PROJECT SUMMARY

Overview:

The idea for The International Brain Station (TIBS) is based on a view of the scientific process as an "upward spiral": a collective effort where each new experiment yields data, which is analyzed, leading to new or refined models that then suggest novel experiments. Historically, data analysis has been kept relatively simple by the small scale of data acquired, but recent advances in experimental techniques have made data analysis much more challenging. Analysis now needs to be performed on multiple, large, heterogeneous data sets with distributed computing. In addition, the information technology (IT) revolution in artificial intelligence and cloud computing is leading towards a "software as a service" model, in which locally installed programs are replaced by Web apps. These forces create a massive opportunity to develop new computational technologies that complement advances in data collection, to accelerate and democratize model building, hypothesis testing, and model refinement. Other disciplines, including molecular genetics, cosmology, and plant biology, have capitalized on this opportunity. In neuroscience, however, 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 a custom spreadsheet, analyzes it using software that he buys and installs on local computers that he updates regularly, and publishes a summary of the results. TIBS will create the possibility of a superior strategy for collaboration and community building: 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.

Intellectual Merit:

Our multidisciplinary team, comprising neurobiologists, engineers, mathematicians, computer scientists, and physicists, will converge to develop and build a robust, comprehensive, open, science-driven, cyber-infrastructure ecosystem that can accelerate a broad spectrum of data-intensive brain science research. We will complement the ecosystem by incorporating fundamental research in mathematics, statistics, and computational science to enable data-driven discovery and decision-making through visualization, modeling, and analysis of large and complex brain-imaging data. These tools will enable scientists of all backgrounds and resources to contribute to understanding the rules of life. Specifically, we will provide the data and tools to enable researchers to predict phenotypic properties of neural circuits based on multiple types of information processed across multiple scales, including genetic, environmental, evolutionary, and biophysical. As an example, the initial data and technology we disseminate will enable anyone with internet access to address the following questions: what is changing and what is preserved in neural circuits with differing individual, developmental, and evolutionary histories across spatial, temporal, and phylogenetic scales?

Broader Impacts:

This project will significantly change the way neuroscience is conducted, and by whom. Advances in neuroscience will no longer be limited by data processing; algorithms and analysis will be shared quickly and easily. Barriers to entry into neuroscience will be significantly lowered along multiple fronts. Some of the most expensive data ever collected will be shared publicly for visualization and exploration via mobile technologies. Professional and citizen brain scientists lacking computational expertise will be able to upload their data, re-run or extend analyses on existing data, and make novel discoveries. Brain science data has properties-multimodal, multifarious, multiscale-similar to other fields that have not yet benefited from the IT revolution. Our infrastructural advances, therefore, can be extended to revolutionize those fields as well. This project will also enhance many outreach activities at our local universities and world-wide. We will run a local high school program, and multiple annual hackathons for college students from near and far. We will sponsor a summer internship program for undergraduates with hearing loss, who will work with students from URM backgrounds. Finally, we will create mobile compliant digital education content to complement our existing online courses, which have netted more than 4 million enrollees, to target STEM students, and educate global citizens.

Overview

Neuroscience is undergoing a technology revolution, enabling data collection at unprecedented rates, precisions, and scales, with the potential to answer age-old questions. These technological changes have also initiated a change in culture. Across disciplines, more data-collection efforts are willing to be open access, from electron microscopy to calcium imaging, to gene expression maps, to multimodal magnetic resonance imaging. 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 [1]. 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. Moreover, the availability of data enables the democratization of neuroscience, because a much larger fraction of the community can participate in the scientific process.

