Breakout Session 5:

National Sleep Research Resource

Dr. Shaun Purcell
Associate Professor, Brigham & Women's Hospital, Harvard Medical School

The National Sleep Research Resource (NSRR)

NIH STRIDES
January 2024

Shaun Purcell
smpurcell@bwh.harvard.edu
http://zzz.bwh.harvard.edu

BWH BRIGHAM AND WOMEN'S HOSPITAL

HARVARD MEDICAL SCHOOL

"The National Sleep Research Resource (NSRR) offers free web access to large collections of de-identified physiological signals and clinical data elements collected in well-characterized research cohorts and clinical trials."

Founded in 2014 by **Dr. S Redline**

Funded by the **National Heart, Blood and Lung Institute** (Resource Grant → NIH contract)

Polysomnography signal data
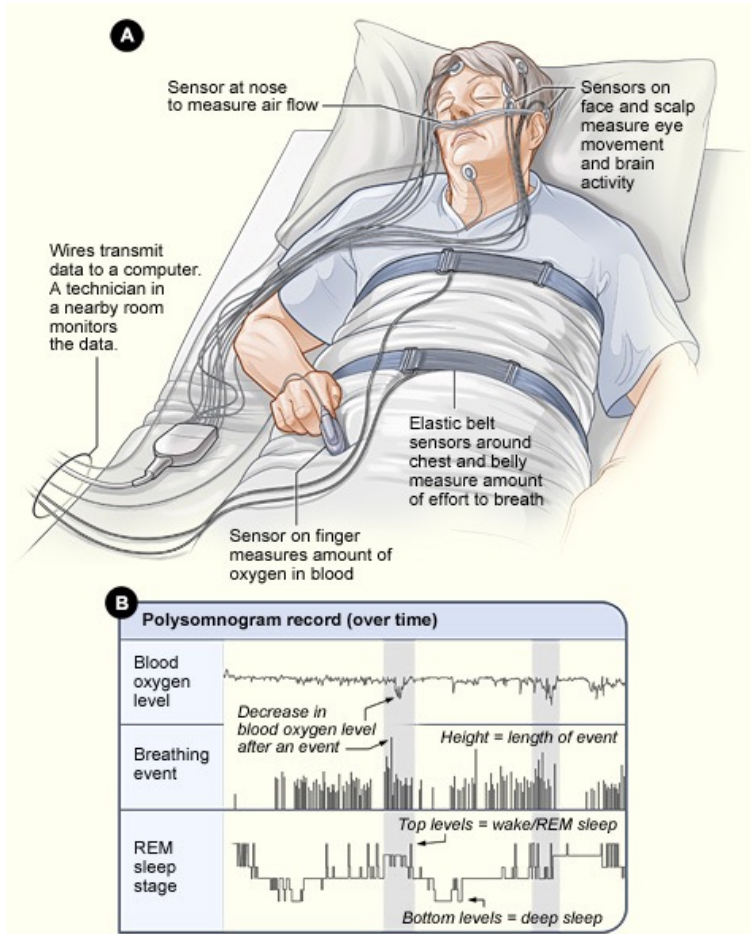
Classically, PSG data often reduced to only ~dozen common metrics
- sleep stage duration, onset latency, apnea hypopnea index, etc

Rich, multi-modal & dynamic raw signal data
- often requires specialized tools/analytic approaches
  (historically, often via in-house *ad hoc* Matlab scripts)

