TO THE CLOUD! A GRASSROOTS PROPOSAL TO ACCELERATE BRAIN SCIENCE DISCOVERY

- Joshua T. Vogelstein; Dept of Biomedical Engineering, Center for Imaging Science, Kavli Neuroscience Discovery Institute, Institute for Computational Medicine, Institute for Data Intensive Engineering and Sciences; Johns Hopkins University
- Brett Mensh, Optimize Science, Janelia Research Campus, UCSF Kavli Institute for Fundamental Neuroscience
- Michael Hausser, Department of Physiology, University College London
- Nelson Spruston, Janelia Research Campus, Howard Hughes Medical Institute
- Alan Evans, Montreal Neurological Institute, McGill University,
- Konrad Kording, Northwestern University
- Katrin Amunts, Institute for Neuroscience and Medicine, INM-1, Research Centre Juelich, Germany, C. and O. Vogt Institute for Brain Research, University Hospital Duesseldorf, University Duesseldorf, Germany and The Human Brain Project (HBP)
- Christoph Ebell, Human Brain Project, EPFL, Geneva, Switzerland
- Jeff Muller, Human Brain Project, EPFL, Geneva, Switzerland
- Martin Telefont, Human Brain Project, EPFL, Geneva, Switzerland
- Sean Hill, Blue Brain Project, EPFL, Campus Biotech, Geneva, Switzerland
- Sandhya P. Koushika, Department of Biological Sciences, Tata Institute of Fundamental Research, Mumbai, India
- Corrado Cali, Biological and Environmental Science and Engineering, KAUST, Thuwal, Saudi Arabia
- Pedro Antonio Valdés-Sosa, Cuban Neuroscience Center/University of Electronic Science and Technology of China
- Peter Littlewood, Argonne National Laboratory
- Christof Koch, Allen Institute for Brain Science
- Stephan Saalfeld, Janelia Research Campus
- Adam Kepecs, Cold Spring Harbor Laboratory
- Hanchuan Peng, Allen Institute for Brain Science
- Gregory Kiar, Dept of Biomedical Engineering, Center for Imaging Science, Kavli Neuroscience Discovery Institute, Institute for Computational Medicine, Johns Hopkins University
- Mu-Ming Poo, Institute of Neuroscience, CAS Center for Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China
- Jean-Baptiste Poline, Henry H. Wheeler Jr. Brain Imaging Center, Helen Wills Neuroscience Institute, University of California, Berkeley
- Michael P. Milham, Center for the Developing Brain, Child Mind Institute. Nathan S. Kline Institute for Psychiatric Research
- Alyssa Picchini Schaffer, Simons Collaboration on the Global Brain, Simons Foundation
- Rafi Gidron, Israel Brain Technologies
- Hideyuki Okano, Department of Physiology, Keio University School of Medicine, Tokyo 160-8582, Japan. RIKEN Brain Science Institute, Laboratory for Marmoset Neural Architecture, Wako-shi, Saitama, Japan
- Vince D Calhoun, The Mind Research Network & The Dept. of Electrical and Computer Engineering, The University of New Mexico
- Miyoung Chun, The Kavli Foundation
- Dean M. Kleissas, Johns Hopkins University Applied Physics Laboratory
- R. Jacob Vogelstein, Intelligence Advanced Research Projects Activity (IARPA)
- Eric Perlman; Center for Imaging Science, Johns Hopkins University
- Randal Burns; Dept Computer Science, Institute for Data Intensive Engineering and Sciences, Johns Hopkins University
- Richard Huganir; Department of Neuroscience, Kavli Neuroscience Discovery Institute The Johns Hopkins University School of Medicine
- Michael Miller; Dept of Biomedical Engineering, Center for Imaging Science, Kavli Neuroscience Discovery Institute, Institute for Computational Medicine, Johns Hopkins University

ABSTRACT

The revolution in neuroscientific data acquisition is creating an analysis challenge. We propose leveraging cloud-computing technologies to enable large-scale neurodata storing, exploring, analyzing, and modeling. This utility will empower scientists globally to generate and test theories of brain function and dysfunction.

