

## **Breakout Session 3: Track B**

# **Implementation of Provenance Metadata on Neuroscience Gateway – A Platform for Neuroscience Software Dissemination**

Dr. Amit Majumdar

*Division Director, Associate Professor, University of California San Diego*

**Presentation Title: Implementation of Provenance Metadata on Neuroscience Gateway – A Platform for Neuroscience Software Dissemination**

**Supplement Award Title: Neuroscience Gateway to Enable Dissemination of Computational And Data Processing Tools And Software**

Amit Majumdar, San Diego Supercomputer Center (SDSC); Department of Radiation Medicine and Applied Sciences; University of California San Diego (UCSD), La Jolla, CA

Co-authors:

Subhashini Sivagnanam, Kenneth Yoshimoto, Steve Yeu, Kai Lin, Scott Sakai, Fernando Garzon, SDSC, UCSD, La Jolla, CA

Satya S. Sahoo, Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine; Department of Neurology, University Hospitals Cleveland Medical Center, Cleveland, OH

Katrina Prantzalos, Dipak Upadhyaya, Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine

# Outline

## ➤ Neuroscience Gateway (NSG)

- Integration of NSG and Open Science Chain (OSC) – Using the Neuro-Integrative Connectivity (NIC) tool for improving AI readiness through provenance metadata
- Conclusion

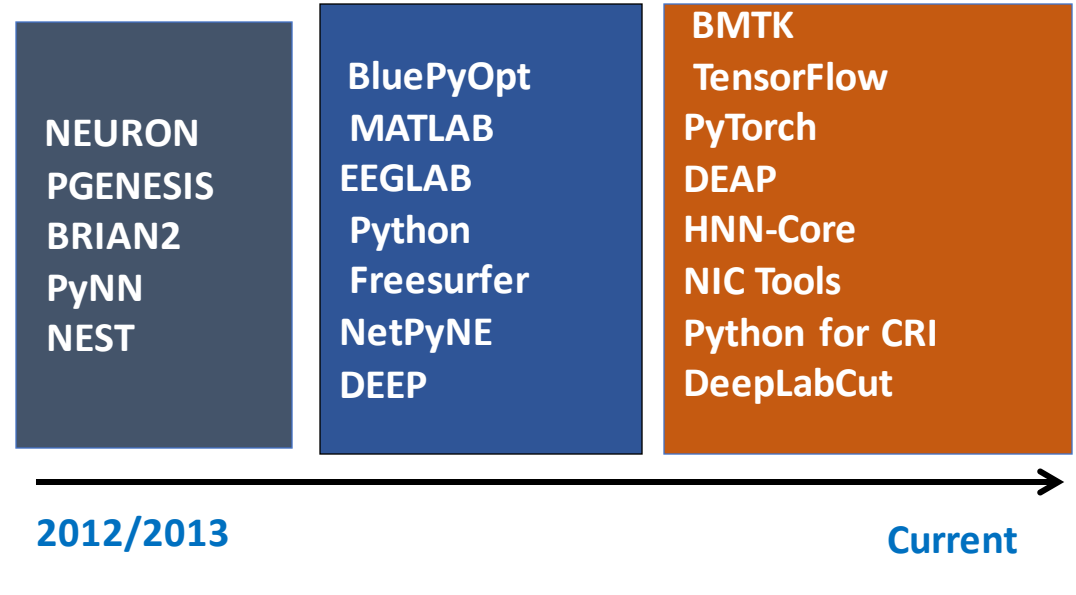
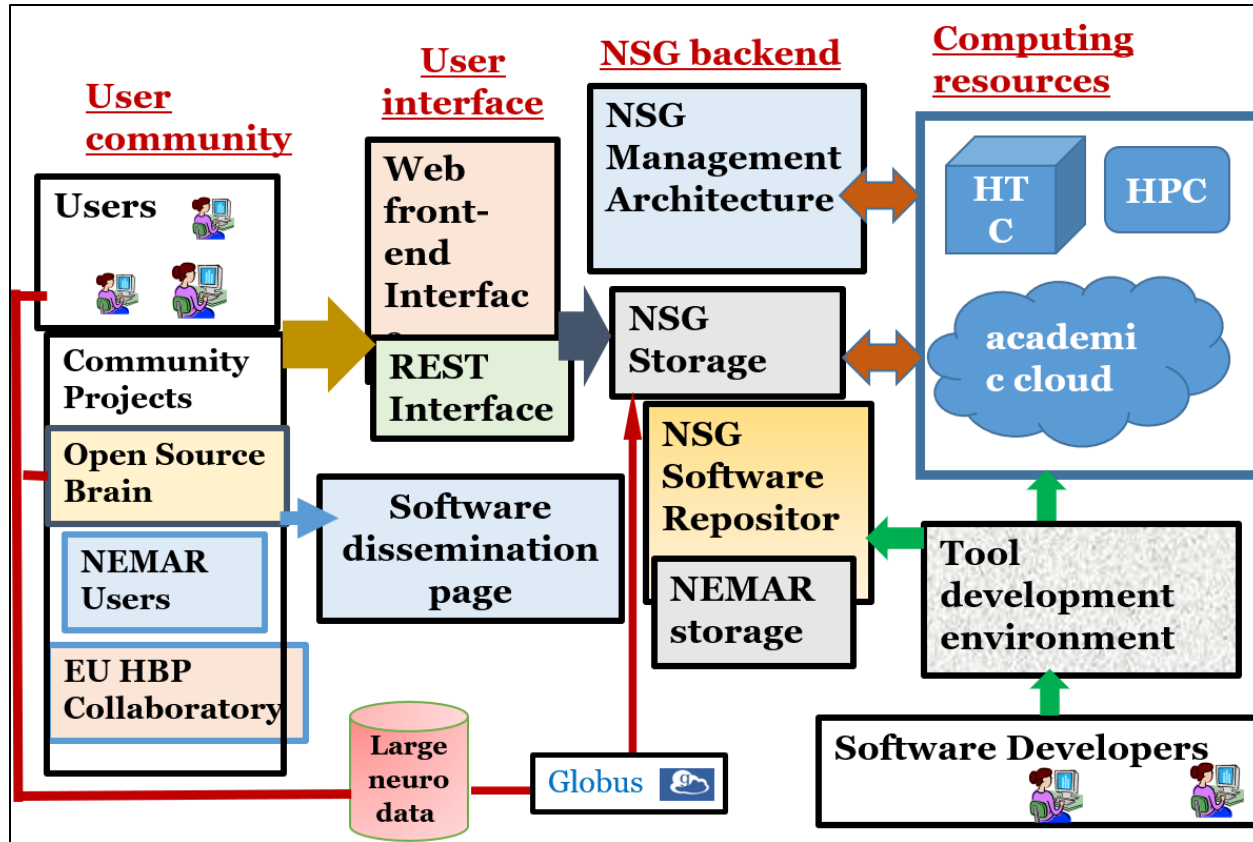
# The Neuroscience Gateway (NSG)

