BWH BRIGHAM AND WOMEN'S HOSPITAL

VE RI TAS HARVARD MEDICAL SCHOOL

Perspectives from a domain specific data repository: *The National Sleep Research Resource*

Susan Redline, M.D.,M.P.H.

Professor of Sleep Medicine
Harvard Medical School
Brigham and Women's Hospital
Beth Israel Deaconess Medical Center
sredline@bwh.harvard.edu

Outline

- Data specific domain: potential and challenges
    - Sleep and Circadian Data
- Goals and organization of the National Sleep Research Resource
- The user community
    - Defining their needs
    - Measuring Impact
- Challenges

Reservoir of sleep data

- **2,800** accredited sleep labs in the U.S.

- **845,569 sleep studies** were performed in 2014
  - Increasing per year

- **~400 MB/study– 340 TB/yr**

- **NIH and industry-funded research sleep studies**

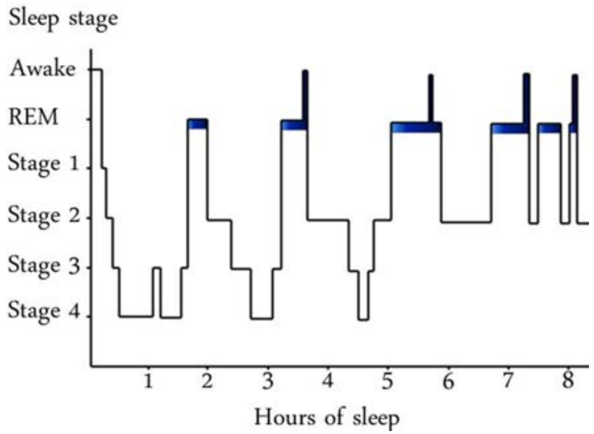
SLEEP HEART HEALTH

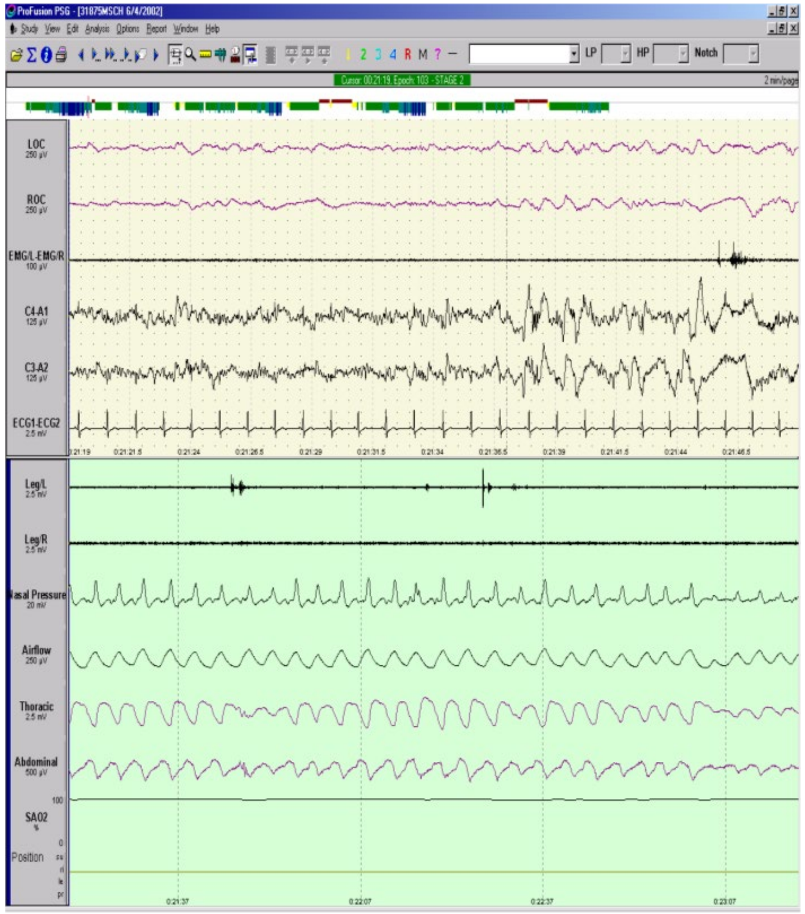
HEARTBEAT
Heart Biomarker Evaluation in Apnea Treatment

FRAMINGHAM HEART STUDY
Three Generations of Dedication
1948 1971 2002

MESA

SOL
Hispanic Community Health Study

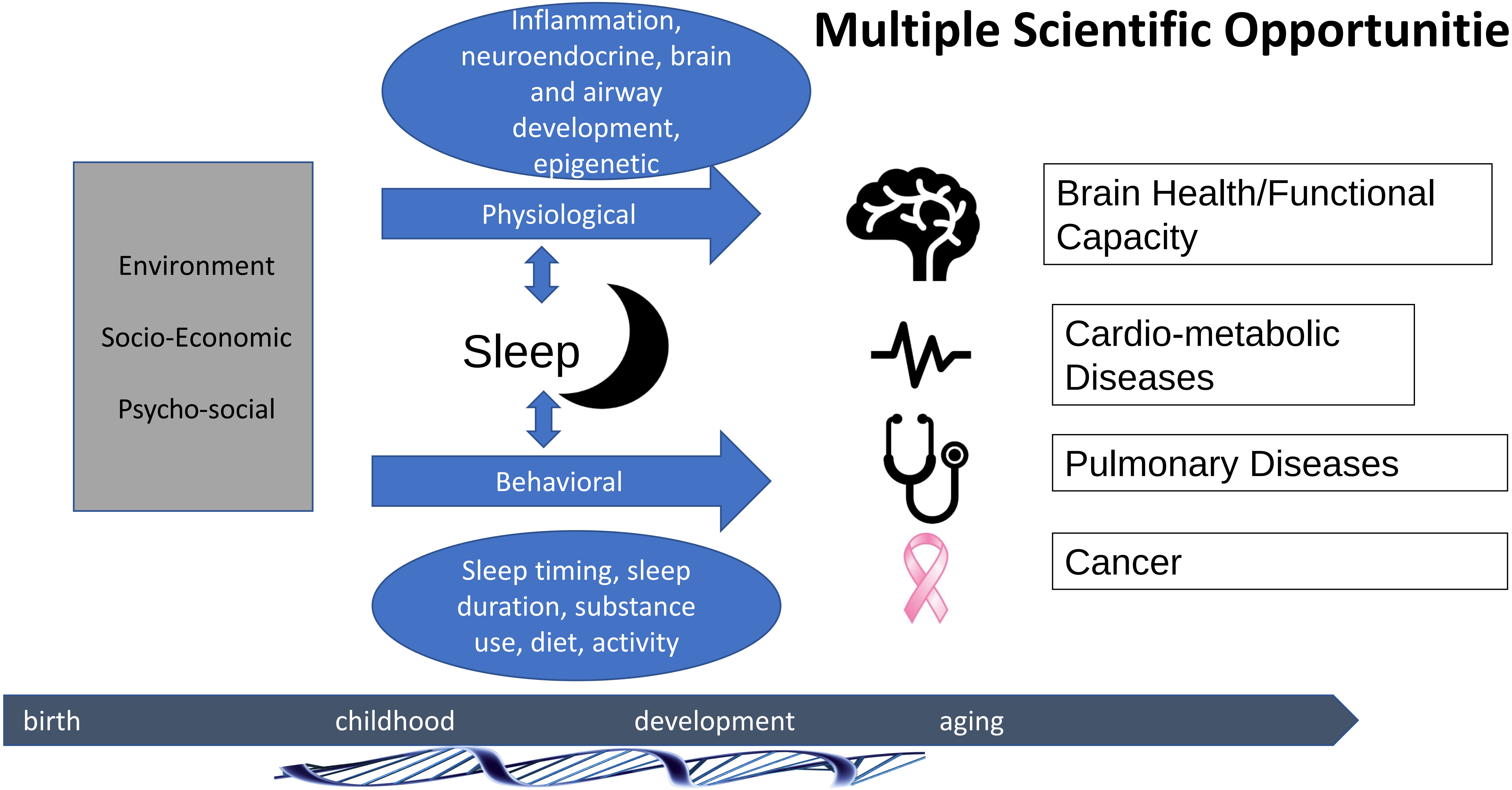




Untapped Data Signatures

- Hours of physiological signals
- Cross-talk between physiological systems
- Temporal and dynamic features
- Days of physiological/behavioral signals
- Multiple ways to annotate