INTRODUCTION

Technological advances from all around the globe [1] are allowing neuroscientists to collect more precise, complex, varied, and extensive data than ever before [2]. How can we maximally accelerate our collective ability to extract meaning from such data? To answer this question, the United States Congress commissioned the National Science Foundation (NSF) to "convene government representatives, neuroscience researchers, private entities, and non-profit institutions" (https://www.congress.gov/congressional-report/113th-congress/house-report/448). The NSF funded two events. The first was a workshop of over 75 individuals from 12 countries and 5 continents that was broadcast live over the internet. Each person was invited to bring a single big idea—one that could have maximal impact, while being both feasible, given existing resources, and universally inclusive. Four ideas emerged as grand challenges for global brain science [3]. A second event was organized to discuss these ideas with a larger (425 participants) and more diverse community, which will be the subject of another article (Bargmann et al, in prep). The goal of this NeuroView is to describe one of the four grand challenges and propose a strategy to overcome it, in order to gather feedback from the larger community. The authors are participants in the first conference who volunteered to hash out these ideas via emails, online documents, conference calls, and in-person visits.

The kernel of the idea is based on a view of the scientific process as an "upward spiral": a collective effort where each new experiment yields data, upon which analysis is performed, leading to new or refined models, which suggest novel experiments (see Figure 1). Historically, the process of data analysis has been kept relatively simple by the small scale of data acquired. But recent advances in experimental technology, such as serial electron microscopy [4], light sheet microscopy [5], and models of the whole human brain at the microscopic level [6] have made data analysis significantly more challenging. While experimental neuroscience is enabling collection of ever larger and more varied data sets, information technology is undergoing a revolution of its own. Commercial development of artificial intelligence and cloud computing innovations are changing the computational landscape [7]. Computing is moving toward "cloudification," a "software as a service" model, in which locally installed software programs are replaced by Web apps. These forces create a massive opportunity to develop new computational technologies that complement advances in data collection in order to accelerate and democratize model building, hypothesis testing, and model refinement.

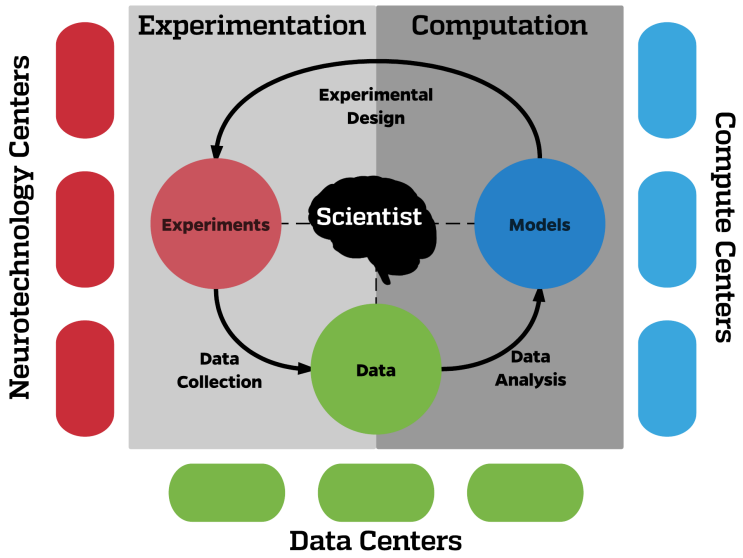

Figure 1: The upward spiral of science.

What Would Change If We Capitalize on this Opportunity?

Consider sending a letter, watching a movie at home, or obtaining reference information. Ten to twenty years ago, to send a letter, we purchased paper, stamps, and envelopes; to watch a movie at home, we rented or purchased a VHS or DVD; to obtain reference information, we bought an encyclopedia and obtained yearly revisions. Today, each of those options is still available and indeed preferred in certain circumstances. However, Web options exist for each activity as well. In each case, we have privacy concerns, bandwidth concerns, and financial concerns. Nonetheless, for many of our daily practices we use these cyber solutions, sometimes putting our most private information in the cloud. The everyday practice of brain science is just beginning to benefit from similar technology development.