- Provides simple and secure access through portal and programmatic services, to run neuroscience modeling and data processing software on high performance, high throughput and accelerators compute resources
- A platform for neuroscience software dissemination
- <http://www.nsgportal.org>
- Free and open to any academic and non-profit researchers worldwide

*NSG catalyzes and democratizes computational and data processing neuroscience research and education for everybody including researchers and students from underrepresented institutions*

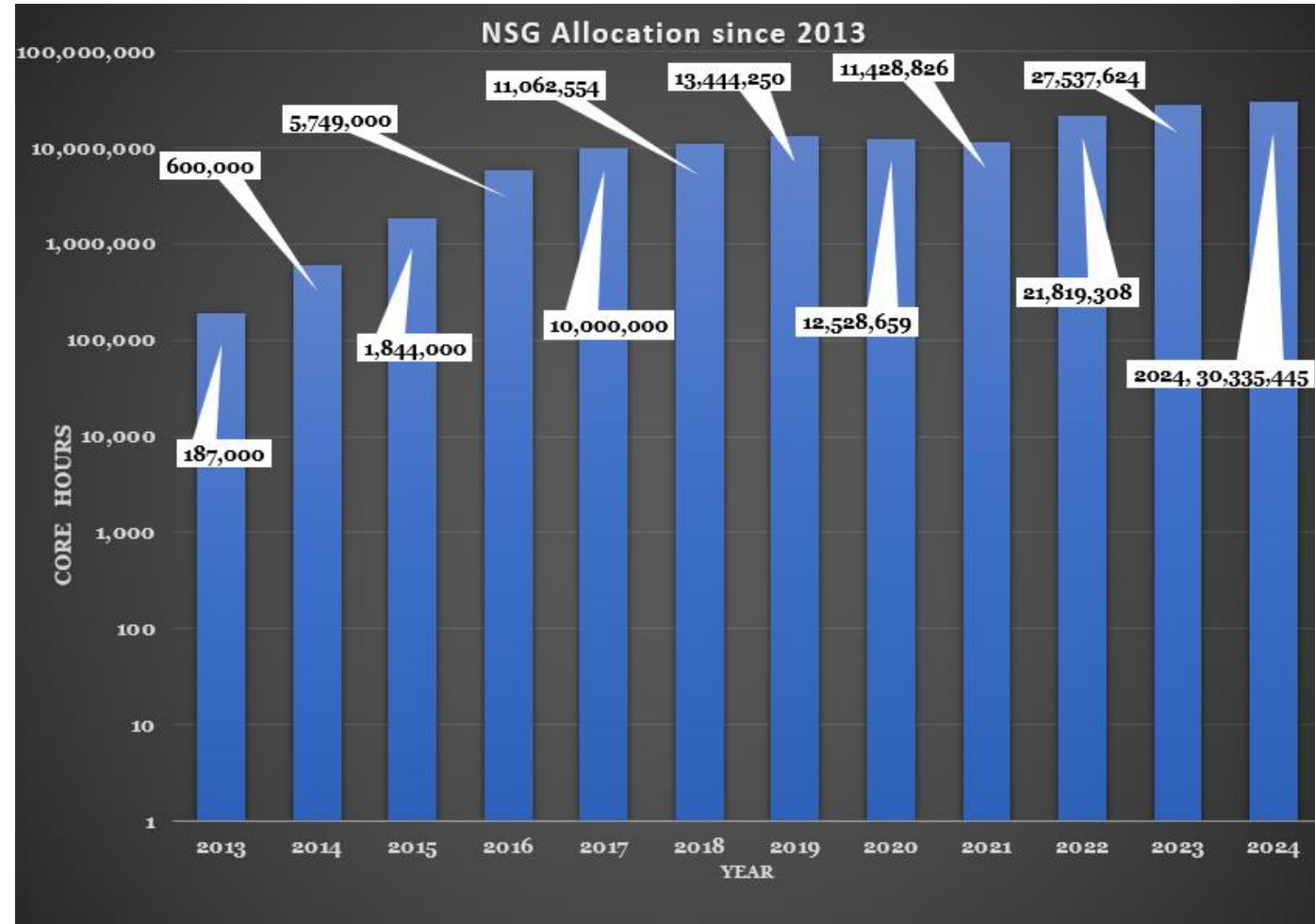
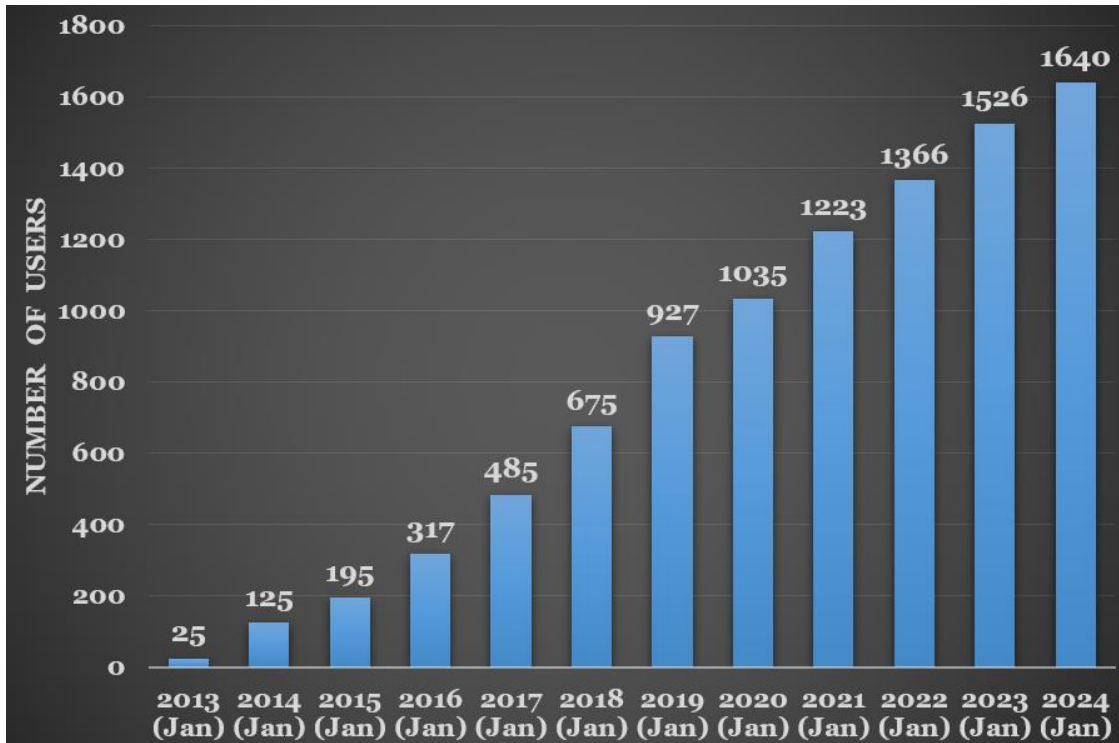
# NSG - Access and Software Dissemination

