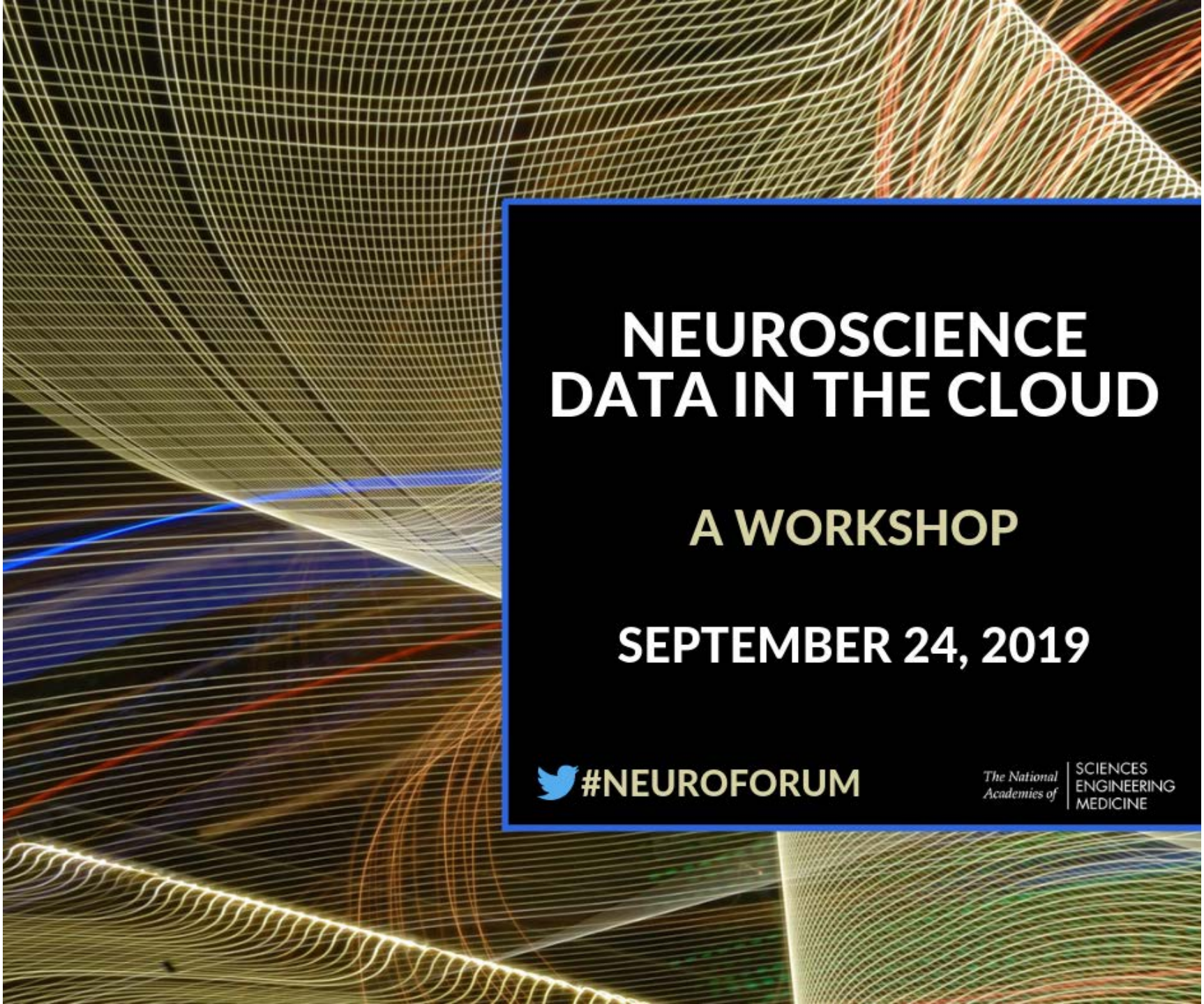
NEUROSCIENCE
DATA IN THE CLOUD
A WORKSHOP
SEPTEMBER 24, 2019
#NEUROFORUM
The National Academies of
SCIENCES
ENGINEERING
MEDICINE

SESSION III: FUTURE DIRECTIONS

Session Objective:

- Synthesize key highlights from the workshop discussions, including identifying next steps and promising areas for future action.