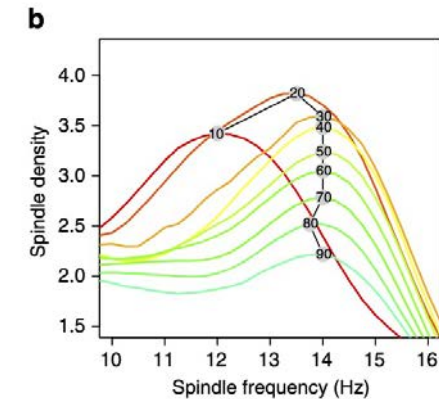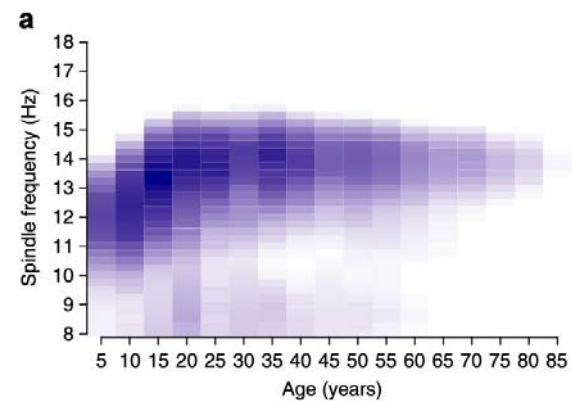Whole night EEG spectrograms & hypnograms for 3 individuals
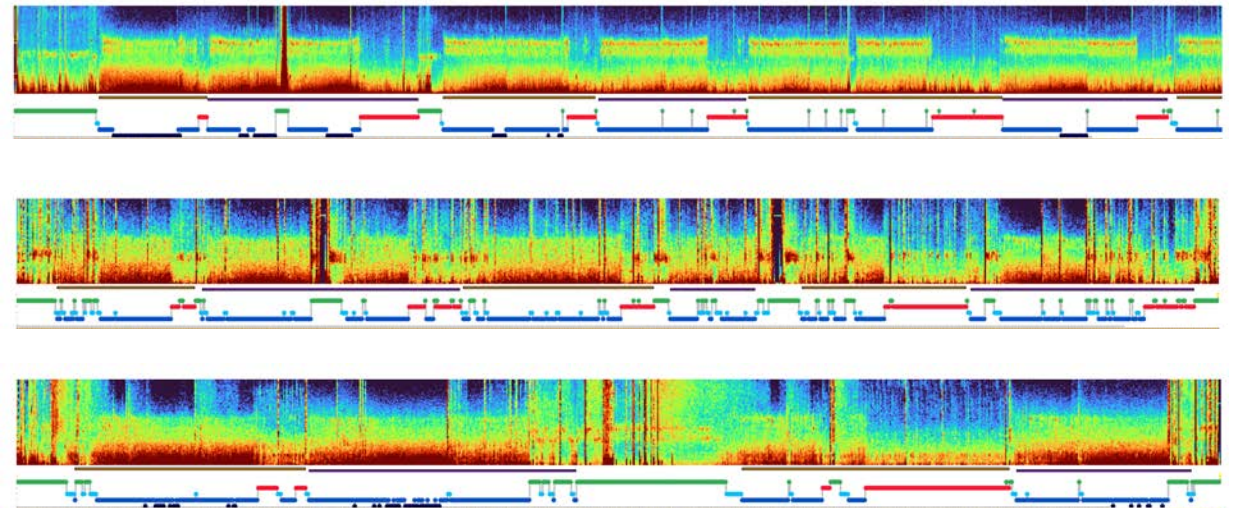
Scaling up sleep science

- NREM EEG microarchitecture

- *N* > 10,000 aged 2-90

- Captures person-to-person variability in markers of sleep & brain function/development
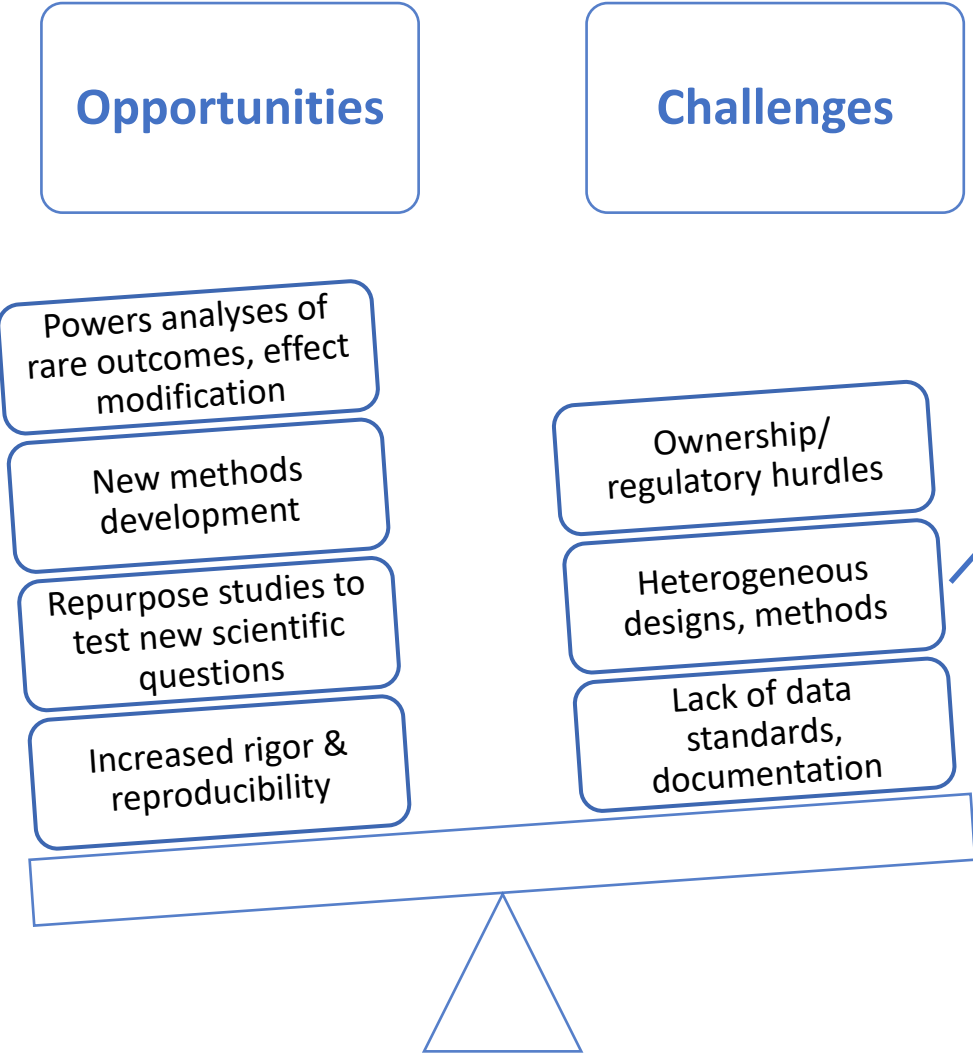
- Opportunities to link to genetic risk

Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource

S.M. Purcell[1,2,3], D.S. Manoach[2,4,5], C. Demanuele[2,4,5], B.E. Cade[6,7], S. Mariani[6,7], R. Cox[2,8], G. Panagiotaropoulou[2,4,5], R. Saxena[9,10,11], J.Q. Pan[12], J.W. Smoller[2,4,13], S. Redline[2,6,7,*] & R. Stickgold[2,8,*]

# Data archiving & aggregation: important but often challenging

Opportunities

Challenges

Powers analyses of rare outcomes, effect modification

New methods development

Repurpose studies to test new scientific questions

Increased rigor & reproducibility

Ownership/ regulatory hurdles

Heterogeneous designs, methods

Lack of data standards, documentation

End user's perspective:

Combining different datasets can be logistically/technically challenging

Subtle biases can be amplified in large datasets

The original studies were likely not specifically designed to answer your current research question

NSRR's attempts to alleviate some of these issues:

- harmonized & well-documented datasets
- tools to work with large datasets
- technical support

Data → Tools → Sharing → Community (cycle diagram)
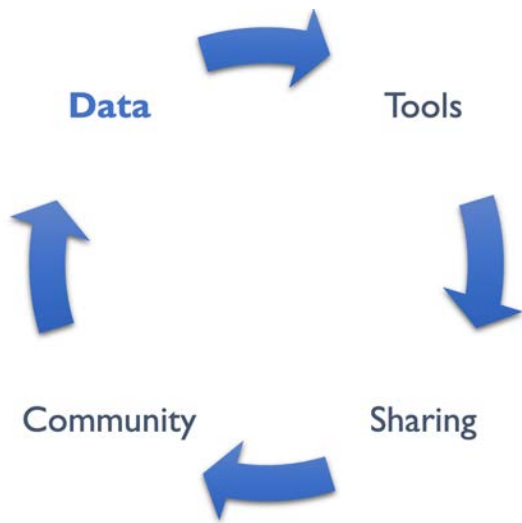
~ 50,000 individuals

> 30,000 full PSGs

~7000 actigraphy studies

2 TB of data shared weekly

4,643 Data Access Use Agreements

> 5,000 defined variables

Ongoing harmonization of variables across studies, mapping to CDEs

Extensive documentation on study design

Data → Tools → Sharing → Community → (cycle)

Raw physiologic signals (EDF) &
annotations on 10,000s of
individuals

In total, ~30 years' worth of
multi-modal sleep signal data

Controlled access to human subjects data

- All data de-identified, but need to adhere to data sharing language in original participant informed consent

- Stream-lined process for DAUA, with proposals reviewed by NSRR staff

- Provide in-house IRB review for users without local IRB

NATIONAL
SLEEP RESEARCH
RESOURCE

ADVANCING SCIENCE GLOBALLY
THROUGH DATA SHARING

- Ongoing **NHLBI contract** to sustain NSRR and facilitate integration with the NHLBI BioData Catalyst (BDC) ecosystem

  - to link sleep data with large amounts of genetics, omics, imaging, bioassays, cognitive, medical and other data available on many of the NSRR cohorts