- NSG Portal: Simple and easy to use web interface
- NSG-R: Programmatic access through RESTful services



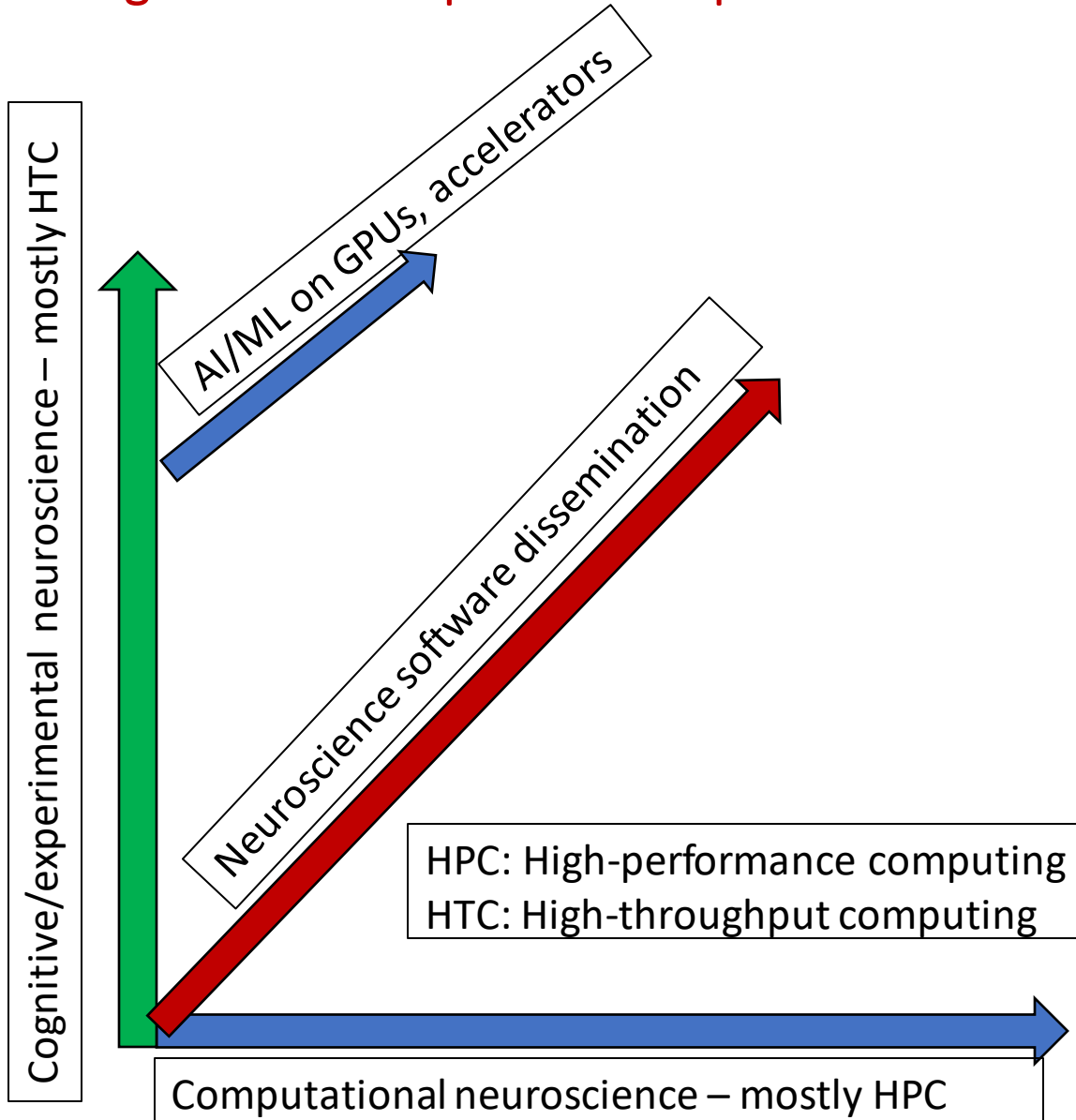
- New tools added based on user needs
- Un-supported/un-used tools are retired
- We plan to add ~70 new neuroimaging software, tools and pipelines