Session Overview

MAGALI HAAS, Cohen Veterans Bioscience (Moderator)

Panelists

TED WILLKE, Intel Labs; Portland State University

DOUGLAS LANDSMAN, National Multiple Sclerosis Society

SILVANA BORGES, Food and Drug Administration

CVB's Criteria for Selecting a Data Commons

45 Requirements				
Volume	Variety	Velocity	Veracity	Value
Availability	Durability	Redundancy	Recoverability	Scalability
Computational Capability (3)	Computational Configurability	Extensibility & Adaptability	Interoperability	Searchability & Retrievability
Seamless Integration	Accessibility	Usability	Import-Export	Storing Capability
Affordability	Migratability	Workflow Controllability	Support Capability	Environment Longevity
Retainability	Sustainability	Data Reproducibility	Organizational Survivability	Protectability
Privacy	Cyber Security	Shareability	Transparency	Reproducibility
Compatibility	Global Compliance	Accounting and Auditability	Flexibility & Elasticity	Analytics Visualization
Domain Focus	Open Source	Geographic Diversity		

10 Areas of Platform Requirements

Data Volume
No. 1

- Petabyte Scale
- (FAIR)
- Data Driven Discovery

Data Variety
No. 2

- Unstructured
- Heterogeneous

No Constraints on Raw Data Type

Data Velocity
No. 3

- High Throughput Approaches

Data flows - streaming, ingest - processing

Data Veracity
No. 4

- Complex Analytics

Best in class analytics and bioinformatics tools, workflows, pipelines

Data Value
No. 5

- Globally - Cloud

Durability, redundancy, survivability, longevity, platform sustainability, recoverability, reproducibility

Computation
No. 6

- Workflows, Pipelines

Finding, computing style, configurability, complex analytics, locality of reference. "One size does not fit all."

Interoperability
No. 7

- Heterogeneous Data

Seamless , APIs, restful services, global , concurrency, standards, open source.

Privacy & Security
No. 8

- Global Compliance
- Governance - auditability

HIPPA HITECH, EU-GDPR, FISMA, NIST, NIH-BD2K, FAIR, GA4GH, VA, DOD, FDA, BIDS, FedRAMP, etc.

Scalability
No. 9

- Performance execution

Linear, network computation

Sustainability
No. 10

- Reusability

Retainability, protectability, survivability, funded, affordable

Compared 105+ Available Platforms

Platforms and Software Technologies

#1. NCI Genomic Data Commons (GDC)	#2. tranSMART Knowledge Management Platform	#3. Informatics for Integrating Biology and the Bedside (i2b2)	#4. Ontario Brain Institute (Brain-CODE)	#5. EU EPILEPSIAE Database
#6. IEEG.org – International Epilepsy Electrophysiology	#7. NSF Cloud Platforms - Computing in the Cloud	#8. NIMH Data Archive - National Institute of Mental Health	#9. MIT "SuperCloud"	#10. HPI Hasso Pattner Institute - Univ of Potsdam
#11. EMC – Pivotal - Large Scale Hadoop Testbed	#12. Perkin Elmer – "Signals"	#13. PMI (Precision Medicine Initiative) New York Genome Center + IBM	#14. The Open Cloud Consortium – Open Science Data Cloud	#15. CG HUB from The Cancer Genome Atlas (TCGA)
#16. Cancer Genome Collaboratory - (Canada)	#17. Blackflynn	#18. "Genome Bridge" – The Broad	#19. IBM Watson Health & IBM Watson Health Cloud	#20. MVP - Million Veterans Program (GenISIS)
#21. Intel PCCSB - Intel Parallel Computing Center Structural Biology	#22. Collaborative Cancer Cloud - Intel	#23. LONI Laboratory of Neuro Imaging - IDA Image and Data Archive (USC)	#24. European Open Science Cloud	#25. ICGC Data Portal
#26. LORIS Longitudinal Online Research & Imaging System (Canada)	#27. Frederick National Laboratory FFRDC SysBioCube	#28. DNAnexus Cloud Based Platform	#29. NCBI National Center for Biotechnology Information	#30. GENISIS Project – Cloud Based
#31. cBio Cancer Genomics Portal	#32. Sage Bionetworks - Synapse	#33. Palantir	#34. BioStorage Technologies	#35. BC Platforms – Federated DB
#36. DART –American College of Radiology	#37. Sentinel	#38. XNAT	#39. BioMart	#40. Ensembl Project
#41. REDCap	#42. INCF	#43. NeuroVault	#44. Shanoir Data Management	#45. COINS
#46. NITR	#47. Vivli	#48. Google	#49. Facebook	#50. NIDB Neuro informatics Database

ACTION PLAN:

Working Groups Proposed:
Informed Consent Templates
Governance – Consensus Principles for Data Policy, Privacy, Access, Use, Disclosure
Common Data Model framework and methods for transformation across versions
Incentive models for a variety of stakeholders
Training, training, training.

Policy & Advocacy Proposed:
Legislation that addresses re-identification

Sustainability Models – infrastructure and funding:
- For data, platforms, software
- Frameworks for types of data to be maintained/future proofing

Best Practices:
- Code-sharing with protections
- Reporting of Incidental findings
- Storage practices (e.g. containerization, cloud vs local)

Approaches into tapping into existing initiatives:
- Catalog of Platforms, Data, Pipelines
- Coordinating Bodies (e.g. NIH, INCF)