Other scientific disciplines have already navigated similar waters with remarkable success. For example, the Sloan Digital Sky Survey (SDSS) changed the daily practice of astronomers and cosmologists [8]. They still have the option to wait six months for telescope time, analyze their data locally on machines they own and maintain, and publish a summary of the results (and many do). Yet there are more accounts in SDSS than there are professional cosmologists. Astronomers can now log in to SDSS, find previously published data, run database queries (a skill they typically did not have prior to SDSS), and publish the queries and results. Similarly, molecular geneticists historically sequenced their own data (using machines that they owned and maintained), analyzed it locally, and published the results. Now, they can outsource the sequencing to avoid owning and maintaining the machines, upload the sequences to a national or international database, quantitatively compare their sequences to previously published

sequences, and then publish their findings. The success of these efforts is evident from the cultural shift of daily practices by many, if not most participants in each field. Both fields resolved issues of data privacy, data ownership, governance, and financial concerns, providing a proof of principle that other scientific disciplines can do the same.

In neuroscience, many of our scientific practices remain based on pre-internet methods. A scientist designs an experiment, collects data, stores it locally, keeps metadata in his head or in some custom spreadsheet, analyzes it using software that hs buys and installs on local computers that he updates regularly, and publishes a summary of the results. We predict that another strategy will be superior for many situations: as the scientist collects data, it gets stored privately or publicly in the cloud, she then selects analyses to occur automatically, having the flexibility to pull from a variety of previously published analyses, and finally publishes entire "digital experiments", containing (some of) the data and the entire analysis pipeline.

What Are the Primary Goals?

We see two key goals that, if achieved, would leverage advances in computing to accelerate brain sciences. *The first goal is to make reproducibility and extensibility of science as easy as possible*, even for small amounts of data or simple data. The current practices of private data storage and siloed analyses make reproducing an analytic result tedious at best and impossible at worst. The steps can include requesting the data, identifying the formats and organization, requesting the code, deciding which functions to run and how, getting all necessary dependencies installed, making sure to use the same software versions, and accessing the same computational hardware. Solutions now exist to mitigate each of these challenges, though they are relatively disparate and unconnected. Data can be uploaded to data repositories (e.g., https://figshare.com/), data standards have been proposed for many domains of brain science (e.g., http://bids.neuroimaging.io/, http://www.nwb.org/), code can be stored in publicly accessible repositories (e.g., https://github.com/), interactive tutorials can be provided (e.g., using http://jupyter.org/), all necessary software dependencies can be easily packaged together (e.g., using https://www.docker.com/), and can be run "in the cloud" (e.g., using http://mybinder.org/) on commercial service providers (e.g., on https://aws.amazon.com/ec2/ or https://cloud.google.com/). Nonetheless, given some new data, it is not obvious where to find reference algorithms or how to connect them to the data. Similarly, given a new model, it is not clear how to find reference data, figure out which standard it is using and then fit it, and determine if others have done the same to allow us to compare and assess the results. In either case, once the data are processed, it remains difficult to keep track of the resulting data derivatives and which version of which code resulted in which outputs. So, although many of the pieces are in place, there is still no unified "glue" that makes everything work together seamlessly. Moreover, each of the above-mentioned tools can be used by some brain scientists, but most tools are designed for data scientists, so the learning curve can be incredibly steep. Ideally there would be a place where brain scientists could find all relevant analyses and data, run each analysis on each dataset, and see a leaderboard comparing performances, without writing any lines of code. Cloud-based solutions simplify reproducibility

and extensibility by essentially eliminating activation energy and extraneous sources of analytic variability.

The second goal is to enable such a system to work with "big data" (i.e., data too large to fit on a workstation). Data are scaling in many domains in brain science, either because individual experiments are large (as in calcium imaging and whole-brain CLARITY imaging), or there are thousands of subjects with gigabytes of data each (as in large-scale human brain imaging projects), or there are millions of time points (as in wearable sensor data). Regardless of source and modality, if it is "medium data" (meaning too large to fit in memory, but small enough to fit on your computer), tasks as simple as visualizing, rotating, and opening the data are challenging using standard tools such as MATLAB, Python, or ImageJ. For big data, the challenges are even larger, as questions of how to store, compress, manage, and archive the data exceed the computational capabilities and resources of most experimental labs. Cloud-based solutions simplify big data analysis due to their inherently scalable nature.

WHAT'S THE BIG IDEA?

We are proposing to design, build, and deploy an instance of "cloud neuroscience", meaning that the data, the code, and the analytic results all live in the cloud together. Cloud neuroscience can be thought of as an operating system, a set of programs that run on it, a file system that stores the data, and the data itself, all designed to run in a scalable fashion and to be accessible from anywhere.

WHAT ARE THE DESIGN CRITERIA?