# NSG: Growth in Number of Users and Supercomputer-time Allocations





# NSG growth and impact in multiple dimensions



- Multiple NSG training/workshops yearly since 2013: at SfN, CNS, CogSci conferences, other
  - Virtual and hybrid during the last few years
- Special training for Hispanic Serving Institution teaching faculties and their collaborators – *in 2021 and 2022; Poster at SfN with Prof. Elba Serrano, NMSU*
- NSG used in classroom teaching
- Continuing since 2011 Research Experience for High School
  - HPC, EEG data analyzed using EEGLAB, and modeling using the NEURON software
- Undergraduate student interns

- Publications <https://www.nsgportal.org/citation.htm>; (since 2013)
  - Neuroscience publications, presentations, posters: **238** (that we know of)
  - Cyberinfrastructure related publications, presentations, posters: **67**
  - Educational projects/publications (MS/PhD thesis) and Training/workshops: **48**

# Outline

- Neuroscience Gateway (NSG)
- **Integration of NSG and Open Science Chain (OSC) - Using the Neuro-Integrative Connectivity (NIC) tool for improving AI readiness through provenance metadata**
- Conclusion



# Open Science Chain (OSC)

<https://www.opensciencechain.org>

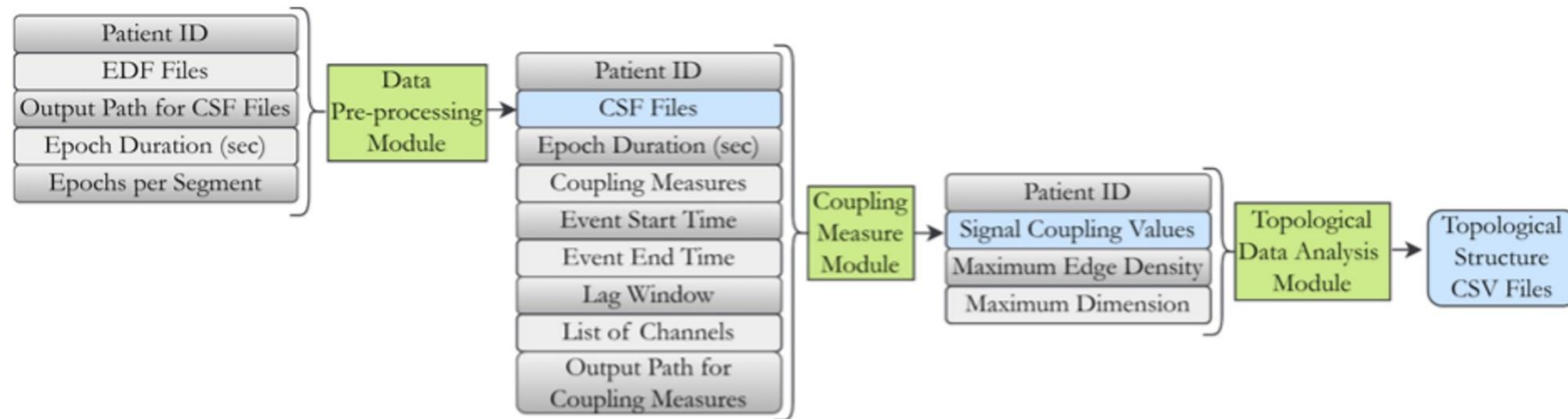
- Uses a consortium blockchain to maintain integrity and provenance associated with published datasets
- Independent verification of datasets while promoting reuse and reproducibility
- Implemented using the open-source Hyperledger Fabric Framework
- Stores the cryptographic hash of the data (e.g. SHA256 checksum) as a manifest in the blockchain along with the metadata
- Actual data are not stored as a part of OSC (scientific data could be large)
- Generates an identifier for the information stored on the blockchain
  - Uniquely ties together metadata elements such as contributor information, location of the data and cryptographic hash of the data
  - 'append only' structure prevents altering or deleting previously entered data
  - Data are therefore verifiable and immutable - essential for reproducibility and audits
- Provides programmatic access to the blockchain a python-based command line utility
- Allows registered external platforms and hubs to connect and use the blockchain

