Hitchhiker's Guide to Using Cloud-Based Resources for Neuroimaging Research

Authors: Deanna Barch, Maryann Martone, Jonathan Cohen, Nita Farahany, Magali Haas, Sean Horgan, David Kennedy, Tara Madhyastha, and Russell Poldrack

Table of Contents

Background

Mission: On September 24, 2019, the National Academies of Sciences, Engineering, and Medicine's Forum on Neuroscience and Nervous System Disorders hosted a workshop titled "Neuroscience Data in the Cloud," that explored the burgeoning use of cloud technology to advance neuroscience research and approaches to addressing current barriers. Following the one-day meeting, it was decided that there would be value in generating an informational guide 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 neuroimaging as an example. Neuroimaging was chosen as this example field as it already has numerous cloud-based infrastructure and tools, but the guidance is meant to be useful for neuroscientific data of all kinds. The action collaborative[1] was charged with producing such guidance to help users and administrators at different levels of expertise access and navigate cloud-based resources for neuroscience research.

Process: We convened a collaborative working group of individuals from the workshop who indicated their interest in participating in such an ongoing effort, attempting to ensure a wide range of representation. The group met monthly starting in February 2020. We began by generating a use case scenario of an early-stage investigator with limited expertise. We then generated an evaluation matrix, comprising different types of concerns and issues that such an investigator would address if they wanted to conduct research that used cloud-based resources. For each of these considerations, we provide a description, a range of values or levels for that consideration (e.g., the researchers skill level in cloud-based computing, level of privacy or security needed) and definitions of those levels. We then proceeded to generate best-case practices for each of those considerations, as well as resources for gathering more information or training, things to avoid, articles, and tools. Finally, we consider the use case against this evaluation matrix to provide an illustration of how these different considerations affect decisions about using the Cloud.

The guide should help researchers and administrators understand:

1. When to use Cloud resources
2. How to use the Cloud effectively

User Scenario

Persona: Jordan is a first-year assistant professor at a midsize R1 (research intensive) university. She received her Ph.D. in cognitive neuroscience in a lab that did primarily small-N task-based fMRI studies, and used relatively traditional approaches to data analysis (lab-based server or laptop) and did not regularly engage in data sharing. However, Jordan did a postdoctoral fellowship in the lab of a mentor

[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.

who was regularly engaged in larger scale studies that sometimes utilized cloud-based computing and engaged in broad data sharing. Jordan learned to do some analyses in the Cloud while a postdoctoral fellow, but was not responsible for setting up any of the infrastructure and was not responsible for setting up data sharing on any projects.

Environment: Jordan's institute has a high performance computing center for use by investigators with expertise in using containerized (i.e., pre-packaged) computing tools, and a data science center with cloud office hours. Further, there is at least one investigator in another department that regularly uses cloud-based computing resources. However, no one in Jordan's department does so. Jordan's institution has been involved in a number of large-scale clinical studies that involved data sharing of clinical data, but has not done so with neuroimaging data in previous studies.

Use case: Jordan would like to conduct a neuroimaging study in which she would recruit approximately 200 individuals from the community to examine the relationship between individual differences in facets of emotion regulation, and associated individual differences in structural and functional connectivity. She is also considering whether she will need a larger N and whether she needs to join or launch a multi-institution study or perhaps whether there are existing data that she can use. Jordan will assess a range of behavioral measures (individual differences in psychopathology, personality and behavior) outside of the scanner, as well as T1- and T2-weighted MRI, resting state fMRI and diffusion MRI. Her team is also considering obtaining DNA samples for whole-genome analysis and the possibility of doing mobile sensing with participants, collecting information about activity, sleep, geolocation, and potentially even scraping information about app usage, frequency of texts and calls, etc.

Jordan would like to establish robust and replicable analysis approaches for all of her data types, and she would like to compare and contrast several different analysis streams. Jordan may also need to develop novel analysis tools to accomplish some of her aims. She wonders whether she should be using the Cloud and if so, how should she go about it? She is planning to share the data through the NIMH Data Archive (NDA) or another repository, but she wants the data to be reusable by other platforms. It is also possible that she will want to follow these individuals longitudinally and conduct follow-up imaging or behavioral studies and make these data available as well.

Why would Jordan use the Cloud?
In general, researchers benefit from using the Cloud when data sets and compute needs become large and are shared. "Large" in this case is a moving target but generally refers to data size and complexity when moving, downloading, storing and working with the data becomes prohibitive over the course of the project. Cloud-based resources allow storage and compute to scale with size and complexity. Sharing data and compute resources also becomes easier as all parties are working off a common platform. Although cloud-based resources can be expensive, with the commercial platforms, costs can be easily shared among parties depending on how they are set up. Nevertheless, the Cloud is not necessarily right for all applications and can lead to high monetary and technological costs if an investigator is not careful.

We developed an evaluation matrix that summarizes major dimensions that can impact whether using the Cloud is right for Jordan. When considering whether to utilize cloud-based resources for her project, Jordan should consider where her project falls on each of the dimensions described below to determine whether the size and goals of the project warrants the use of cloud-based resources and whether her lab or institution has the resources necessary to help her navigate the complexities of cloud-based resources.

Note that in our evaluation matrix, we are generally considering the use of Cloud resources that provide both storage and compute resources. We are not considering popular cloud-based storage options such as Dropbox, Box, Google Drive or iCloud, unless explicitly noted.

A living version of this matrix is available at the International Neuroinformatics Coordinating Facility (INCF)where comments can be provided at https://training.incf.org/cloud-based-computer-matrix. A working group has been established at INCF to handle updates and moderate discussions.

Evaluation Matrix

Dimension	Description	Value set
Researcher skills	What computational skills and data handling skills does the researcher have?	Low, Medium, High
Number of institutions	The more institutions involved, the greater the challenge for coordination and consistency of control over the data and tools. Additional institutions means additional complexities with data use agreements, HIPAA compliance, 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
Access to computational resources and expertise	Access to expertise within a computer science department or data center and degree of services provided by a IT services and/or data science center.	Low, Mixed, High
Data size	# of subjects, # of files per subject, and size of files; to be downloaded or not	Yes, No
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, Medium, High
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, Medium, High
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, Medium, High
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, No
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, Medium, Long
Costs	How many direct costs for compute, 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 you are generating and the output; cost of complying and using standards; and cost of compute.	Low, Medium, High
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
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, Medium, High

Use Case

In the following section, we evaluate the user scenario against the matrix in Table 1. We provide definitions for each of the value sets and situate the user scenario within the matrix.