First and foremost, the design and construction should be organic, grassroots, and open source, to ensure that it remains intimately connected to the needs of all scientific citizens. Over 100,000 people attend annual conferences, including neuroscience, psychology, psychiatry, and neurology. This is a massive human capital resource, so the system should enable contributions from any of them, regardless of background or resources. Thus the system needs to support data and workflows of all kinds, regardless of modality, complexity, or scale—including raw data, derived data, and metadata. Doing so would also further democratize brain sciences, opening the door to the additional 3.5 billion people with mobile broadband access who could contribute if given the opportunity. Encouraging and supporting such involvement motivates an emphasis on ethical standards and cultural sensitivities. Moreover, millions of hours and billions of dollars have been spent developing brain science resources, including vast quantities of data, algorithms, and models. The system should build upon such work. Because different people have different preferences, access controls should be flexible enough to satisfy everyone's needs. For resources that are open, reproducing and extending prior work should be "turn-key", allowing researchers to "swap-in" different datasets or algorithms as desired. Industry is making tremendous headway in this regard, including digital notebooks to keep track of all analyses, software containers to ease the burden of installing and configuring software, and Web-services that dynamically provide computational resources as needed. To the extent possible, we should

leverage these resources and engage with non profit, institutional, and corporate partners to express our domain-specific needs. The design should be highly adaptive, to capitalize on rapid advances from within and outside brain sciences, and of course open-source with permissive licenses. And the entire system should be able to run not just in a single commercial cloud, but also on other clouds, national resources, institutional clusters, local workstations and laptops, to enable maximal portability and utility. Perhaps most importantly, the system should be universally useful, helping to answer the grand challenges of brain science while facilitating much greater participation in the scientific process.

The motivation underlying this endeavor is to accelerate the scientific process by improving the experience of doing brain science. Thus the community can determine the worst pain points in our process and design solutions around them. For example, if looking at data is the largest bottleneck, then one could use a cloud-based visualization app (like Google Maps, CATMAID, or NeuroDataViz). On the other hand, if the largest bottleneck is getting data into a common format before running analyses, then you would benefit from having all the data stored in a format with a standardized application programming interface (API) so every dataset can be accessed in the same way. In other words, it is time for the scientific community to prioritize the user experience to focus the subsequent software development.

How Might we Achieve It?

In this section we propose a potential design of the constituent components that could comprise an instance of cloud neuroscience (see below Figure). The required elements can be divided into into five categories: Data, Infrastructure, Apps, Algorithms, and Education. The goal of breaking down the problem this way is to ensure that all brain scientists, professional and citizen alike, can contribute to and benefit from the system. Crucial to success will be tight integration across components, each of which is described in some detail below. Some brain scientists are able to span the full range from design to analysis including: design and run experiments, analyze data, make discoveries, and even write articles. Such polymaths can seamlessly alternate between different roles. Others might be highly skilled in software engineering, but not data collection. To ensure that all brain scientists can contribute to this effort, we have organized types of activities according to the “role” of the individual performing those activities. These roles are not meant to be prescriptive, rather, they serve to help guide scientists to the kinds of contributions they could make. (see Box I for detailed description of the roles).