# Neuro-integrative Connectivity (NIC) Tool

- A compositional workflow-based tool that analyzes brain functional connectivity patterns in neurological disorders using electrophysiological signal data such as EEG recordings
- NIC tool available on NSG for users
- Uses a modular software architecture to support end-to-end EEG data analysis
  - JSON-based data format for efficient analysis, computation of coupling measures representing interactions between brain regions, and the use of algebraic topology methods to characterize brain interaction patterns
- Data pre-processing module: transforms EDF data to the Cloud wave Signal Format (CSF)
  - JSON-based CSF file format is self-descriptive, containing study metadata, channel-specific metadata, clinical event annotations and fragments of signal data; enhances human readability; allows distributed storage; efficient access; integration with parallel processing
- Signal coupling measure computation module: computes quantitative measures of the coupling between signal recordings from different electrodes
  - Coupling measures (non-linear correlation coeff, phase coherence, linear correlation coeff) are stored in text files for subsequent analysis
- Topological data analysis module: uses coupling measure values to generate algebraic topology structures
  - Generates algebraic topology structures such as simplicial complexes representing high-dimensional interaction patterns using persistent homology methods
  - Generates a single output file with values corresponding to the birth, death and dimension values of algebraic topology structures
  - Subsequently analyzed using statistical and ML methods

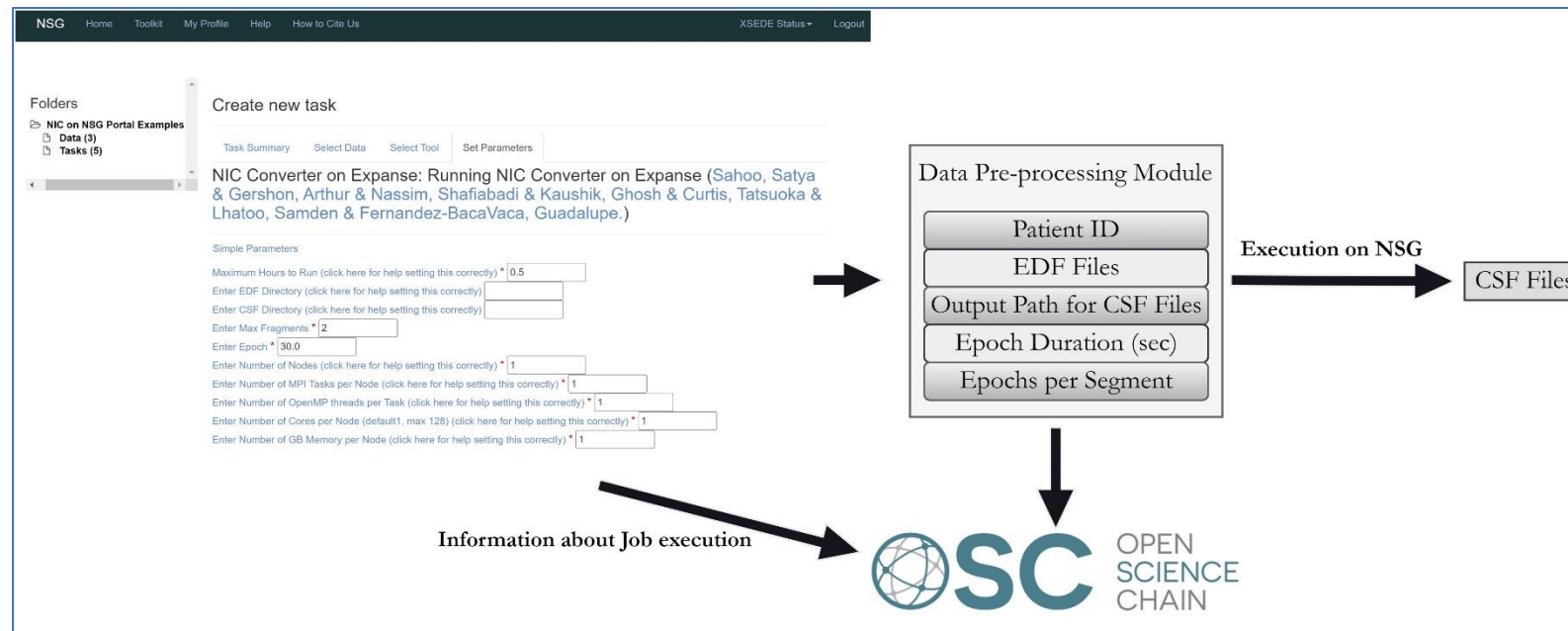
## Provenance Metadata in NIC Tool

- NIC workflow tool involves the recording of unique provenance metadata elements to enable
  - accurate recording of the context of each experimental study
  - subsequent reproducibility and compliance with FAIR guidelines for neurological studies in the future
- Figure illustrates various metadata associated with each NIC module
  - Each NIC module (green) has its own distinct set of metadata requirements (grey) to generate output (blue).