Multiple Scientific Opportunities



Inflammation, neuroendocrine, brain and airway development, epigenetic

Physiological

Environment

Socio-Economic

Psycho-social

Sleep

Behavioral

Sleep timing, sleep duration, substance use, diet, activity

Brain Health/Functional Capacity

Cardio-metabolic Diseases

Pulmonary Diseases

Cancer

birth

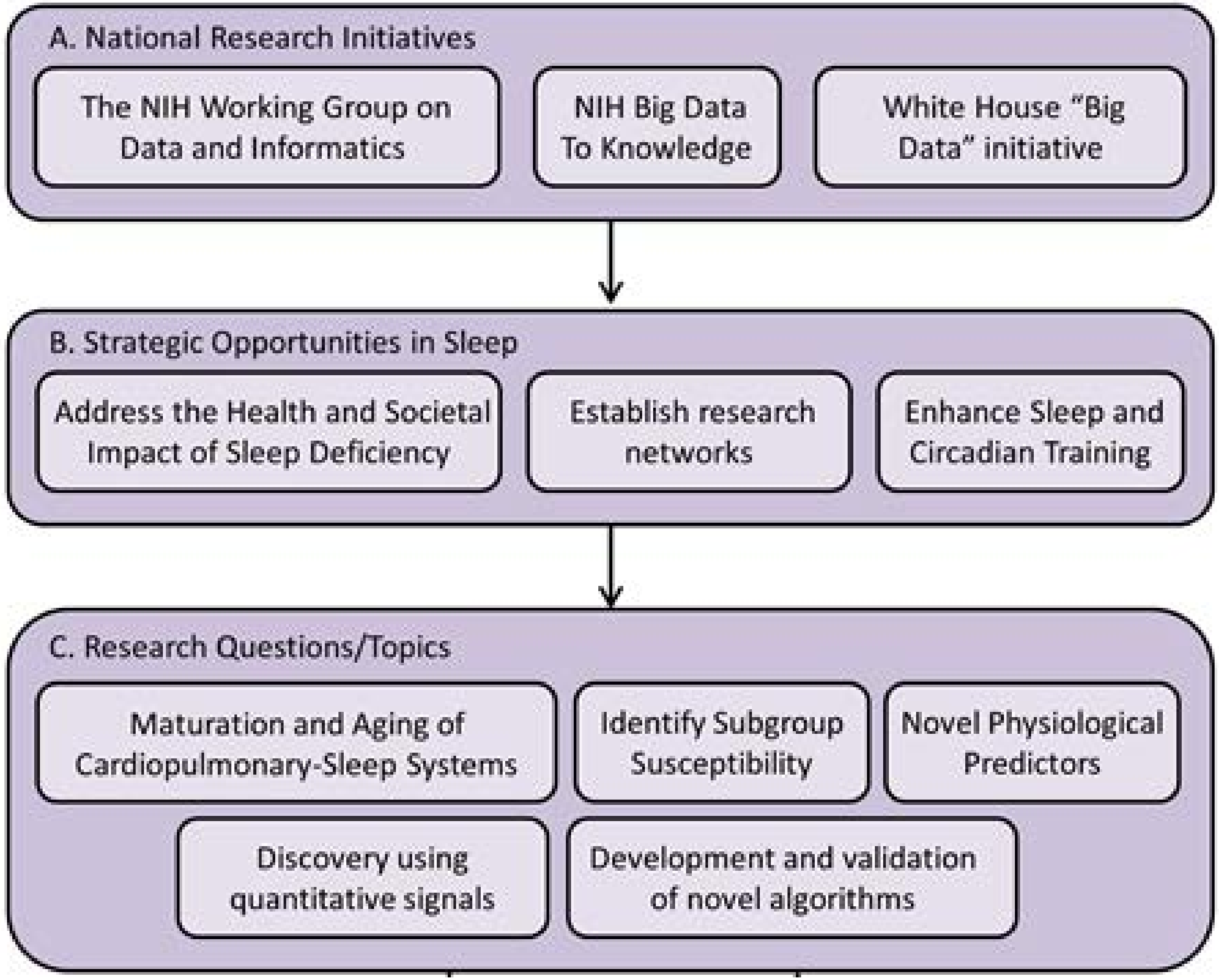
childhood

development

aging

Critical timepoints; Cumulative risk models

Sleep / Circadian Big Data Opportunities



A. National Research Initiatives

- The NIH Working Group on Data and Informatics
- NIH Big Data To Knowledge
- White House "Big Data" initiative

B. Strategic Opportunities in Sleep

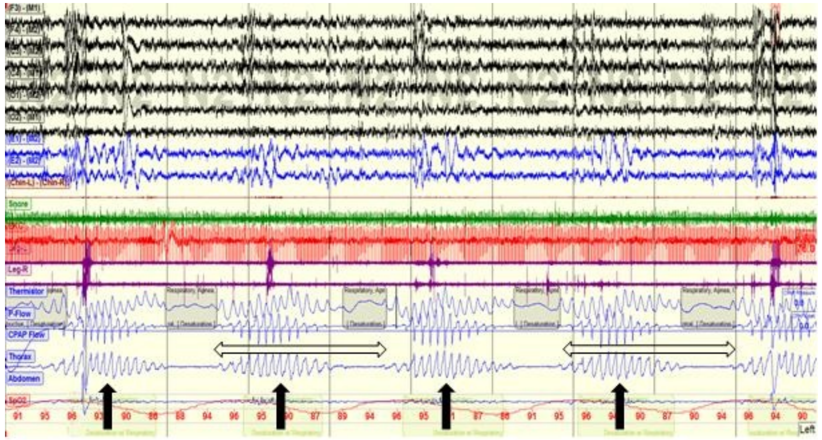
- Address the Health and Societal Impact of Sleep Deficiency
- Establish research networks
- Enhance Sleep and Circadian Training

C. Research Questions/Topics

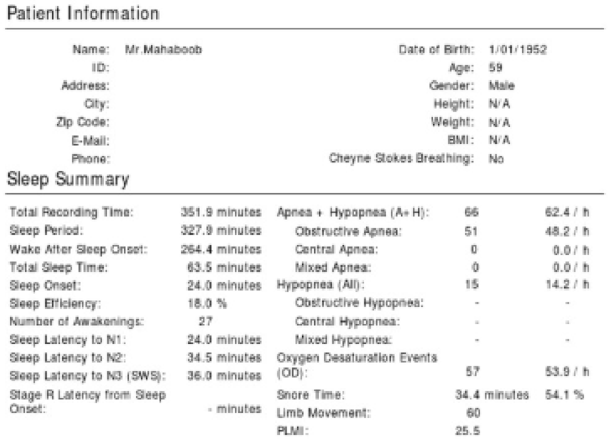
- Maturation and Aging of Cardiopulmonary-Sleep Systems
- Identify Subgroup Susceptibility
- Novel Physiological Predictors
- Discovery using quantitative signals
- Development and validation of novel algorithms

