputs to BIDS-Derivatives format
  - Provides powerful visualizations :
    - to help identify problems
    - allowing glass-box access to help train researchers



Esteban et al. 2019

BIDS apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods

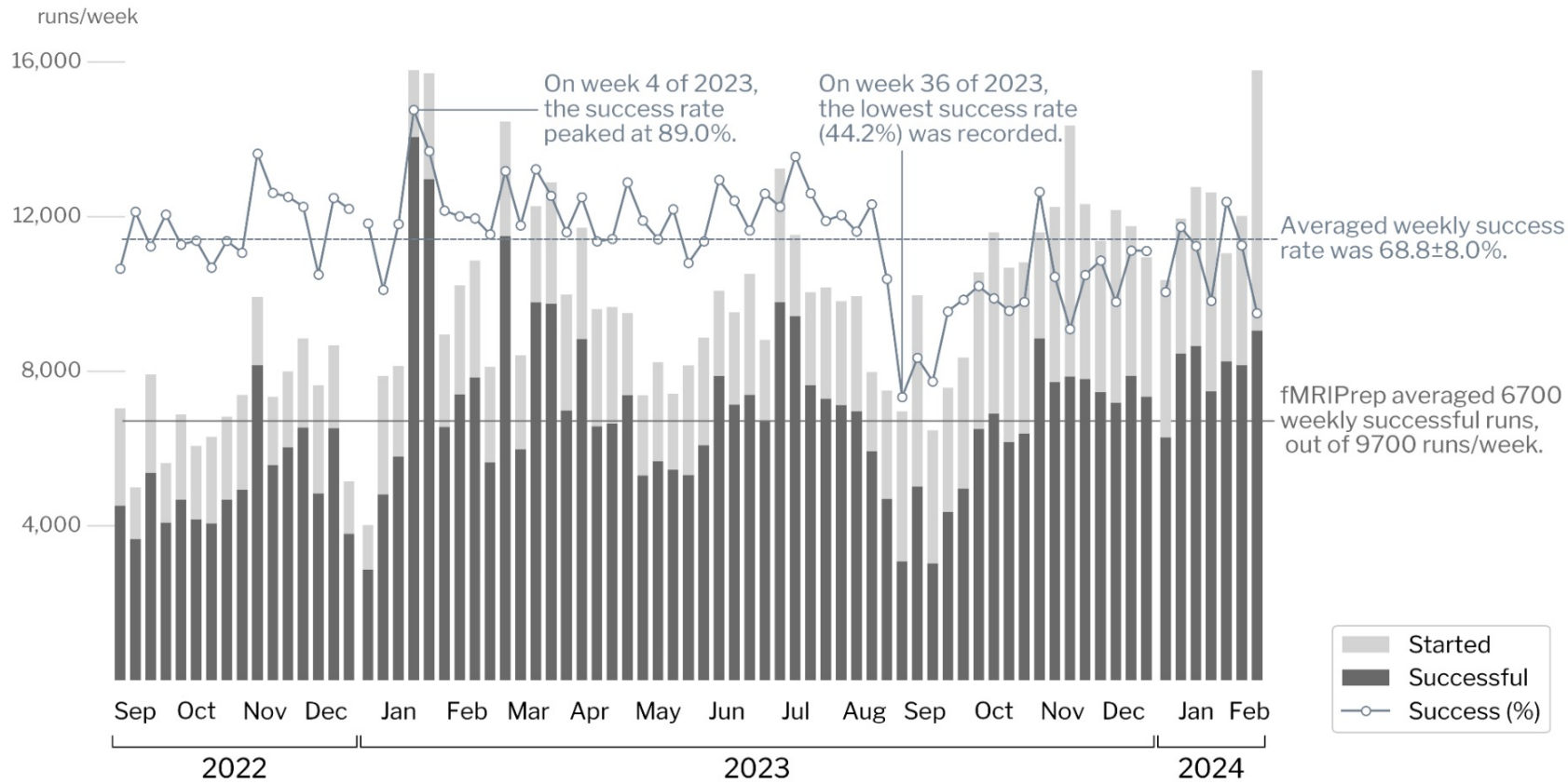
Krzysztof J. Gorgolewski[1]*, Fidel Alfaro-Almagro[2], Tibor Auer[3], Pierre Bellec[4,5], Mihai Capotă[6], M. Mallar Chakravarty[7,8], Nathan W. Churchill[9], Alexander Li Cohen[10], R. Cameron Craddock[11,12], Gabriel A. Devenyi[7,8], Anders Eklund[13,14,15], Oscar Esteban[1], Guillaume Flandin[16], Satrajit S. Ghosh[17,18], J. Swaroop Guntupalli[19], Mark Jenkinson[2], Anisha Keshavan[20], Gregory Kiar[21,22], Franziskus Liem[23], Pradeep Reddy Raamana[24,25], David Raffelt[26], Christopher J. Steele[7,8], Pierre-Olivier Quirion[15], Robert E. Smith[26], Stephen C. Strother[24,25], Gaël Varoquaux[27], Yida Wang[6], Tal Yarkoni[28], Russell A. Poldrack[1]