# NSG-OSC Integration Using the NIC Tool

- Stored the standardized metadata information associated with the data generated by the NIC tool into the OSC
- Motivation to study the reuse and reproducibility of the application
- Used the existing OSC command line utility to store and update the metadata
- NSG-NIC-OSC integration is the first test of the design and will enable such integration for other NSG tools
- Metadata associated with a tool running on supercomputers via NSG:
  - Number of servers, number of processors, estimated time to run on the supercomputers and name of the input file
  - Specific tools will have their specific parameters tied to the tool

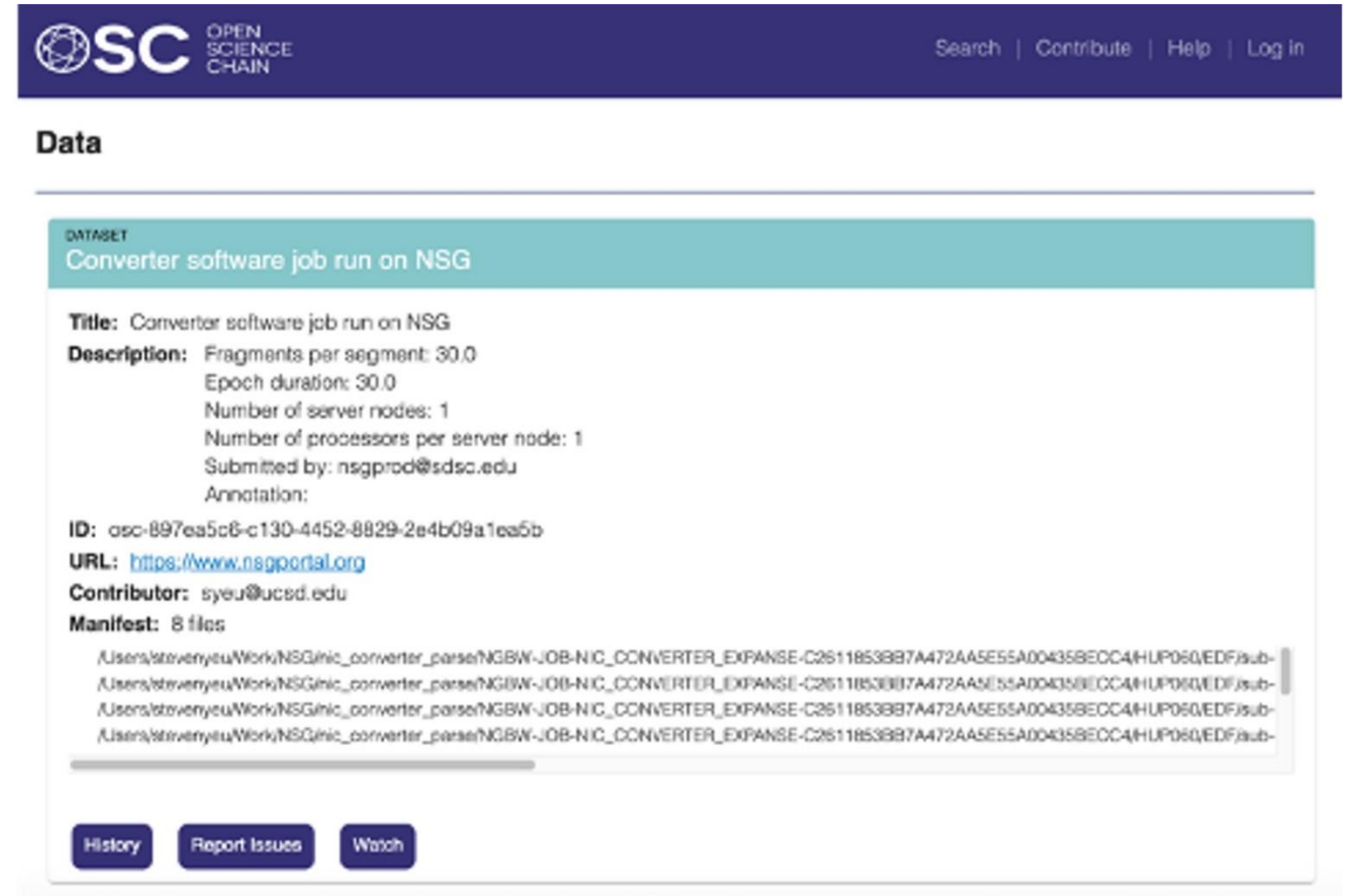


## Development Work for NSG-OSC Integration Using the NIC Tool

- A new table was introduced within NSG's MySQL database
  - Captures essential data elements, including the job identifier (job id) for the application, the status of job execution (such as 'completed', 'queued' or 'running') within the NIC module and the unique OSC-ID
  - OSC-ID is generated and logged upon the successful addition of the corresponding information associated with running the NIC tool within NSG framework to the OSC blockchain
- A programmatic module was developed to facilitate the transfer of relevant metadata as an artifact to the OSC blockchain
- Automated scripts were developed to invoke the submission process to OSC using the command line utility
- Metadata associated with both the NIC tool and the job are captured
- The cryptographic checksum of the data used in the experiment was calculated and appended to this artifact.
- The OSC command line utility initiates connection to OSC and transfers the information to be stored in the blockchain
- The chain code of the blockchain can identify duplicate entries in which case the transaction is rejected and the status is updated in NSG
- Upon successful inclusion in the blockchain, an OSC-ID is generated and sent back to NSG which is then added to the MySQL table
- When the data or metadata evolve for the same experiment, updates are identified and sent to the blockchain, thus maintaining the provenance of the data or metadata of the experiment

## Data and Metadata Stored in Blockchain

- Additional provenance metadata information includes (i) specific file path where the associated files are located, (ii) the cryptographic hash integrity information of the input and out-put files, (iii) the contributor email, (iv) information related to running the job on the HPC resource such as the number of processors used, (v) the version of the NIC software is also included in the transmission
- OSC provide customized metadata fields
- Using OSC's search API, information already stored in the blockchain can be extracted for independent verification and future reuse
- Future plan includes to integrate the search API with the NSG portal which will allow researchers to search for datasets or workflow processes that have been run on NSG for reuse