Researcher Skills

Description: What computational skills and data handling skills does the researcher have?

Value Set Definitions:

- **Low**: Researcher has basic familiarity with neuroimaging tools and workflows in a local environment, but little or no experience with cloud-based computing. Researchers who would label themselves as low on this dimension should consider carefully whether they have the time and resources needed to develop the necessary expertise in order to use cloud-based resources for their projects. Even though the research group may not have the necessary expertise now, the group will want to think about whether they need to develop this expertise for their future research efforts.
- **Medium**: Researcher has good computational and data skills but only modest cloud-based computing experience.
- **High**: Researcher has computational and data skills; has cloud-based computing experience.

Value of Use Case Example: *Medium* - Jordan did a postdoctoral fellowship in the lab of a mentor who was regularly engaged in larger scale studies that sometimes utilized cloud-based computing and engaged in broad data sharing. Jordan learned to do some analyses in the Cloud while a postdoctoral fellow, but was not responsible for setting up any of the infrastructure and was not responsible for setting up data sharing on any projects.

Discussion of Use Case Example: The researcher is fortunate that she has some experience with working in the Cloud, but probably not enough to avoid common mistakes without additional training and access to those with more expertise. Therefore, it is critical that Jordan increase her level of training and familiarize herself with best practices and what is available as whether or not she uses the Cloud now, she may need to do so in the future.

Best Practices:

- Utilize available campus resources on cloud-based computing/data science.
- Determine to what extent you, the researcher, will need to provide system administration for the Cloud as opposed to having someone else take this on.
- Look for existing tooling that meet the needs and skill level of the researcher (i.e., look at recent papers).
- Participate in training courses for neuroscience data in the cCoud.
- Look within funding agencies for training or informational researcher resources on cloud-based computing.

Things to Avoid:

- “Hidden” costs and risks: educate yourself! Remember that things you may not be used to paying for (e.g., CPUs you own but do not use, may cost you in a cloud-based environment).
- Reinventing the wheel: take advantage of what is already available.

- Don't go it alone.
- Don't be afraid to ask the Cloud provider for help/support/money, including coverage of any recurring fees.
- Focusing only on the short term: think about how your future needs, both for this study and future studies, and whether you need to invest in additional training now.

See Also:

- Costs
- Access to Computational Resources and Expertise

Resources and Tools:

- INCF training space: Growing, centralized resource for training materials in neuroinformatics
    - Current offerings relevant to the Cloud.
- NeuroHackademy: Summer school in neuroimaging and data science
    - See NeuroHackademy/: Cloud resources for neuroimaging
- CloudBank: A Cloud access entity that will help the NSF-supported computer science community access and use public Clouds for research and education by delivering a set of managed services designed to simplify access to public Clouds. Educational and training materials are available to all.
    - Getting started using the Cloud
    - Training materials
- STRIDES: NIH program that includes educational materials, training opportunities and other resources
- ReproNim Training Materials: ReproNim's (A Center for Reproducible Neuroimaging Computation) general training materials. Cloud examples in the "How would ReproNim do *That*?" series of documents.
- Neuroimaging Informatics Tools and Resources Collaboratory Computational Environment (NITRC-CE): NITRC-CE is an on-demand, virtual computing platform designed for neuroimaging researchers, incorporating many neuroimaging tools, and deployable on the Amazon Web Services (AWS) Elastic Cloud Computing (EC2) environment. The User Guide provides detailed instructions for using the NITRC-CE for cloud-based computing.

Relevant Articles:

- Science in the Cloud (SIC): A use case in MRI connectomics
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467033/
- Heads in the Cloud: A primer on neuroimaging applications of high performance computing
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4896536/
- Running neuroimaging applications on Amazon Services: How, when and at what cost
    - https://www.frontiersin.org/articles/10.3389/fninf.2017.00063/full

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Number of Institutions

Description: The more institutions involved, the greater the challenge for coordinating and ensuring consistency of control over the data and tools. Additional institutions means additional complexities with data use agreements, HIPAA rules, IRB approvals and intellectual property, as well as ensuring consistency of standards for data storage and calibration of associated instruments used for evaluations and tools used for data analyses.

Value Set Definitions:

- **Single Institution:** All key members of the research team are at the same institution
- **Multiple Institutions:** Key members of the research team are at more than one institution

Value of Use Case Example: *Single Institution* - Jordan would like to conduct a study in which they would recruit between 200 individuals from the community to examine the relationship between individual differences in facets of emotion regulation, and associated individual differences in structural and functional neural connectivity. She is also considering whether she will need a larger N, whether she needs to join or launch a multi-institution study, or whether there may be existing data that she can use.

Discussion of Use Case: While in the past, Jordan may have been able to justify the use of 200 subjects, given concerns about reproducibility a power analysis may show that she needs to increase the number of subjects in order to have confidence in any reported effects. It is not inconceivable that the number may be > 1000. Therefore, she will have to consider whether she needs to either lead or participate in a multi-institution study. Alternatively, she may be able to augment or perform her study with existing publicly available data. When using the Cloud, designing a multi-institutional study can add several levels of complexity, as policies, practices, and cloud access may vary across institutions. If institutions are international or comprise a mix of types, e.g, academic, commercial, governmental, the situation may be even more complex, as different countries have different privacy laws regarding use of the Cloud. Different types of partners may also have different requirements.

Best Practices:

- Single institution
    - Ensure that all team members adhere to the institution's relevant policies and practices.
- Multi-institution: Using lead institution as a starting and reference point and address the following
    - Generate an inventory of all institutions' relevant policies and practices.
    - Generate a data use/sharing policy across institutions before onset of data collection.
    - Consider whether different institutions within the project have different objectives, and whether and how those differences may impact the policies and practices for the project as a whole.
    - Pick a reference set of policies and practices, and identify any deviations of the guidelines applicable to one or more participating institutions that are more or less stringent than the reference set, and then determine whether institutions that are more lenient in those policies or practices can comply with the most stringent ones.
    - Negotiate failures to converge on the most stringent set, possibly using community-wide resources (such as this matrix!!) as additional reference points.
    - Distribute final guidelines to all team members.
    - The lead investigator at each institution should ensure adherence to guidelines at that institution.