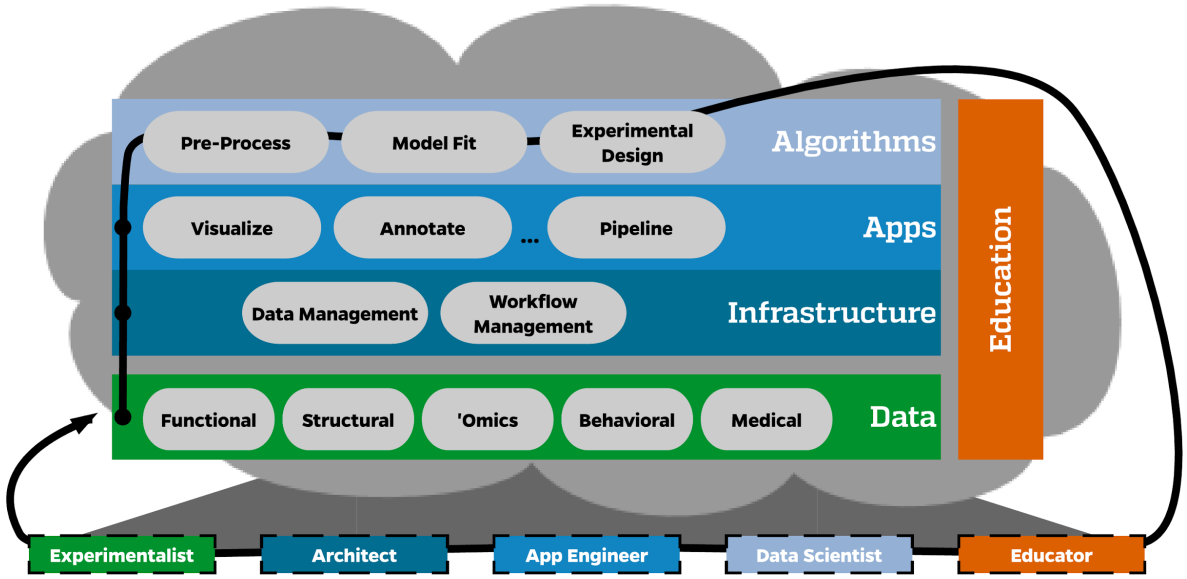

Figure 2: Schematic of the five proposed components. An individual can adopt any or all of the five roles (dashed rectangles). For each component, the content is generated by individuals in one of the five roles who interacts with the cloud services.

Data

The data component is intended to mitigate difficulties with storing and accessing data, regardless of the modality, scale, or complexity of the data. Anybody would be able to upload raw data, derived data, and metadata as it flows off the sensors and dynamically control access. Functionality would build on and incorporate existing brain-science data repositories [9–13], as well as more general services (e.g., FigShare). Therefore, the technical challenges for small and large data storage and access, for the most part, already have reasonable solutions for many data types. The remaining challenges are to further lower the barrier to entry, making data upload and access easier, especially for multi-terabyte datasets. Data contributions will be able to come from anyone and could be stored in a variety of accessible places to minimize transfer cost and time. Access controls would enable scalable sharing with minimal effort. Storage costs would be the responsibility of the data provider if the data are private; if public, others could financially contribute. In either case, economies of scale would reduce storage costs, and we would work with commercial clouds and national infrastructures to offset costs to the extent possible. The data-storage formats would allow visualization and analysis at scale.

Data contribution would be desirable and possible from any lab, regardless of their financial resources or locations. For example, some methods are relatively inexpensive, such as EEG, fNIRS, and wearable technologies. Moreover, certain important subpopulations are better represented in less wealthy countries, enabling unique contributions from those places. If the same measures are included in more expensive projects, analysis bridges could be established between the datasets. This would enhance translational research at a global scale.. These

factors would lead to important collaborations in which less wealthy countries could influence the content and usefulness of this effort [14].

Data types would include raw, derived, and metadata (see Box II for additional details). Raw data includes data from any kind of experiment, including anatomical, physiological, behavioral, (epi-) genetic, and medical data. Every experiment will be given a unique data identifier. Medical data will be given special attention to ensure compliance with national guidelines for patient privacy. Each data type will yield a wide diversity of derived data, including summary statistics, matrices, networks, shapes, and more. Associated with each entry is a collection of metadata, including a community-driven controlled vocabulary, as well as custom ad hoc fields. Metadata on the derived data will include detailed provenance history. The system would be seeded with existing reference datasets spanning spatial, temporal, and phylogenetic scales, including data from the Human Brain Project, the Human Connectome Project, the Allen Institute for Brain Science's data portal, IARPA's MICrONs program, and more.

INFRASTRUCTURE

The Infrastructure component is intended to mitigate difficulties in finding data or tools, linking them together, installing software, managing computers, and reproducing and extending results. When the infrastructure is operational, much of the scientific process can be conducted from a tablet or smartphone, replacing the need to buy and maintain high-power computers or keep software up-to-date. The infrastructure is essentially the operating system upon which all the services would run, akin to NeuroDebian [15], but designed specifically for the cloud. This virtual operating system will run in the commercial cloud, on institutional resources, national centers, or local workstations, regardless of hardware configuration (e.g., Mac, Windows, Linux, etc.) The software could be designed and written by a small and distributed team of architects to facilitate design decisions considering diverse use cases.