Dean D; Sleep 2014

Challenges in sleep data analysis



- **Data sets heterogeneous, some poorly annotated and difficult to harmonize**
  - Different collection protocols, lack of standardized montages, variable scoring
  - Lack of accepted sleep ontologies/variable vocabularies
    - Summary data and raw signals



Patient Information

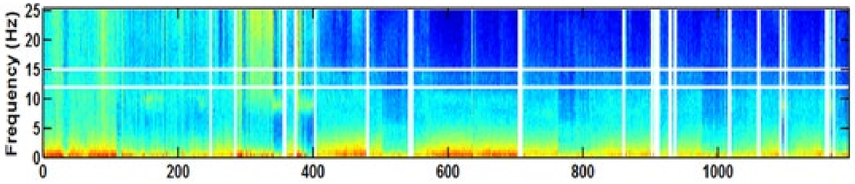
Name:	Mr.Mahaboob	Date of Birth:	1/01/1952
ID:		Age:	59
Address:		Gender:	Male
City:		Height:	N/A
Zip Code:		Weight:	N/A
E-Mail:		BMI:	N/A
Phone:		Cheyne Stokes Breathing:	No

Sleep Summary

Total Recording Time:	351.9 minutes	Apnea + Hypopnea (A+H):	66	62.4 / h
Sleep Period:	327.9 minutes	Obstructive Apnea:	51	48.2 / h
Wake After Sleep Onset:	264.4 minutes	Central Apnea:	0	0.0 / h
Total Sleep Time:	63.5 minutes	Mixed Apnea:	0	0.0 / h
Sleep Onset:	24.0 minutes	Hypopnea (All):	15	14.2 / h
Sleep Efficiency:	18.0 %	Obstructive Hypopnea:	-	-
Number of Awakenings:	27	Central Hypopnea:	-	-
Sleep Latency to N1:	24.0 minutes	Mixed Hypopnea:	-	-
Sleep Latency to N2:	34.5 minutes	Oxygen Desaturation Events		
Sleep Latency to N3 (SWS):	36.0 minutes	(OD):	57	53.9 / h
Stage R Latency from Sleep Onset:	- minutes	Snore Time:	34.4 minutes	54.1 %
		Limb Movement:	60	
		PLMI:	25.5	

- **Limited data types**
  - Focus on summary data
    - Untapped potential of advanced signal processing/machine learning

- **Few “open” sources of well-defined signals, linked covariates, and analysis tools**



Gaps in data access and appropriate tools

- **Many web portals have a..**
  - Limited ability to query and visualize data
  - Limited ability to directly access data
  - Limited ability to access tools for visualizing and processing data
- **Large data analytics**
  - High data storage/egress costs
- **Access/download procedures**
  - Concerns over privacy/security
- **“Sandboxes” needed for**
  - Collaboration, promote documentation (transparency/reproducibility)
- **Barriers to users unfamiliar with dataset or dependencies on others**

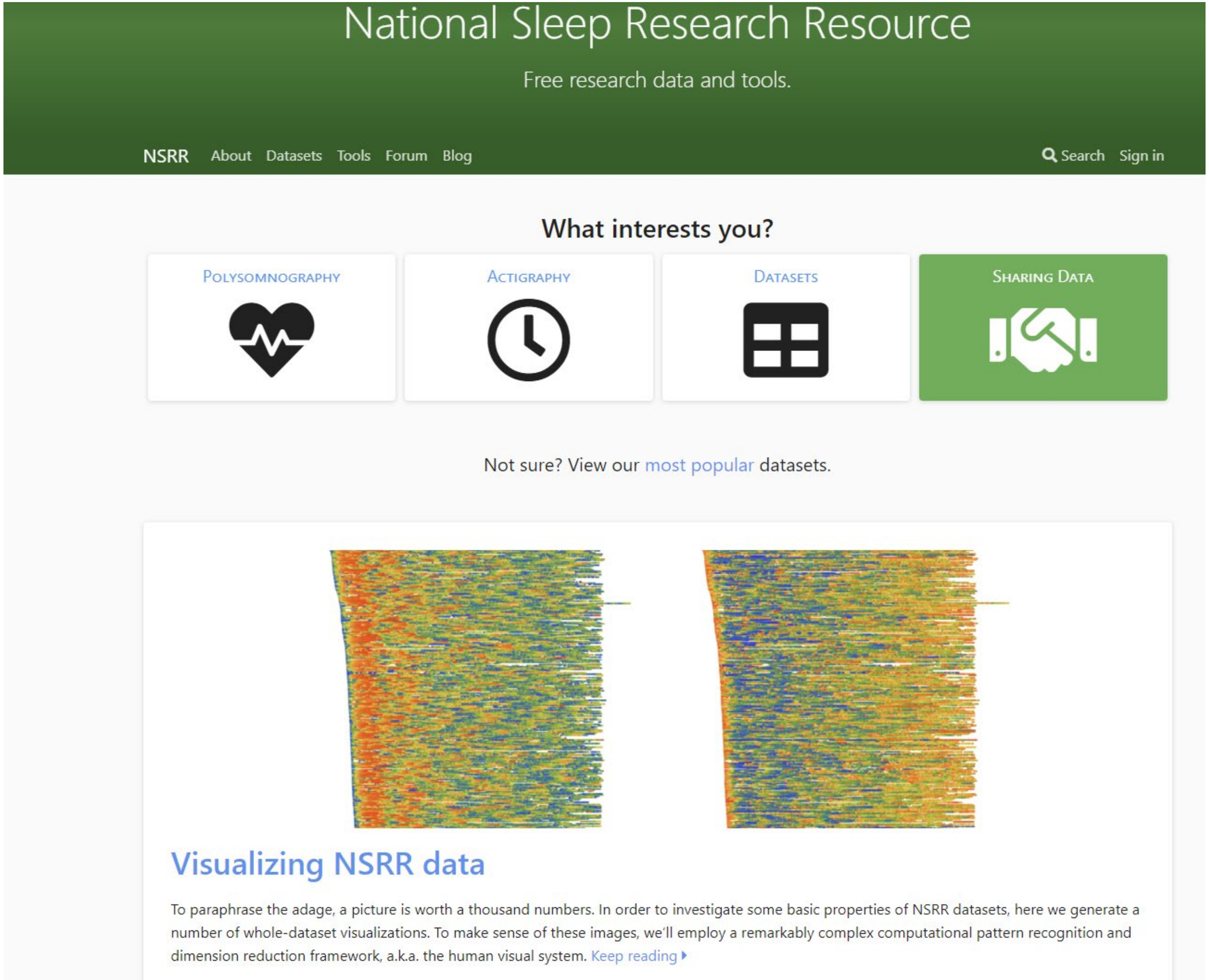
National Sleep Research Resource: sleepdata.org (2014-)

Provide users web-based tools to assist with preliminary exploration of data within and across data sets and identify subsets of data most useful using clearly mapped terms

Community resource to deposit and access “raw” or complex primary data (physiological signals), including processed physiological signals

Provide users access to a hub of tools for processing physiological signals as well as a resource to support communications among sleep researchers

Partner with and link to other resources, such as BioLINCC and dbGAP (BioData Catalyst)



National Sleep Research Resource

Free research data and tools.

