Using Cloud-Based Resources for Neuroimaging Research: A Practical Approach

Deanna M. Barch, PhD, Washington University in St. Louis; **Sheena M. Posey Norris, MS,** The National Academies of Sciences, Engineering, and Medicine; and **Maryann E. Martone, PhD,** University of California, San Diego, and the International Neuroinformatics Coordinating Facility

July 19, 2021

Introduction

The scale and scope of human and animal neuroscience research has been increasing exponentially over the past decade. This growth has manifested both as increases in the number of participants in many studies, as well as an increase in the volume and types of data collected from each individual [1,2,3]. Many of these efforts have been enabled by the ability to use "cloud-based" tools for storage and computation. By cloud-based tools, the authors of this manuscript mean storage, computational resources, and programs that are available to a wide array of users on demand via the internet through a particular provider's cloud-based servers. Initially, the use of such resources required extensive expertise held by relatively few researchers and few institutions. While the use of these tools still requires a level of knowledge and expertise that is not necessarily widespread, the tools have become much more accessible, and a growing number of investigators are interested in harnessing their power in support of their research. However, many barriers and consequences for misuse still exist as these tools are used to support human and animal neuroscience research. As with many new technologies, investigators have a tendency to use the cloud like they use their local computer and storage resources. Misuse can lead to inefficiencies, extra costs, and sometimes unwitting security or privacy violations because different policies and costs accrue with the use of the cloud. For example, rather than working with data directly in the cloud, researchers may continue to download copies of data to their local drives, not realizing that there are costs associated with downloading files from commercial clouds.

To better understand both the strengths of and barriers to appropriate use of such technology, the National Academies of Sciences, Engineering, and Medicine's Forum on Neuroscience and Nervous System Disorders hosted a workshop on September 24, 2019, entitled "Neuroscience Data in the Cloud" [4]. This workshop explored the burgeoning use of cloud technology to advance neuroscience research and approaches to addressing current barriers [5].

Although the workshop highlighted great strengths in the use of cloud-based tools and the progress that has been made to date, numerous barriers and challenges remain for many researchers to move into this space. Based on discussions at the workshop, it seemed clear that there would be value in generating an informational resource for investigators and administrators in the field at different levels of experience for understanding, accessing, and successfully using cloud-based tools in support of neuroscience research, using human neuroimaging as an example. Human neuroimaging was chosen as it already has numerous cloud-based infrastructure and tools, but the resource is meant to be useful for neuroscientific data of all kinds.

Developing a Guide for Neuroscience Data in the Cloud

To explore how such resources might be organized, a collaborative working group came together, comprised of interested individuals from the workshop. The Action Collaborative[1] on Neuroscience Data in the Cloud (see Acknowledgements for a list of members) included a diverse group of individuals with a wide range of expertise in cloud-based tools as well as legal and ethi-

1 The Collaborative is an ad hoc activity convened under the auspices of the Forum on Neuroscience and Nervous System Disorders at the National Academies of Sciences, Engineering, and Medicine (the National Academies). The work it produces does not necessarily represent the views of any one organization, the Forum, or the National Academies, and is not subjected to the review procedures of, nor is it a report or product of, the National Academies.

cal issues surrounding the use of cloud-based technology. Members of the action collaborative produced a guide (https://training.incf.org/cloud-based-computer-matrix) that could be used by investigators to make decisions about whether or not to use the cloud for their research and to provide guidance on how to use the cloud effectively.

Use Case Scenario and Evaluation Matrix

To provide useful examples for the field, the guide offers a use case scenario of an early-stage investigator with limited expertise using cloud-based technology resources and the types of informed choices the investigator would have to make across many different dimensions.

The guide includes an evaluation matrix (see *Table 1*), comprising different types of concerns and issues that an investigator would address, including dimensions relevant to the size and scope of the study (e.g., number of participants, amount of data per participant, length of study), but also considerations related to the type of data being collected (e.g., privacy and data sharing), the expertise and financial resources available to the investigator through the home institution, the number of institutions involved in the project, and requirements or desires in regard to data sharing and longevity of the data.

For each dimension in the matrix, a description, range of values or levels (e.g., the researcher's skill level in cloud-based computing, level of privacy or security needed) and definitions of those levels is provided. Not all of the dimensions are technical. Issues involved in gaining institutional approval for cloud-based studies and the impact of involving multiple institutions in a cloud-based study is an example (see *Box 1*). The matrix also includes information about options and choices for each of the considerations, as well as resources for gathering more information or training, things to avoid, relevant articles, tools, and user stories. By providing different value sets for each dimension in this guide, researchers will be able to consider their own use cases and evaluate them against the matrix. These value sets are not intended to encourage an investigator either to use or not use cloud-based resources. Instead, through this process, the goal is for researchers to gain a better understanding of how different levels of these values sets impact the use of the cloud for neuroscientific data and the resources available to them for effective and responsible cloud use. While the evaluation matrix provides an overview of those decision points, detailed information on the suggested next steps is provided in the full guide.