The Infrastructure could be composed of two core sub-components. First, a *Data Management System* would store and organize all the data. This could include managing access, assigning digital object identifiers (DOIs), and supporting common data formats, and would be easily extensible to new or custom formats. Data could also be compressed with or without loss, as desired by the contributor. Technically, data would be stored in a set of databases optimized for different brain science use cases. Second, a *Workflow Management System* would store and organize analyses, leveraging existing Web services such as Github and continuous integration to the extent possible. This would enable "digital experiments", including all stages of data processing. Crucially, such experiments could be done on different hardware platforms, applied to different data (by merely swapping the DOI), or use different algorithms (a similarly simple modification). All infrastructure services would have easy to use APIs to maximize utility and extensibility.

APPS

The Apps component is intended to mitigate difficulties in maintaining software versions, paying for software, and finding tools appropriate to run on data. Apps are the programs that run on the

system, akin to tools like Dropbox (to upload/download), Google Maps (to visualize), Pubmed Central (to search for information), BLAST (to compare your data with other data), and pipelines (to process your data). Apps can be developed by anybody with minimal programming skills, due to the careful design of the APIs in the infrastructure. A specification would be formalized and quality standards agreed upon by the community of users to publish apps in the open app marketplace. Different apps would be designed for users with different backgrounds, roles, and goals. For example, apps targeted at people in the *experimentalist* role could include features to enable uploading, downloading, and managing access without having to learn the APIs. On the other hand, apps targeted at people in the *data analysis* role could include pre-processing data, fitting models, testing hypotheses, plotting results, and running digital experiments. General purpose apps would include tools to visualize, manipulate, and manually annotate data. These general-purpose apps enable a much broader community of users to participate in the scientific process, including those without extensive technical training or financial resources.

ALGORITHMS

The Algorithms component is intended to mitigate difficulties in analyzing data with increasing scale or complexity. Recent advances in artificial intelligence, including distributed machine learning libraries and deep learning, could be leveraged here. Algorithms operate on simulated, measured, or derived data to produce transformed representations or summary statistics of the data. Algorithms can be written by anybody with minimal data-science skills, including many current brain scientists, without knowledge of this proposed system (unlike Apps). Algorithms are essentially “wrapped” in Apps to run, and therefore inherit many of the conveniences of the system. We partition algorithms into three different types. *Scalable data processing algorithms* can be applied to a wide variety of kinds of data. These will be easily daisy-chained together to obtain pipelines, which can similarly be adapted to apply different algorithms or data. Because algorithms will be applied more generally to less familiar data, or less familiar algorithms will be applied to familiar data, *quality assessment* will be particularly important. This would include both qualitative dashboards providing figures and quantitative metrics to evaluate and compare performances along different metrics. Finally, to optimize resources and avoid duplicating efforts across labs, experiments will need to be useful for a large number of people. *Experimental design* will therefore be a key algorithmic component as well.

EDUCATION

Just like there is a learning curve when switching from Windows to Mac, so too switching from current practices to this system will involve a learning curve. Therefore, the success of this endeavor will depend on extensive educational material, including documentation, tutorials, online courses, hackathons, workshops, and summer courses. All the content will be designed to complement existing educational resources, such as Coursera courses. The variety of educational resources would reflect the backgrounds and skills of the user and contributor communities, with the goal of universal access. Because of this variety, community-driven cultural sensitivity guidelines would be posted for all contribution types.

Discussion

Here we describe an immediately actionable grassroots proposal to marry recent advances in neurodata acquisition with scalable cloud-computing in order to accelerate the process of discovery by scientists independent of how well-resourced they are. There are several mechanisms by which Cloud Neuroscience may yield benefits. Global collaborations may become much simpler and therefore more prevalent. Open science may be facilitated, and the barriers and benefits to conducting open science may become more transparent by virtue of the design. Many models can be tested on the same dataset, and individual models can be subjected to greater diversity of data-based reality-checks. In the near-term, any effort that generates reference data of interest to a large segment of the community can benefit from Cloud Neuroscience. One example is the upcoming ~10 petabytes from the IARPA MICrONS program.