- Address at the outset who owns which data and how this interacts with data sharing policies, both among the participating institutions and with outsiders (as outlined along other dimensions in this matrix).
- Ideally, all institutions should use the same cloud-based service/repository for the project, and one that is standard and appropriate for the type of data being collected (e.g., dedicated to neuroimaging NITRC IR NIMH Data Archive [NDA]). If project- and/or institution-specific one(s) are preferable/necessary (e.g., for pilot / interim results and/or if there are special needs and/or restrictions, such as HIPAA restrictions, project-specific meta-data requirements, etc., that are not met by existing standard services or repositories), then these should also be identified at the outset, and policies/practices established for how the affected information will eventually be migrated to the project-wide platform(s).

Things to Avoid:

- Do not assume that the lead institution is the **only one** responsible for determining and overseeing policies and practices related to the study.
- Do not assume that all institutions have the same policies, practices, or even objectives.
- Do not assume that each institution can operate independently.
- Do not assume that the lead investigator at each institution is solely responsible for policies and practices regarding the part of the study at that institution alone and does not need to coordinate with the other involved institutions.

See Also:

- Software/Pipelines
- Privacy
- Security
- IRB Experience with Neuroimaging and Cloud Data

Resources and Tools:

- ReproNim Statistics Module: including power analysis to determine required study size

Relevant Articles:

- Security and privacy requirements for a multi-institutional cancer research data grid: an interview-based study
  - https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-9-31
- Guidelines for Multi-Institutional/Collaborative Research
  - https://journals.lww.com/academicmedicine/Fulltext/2015/03000/Guidelines_for_Multi_Institutional_Collaborative.33.aspx
- Establishing a Multi-Institutional Quality and Patient Safety Consortium Collaboration Across Affiliates in a Community-Based Medical School
  - https://journals.lww.com/academicmedicine/Abstract/9000/Establishing_a_Multi_Institutional_Quality_and.97145.aspxEstablishing

User story: A real-time fMRI scanning project was carried out collaboratively between two major research universities, one with a medical center and the other without. This involved scanning at the institution with the medical center using protocols designed by the other, and with data delivered to a cloud-based ‘engine’ for real-time analysis and feedback. In order to

accomplish this, a variety of obstacles had to be hurdled. The respective institutional IRB's had to understand and approve the protocol. Network connections had to be developed in accordance with privacy protections (which were warranted by both institutions as well as the Cloud vendor). Technical experts from both institutions had to arrive at a common understanding of what was to be accomplished, who owned what aspects of the data, and ultimately deliver a reproducible prototype that allowed the work to begin. This process took roughly a year, much of which was occupied by efforts to coordinate the independent policy reviews carried out by the various units at each institution.

Access to Computational Resources and Expertise

Description: Access and degree of services provided by a computer science department and/or data science center at the investigator's institution.

Value Set Definitions:

- **Low**: Researcher has access to few institutional resources. Researchers who would label themselves as low on this dimension should consider carefully whether they have the time and resources needed to develop the necessary expertise in order to use cloud-based resources for their projects.
- **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.

Value of Use Case Example: *Mixed* - Jordan's institute has a high-performance computing center for use by investigators with expertise in using containerized (i.e., pre-packaged) computing tools, and a data science center with "cloud office hours", i.e., a dedicated help resource that can answer her questions about using the Cloud. Further, there is at least one investigator in another department that regularly uses cloud-based computing resources. However, no one in Jordan's department does so. Jordan's institution has been involved in a number of large-scale clinical studies that involved data sharing of clinical data, but has not done so with neuroimaging data in previous studies.

Discussion of use case: Based on the resources available to Jordan, she should reach out to colleagues with specific expertise in cloud-based neuroimaging or consider taking a training course that provides training in cloud-based use of neuroimaging tools.

Best Practices:

- Look to campus high-performance computing resources as a first start, if available.
- If not, look to the Computer Science department for cloud-based computing courses.
- Understand how others in the same field of research use cloud-based computing, and leverage their experience whenever possible.
- Work with the Cloud vendor to understand options, costs, and programs to support your work.

Things to Avoid:

- Do not assume you know what resources are available without checking. You need to do due diligence to check on resources.
- Do not be limited by the computing power of your laptop. There are many resources that will get you computing for free or a low cost.
- Do not invest in solutions that do not have a clear track record and evidence of some longevity.

See Also:

- Software/Pipelines

Resources and Tools:

- NeuroHackacademy: Summer school in neuroimaging and data science
    - See NeuroHack Academy/: Cloud resources for neuroimaging

- CloudBank: A cloud access entity that will help the NSF-supported computer science community access and use public clouds for research and education by delivering a set of managed services designed to simplify access to public clouds. Educational and training materials are available to all.
    - Getting started using the Cloud
    - Training materials
- STRIDES: NIH program that includes educational materials, training opportunities and other resources
- Neuroimaging Informatics Tools and Resources Collaboratory Computational Environment (NITRC-CE): NITRC-CE is an on-demand, virtual computing platform designed for neuroimaging researchers, incorporating many neuroimaging tools, and deployable on the Amazon Web Services (AWS) Elastic Cloud Computing (EC2) environment. The User Guide provides detailed instructions for using the NITRC-CE for cloud computing.
- Jetstream Cloud Resources: NSF-supported project led by the Indiana University Pervasive Technology Institute (PTI) designed for those who have not previously used high performance computing and software resources. The system is particularly geared toward 21st-century workforce development at small colleges and universities – especially historically black colleges and universities, minority serving institutions, tribal colleges, and higher education institutions in EPSCoR States.
- Cloud Carpentry for Genomics:on-line course materials for a Data Carpentry course providing practical training in understanding and using the Cloud for analysis

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Data Size

Description: Number of participants, number of files per participant, and size of files; number of copies of files, and whether the data will be downloaded or not.

Value Set Definitions: Yes or no, based on whether or not the size of data are sufficient (>= terabytes) to warrant pushing to cloud

Value of Use Case Example: *Yes* - The size of the data for the proposed study will be sufficiently large to warrant the use of the Cloud.

Discussion of Use Case: A modest amount of data may not warrant the use of cloud-based resources unless there are other considerations that warrant the use of the Cloud, such as a lack of resources at the investigators home institution, the need to coordinate data collection/processing across multiple sites, or the need to share data in a way that cannot be supported by the home institution.