Table 1 | Evaluation Matrix

Dimension	Description	Value set
Researcher skills	What computational skills and data handling skills does the researcher have?	• Low (basic familiarity with neuroimaging tools and workflows in a local environment, but little or no experience with cloud-based computing) • Medium (good computational and data skills, but only modest cloud-based computing experience) • High (computational and data skills; has cloud-based computing experience)
Number of institutions	The more institutions involved, the greater the challenge for coordination and consistency of control over the data and tools. Additional institutions mean additional complexities with data use agreements, HIPAA compliance, institutional review board (IRB) approvals and intellectual property, as well as more technical factors, such as standards for data storage and calibration of tools.	• Same institution • Multi-institution

Published July 19, 2021

Dimension	Description	Value set
Access to computational resources and expertise	Access to expertise within a computer science department or data center and degree of services provided by information technology services and/or data science center.	• Low (access to a few institutional resources) • Mixed (good neuroimaging expertise, but little institutional computer or data science support; or good computational expertise and resources but little neuroimaging expertise) • High (good neuroimaging expertise and strong institutional computer and data science support using cloud computing)
Data Size	# of subjects, # of files per subject, and size of files; to be downloaded or not	• Yes, the size of data are sufficient (>= terabytes) to warrant pushing to cloud • No, the data size is small and may not require the cloud
Data complexity/ scope	Number of modalities and data types; dimensions of these data (e.g., different licenses; identifiability; and different sharing, IRB, and HIPAA regulations)	• Low (a limited number of neuroimaging data types) • Medium (multiple structural and functional neuroimaging types coming from multiple sources covered by different licenses) • High (multiple structural and functional neuroimaging types as well as other data types, such as behavioral data and/or sequence data)
Number of copies	Will all data be accessed through a single centralized storage (e.g., cloud), or are local copies required? Multiple copies can lead to issues with data integrity, versioning, and archival storage.	• 1 • >1
Privacy	Protections for human subjects or other types of access control. Note that this will interact with the data complexity issue because privacy concerns may change as more and more data types accrue.	• Low (anonymized data with no PHI) • Medium (de-identified data—no special access controls) • High (identifiable data with substantial PHI)
Security	Issues include different regulatory policies that would govern compliance and what the archive already has in place (e.g., NIH Authority to Operate)	• Low (ISO 27001) • Medium (FISMA/FedRAMP moderate; NIST 800.53 rev4) • High (FISMA/FedRAMP high or data residency & exfiltration controls [in/out])

Dimension	Description	Value set
Data generation sources	Will all data be generated by the institutions involved in the study, or will some come from outside parties (e.g., wearable devices)?	• Yes (at least some data will come from outside sources) • No (all data will be generated by the institutions involved in the study)
Length of study	Longer studies may necessitate the use of multiple scanner protocols over time or analysis strategies may change. Issues include complexities of managing a length of study and length of time data required to be stored as well as data and software versioning.	• Short (data collected over relatively short time [e.g., 1–2 years] and no need for active storage post study completion) • Medium (either data collected over longer time period [3-5 years] and/or need for longer active storage post study completion [e.g., 3-5 years]) • Long (longitudinal study over many years [e.g., 5+ years] and/or long-term active storage post study completion [e.g., 5+ years])
Costs	How many direct costs for computing, storage, network costs are borne by the researcher? Issues include both short-term (while doing the study and analysis) and long-term costs for storage; cost of curation and organizing data, both the data the researcher is generating and the output; cost of complying and using standards; and cost of compute.	• Low (relative low costs [$10,000 or less]) • Medium (greater than $10,000, but less than $25,000) • High ($25,000 or more)
Existing data	If the researcher uses other datasets in the study, then they must understand the conditions for data reuse and sharing of derivative results (e.g., the Adolescent Brain Cognitive Development [ABCD] study has high restrictions on re-release options).	• Yes • No
Software/pipelines	Does the researcher have to develop their own cloud-compliant tools/analysis pipelines?	• No • Yes, but only a few • Yes, and it is many
Degree of data sharing	Will the data be shared with others/made public? If so, will there be any restrictions on access or usage of the shared data?	• Public, controlled access • No sharing
Submission to third party repository	Will the data be deposited in a third-party repository? What are the requirements of the repository?	• Yes • No
IRB experience with neuroimaging and cloud-based data	Does the IRB have familiarity with issues surrounding sharing data in a cloud-computing environment?	• Yes • No