The screenshot displays the OSC interface. At the top, there is a dark blue header with the OSC logo (a globe with 'SC' and 'OPEN SCIENCE CHAIN' text) on the left, and navigation links 'Search | Contribute | Help | Log In' on the right. Below the header, the word 'Data' is prominently displayed. The main content area features a teal header for the dataset 'Converter software job run on NSG'. The dataset details include: Title: Converter software job run on NSG; Description: Fragments per segment: 30.0, Epoch duration: 30.0, Number of server nodes: 1, Number of processors per server node: 1, Submitted by: nsgprod@sdsu.edu, and Annotation: (empty). The ID is 'osc-897ea5c8-c130-4452-8829-2e4b09a1ea5b', the URL is 'https://www.nsgportal.org', and the contributor is 'syeu@ucsd.edu'. The manifest lists 8 files with their full paths. At the bottom of the dataset card, there are three buttons: 'History', 'Report Issues', and 'Watch'.

Sivagnanam, S., Yeu, S., Lin, K. et al. Towards building a trustworthy pipeline integrating Neuroscience Gateway and Open Science Chain. Database (2024) Vol. 2024: article ID baae023; DOI: <https://doi.org/10.1093/database/baae023> (accepted)

# Outline

- Neuroscience Gateway (NSG)
- Integration of NSG and Open Science Chain (OSC) - Using the Neuro-Integrative Connectivity (NIC) tool for improving AI readiness through provenance metadata

## ➤ Conclusion



## Conclusion

- We presented integrating NSG and OSC platforms to support reproducibility in neuroscience research using novel blockchain techniques with the NIC workflow tool
- Work is ongoing to include such metadata capture of other NSG tools such as the computational neuroscience software NEURON
- NEURON is one of the most used computational neuroscience tool used for modeling of single and network of neurons
- We plan to explore the possibility of using the information stored in the blockchain to identify analogous files used in creating neuronal models or in EEG studies
- We will also explore integration of citation data that can help researchers track and reference prior work more efficiently, fostering a culture of knowledge sharing and intellectual collaboration
- Other science gateways in phylogenetics, Cryo-electronmicroscopy uses the same gateway software framework as NSG and there is potential for OSC integration

Acknowledgement: NIH NIBIB 3U24EB029005-04S1 (UCSD, Case Western Reserve University), 2022-2023 NOT-OD-22-067 – FY2022 Request for ODSS Funds to Support Collaborations to Improve the AI/ML Readiness of NIH-Supported Data (FY22 AI-Readiness program) OFFICE OF THE DIRECTOR, NIH. NSF grants—1935749, 1840218, 2114202. This work was also supported in part by the grants from the NIH: R01DA053028, the US Department of Defense (DoD) grant W81XWH2110859, and the Clinical and Translational Science Collaborative of Cleveland, which is funded by the NIH, National Center for Advancing Translational Sciences, Clinical and Translational Science Award grant, UL1TR002548.