Best Practices:

- Do not maintain copies of data if not necessary (see number of copies, below).
- Store data so that it can be queried/explored (see number of copies, below).
- Organize data to make it easy to optimize cost of cloud storage by using different storage classes (set up rules so this can be done automatically). AWS recently added intelligent tiering to make this even easier; see https://aws.amazon.com/blogs/aws/s3-intelligent-tiering-adds-archive-access-tiers/
- Separate derived products from archival/raw data (using different folders, for example).
- Use consistent naming conventions across projects (see BIDS format below).
- Raw data should be saved with read-only permissions to avoid accidental changes or deletion.
- Focusing on the short term: Consider how data sizes and ingress/egress may change over the course of the study. If you plan to remove your data from the Cloud at the end of the project, it is a good idea to reserve money in the budget for that in advance.

Things to Avoid:

- Do not group all files into a single archive (e.g., tarbal)l per participant.
- Do not duplicate raw data across researchers working on the same project. Consider a shared raw data repository.

See Also:

- Data Complexity/Scope
- Number of Copies

Resources and Tools:

- BIDS format for data organization
- AWS user guides for Cloud set up
    - https://aws.amazon.com/blogs/aws/s3-intelligent-tiering-adds-archive-access-tiers/
    - https://docs.aws.amazon.com/AmazonS3/latest/user-guide/create-lifecycle.html

Relevant Articles:

- The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments
    - https://www.nature.com/articles/sdata201644
- MEG-BIDS, the brain imaging data structure extended to magnetoencephalography
    - https://www.nature.com/articles/sdata2018110

User Story: Cloud providers typically charge fees for moving data out of the Cloud, between regions, and between certain types of storage classes (for example, moving data out of archival storage may incur a fee). It is important to know the fee structure for these operations to avoid unexpected bills. If you plan to remove your data from the Cloud at the end of the project, it is a good idea to reserve money in the budget for that in advance.

Data Complexity/Scope

Description: Number of modalities and data types; dimensions of these data, e.g., different types of data that requires different types of licenses, identifiability, and different sharing, IRB and HIPAA controls.

Value Set Definitions:

- **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

Value of Use Case Example: *High* - Jordan will assess a range of behavioral measures (individual differences in psychopathology, personality and behavior) out of the scanner, as well as T1, T2, resting state and diffusion imaging. They are also considering obtaining DNA samples and the possibility of doing mobile sensing with participants, collecting information about activity, sleep, geolocation, and potentially even scraping information about app usage, frequency of texts and calls, etc.

Discussion of Use Case: The acquisition or use of only a single data type, especially if not a large amount of data, may not warrant the use of cloud-based resources unless there are other considerations that warrant the use of the Cloud, such as a lack of resources at the investigators home institution, the need to coordinate data collection/processing across multiple sites, or the need to share data in a way that cannot be supported by the home institution. In Jordan's case, the high complexity/scope of the data she proposes to collect may make this project appropriate for the use of cloud-based resources.

Best Practices:

- Organize data with sufficient metadata (information about interpretation and provenance) to be accessed/queried programmatically (a centralized data repository, or a data lake)[2] and also to be readable/understandable by humans.
- Generate metadata from pipelines.
- Consider using a standardized data format, such as the BIDS data format.
- Consider using file formats that can be organized and queried with SQL and easily parsed, such as CSV.
- Use consistent naming conventions across projects and datatypes. Your life will be much easier if you adopt practices of wherever your data is going to end up or whatever the tools expect.

Things to Avoid:

- Thinking that you are going to "organize the data later" - it is best to be built into the pipelines from the start.
- Do not keep metadata separately from data, assuming you will be able to integrate later.

See Also:

- Degree of Data Sharing
- Software/Pipelines

Resources and Tools:

[2] A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.-AWS

- BIDS format for data organization

Relevant Articles:

- The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments
    - https://www.nature.com/articles/sdata201644
- MEG-BIDS, the brain imaging data structure extended to magnetoencephalography
    - https://www.nature.com/articles/sdata2018110

User Story: A researcher was trying to conduct a metaanalysis of three datasets, each with multiple spreadsheet manifests that stored the metadata for different types of data (imaging and cognitive). It was very difficult to collect all the missing pieces (scanner parameters, correction of inconsistent identifiers) and have confidence that the final data set was correct.

Number of Copies

Description: Will all data be accessed through a single cloud instance or are local copies required? Issues include data integrity, versioning, and archival storage.

Value Set Definitions:

- **1:** Single copy + backups, no downloading
- **>1 copy:** Data stored in cloud but must maintain single version of record for legal, ethical, or technical reasons
- **>1 copy+**: Copies can be made and distribution rights can be granted

Value of Use Case Example: *>1 copy+* - Jordan is planning to share the data through NDA or another repository, but they want the data to be reusable (i.e., interoperable) by other platforms.

Discussion of Use Case: Jordan will benefit from planning prospectively for the cost of sharing data and providing multiple copies both through the NDA and other repositories, and will need to ensure sufficient resources.

Best Practices:

- Consider using a repository or cloud service that helps track data versions. For example, Amazon S3 and Dropbox (see other examples below) track data versions. Consider the level of HIPPA compliance for each repository and whether it meets your needs.
- Identify ways to explore data without downloading it (either open summaries, or platforms that support authentication and querying of data) Cloud providers will keep snapshots of version
- Work with cloud vendors to see if they can host valuable data for free
- Consider whether you really need backups given the durability of the Cloud storage (i.e., many cloud providers have very robust redundant storage already, so data may be sufficiently protected through permissions and versioning)
- Make sure any policies on copies will allow necessary computing (permanent vs transient copies).

Things to Avoid:

- Do not store unnecessary copies.

See Also:

- Data Size
- Costs

Resources and Tools:

- Cloud storage services often used by academics: Check whether your institution has an institutional account for these services. Note that these are storage only and do not include compute.
    - Amazon Public AWS Datasets
    - Box
    - Dropbox
    - Google Drive
    - iCloud

- Datalad: Cutting-edge tool for data versioning and provenance tracking, including in cloud storage.

- Institutional Repositories: Many institutions offer cloud-based storage for their constituents

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User story: An IT department took to heart the requirement that data should not be copied. This was implemented in software design and policies in a solution that assumed researchers would only access object storage. Unfortunately, object storage (see https://en.wikipedia.org/wiki/Object_storage for definition) does not support portable operating system interface (POSIX) file system operations that are used by most research software. Furthermore, it precluded copying data to local storage to process it using a compute cluster. It is critical to think about the entire research workflow and make sure that policies do not hinder natural data use.

Privacy

Description: 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 accrue.