Dimension	Description	Value set
Informed consent data sharing coverage	The degree of sharing and use allowed by informed consent. Issues include the type of repository to which data can be shared, the nature of data use agreements requirements, and the degree or re-release allowed and whether the data must be de-identified or anonymized.	• Low (does not allow any data sharing) • Medium (allows data sharing under restricted conditions) • High (allows broad and open data sharing)

SOURCE: Barch, D., M. E. Martone, J. Cohen, N. Farahany, M. Haas, S. Horgan, D. Kennedy, T. Madhyastha, and R. Poldrack. 2021. *Hitchhiker's guide to using cloud-based resources for neuroimaging research.* Work licensed under a CC BY 4.0 license. Available at: https://training.incf.org/sites/default/files/CloudBasedComputer_MATRIX.pdf (accessed July 7, 2021).
NOTE: (a) FISMA = Federal Information Security Moderation Act; (b) FedRAMP = Federal Risk and Authorization Management Program; (c) HIPAA = Health Insurance Portability and Accountability Act; (d) ISO = International Organization for Standardization; (e) NIH = National Institutes of Health; (f) NIST = National Institute of Standards and Technology; (g) PHI = protected health information

Future Directions of a Living Resource

Cloud-based technologies and approaches are constantly evolving; therefore, this resource is designed to be a living document that is updated and modified as the space of cloud-based tools shifts and grows and as the concerns and considerations change, updated by members of the field with experience in this domain. To support this, a living version of the evaluation matrix is available at the International Neuroinformatics Coordinating Facility (INCF.org) [6] where the documents will be available online and comments can be provided (https://training.incf.org/cloud-based-computer-matrix). The hope is that members of the community interested in this work would be willing to contribute to the comprehensiveness of this guide by sharing additional resources, best practices, and user stories. A working group has been established at INCF to handle updates and moderate discussions to help ensure that this guidance can be as helpful as possible to investigators who wish to engage with cloud-based tools, or who are already operating in this space but wish to gain greater knowledge or share the knowledge that they have gained.

References

1. Harms, M. P., L. H. Somerville, B. M. Ances, J. Andersson, D. M. Barch, M. Bastiani, S. Y. Bookheimer, T. B. Brown, R. L. Buckner, G. C. Burgess, T. S Coalson, M. A. Chappell, M. Dapretto, G. Douaud, B. Fischl, M. F. Glasser, D. N. Greve, C. Hodge, K. W. Jamison, S. Jbabdi, S. Kandala, X. Li, R. W. Mair, S. Mangia, D. Marcus, D. Mascali, S. Moeller, T. E. Nichols, E. C. Robinson, D. H. Salat, S. M. Smith, S. N. Sotiropoulos, M. Terpstra, K. M. Thomas, M. Dylan Tisdall, K. Ugurbil, A. van der Kouwe, R. P. Woods, L. Zöllei, D. C. Van Essen, and E. Yacoub. 2018. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *NeuroImage* 183: 972-984. https://doi.org/10.1016/j.neuroimage.2018.09.060.
2. Miller, K. L., F. Alfaro-Almagro, N. K. Bangerter, D. L. Thomas, E. Yacoub, J. Xu, A. J. Bartsch, S. Jbabdi, S. N. Sotiropoulos, J. L. R. Andersson, L. Griffanti, G. Douaud, T. W. Okell, P. Weale, I. Dragonu, S. Garratt, S. Hudson, R. Collins, M. Jenkinson, P. M. Matthews, and S. M. Smith. 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* 19(11): 1523-1536. https://doi.org/10.1038/nn.4393.
3. Casey, B. J., T. Cannonier, M. I., Conley, A. O. Cohen, D. M., Barch, M. M. Heitzeg, M. E. Soules, T. Teslovich, D. V. Dellarco, H. Garavan, C. A. Orr, T. D. Wager, M. T. Banich, N. K. Speer, M. T. Sutherland, M. C. Riedel, A. S. Dick, J. M. Bjork, K. M. Thomas, B. Chaarani, M. H. Mejia, D. J. Hagler Jr., M. Daniela Cornejo, C. S. Sicat, M. P. Harms, N. U. F. Dosenbach, M. Rosenberg, E. Earl, H. Bartsch, R. Watts, J. R. Polimeni, J. M. Kuperman, D. A. Fair, A. M. Dale, and ABCD Imaging Acquisition Workgroup. 2018. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience* 32: 43-54. https://doi.org/10.1016/j.dcn.2018.03.001.
4. National Academies of Sciences, Engineering, and Medicine (NASEM). 2019. *Neuroscience Data in the Cloud: A Workshop.* Available at: https://www.nationalacademies.org/event/09-24-2019/neuro-

Box 1 | IRB Experience with Neuroimaging and Cloud Data/Computing

Description: Does the IRB have familiarity with issues surrounding sharing data in a cloud computing environment?