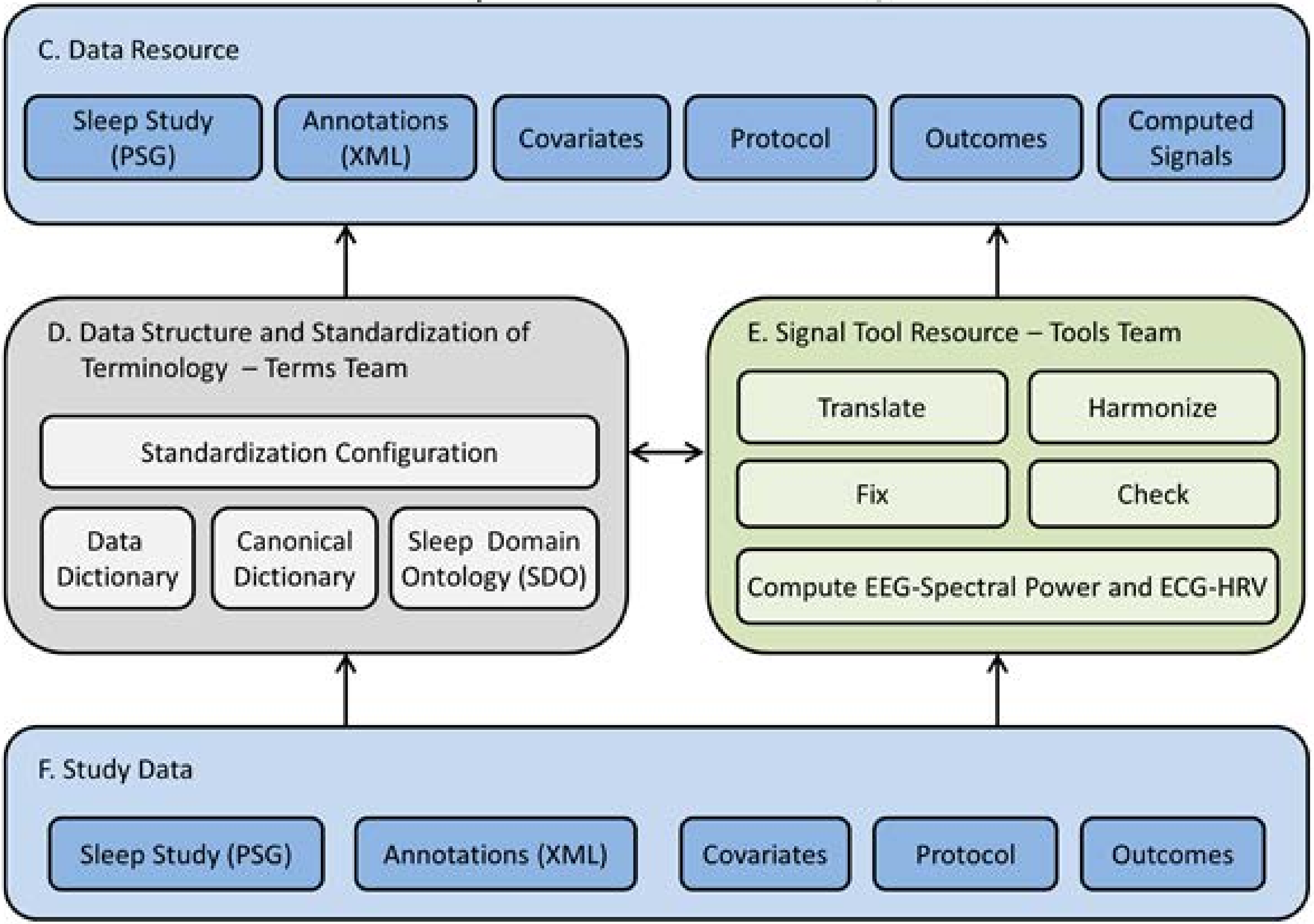
What interests you?

Not sure? View our most popular datasets.

Visualizing NSRR data

To paraphrase the adage, a picture is worth a thousand numbers. In order to investigate some basic properties of NSRR datasets, here we generate a number of whole-dataset visualizations. To make sense of these images, we'll employ a remarkably complex computational pattern recognition and dimension reduction framework, a.k.a. the human visual system. Keep reading

National Sleep Research Resource: Sleepdata.org



C. Data Resource

Sleep Study (PSG) | Annotations (XML) | Covariates | Protocol | Outcomes | Computed Signals

D. Data Structure and Standardization of Terminology – Terms Team

Standardization Configuration

Data Dictionary | Canonical Dictionary | Sleep Domain Ontology (SDO)