Value Set Definitions:

- **Low**: anonymized data with no PHI
- **Medium**: de-identified data-no special access controls
- **High**: Identifiable data with substantial PHI

Value of Use Case Example: *High* - Jordan will assess a range of behavioral measures (individual differences in psychopathology, personality and behavior) out of the scanner, as well as T1, T2, resting state and diffusion imaging. They are also considering obtaining DNA samples and the possibility of doing mobile sensing with participants, collecting information about activity, sleep, geolocation, and potentially even scraping information about app usage, frequency of texts and calls, etc.

Use Case Discussion: Jordan will need to make sure that whatever cloud-based computing or storage options she uses will provide adequate privacy protection for the type of data she is proposing to collect and use. She will also need to ensure prospectively that her consent allows her to share all of the data she wishes to share in the ways she wishes to share it using the platforms she wishes to use, with consideration of what types of data might be required in the future or what uses or combinations of data future users might need.

Best Practices:

- Go to the local IRB and align your consent with what is possible in your local environment and be sure that the practices with regards to policy align with informed consent policies from your institution.
- Consult with the privacy officer on campus.
- Consider "right to be forgotten" policies, e.g., the GDPR, CCPA. **Right to be forgotten** refers to the **right** to have private information about a person be removed from Internet searches and other directories under some circumstances.
- Have an explicit policy on how to deal with privacy issues.
- Understand the level of privacy you must maintain (PHI). The more dimensions you collect, the higher the chance of reidentification.
- Many believe that brain data can potentially be used to identify individuals and ought to be treated as such.
- Ensure that your data collection procedures do not unintentionally embed PHI in the files. For example, make sure that no participant name or birthdate information is used for registering participants at the scanner, as this data can become embedded in dicom files.
- Ideal if all operations are performed in the Cloud.
- Be careful about downloading protected data onto your own systems, and if you do, ensure that they are secure and compliant with relevant privacy rules.
- Stay abreast of the changing regulatory environment.

Things to Avoid:

- Failure to ensure that the consent made with the participants meets your data analysis/sharing needs
- Assume that the policies you adhered to at the start of the project remain relevant throughout the project

- Making copies and using them external to the system

See Also:

- Number of Copies
- Security
- Length of Study
- IRB Experience with Neuroimaging and Cloud Data
- Informed Consent/Data Sharing Coverage

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/
- Ethical Wearables: White paper describing an ethical framework for wearable devices
- Right to be forgotten:The **right to be forgotten** is the **right** to have private information about a person be removed from Internet searches and other directories under some circumstances.)
- ENIGMA Process for protecting privacy

Relevant Articles:

- Privacy Challenges to the Democratization of Brain Data, iScience
    - https://www.sciencedirect.com/science/article/pii/S2589004220303199?via%3Dihub

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Security

Description: Issues include different regulatory policies that would govern compliance and what the archive already has in place; NIH Authority to Operate; FedRAMP and FISMA [moderate, high] requirements/capabilities

Value Set Definitions:

- **Low**: ISO 27001
- **Medium**: FISMA/FedRAMP moderate; NIST 800.53 rev4
- **High**: FISMA/FedRAMP high or data residency & exfiltration controls (in/out)

Value of Use Case Example: *Medium/High* - Platform will have to comply with appropriate security standards which may change depending on what types of data Jordan ends up collecting.

Discussion of Use Case: Not all Cloud environments are HIPAA compliant and even if they are, Jordan is still responsible for setting it up and using it properly. Thus, it is critical that Jordan understand the relevant security issues herself, or partner with the experts on her campus to ensure that she has the appropriate security controls. It will also be important for Jordan to regular review and update security protocols as there may be changes in relevant requirements and policies over time. Again, this can be accomplished in part by continued engagement with experts on campus or elsewhere.

Best Practices:

- Should contact the University IT, Info Security or CIO, and determine what practices are in place for their university
- Consult experts when filling out any requirements rather than doing it yourself
- Document what services are in place for the study
- Consider using an existing cloud platform that provides a secure environment for neuroimaging rather than setting up your own
- Make sure that your local IT understands how to create a secure cloud environment by asking specific questions, e.g., do you have support for creating a FISMA moderate compliant environment in X cloud platform?
- Consult any policies from the funding source around security
- In the absence of clear security policies from either source, return to the data owner for guidance as you should not move forward without clear policies.
- Understand what level of security is required
- Plan for regular compliance review as the project evolves
- Since shared compute resources are built upon a specific OS, and OS's have security implications, it's important to keep your shared cloud compute resources with up-to-date OS security patches.

Things to Avoid:

- Do not have a graduate student be in charge of security.
- Do not assume that you know the answer without consulting with the relevant agencies/resources: security policies and guidelines change all the time; need to confirm that you are right.
- Avoid being shortsighted—establish a framework that is agile.
- Do not assume that the Cloud environment handles all your security concerns (e.g., a malicious web browser extension could send data in the browser anywhere).
- Do not assume that additional security features are free or turned on by default.

- Ensure that the settings you choose don't create avenues for data access that you didn't intend (e.g., external IPs that you didn't plan for).

See Also:

- Privacy
- Data Generation Sources

Resources and Tools:

- Understanding Different Levels/Layers of Security
    - Google Cloud FEDRAMP
    - AWS FEDRAMP
    - AZURE FEDRAMP
    - ISO/IEC standard for information security
    - Terra/FireCloud Security Posture
- Cloudbank: Introduction to Cloud Security
- Security Trainings
    - NIH Information Security And Information Management Training Courses

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User Stories: A researcher accessed data from a European Biobank that requires data to remain in the EU. The researcher used a public cloud to create a virtual machine to access that data but didn't change the default configuration for their compute & storage resources, which in this cloud was for the US region (Iowa). The researchers therefore were in violation of the policy and had to spend much time proving that they were in compliance. There are penalties for being out of compliance. https://healthitsecurity.com/news/uw-medicine-hit-with-lawsuit-for-breach-impacting-974k-patients

A researcher developed a visualization tool for analyzing some controlled-access data. When trying to share that tool with some colleagues, they used another tool to set up a tunnel to the VM inside their network that exposed an external IP to the public. On that VM there were private keys to controlled-access data that any malicious user could easily locate. The researcher also bundled up these tools into a docker image for other researchers to use and deploy within their networks. https://techbeacon.com/security/hackers-love-docker-container-catastrophe-3-2-1

Data Generation Sources

Description: Will all data be generated by the institutions involved in the study or will some come from outside parties, e.g., wearables?