Value set definitions:

- **Yes**, the institution or the investigator does have IRB experience with neuroimaging and cloud data/computing.
- **No**, the institution and the investigator do not have IRB experience with neuroimaging and cloud data/computing.

Value of use case example: *No*, the researcher's institution does not have a history of IRB experience with neuroimaging and cloud storage/computing.

Discussion of use case: The researcher will need to ensure that their institution's IRB consults with more experienced institutions or will need to gather suggestions from other institutions with more experience to ensure that the IRB can appropriately evaluate and advise on issues surrounding analysis and sharing in a cloud computing environment.

Best practices:

- Engage the IRB in a discussion about data-sharing approvals at the start of the project.
- Identify a colleague at an institution with good experience with data sharing in a cloud environment to determine if you can provide a consultant from their IRB to your IRB.
- Provide your IRB with sample copies of approved consent forms and procedures from other institutions and projects that have successfully engaged in data sharing in a cloud environment.

Things to avoid: Avoid having your IRB create a policy or set of procedures not in line with the broader community.

See also (within the guide):

- Informed Consent/Data Sharing Coverage
- Degree of Data Sharing

Resources and tools:

- Open Brain Consent: portable consent forms specifically for sharing human neuroimaging data, developed by the Open Brain Consent Working Group (preprint: https://psyarxiv.com/f6mnp/)

SOURCE: Barch, D., M. E. Martone, J. Cohen, N. Farahany, M. Haas, S. Horgan, D. Kennedy, T. Madhyastha, and R. Poldrack. 2021. *Hitchhiker's guide to using cloud-based resources for neuroimaging research.* Work licensed under a CC BY 4.0 license. Available at: https://training.incf.org/sites/default/files/CloudBasedComputer_MATRIX.pdf (accessed July 7, 2021).

Published July 19, 2021

science-data-in-the-cloud-a-workshop (accessed May 18, 2021).
5. NASEM. 2020. *Neuroscience Data in the Cloud: Opportunities and Challenges: Proceedings of a Workshop.* Washington, DC: The National Academies Press. https://doi.org/10.17226/25653.
6. Abrams, M. B., J. G. Bjaalie, S. Das, G. F. Egan, S. S. Ghosh, W. J. Goscinski, J. S. Grethe, J. H. Kotaleski, E. T. W. Ho, D. N. Kennedy, L. J. Lanyon, T. B. Leergaard, H. S. Mayberg, L. Milanesi, R. Mouček, J. B. Poline, P. K. Roy, S. C. Strother, T. B. Tang, P. Tiesinga, T. Wachtler, D. K. Wójcik, and M. E. Martone. 2021. A Standards Organization for Open and FAIR Neuroscience: the International Neuroinformatics Coordinating Facility. *Neuroinformatics*. https://doi.org/10.1007/s12021-020-09509-0.

DOI

https://doi.org/10.31478/202107b

Suggested Citation

Barch, D. M., S. M. Posey Norris, and M. E. Martone. 2021. Using cloud-based resources for neuroimaging research: A practical approach. *NAM Perspectives.* Commentary, National Academy of Medicine, Washington, DC. https://doi.org/10.31478/202107b.

Author Information

Deanna M. Barch, PhD, is professor and chair of the Department of Psychological & Brain Sciences, and the Gregory B. Couch Professor of Psychiatry at Washington University in St. Louis. **Sheena M. Posey Norris, MS**, is a program officer on the Board on Health Sciences Policy at the National Academies of Sciences, Engineering, and Medicine; **Maryann E. Martone, PhD**, is professor emerita at the University of California, San Diego, and Chair of the Governing Board at the International Neuroinformatics Coordinating Facility (INCF).

The authors are members and/or staff of the Action Collaborative on Neuroscience Data in the Cloud, an ad hoc activity convened under the auspices of the Forum on Neuroscience and Nervous System Disorders at the National Academies of Sciences, Engineering, and Medicine.

Acknowledgments

The authors wish to thank the members of the Action Collaborative on Neuroscience Data in the Cloud for their tremendous contributions to this practical guide. The members are **Jonathan Cohen**, Princeton University; **Nita Farahany**, Duke University; **Gregory Farber**, National Institute of Mental Health; **Magali Haas**, Cohen Veterans Bioscience; **Sean Horgan**, Verily Life Sciences; **David Kennedy**, University of Massachusetts Medical Center; **Tara Madhyastha**, Amazon Web Services; and **Russell Poldrack**, Stanford University.

In addition, we wish to thank **Clare Stroud** at the National Academies for her support and guidance during the work of the Action Collaborative.

Conflict-of-Interest Disclosures

Maryann Martone is a founder and has equity interest in SciCrunch, a tech start-up out of the University of California, San Diego, that develops tools to support rigor and reproducibility used in scientific publishing.

Correspondence

Questions or comments should be directed to Deanna Barch at dbarch@wustl.edu.

Disclaimer