  - to transition away from our locally-hosted data-download model

  - to foster modern cloud-based, reproducible workflows and analysis capabilities

- **STRIDES** supplement: pending BDC-readiness for data ingestion, to support interim transfer of NSRR data to the cloud, enabling cloud-savvy users direct access

NATIONAL SLEEP RESEARCH RESOURCE

ADVANCING SCIENCE GLOBALLY THROUGH DATA SHARING

- Configured a secure AWS-STRIDES environment
- Custom-built authentication/authorization system for controlled access
- Python client library for users to pull data from S3 buckets to EC2 instances
- Migrated all Cleveland Family Study data

- Bottlenecks encountered:
  - **regulatory approvals** required to transfer key cohorts from on-premise servers to NIH-administered STRIDES AWS environment
  - non-trivial work with NIH, depositors and parent cohorts to amend original data hosting agreements
  - unable to use STRIDES credits on AWS accounts administered by our local institution
  - **lessons learned:** moving forward broader, host-agnostic data use agreement templates for new studies
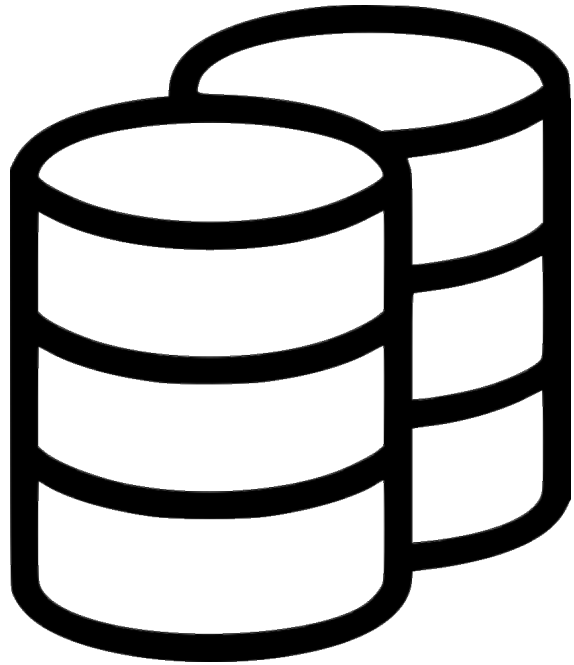
Data need tools

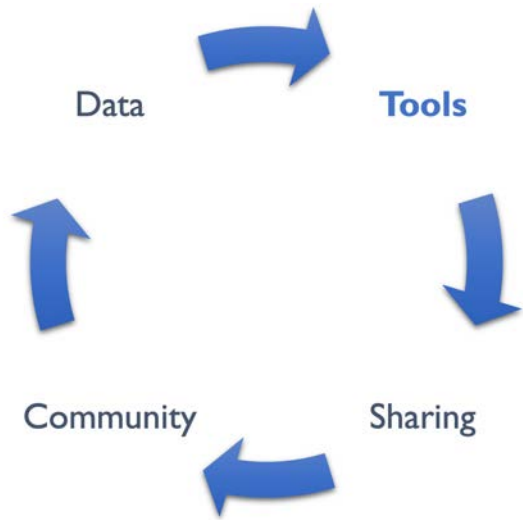What can we learn from these data?

"Reference-driven analysis"

Using old data to inform the analysis of new data

Making tools accessible

Sharing tools as well as data

DATA

TOOLS

Many published results cannot be reproduced
 - open tools and data important for reproducibility
 - but often not transparently reported or shared

Literature review of all sleep methods/algorithms papers published in 2019:

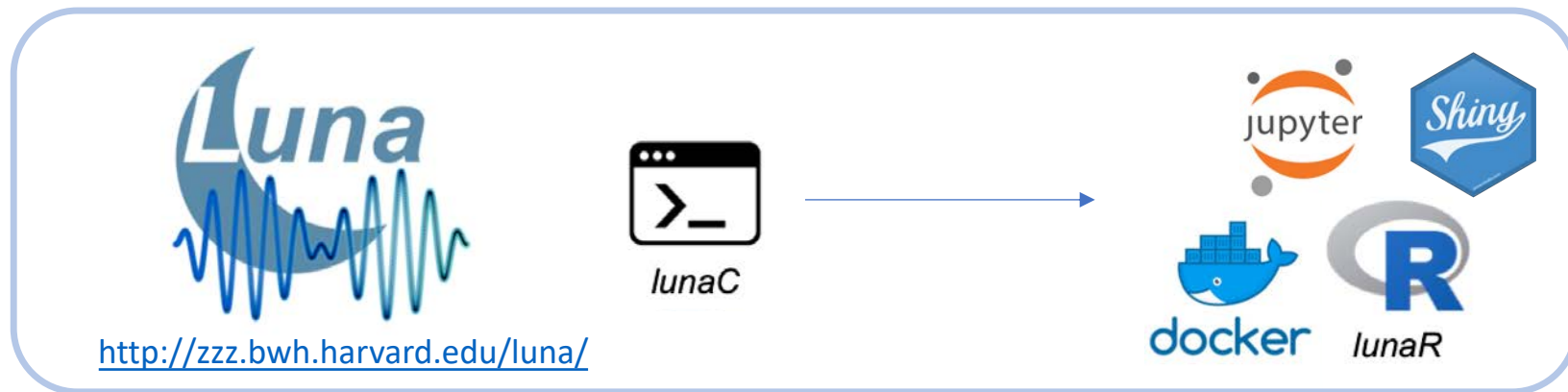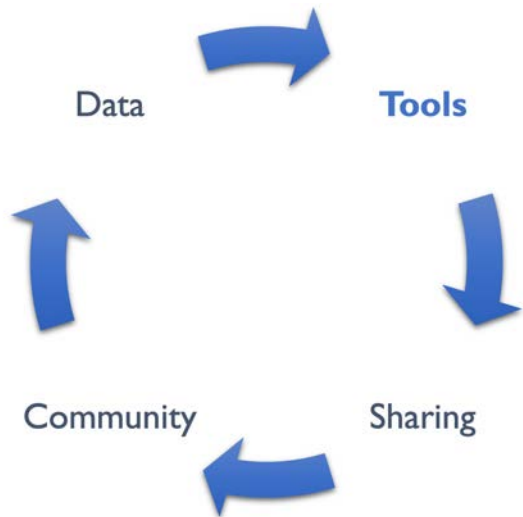80% of relevant papers did not publish either:
   1) code, 2) an example or 3) any data

Of 5 articles that supplied all of these, 4 were published in journals that explicitly ask for code and/or raw data