E. Signal Tool Resource – Tools Team

Translate | Harmonize

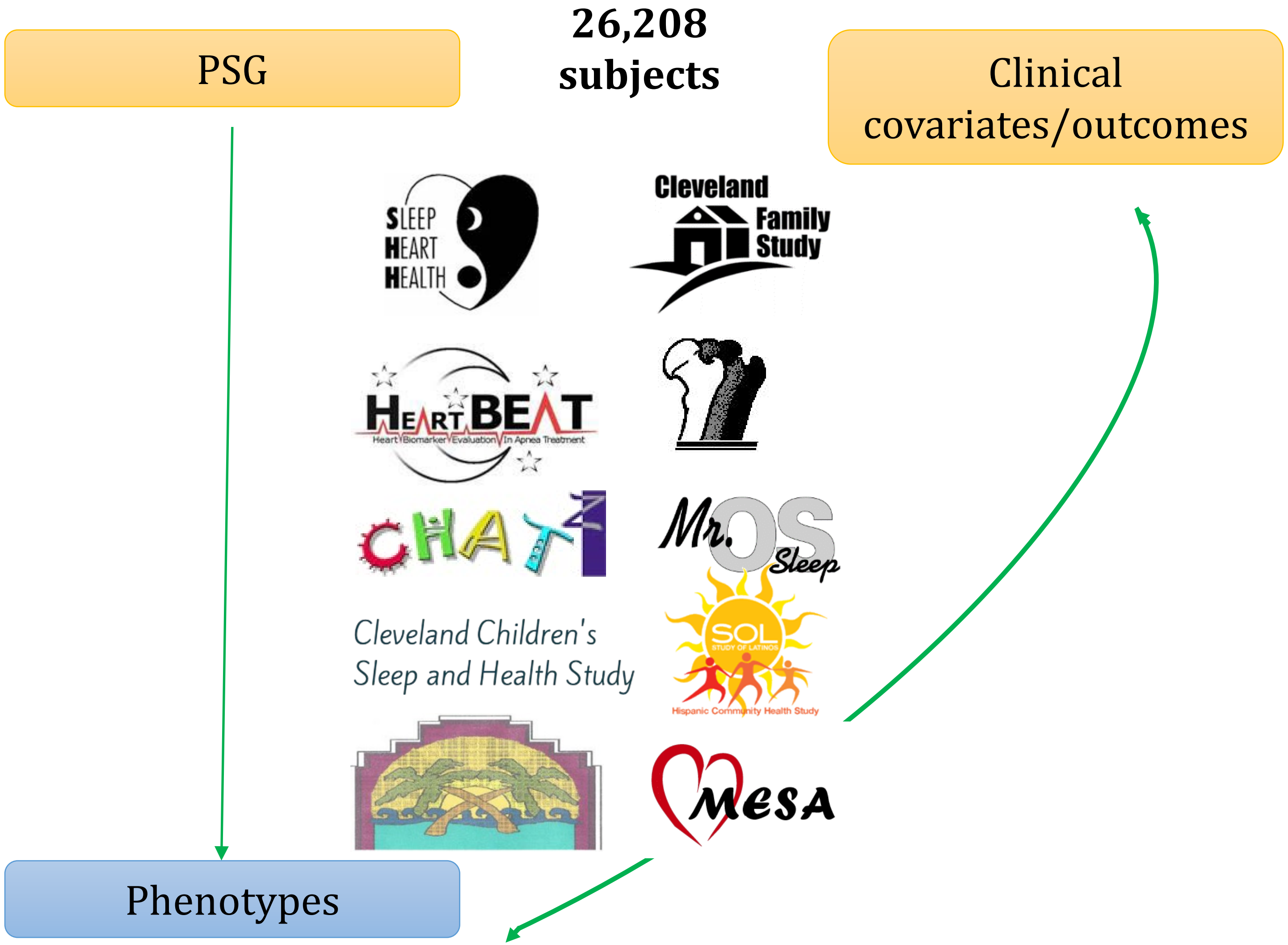
Fix | Check

Compute EEG-Spectral Power and ECG-HRV

F. Study Data

Sleep Study (PSG) | Annotations (XML) | Covariates | Protocol | Outcomes

Data Integration



PSG

26,208 subjects

Clinical covariates/outcomes

Phenotypes

SLEEP HEART HEALTH
Cleveland Family Study
HeartBEAT
Heart Biomarker Evaluation in Apnea Treatment
CHAT
Mr.OS Sleep
Cleveland Children's Sleep and Health Study
SOL
STUDY OF LATINOS
Hispanic Community Health Study
MESA

Available Data

31,580 EDFs from 27,151 subjects

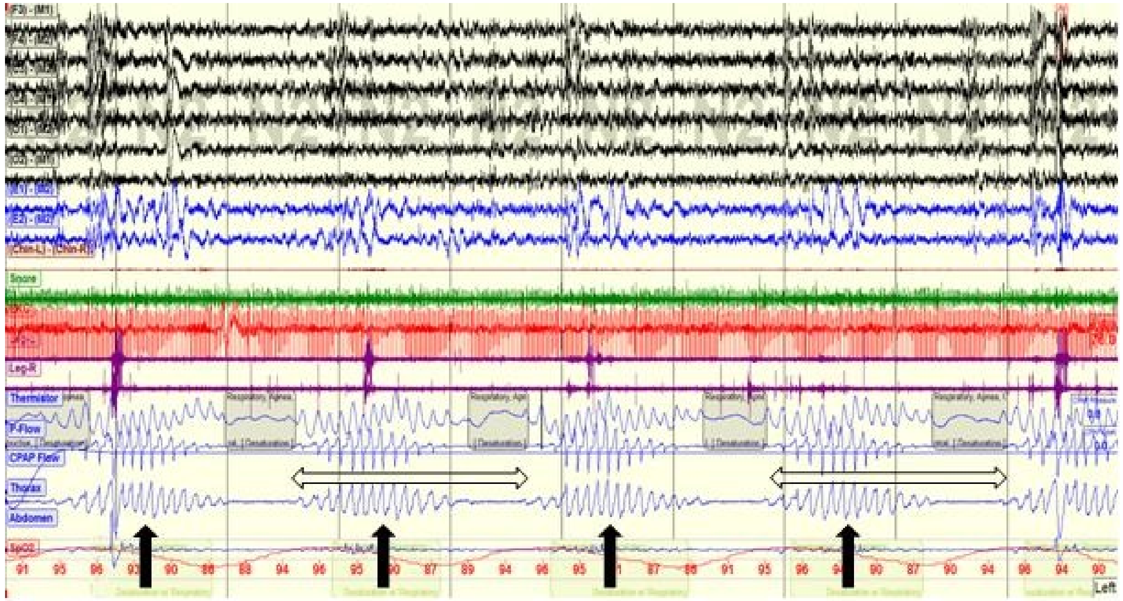
19,235 PSGs with EEG or ECG spectral analysis results