Value Set Definitions:

- **Yes**: At least some data will come from outside sources
- **No**: All data will be generated by the institutions involved in the study

Value of Use Case Example: *Yes* - Wearable device or other type of data stored on some other cloud service (see also existing data); from device to mobile phone app (via Bluetooth) to vendor cloud (via WIFI or cellular network); some devices can communicate with the vendor cloud directly via WiFi or even cellular network (like Apple Watch).

Discussion of Use Case: Jordan will need to ensure that any cloud-based computing and/or storage resources that she chooses to use will both be capable of handling these outside data sources, and will provide the needed level of privacy and security for these data sources.

Best Practices:

- Where possible use searchable and indexable formats, such as:
    - Json
    - Csv
    - xlsx/xls
    - txt
    - Determine whether and how harmonization across different devices and apps for a seemingly common data stream is possible and what work will be required to achieve it.
- Determine data threshold to find signal in the data to be considered statistically significant or meanginful (e.g. power analysis, etc.)?
- Keep the source data. It is important to maintain original data downloaded from device/vendor cloud without any pre-processing so any errors in the processing algorithms can be fixed in the future

Things to Avoid:

- Missing data/metadata (organize your data collection from the beginning to ensure inclusion)
- Using non-standard data formats (e.g. PDF, Word)

See Also:

- Existing Data

Resources and Tools:

- Raw Data Repositories
    - NDA
    - OpenNeuro
    - DANDI
    - DABI
    - NEMAR
- Clinical Trials and Mobile Technologies: https://feasibility-studies.ctti-clinicaltrials.org/resources

- Platforms for remote assessment
    - Radar-base (Remote Assessment of Disease And Relapses): An open source platform to leverage data from wearables and mobile technologies. The main focus of RADAR-base is seamless integration of data streams from various wearables to collect sensor data in real time and store, manage and share the collected data with researchers for retrospective analysis
    - Elektra Academy: Free curriculum, resources, and research for organizations considering the incorporation of remote health monitoring.

- **Relevant Articles:**
    - Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation
        - https://pubmed.ncbi.nlm.nih.gov/27179606/

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Length of Study

Description: Longer studies may necessitate the use of multiple scanner protocols over time or analysis strategies may change. Issues include the complexities of managing a length of study and length of time data required to be stored as well as versioning.

Value Set Definitions:

- **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)

Value of Use Case Example: *Long* - It is also possible that they will want to follow these individuals longitudinally and conduct follow-up imaging or behavioral studies and make these data available as well.

Discussion of Use Case: A potential challenge to the use of cloud computing for long studies is that price and features of cloud computing may change over time, leading to unexpected costs or technical challenges. This is particularly the case if the researcher takes advantage of features that are specific to a particular cloud provider; this can lead to them being locked into that provider.

Best Practices:

- Develop a plan in advance for technical change over time (e.g. what do you do when BIDS 2.0 comes out?)
- Pay off technical debt gradually over time (e.g. regularly check to ensure that any custom software is compatible with the latest version of the language and important libraries).
- Understand the long-term support options for your operating system and analysis software.
- Develop and actively maintain documentation for data and analysis procedures.
- Develop a plan for archiving of unused data to less expensive storage platforms or less expensive tiers of data storage or archiving.

Things to Avoid:

- Waiting until incompatibilities or changes in platforms occur to develop a plan to deal with them

See Also:

- Data Size
- Data Complexity/Scope

Resources and Tools:

- fMRIPrep: Versioning and Long-Term Support

Relevant Articles:

- Cold Storage Data Archives: More Than Just a Bunch of Tapes
    - https://arxiv.org/pdf/1904.04736.pdf

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Costs

Description: How many direct costs for compute, 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 you are generating and the output; cost of complying and using standards; and cost of computing.

Value Set Definitions:

- **Low**: Relative low costs ($10,000 or less)
- **Medium**: Greater than $10,000, but less than $25,000
- **High**: $25,000 or more

Value of Use Case Example: *High* - The amount of data and length of store and compute demands are likely to be considerable.

Discussion of Use Case: Jordan will either need to ensure prospectively that her budget includes sufficient resources for all planned cloud computing and storage costs, or determine whether she will have sufficient budget for the duration of the project and all needs before embarking on the use of cloud computing and storage. If her budget did not initially cover these costs, there may be additional institutional or federal funds available to do so that she could pursue. It will be particularly important for Jordan to plan for costs that will be needed throughout the life of the project, including any longer term archiving or sharing costs.

Best Practices:

- Use available tools to estimate storage and compute costs of the project.
- Plan for ongoing costs.
- Determine whether subsidies are available from the institution, granting agencies or other sources.
- Remember that commercial cloud costs are “Pay as you go.”
- Estimate on the high side—include a cushion

Things to Avoid:

- Paying to store additional copies of data
- Not taking advantage of archival/"cold" storage
- Not accounting for network costs associated with copying data
- Not accounting for access costs or not taking advantage of "requester pays" capabilities
- Forgetting to turn off machine you are not using
- Not accounting for the free-tier of compute resources

See Also:

- Number of Copies
- Length of Study

Resources and Tools:

- How to control cloud costs: Use cases and discussion by Terra
- Cloudbank: Cost estimation

- Cost Calculators
    - https://calculator.s3.amazonaws.com/index.html
    - https://cloud.google.com/products/calculator
    - https://azure.microsoft.com/en-us/pricing/calculator/
    - https://docs.aws.amazon.com/AmazonS3/latest/dev/RequesterPaysBuckets.html
- Life Cycle Decisions for Biomedical Data: National Academy of Sciences, Engineering, and Medicine Report (see in particular Chapter 4 and Appendix E)
- NIH STRIDES: STRIDES is the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative. It "allows NIH to explore the use of cloud environments to streamline NIH data use by partnering with commercial providers. NIH's STRIDES Initiative provides cost-effective access to industry-leading partners to help advance biomedical research. These partnerships enable access to rich datasets and advanced computational infrastructure, tools, and services." Through this program NIH-funded researchers with an active NIH award may take advantage of the STRIDES Initiative for their NIH-funded research projects. The STRIDES Initiative provides:
    - Favorable pricing on computing, storage, and related cloud services
    - Access to training for researchers, data owners, and others to help ensure optimal use of available tools and technologies
    - Access to professional service consultations and technical support from the STRIDES Initiative partners
- Check whether institutions have policies regarding use of the Cloud that can affect costs. For example, academic discounts, waivers for indirects for cloud computing, e.g., see https://itconnect.uw.edu/research/waiver/

Relevant Articles:

- Running Neuroimaging Applications on Amazon Web Services: How, When, and at What Cost?
    - https://docs.aws.amazon.com/AmazonS3/latest/dev/RequesterPaysBuckets.html