research:
a call for action

Purcell & Sofer (2024) SLEEP.
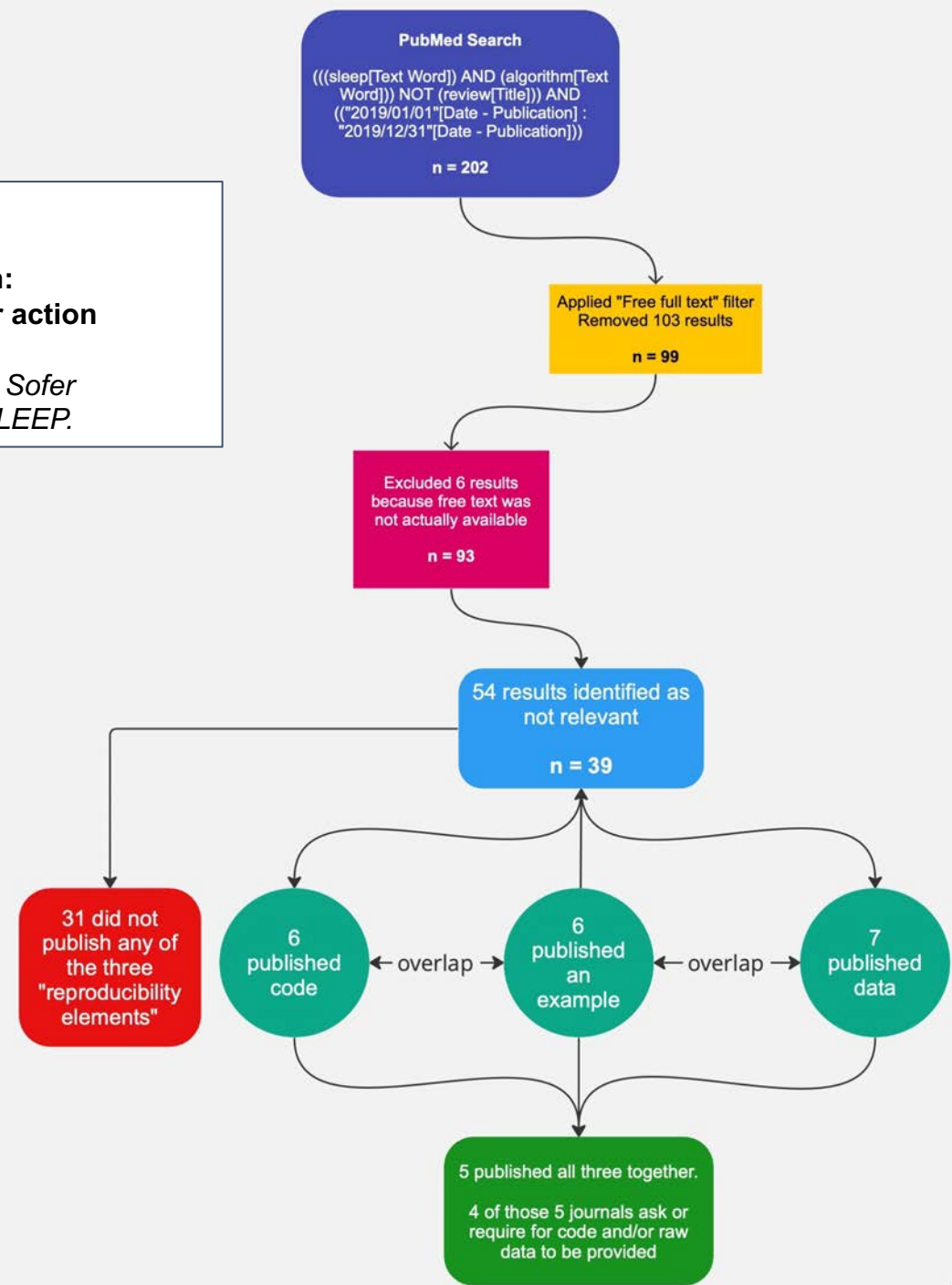
1) Making in-house tools more accessible
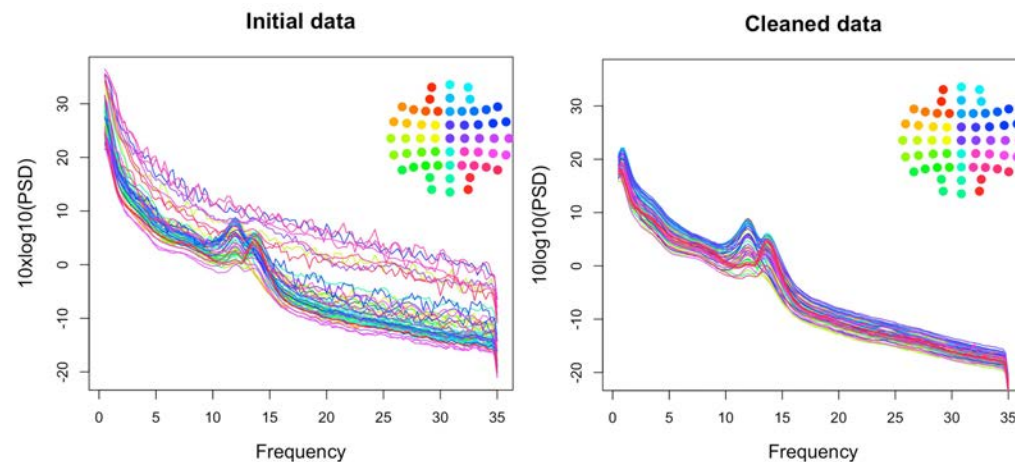
http://zzz.bwh.harvard.edu/luna/