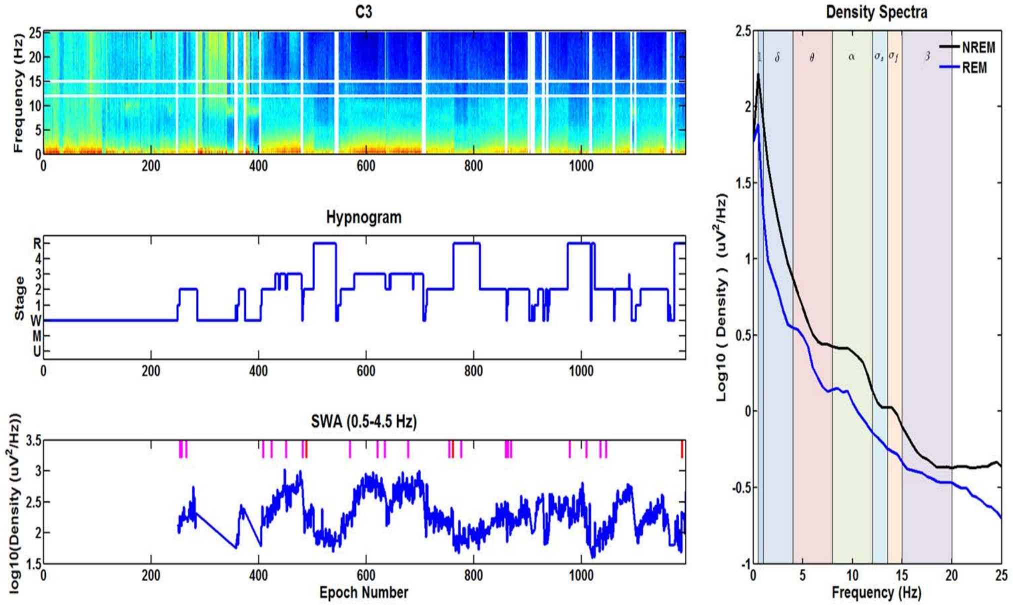
4,064 actigraphy files

5,324 terms annotated to structured definitions

4,681 with provenance attributes

Quantitative Signal Analysis





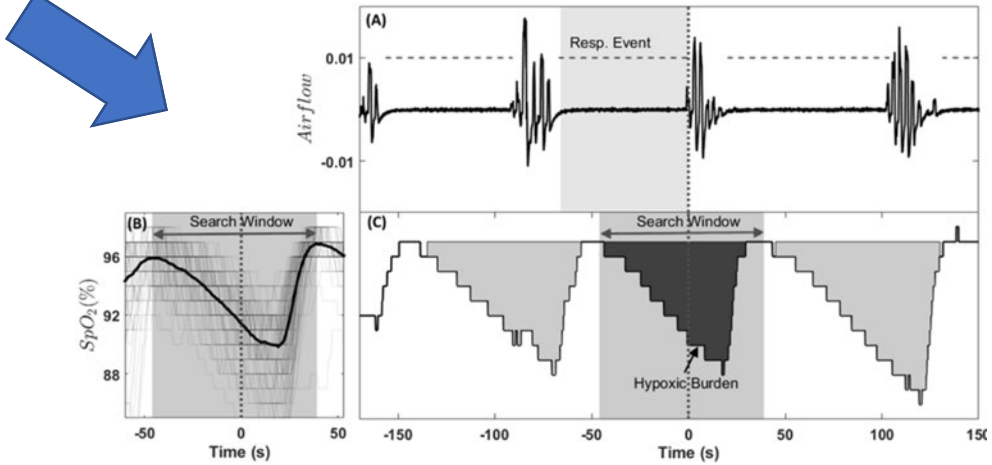
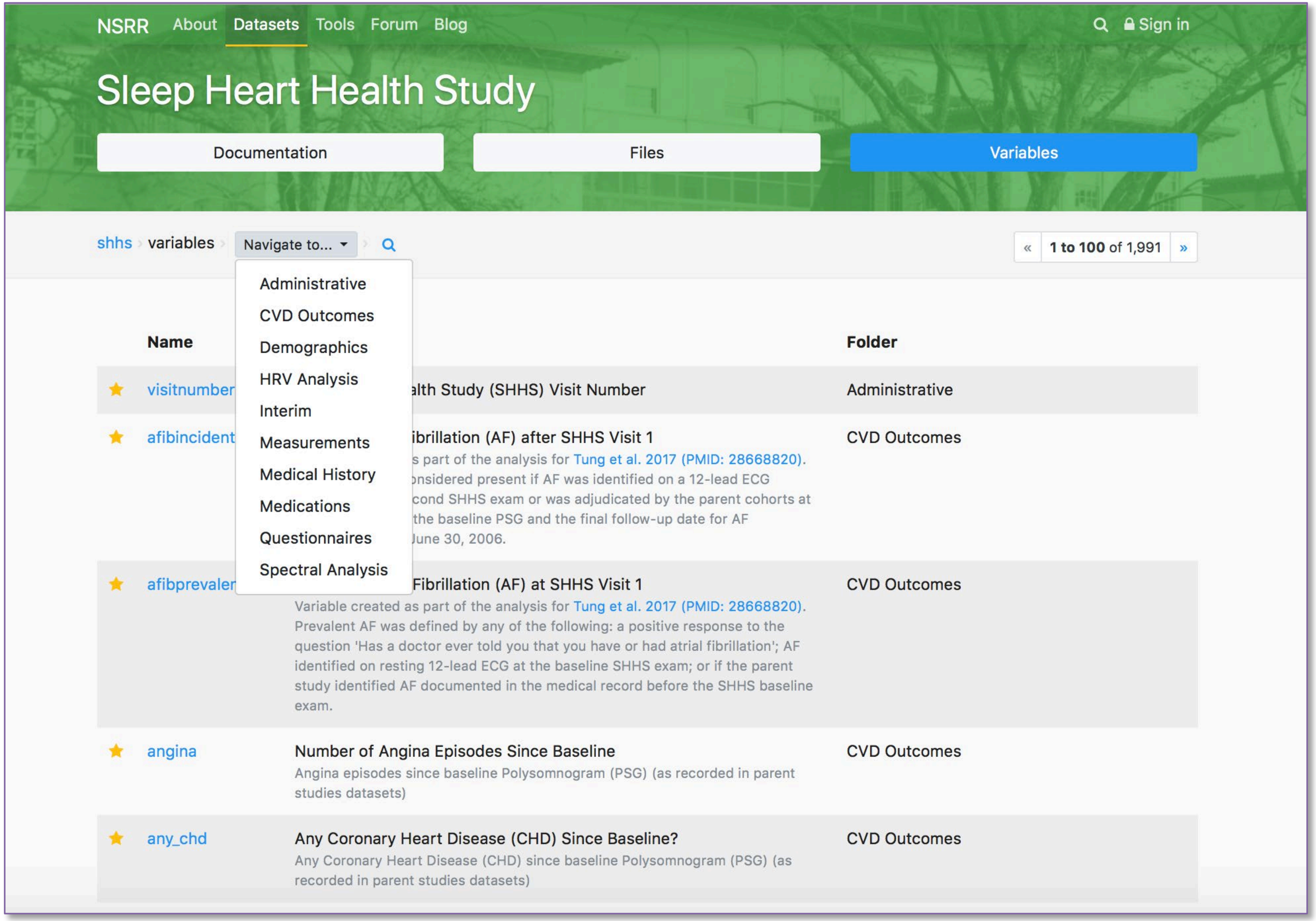


Figure 1: Example of hypoxic burden calculation for an individual respiratory event (Resp. Event). Panel A