Several potential criticisms are worth addressing, and many details need to be fleshed out. Privacy concerns for human data will require careful additional thinking so that best practices of anonymization and security can implemented—precedent is provided by ongoing large research initiatives (e.g., [16–18]). A viable financial model will be required. Potential partners include national laboratories that could contribute computing and storage resources, or companies interested in providing cloud-based Web-services for specific scientific subdomains. Return on investment must be considered. Cosmology, molecular genetics, and plant biology (see http://www.cyverse.org/) are existence proofs that when designed well, such resources can yield dramatic and positive impact on the field. Other cloud-computing neuroscience efforts that focus on the human brain are already underway, such as CBRAIN [19] (human brain imaging) and the Human Brain Project (human brain modeling). Such efforts are important; the proposed project has been designed to leverage the developments from those projects, and extend them to address a greater diversity of brain science questions, species, data modalities, and functionalities.

The above plans and challenges suggest immediately actionable next steps. A field engineer has been appointed to develop a survey to determine which existing resources are most useful (pooling information from places like https://github.com/ and https://www.nitrc.org/) and what new resources would be most useful. A software engineer has agreed to contribute significant effort towards building a “Neuroscience as a Service” framework (the virtual operating system and apps described above) based upon existing related services. They will begin formalizing minimal specifications for all resources. We have also obtained private seed funding to hire an additional senior software engineer. To gather community feedback, we will be monitoring https://neurostars.org/ for any posts that contain the tag “neurocloud”. Next, sustainable governance, funding, and advisory models will be devised.

Pablo Picasso famously quipped, “Every child is an artist. The problem is how to remain an artist once we grow up.” We believe that every child—and adult—is a scientist too. The enormous potential of scientists who currently lack access to the world’s best brain science resources can be realized. In the last centuries, we have democratized nations. In this century, we can democratize science.

REFERENCES

1. Grillner S, Ip N, Koch C, Koroshetz W, Okano H, Polachek M, et al. Worldwide initiatives to advance brain research. Nat Neurosci. 2016;19: 1118–1122. doi:10.1038/nn.4371

2. Sejnowski TJ, Churchland PS, Movshon JA. Putting big data to good use in neuroscience. Nat Neurosci. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014;17: 1440–1441. doi:10.1038/nn.3839

3. Vogelstein JT, Amunts K, Andreou A, Angelaki D, Ascoli G, Bargmann C, et al. Grand Challenges for Global Brain Sciences [Internet]. 2016. Available: http://arxiv.org/abs/1608.06548

4. Denk W, Horstmann H. Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. PLoS Biol. 2004;2: e329. doi:10.1371/journal.pbio.0020329

5. Weber M, Mickoleit M, Huisken J. Light sheet microscopy. Methods Cell Biol. 2014;123: 193–215. doi:10.1016/B978-0-12-420138-5.00011-2

6. Amunts K, Lepage C, Borgeat L, Mohlberg H, Dickscheid T, Rousseau M-E, et al. BigBrain: An Ultrahigh-Resolution 3D Human Brain Model. Science. 2013;340: 1472–1475. doi:10.1126/science.1235381

7. The future of computing. In: The Economist [Internet]. 2016 [cited 12 Oct 2016]. Available: http://www.economist.com/news/leaders/21694528-era-predictable-improvement-computer-hardware-ending-what-comes-next-future

8. Kent SM. Sloan digital sky survey. Science with astronomical near-infrared sky surveys. Springer; 1994; Available: http://link.springer.com/chapter/10.1007/978-94-011-0946-8_6

9. Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, et al. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 2013;7: 12. doi:10.3389/fninf.2013.00012

10. Crawford KL, Neu SC, Toga AW. The Image and Data Archive at the Laboratory of Neuro Imaging. Neuroimage. 2016;124: 1080–1083. doi:10.1016/j.neuroimage.2015.04.067

11. Teeters JL, Harris KD, Millman KJ, Olshausen BA, Sommer FT. Data sharing for computational neuroscience. Neuroinformatics. 2008;6: 47–55. doi:10.1007/s12021-008-9009-y

12. Ascoli GA, Donohue DE, Halavi M. NeuroMorpho.Org: A Central Resource for Neuronal Morphologies. J Neurosci. 2007;27: 9247–9251. doi:10.1523/JNEUROSCI.2055-07.2007

13. Burns R, Lillaney K, Berger D, Deisseroth K, Kazhdan M, Szalay AS, et al. The Open Connectome Project Data Cluster: Scalable Analysis and Vision for High-Throughput Neuroscience. Scientific and Statistical Database Management. 2013. Available: http://arxiv.org/abs/1306.3543

14. Neuroinformatics Collaboratory. In: Neuroinformatics Collaboratory [Internet]. [cited 13 Oct 2016]. Available: http://www.neuroinformatics-collaboratory.org/