2) Sharing pipelines as well as data

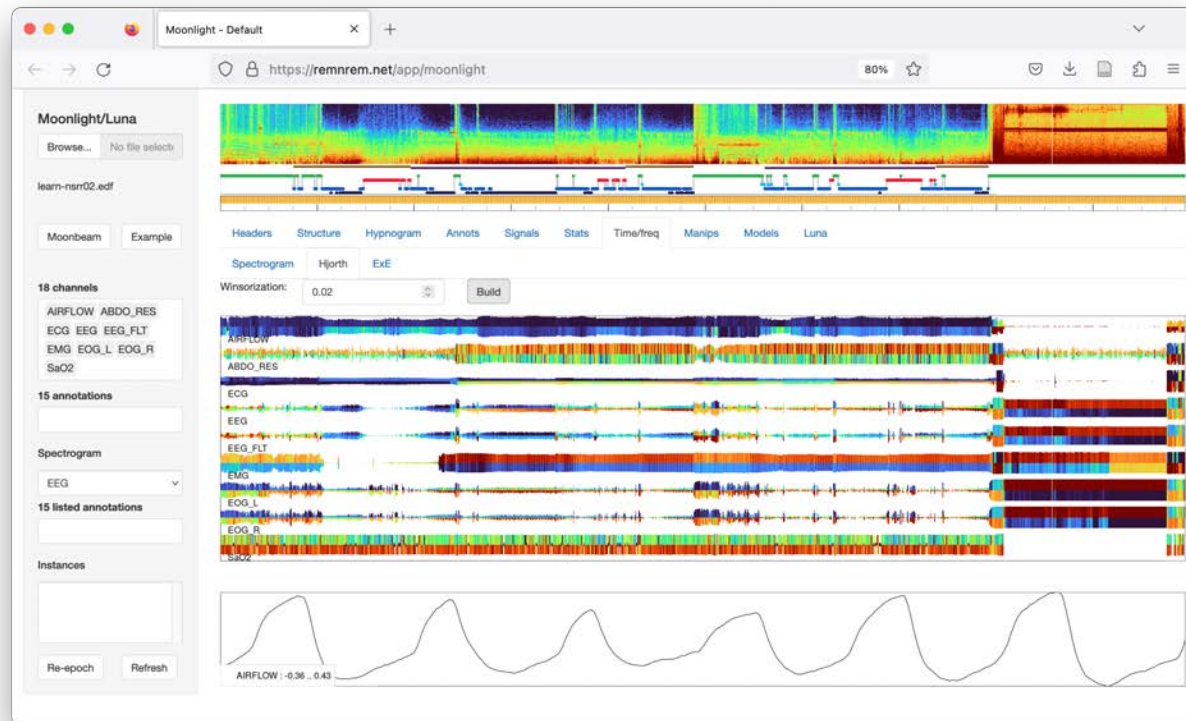**NAP:** NSRR Automated Pipeline

Harmonize/flag issues w/ labels, referencing, units, sample rates, filtering, polarities, corrupt signals, artifact, non-standard channels, (automated) staging alignment, etc

Uniform annotation format
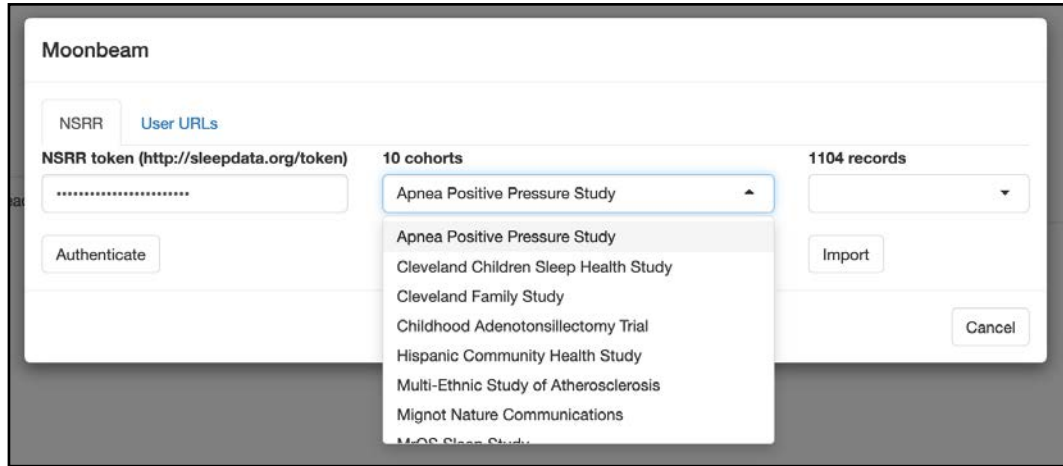
Cloud-based apps for sleep signal data

Moonlight

Hypnoscope

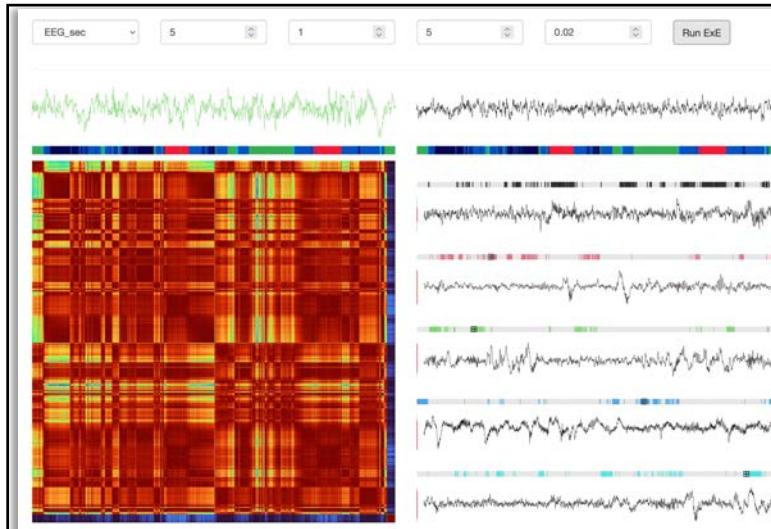# Alternative ways to access NSRR data

Moonbeam

Clinical, demographic & other phenotypic data

Signal data & tools

Sharing models as well as data & tools

- Automated sleep staging

- Biological age (BA) prediction

- Disease risk/status

- etc.