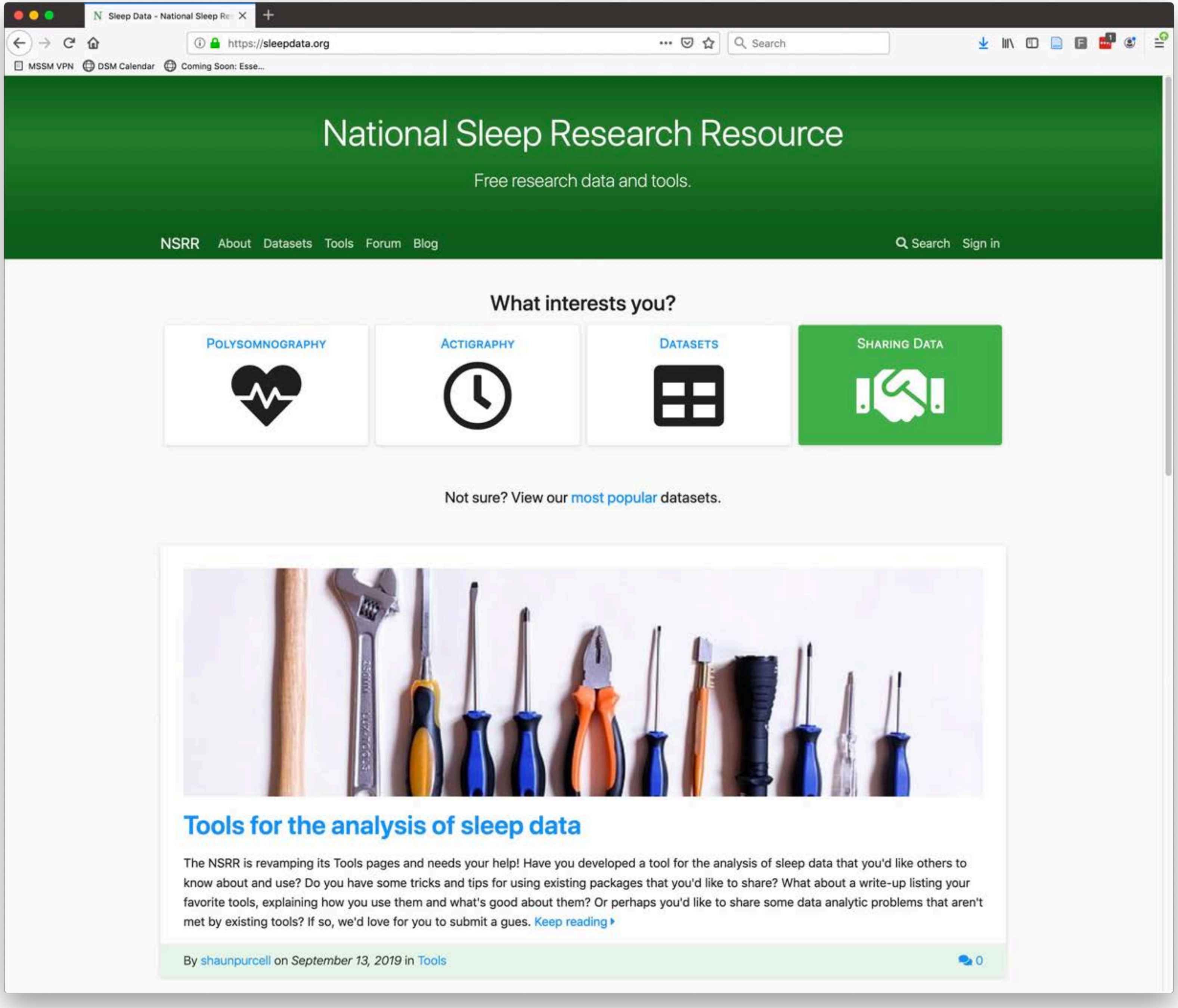
Search across 1,000s of variables…



Sleep Heart Health Study

	Name		Folder
★	visitnumber	alth Study (SHHS) Visit Number	Administrative
★	afibincident	ibrillation (AF) after SHHS Visit 1 s part of the analysis for Tung et al. 2017 (PMID: 28668820). onsidered present if AF was identified on a 12-lead ECG cond SHHS exam or was adjudicated by the parent cohorts at the baseline PSG and the final follow-up date for AF June 30, 2006.	CVD Outcomes
★	afibprevaler	Fibrillation (AF) at SHHS Visit 1 Variable created as part of the analysis for Tung et al. 2017 (PMID: 28668820). Prevalent AF was defined by any of the following: a positive response to the question 'Has a doctor ever told you that you have or had atrial fibrillation'; AF identified on resting 12-lead ECG at the baseline SHHS exam; or if the parent study identified AF documented in the medical record before the SHHS baseline exam.	CVD Outcomes
★	angina	Number of Angina Episodes Since Baseline Angina episodes since baseline Polysomnogram (PSG) (as recorded in parent studies datasets)	CVD Outcomes
★	any_chd	Any Coronary Heart Disease (CHD) Since Baseline? Any Coronary Heart Disease (CHD) since baseline Polysomnogram (PSG) (as recorded in parent studies datasets)	CVD Outcomes

Variables logically grouped by type



National Sleep Research Resource

Free research data and tools.

What interests you?

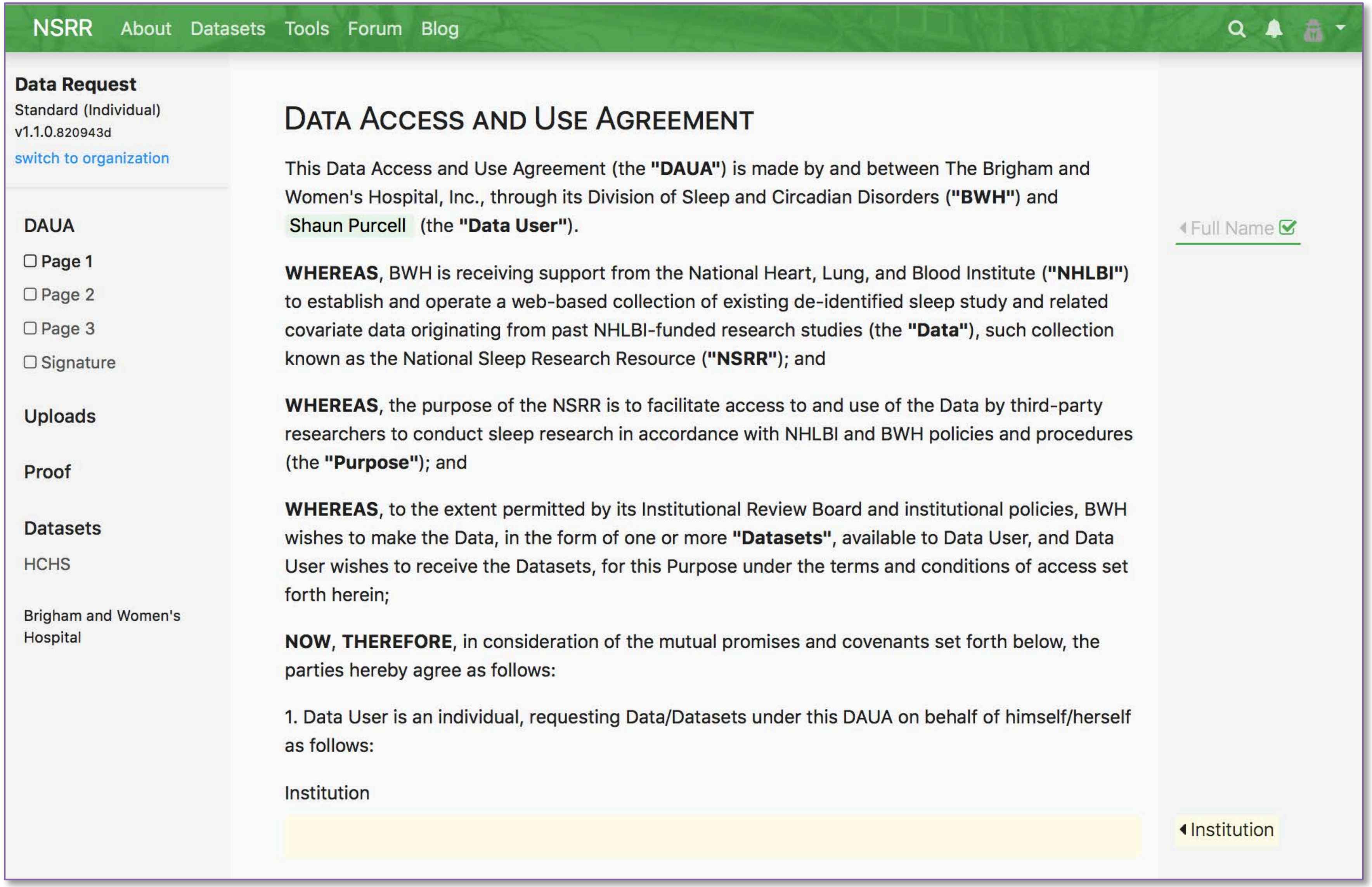
Not sure? View our most popular datasets.

Tools for the analysis of sleep data

The NSRR is revamping its Tools pages and needs your help! Have you developed a tool for the analysis of sleep data that you'd like others to know about and use? Do you have some tricks and tips for using existing packages that you'd like to share? What about a write-up listing your favorite tools, explaining how you use them and what's good about them? Or perhaps you'd like to share some data analytic problems that aren't met by existing tools? If so, we'd love for you to submit a gues. Keep reading ▸

By shaunpurcell on *September 13, 2019* in Tools 0

Easy, but not uncontrolled, access to data…



Data Request
Standard (Individual)
v1.1.0.820943d
switch to organization

DAUA

□ Page 1
□ Page 2
□ Page 3
□ Signature

Uploads

Proof

Datasets

HCHS

Brigham and Women's Hospital

Data Access and Use Agreement

This Data Access and Use Agreement (the **"DAUA"**) is made by and between The Brigham and Women's Hospital, Inc., through its Division of Sleep and Circadian Disorders (**"BWH"**) and Shaun Purcell (the **"Data User"**).

Full Name ✓

WHEREAS, BWH is receiving support from the National Heart, Lung, and Blood Institute (**"NHLBI"**) to establish and operate a web-based collection of existing de-identified sleep study and related covariate data originating from past NHLBI-funded research studies (the **"Data"**), such collection known as the National Sleep Research Resource (**"NSRR"**); and

WHEREAS, the purpose of the NSRR is to facilitate access to and use of the Data by third-party researchers to conduct sleep research in accordance with NHLBI and BWH policies and procedures (the **"Purpose"**); and

WHEREAS, to the extent permitted by its Institutional Review Board and institutional policies, BWH wishes to make the Data, in the form of one or more **"Datasets"**, available to Data User, and Data User wishes to receive the Datasets, for this Purpose under the terms and conditions of access set forth herein;