15. Halchenko YO, Hanke M. Open is Not Enough. Let's Take the Next Step: An Integrated, Community-Driven Computing Platform for Neuroscience. Front Neuroinform. 2012;6: 22. doi:10.3389/fninf.2012.00022

16. Sarwate AD, Plis SM, Turner JA, Arbabshirani MR, Calhoun VD. Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. Front Neuroinform. 2014;8: 35. doi:10.3389/fninf.2014.00035

17. Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J Magn Reson Imaging. 2008;27: 685–691. doi:10.1002/jmri.21049

18. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17: 124–130. doi:10.1136/jamia.2009.000893

19. Das S, Glatard T, MacIntyre LC, Madjar C, Rogers C, Rousseau M-E, et al. The MNI data-sharing and processing ecosystem. Neuroimage. 2016;124: 1188–1195. doi:10.1016/j.neuroimage.2015.08.076

BOXES

BOX 1: ROLES

We enumerate six different roles for participants. Note that these are not characterizing individuals, rather, roles that any individual play. Roles differ in their degree of interest and expertise in various aspects of the scientific process, all of which are important.

Experimentalist A person in this role is acquiring data. This includes activities such as recruiting subjects and specifying inclusion guidelines (for human studies), experimental setup, subject care, data acquisition, as well as some aspects of data management and quality control. In this role, a person has extensive knowledge of the experiment details, though computational acumen can be quite modest.

Architect A person in this role is developing the Neuroscience as a Service component. She will be a professional software engineer working collaboratively on open source repositories, possibly co-localized.

App Engineer A person in this role is writing apps. These apps might wrap algorithms written by themselves or others. This person knows and implements best practices of software development for science, including proper documentation.

Data Scientist A person in this role is writing and running algorithms. These algorithms might serve any step of the scientific process. Data scientists have a wide variety of computational backgrounds, including engineering, physics, mathematics, statistics, and computer scientists.

Scientific User A person in this role is using tools to analyze and understand the data. This can take many forms, ranging from looking at images and figures generated directly from the data acquisition system, to fitting statistical models and combining multiple disparate datasets. In this role, computational acumen is not required. Familiarity with the data, experimental details, etc., can also be widely varying.

Educator A person in this role is either creating or presenting educational content, including documentation, tutorials, and Massive Online Open Courses, as well as running workshops, hackathons, and summer courses.

BOX II: TYPES OF BRAIN SCIENCE DATA

Functional data is fundamentally temporal and dynamic. Whether it is univariate or multivariate, the standard operations to apply include zooming in time, subsampling, smoothing, and converting to other domains such as the fourier domain. Functional data also have a spatial domain, which links them to Structural data. The subdivision between functional and structural data may be, for some data, ambiguous.

Structural data is fundamentally spatial in nature, include two-dimensional (2D) images, 3D volumes, and 4D and 5D hypervolumes for multispectral and/or time-varying data (spatiotemporal data, such as fMRI and calcium imaging, are both structural and functional). This can include structural images, as well as sparse fluorescent images, gene expression maps, etc. Standard operations for these data include different compression algorithms, downloads of volumes of arbitrary sizes and shapes, maximum projections, averages, and more.

'Omics data is sequential and categorical, including the genome, epigenome, metabolome, and microbiome. Standard queries for genetic data include sequence compression, alignment, and comparisons. 'Omics data may also have a spatial domain (e.g., gene expression data).

Behavioral data can be of several different types. For example, behavior can be captured via video capture (e.g., behavioral observation of children during play), time series of task events during physiological measurements, questionnaires (e.g., symptom checklists), performance testing instruments (e.g., such as the NIH Toolbox), and other devices (e.g., actigraphy, voice recorders). Each data has unique qualities and, therefore, functionality.

Medical data includes all electronic health data, including semi-structured text. It is amongst the most challenging of data types to aggregate, as until recently, the vast majority of the field has relied on paper charts or poorly structured electronic health record (EHR) systems. Fortunately, regulatory and funding agencies are incentivizing the widespread use of EHR's, as well as common data elements that are more amenable to data aggregation for the purposes of discovery science (e.g., the eMerge Network). Additionally, informatics frameworks are being developed to safely link disparate EHR data (e.g., https://www.i2b2.org/), and calls for the creation of open API's are gaining attention.