Hours of raw EEG
+ chronological age (CA) = 62 yrs

BA = 67.7 yrs
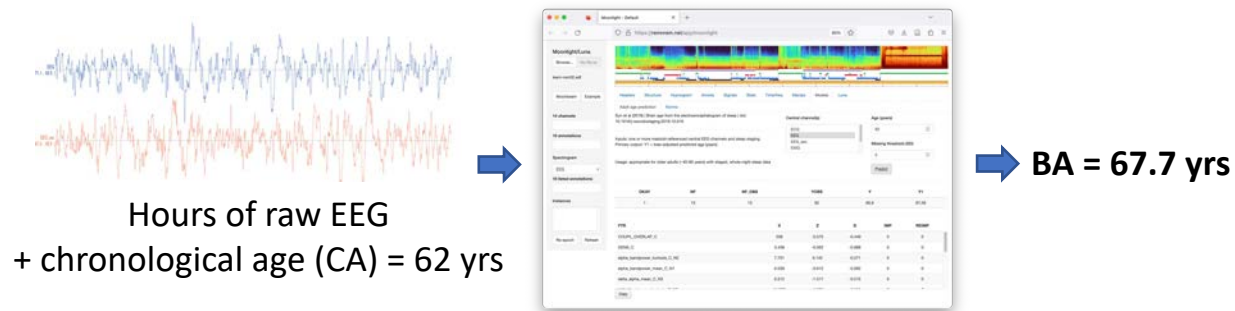
"Soup-to-nuts" cloud-based analyses
i.e. staging, QC, feature extraction, missing data imputation & prediction
(but with intermediate steps still visible/modifiable)
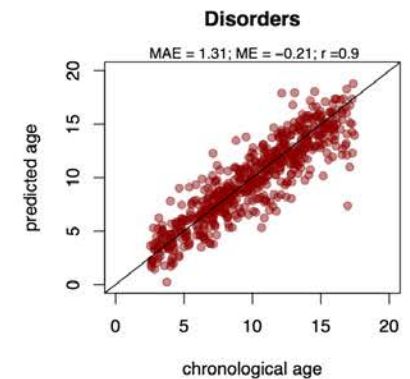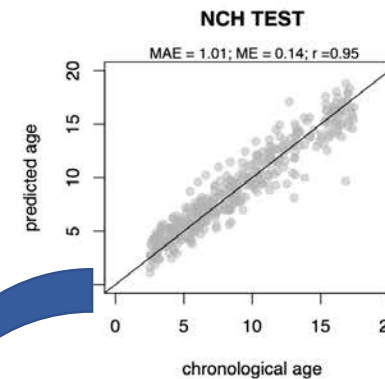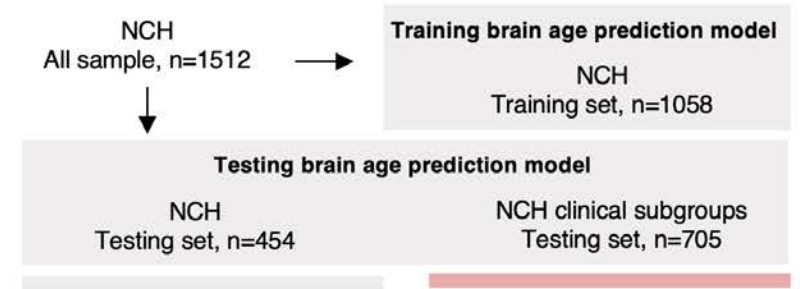
Kozhemiako et al (2023) Mapping Typical and Altered Neurodevelopment with Sleep Macro- and Micro-Architecture. *NeuroImage: Clinical*

r(CA,BA) = 0.95

Δ = BA – CA tracks with diagnostic status for several neurodevelopmental delays

Summary

- Primary goal: to make NSRR data and tools more accessible

- For controlled-access data, cloud environments can bring additional layers of regulatory and administrative challenges, especially when working with legacy datasets and a still-evolving broader ecosystem

- NSRR ingestion into BDC now on a smoother path
  - uploading new cohorts
  - sleep-specific CDEs
  - integration of NSRR tools, e.g. via Seven Bridges workflows

Acknowledgements

NSRR

http://sleepdata.org

Whole night delta-band spectral power from 1000 MESA participants