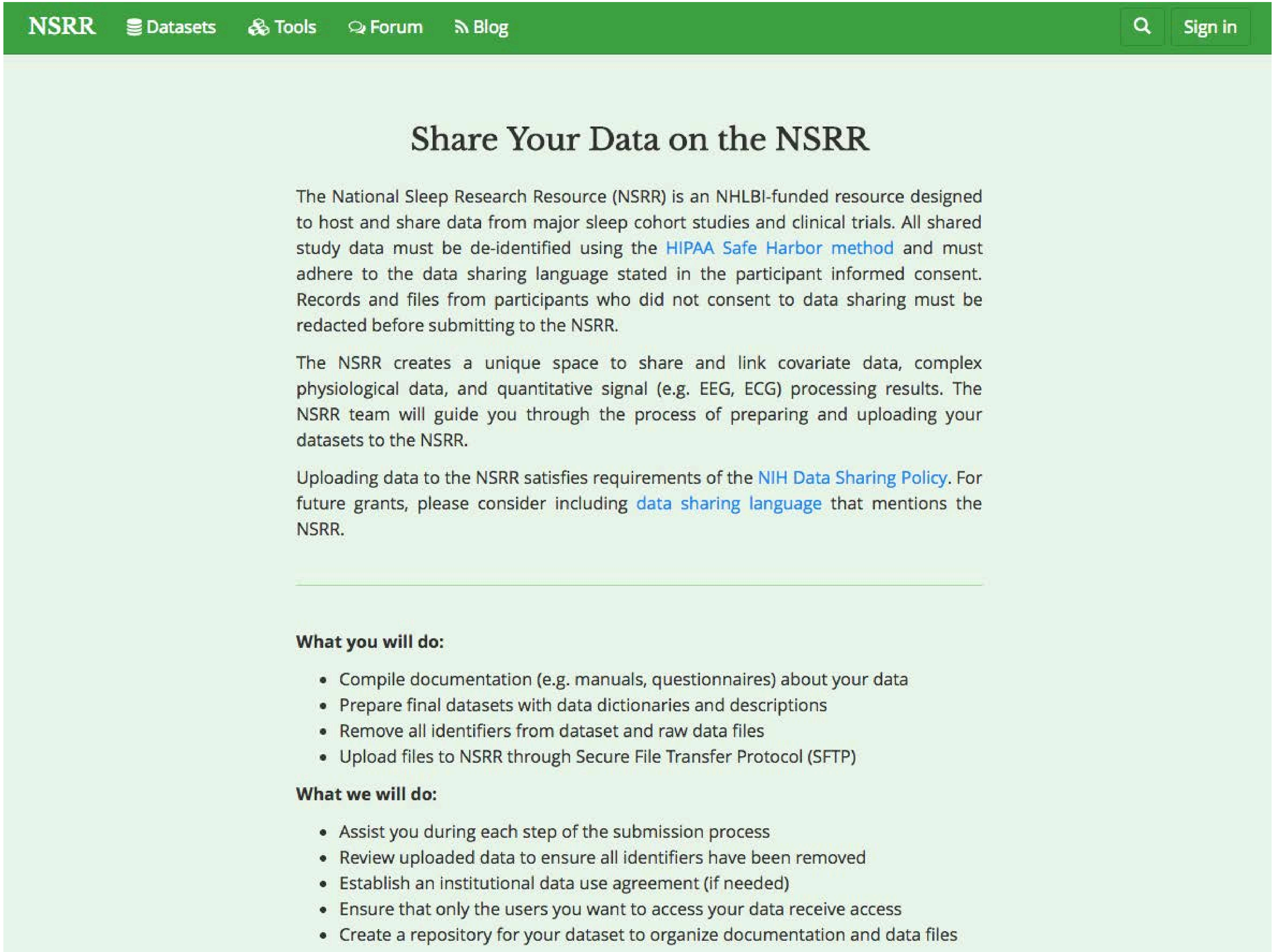
NOW, **THEREFORE**, in consideration of the mutual promises and covenants set forth below, the parties hereby agree as follows:

1. Data User is an individual, requesting Data/Datasets under this DAUA on behalf of himself/herself as follows:

Institution

Institution

Interface for DAUA and IRB approval required for data access



Share Your Data on the NSRR

The National Sleep Research Resource (NSRR) is an NHLBI-funded resource designed to host and share data from major sleep cohort studies and clinical trials. All shared study data must be de-identified using the HIPAA Safe Harbor method and must adhere to the data sharing language stated in the participant informed consent. Records and files from participants who did not consent to data sharing must be redacted before submitting to the NSRR.

The NSRR creates a unique space to share and link covariate data, complex physiological data, and quantitative signal (e.g. EEG, ECG) processing results. The NSRR team will guide you through the process of preparing and uploading your datasets to the NSRR.

Uploading data to the NSRR satisfies requirements of the NIH Data Sharing Policy. For future grants, please consider including data sharing language that mentions the NSRR.

What you will do:

- Compile documentation (e.g. manuals, questionnaires) about your data
- Prepare final datasets with data dictionaries and descriptions
- Remove all identifiers from dataset and raw data files
- Upload files to NSRR through Secure File Transfer Protocol (SFTP)

What we will do:

- Assist you during each step of the submission process
- Review uploaded data to ensure all identifiers have been removed
- Establish an institutional data use agreement (if needed)
- Ensure that only the users you want to access your data receive access
- Create a repository for your dataset to organize documentation and data files

Assessing Impact

- **User Base**
  - Register/Access data

- **Products**
  - Use/publish data
  - Contribute data
  - Discoveries/new tools
  - Support new grants
  - Training

- **Engagement**
  - Interactive user community- collaboration, blogs, etc

Assessing Impact: Access Data

- **6,041 registered users**
    - **1793 approved DUAs**

- Over **13 million files** downloaded, over **321 TB** of data
    - **2 TB** data per week

- **Ease of access**
    - Time interval from access to approval
        - User-friendly on-line DUA

Assessing Impact: Use Data/Publish Results

- Publications
  - Epidemiological associations
  - Discovery/replication
    - New signals/Associations
  - Machine learning
  - Algorithm development/validation
- Tracking difficult
  - DUA: Cite grant / resource
  - NEED: Datasets as “citable” object
  - Track “Impact Factor” of resource

Assessing Impact: Grants

- Training grants (NIH, AHA, AASM), R21s, RO1s

Assessing Impact: Training

- Multiple levels
    - High school– post graduate

    - For example,
        - > 100 Georgia Tech students: capstone project
        - OSHU Data Wrangling courses/workbooks
        - Basis for Harvard ML course
        - Resource for a biostatistics book/course

Assessing Impact: Contribute Data

- **New contributors**
  - Individuals
  - New cohorts: ~15 new cohorts identified
  - NIH (NIMH; E.S.P)

- Incentives for data sharing
- Reducing “friction”
  - Regulatory
  - Data structure/documentation

Assessing Impact: Contribute CDEs

- > 5,000 variables mapped to standards
  - ICSD-3, NIH CDE
- >4000 variables mapped to provenance data
  - Bioportal CMS (wiki)

Summary: Challenges/Needs

- Systematize citation process/Orchid registrations
  - Data resource impact factor?
- Link NIH funded grants for secondary data to sources?
- Trainee impact
  - Inventories of courses/books/trainee grants
- Data and tool contributions
  - Publish/highlight attributions

“Dreams of the NSRR”

“Data Repurposing”

Community: tools & support

Make data accessible

Novel phenotypes

Spans physiological domains

Forward & reverse genetics

Demographically diverse populations

Heterogeneity

Healthy & patient populations

Attrib: S Purcell

Acknowledgements

- Brigham and Women’s Hospital
  - Dennis Dean
  - Matthew Kim
  - Sara Mariani
  - Daniel Mobley
  - Remo Mueller
  - **Shaun Purcell**
  - Michael Rueschman
  - Ying Zhang
- NHLBI
- University of Kentucky
  - GQ Zhang
  - Satya Sahoo
  - Licong Cui
- Beth Israel Deaconess Medical Center
  - Ary Goldberger
  - Madalena Costa