BIDS App: A containerized software application

The container encapsulates all software and dependencies
- allows better reproducibility

Telemetry provides information about usage and user success



Publish your computer code: it is good enough

Freely provided working code — whatever its quality — improves programming and enables others to engage with your research, says **Nick Barnes**.

14 OCTOBER 2010 | VOL 467 | NATURE | 753

If your code is not good enough to share, then why is it good enough to publish the results that rely upon it?

Summary

- Data needs to be findable, accessible, interoperable and reusable (FAIR) to be useful
- Sharing data in a standardized format improves reusability
- When standardized data is shared, it is easier to develop code and software that targets this structure,
  and this can be made fully reproducible with containers.

Interactive group activity: Optimizing Neuroimaging Pipelines and workflows

Interactive group activity: Optimizing Neuroimaging Pipelines

Objectives

- Identify common reproducibility challenges in neuroimaging workflows.
- Propose practical strategies to improve pipeline transparency and reliability.

Format

- Work in small groups (3–5 people).
- Each group gets a project scenario.
- Discuss for 10–12 minutes
  - Where are the reproducibility risks? What strategies/tools would improve it?
- Report back (2 min per group).

Interactive group activity: Optimizing Neuroimaging Pipelines

Scenario A: fMRI + Behavioral Task

- **Study:** Cognitive control task, 30 participants, fMRI scanning.
- **Pipeline:** Data preprocessed manually in SPM, custom MATLAB scripts for stats.
- **Issues observed:**
  - Motion correction parameters chosen ad hoc.
  - No version control for scripts.
  - Data stored on personal laptop.

Questions for discussion:

- How would you standardize preprocessing?
- What tools/databases could help ensure reproducibility?
- How should data and code be shared?

Scenario B: EEG Study on Sleep

- **Study:** Overnight EEG on 15 participants, investigating sleep spindles.
- **Pipeline:** In-house scripts written over 3 years by different lab members.
- **Issues observed:**
  - Different preprocessing pipelines used across participants.
  - Metadata (electrode placement, sleep stage scoring) missing for some datasets.
  - Results stored as Excel files without raw data links.

https://github.com/openneuropet/outreach/blob/main/D3A2025/group_activity.pdf

Interactive group activity: Optimizing Neuroimaging Pipelines

https://www.menti.com/al4njmm23exz

Practical session: Neuroinformatics Tools & Database Management

Data sharing

Tools for conversion
https://bids.neuroimaging.io/tools/converters.html

Tools for validation
https://bids.neuroimaging.io/tools/validator.html

Databases
https://openneuro.org/

Preprocessing

BIDS applications
https://bids.neuroimaging.io/tools/bids-apps.html

Tools for handling data
https://bids.neuroimaging.io/tools/others.html

Visualization & analysis

Visualization
https://nilearn.github.io/dev/index.html

Analysis
https://nipype.readthedocs.io/en/latest/

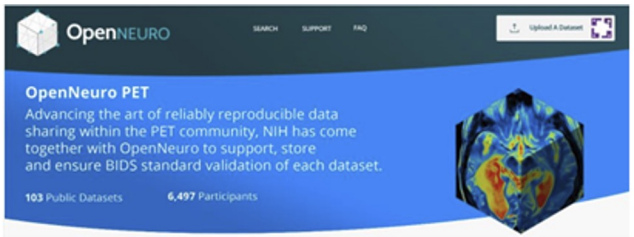
Discussion & wrap-up

Key messages

- Standardize your workflows if possible.
- Use open-source tools & formats.
- Document everything (metadata, pipelines, decisions).

Thank you for your attention

Georgia State University



novo nordisk fonden

NIH National Institute of Mental Health

Athinoula A. Martinos Center For Biomedical Imaging

Harvard–MIT Health Sciences & Technology

MASSACHUSETTS GENERAL HOSPITAL

Stanford University

University of Copenhagen

REGION H

Rigshospitalet

Different levels of variability in post-processing pipelines

- Different end-to-end pipelines [Carp 2012]
- Different software version [Gronenschild 2012]
- Different operating system [Glatard 2015]

- Different **post-processing software packages** [Bowring 2019, 2021]