User Story: "The committee heard from researchers about their ability to 'experiment' with data-intensive computations, at no additional cost to them, when data resources were hosted by their research institutions. However, when their data were moved to a commercial cloud, the same levels of experimentation resulted in unexpected and large computational bills at the end of the month. Once the cost consequences of their behaviors became transparent (requiring compute bills to be sufficiently granular)—and especially when they were responsible for some or all of those costs—the researchers learned to be more thoughtful and efficient. For example, they began to pilot their analyses before performing them on full data sets. Making people responsible for their costs, helping them understand that their actions generate costs for someone, and providing appropriate training might help reduce resource consumption with more efficient workflows. The information resource platform manager might develop a compelling narrative to alert researchers to storage and computational cost structures and the empowering benefits to researchers of forecasting their costs (Chodacki, 2019). The narrative could properly stress the researchers' larger responsibility to the research community."-*National Academy of Sciences, Engineering, and Medicine Report: Life Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs, Box 3.3, pg 53*

Existing Data

Description: Will the researcher use other datasets in the study (e.g., ABCD has high restrictions on re-release options, but HCP has more open re-release options).

Value Set Definitions:

- **No**: No other datasets will be used
- **Yes**: but the data has no re-release restrictions
- **Yes**: and the data has re-release restrictions

Value of Use Case Example: *Maybe* - Given that Jordan may not be able to obtain enough subjects to test her hypothesis, she should consider whether she can use an existing dataset.

Discussion of Use Case: Jordan will need to determine whether incorporation of other data sets will either necessitate the use of Cloud resources (i.e., that is the only pathway for use), or whether they put any limits on the use of Cloud resources (e.g., privacy or security constraints).

Best Practices:

- Familiarize yourself with sharing, privacy, and security requirements of any existing data sources that you will use.

Things to Avoid:

- Data of questionable quality
- Excessive variance in acquisition/analysis method across data

See Also:

- Data Size
- Data Complexity/Scope

Resources and Tools:

- Raw Data Repositories
    - NDA
    - OpenNeuro
    - DANDI
    - DABI
    - NEMAR
    - Human Connectome Data
    - 1000 Functional Connectomes
- Processed data archives:
    - NeuroVault: NeuroVault is a public repository of unthresholded statistical maps, parcellations, and atlases of the brain. It complements the raw data stores listed above by providing a host for the raw derived (but unthresholded) statistical maps that accompany published (usually static) thresholded example images. This greatly facilitates metaanalysis from the complete brain space in contrast to foci of activation-based metaanalysis that the typically published (thresholded) tabular results support.
    - ReproLake: The ReproLake is the ReproNim publically accessible neuroimaging metadata store. It hosts metadata that facilitates search and discovery of information based upon experiment and acquisition details, analysis results, processing workflows, etc.

- Preprocessed Connectomes Project: : The goal of the Preprocessed Connectomes Project is to systematically preprocess the data from the 1000 Functional Connectomes Project (FCP) and International Neuroimaging Data-sharing Initiative (INDI) and openly share the results. This effort greatly reduces the redundancy of the preprocessing that typically would have to be performed by each investigator accessing a particular dataset.
-

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Software/Pipelines

Description: Does the researcher have to develop their own cloud-compliant tools/analysis pipelines?

Value Set Definitions:

- **Low**: All necessary tools/analysis pipelines are already available and operating in the Cloud computing environment.
- **Medium**: Some, but not all, of the necessary tools/analysis pipelines are already available and operating in the Cloud computing environment.
- **High**: None of the necessary tools/analysis pipelines are already available and operating in the Cloud computing environment and all must be developed by the researcher.

Value of Use Case Example: *Medium* - Jordan can use some existing tools; although, she may need to adapt some of her existing tools and pipelines so they work in the Cloud.

Discussion of Use Case: Efficient use of cloud computing resources requires a different set of skills and knowledge regarding the implementation and orchestration of software execution on large datasets.

Best Practices:

- Use existing, published and established validated pipelines for cloud computing.
- Ensure software & pipelines are designed to be cost-efficient for cloud computing.
- Develop and actively maintain documentation for all cloud procedures.
- Use containerized software packages with specific versioning.

Things to Avoid:

- Using homegrown software written by a postdoc.

See Also:

- Access to Computational Resources and Expertise

Resources and Tools:

- Using containers (Docker/Singularity) in science
    - A set of tutorials from the 2016 Neurohackweek (https://neurohackademy.org/neurohack_year/2016/)
- Docker
    - This is a container system that can be used on one's local machine.
    - It can also be used to develop containerized analysis that can be run on high-performance computing systems using Singularity.
    - DockerHub is a resource that provides a collection of popular computational tools provided within ready to use Docker containerized environments.
- Singularity
    - This is the container system used by most high-performance computing systems
    - SingularityHub and ReproNim/containers are resources that provide a collection of popular computational tools provided within ready to use Singularity containerized environments.
- Pipelines
    - Nipype: A workflow management tool for neuroimaging analysis.

- Neurodocker: A tool that generates custom Dockerfiles and Singularity recipes for neuroimaging and minimizes the size of existing containers.
- fMRIPrep: A BIDS-App that provides robust preprocessing for fMRI data. It has a long-term support (LTS) version that guarantees it to function with consistent results for at least 4 years. It can be run using Docker or Singularity containers providing easy cloud deployment. Outputs are stored according to the BIDS-Derivatives standard for ease of reuse.
- Software for deployment of computations to the Cloud (more here in https://github.com/meirwah/awesome-workflow-engines):
  - Cloudknot: a Python library designed to run existing Python code in AWS Batch
  - Caliban A tool for developing research workflow and notebooks in an isolated Docker environment and submitting those isolated environments to Google Compute Cloud.
  - Batchit: simple jobs submission via command-line for AWS batch
  - NextFlow: A workflow management system that can be used to scale compute workflows in cloud systems (see this example).

Relevant Articles:

- Cloud computing applications for biomedical science
  - https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006144
- Analysis of task-based functional MRI data preprocessed with fMRIPrep
  - https://rdcu.be/ca108

User Story: A researcher transferred their data and existing code base from a local computer to a cloud system so that they could process a large number of datasets. The code is primarily composed of Python programs written by the students in the laboratory. The postdoc running the project arranged to create a number of virtual machine instances on the Cloud computer, and then manually installed the software on each system. They then logged into each machine and manually started the analysis on a different subset of the data. Because the code was not built for parallel processing, and was executed directly on the system rather than using a workflow management system, most of the computational power of the virtual machines sat idle when the jobs were running, leading to much greater computing expenses than if they had been implemented using tools that support parallel execution of workflow components.

Degree of Data Sharing

Description: Will the data be shared with others/made public?

Value Set Definitions:

- **Public**: open license
- **Controlled access**: Shared with public but account or permission required
- **No sharing**: accessible only to researcher

Value of Use Case Example: *Controlled access* - Jordan plans to submit to the NIMH Data Archive.

Discussion of Use Case: Currently the use of cloud computing resources neither facilitates or impedes the sharing of data with the NIMH Data Archive. However, it will facilitate submission to the NIMH Data Archive if Jordan ensures that any formats or processing streams used are optimally compatible with what will be required by the NIMH Data Archive. Jordan should consult with the NIMH Data Archive to determine whether they are implementing any particular transfer pathways that have any particular requirements.

Best Practices:

- Before you set up your data acquisition pipelines, determine if templates or protocols that match your goals already exist in the intended database (e.g., NDA) that you can use to facilitate data harmonization.
- Ensure that you use a consent form broad enough to allow the intended data sharing and use for research questions other than those driving the initial data acquisition.

Things to Avoid:

- Avoid reinventing the wheel and having to engage in time-intensive post-hoc harmonization if it can be avoided.
- Avoid methods that require you to remove PHI post hoc before data sharing.
- Avoid using a consent form that limits data sharing and reuse for novel research questions.

See Also:

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

Resources and Tools:

- NIMH Data Archive (NDA): Cloud-based repository that makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science. A useful cost calculator can be found here: https://nda.nih.gov/contribute_cost_estimation.html
- Human Connectome Project
- NiPype: a Python project that provides a uniform interface to existing neuroimaging software and facilitates interaction between these packages within a single workflow.
- Open Neuro: Open data repository for sharing MRI, MEG, EEG, iEEG, and ECoG data

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Submission to Third-Party Repository

Description: Will the data be deposited in a third-party repository? What are the requirements of the repository?

Value Set Definitions:
- **Yes**
- **No**

Value of Use Case Example: *Yes* - Jordan has already decided to use the NIMH Data Archive as her third-party repository.

Discussion of Use Case: Jordan has chosen to use the NIMH Data Archive because she wants to analyze her newly acquired data along with data already in NDA. Sharing data in NDA will also bring her into compliance with the data sharing expectations should this research be funded by NIMH (NOT-MH-19-033). Jordan has also noted that NIMH expects a certain set of clinical measures to be collected for all of their funded studies (NOT-MH-20-067), so she is planning on adding those measures when she collects data.

One additional factor in Jordan's decision to use NDA is the "study" feature in that repository. NDA creates "collections" of data that are related to a single research program. Generally, collections are associated with a single grant award, either from NIMH, one of the other NIH Institutes and Centers using the repository, or from non-federal funding groups. The study feature will allow Jordan to create a data set containing the exact data set used in a publication. The data in a study can all come from her collection or can come from more than one NDA collection. Studies allow others to run different data analysis pipelines on the exact data Jordan has already published on.

All of the data in NDA is stored in a commercial cloud provider with appropriate security. Data access is restricted qualified investigators as determined by a Data Access Committee (DAC) convened by NIMH. The DAC also checks whether users requesting access are planning on using the data in a way that is consistent with the informed consent. In addition to storing the data, NDA provides limited cloud computational credits. That feature of the archive should be useful to a user like Jordan.

Best Practices:
- Choose a third-party data archive that is appropriate for the data being measured. Things to think about are data sharing expectations from the funder, whether similar data are already available in the archive, the funding stability of the archive, and whether the data will have appropriate security and access control.
- Choose a third-party data archive that makes the data as widely available as possible. If Jordan's data were less sensitive, a repository like OpenNeuro would make the data more easily available.

Things to Avoid:
- Hosting data on your own website

See Also:
- Informed Consent/Data Sharing Coverage
- Degree of Data Sharing

Resources and Tools:
- NIMH Data Archive (NDA): Cloud-based repository that makes available human subjects data collected from hundreds of research projects across many scientific domains. NDA provides infrastructure for sharing research data, tools, methods, and analyses enabling collaborative science.
- Open Neuro: Open data repository for sharing MRI, MEG, EEG, iEEG, and ECoG data

-

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

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* - Jordan's institution does not have a history of IRB experience with neuroimaging and cloud storage/computing.

Discussion of Use Case: Jordan 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 in order 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:

- 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/)

Relevant Articles:

Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Informed Consent/Data Sharing Coverage

Description: 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.

Value Set Definitions:

- **Low**: The informed consent does not allow any data sharing.
- **Medium**: The informed consent allows data sharing under restricted conditions.
- **High**: The informed consent allows broad and open data sharing.

Value of Use Case Example: *NA* - Jordan is just starting her project and can work prospectively to ensure that the consent allows the degree of sharing that she wishes for, which should be "High."

Discussion of Use Case: Whether or not the informed consent allows data sharing likely does not have an impact on the use of cloud-based resources for analysis and storage of the data if Jordan does not wish to share the data. However, given that Jordan would like to be able to share data, it is critical that they ensure that the consent allows for broad data sharing.

Best Practices:

- If you are still engaged in data collection, consider modifying the consent form to allow data sharing without restrictions, though this will not apply to already consented individuals
- Identify databases for sharing that can meet your necessary restrictions-If you will be sharing directly from your own cloud-based platform, develop a data use agreement that outlines the restrictions

Things to Avoid:

- Avoid using a consent form with more restrictive sharing and re-use than needed for your project

See Also:

- IRB Experience with Neuroimaging and Cloud Data
- 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/)

Relevant Articles:

- Ethical aspects of data sharing and research participant protections
    - https://pubmed.ncbi.nlm.nih.gov/29481107/
- The ethics of secondary data analysis: Considering the application of Belmont principles to the sharing of neuroimaging data
    - https://pubmed.ncbi.nlm.nih.gov/23466937/

User story: Suggestions are welcomed via the open discussion board on the INCF Training Space located here: https://training.incf.org/cloud-based-computer-matrix.

Acknowledgements: The committee would like to thank Dr. Ariel Rokem and Dimitri Papodapoulos for helpful comments and suggestions. We thank Dr. Greg Farber for his contributions as part of the Neuroscience Action Collaborative Working Group and Ms. Sheena Posey Norris for her excellent support.