We propose to lower the barrier to connecting data to analyses and models by providing a coherent cloud computational ecosystem that minimizes current bottlenecks in the scientific process. The International Brain Station (TIBS) will provide a computational infrastructure to support brain sciences across spatial, temporal, phylogenetic, data, and hardware scales. The gap between data collection and modeling is too large to tackle all at once. Therefore, we address a subset of the problem, building the foundation for a larger-scale investment to complete the work. Specifically, we propose to enable answers to the following questions using brain imaging data: *what is changing and what is preserved in neural circuits with differing individual, developmental, and evolutionary histories across spatial, temporal, and phylogenetic scales?* This is but one canonical proof-of-concept example to demonstrate the potential power of this approach, which can be generalized to other questions and eventually other disciplines. The development of TIBS will include four tasks: (i) intaking, processing, curating, and disseminating data from eight different species across five modalities spanning nanometer to millimeter and kilohertz to Hz; (ii) developing and deploying a software ecosystem, which we call "Neuroscience as a Service", to enable searching, exploring, analyzing, and modeling these data via mobile technologies; (iii) designing mathematically principled, statistically justified, and computationally tractable algorithms for all stages of scientific discovery applicable across modalities and scales; and (iv) providing educational content to support our tools and technology.

Current Challenges and Limitations

TIBS will use best practices from brain sciences, computational sciences, mathematical and statistical sciences, to address the below challenges.

1. Big Data. We categorize data as little (fits in memory), medium (fits in a workstation's storage), and big (exceeds workstation capacity). *Tasks as simple as visualizing, rotating, and opening medium data are challenging using tools like MATLAB, Python, or ImageJ. For big data, storage, compression, management, and archiving exceed the capabilities and resources of most experimental labs.* Other disciplines have faced similar challenges. The Sloan Digital Sky Survey (SDSS) is a two-dimensional (2D) spatial database of images of the night sky that allows researchers to run queries on previously published data. Although there are only about 10,000 astronomers [2], SDSS already has >5,000 citations in 15 years and >200,000 queries [3]. Neuroscience data, however, is 3D, 4D (multispectral or multitemporal), or 5D (both), and therefore an SDSS-type computational ecosystem is not sufficient. Keller et al. [4] designed a new format to store large data files, but did not incorporate a flexible application program interface (API) or support multiple modalities. TIBS will be able to store 2D, 3D, 4D, and 5D data, no matter how large.

2. Private Data. Access to most datasets is now limited to lab personnel, even when others could productively collaborate. *But sharing data with collaborators costs time and resources.* Small data can be shared by email or Dropbox. For larger data, a researcher could mail hard drives in a specialized format with instructions. Neuroscience data repositories include openfmri [5], LONI Image and Data Archive [6], International Data Sharing Initiative [7], the Montreal Neurological Institute data-sharing and processing ecosystem [8], CRCNS data sharing [9], and NeuroData [10]. Only NeuroData can store multi-terabyte

datasets, but it lacks sufficient scalability and functionality to have a much broader impact. TIBS will easily make any types of datasets open access.

3. Disorganized Data. Data are typically stored internally in some custom format. For another person to use it, the experimentalist could either provide detailed instructions, or convert it to a standard format, such as Neurodata without Borders [11] or Brain Imaging Data Structure [12]. *But this conversion does not directly benefit the experimentalist, and can be quite time-consuming.* TIBS will eliminate this step by automatically organizing data upon acquisition.

4. Private Code. Analyzing digital data requires code, which is often private. *This limits its utility to people with connections to its authors, thus severely limiting its reach.* Github, Bitbucket, and related sites are popular geretic code repositories, and there are neuroscience-specific code repositories, such as NITRC, but code simply in repositories has no guarantees. Code repositories for specific languages, such as PiPy for python and CRAN and bioconductor for R, have guidelines, such as easy installation, but running these codes on any dataset can be cumbersome or impossible. The TIBS code will be open source and pre-configured to run across different hardware and software platforms.

5. Repeatable Analysis. Within or across labs, researchers often desire to repeat an analysis pipeline. *The lack of repeatable analysis stalls scientific progress by requiring investigators to waste time reimplementing previous results.* Repeatability is plagued by imperfect practices, many of which are errors in reporting. For example, data is stored in a specialized format, or code is not documented. An entire pipeline includes lots of code snippets, each with their own flags and parameters. Moreover, software dependencies and versions must all be the same, or the answers may be different [13]. Worse, code sometimes depends on specific hardware configurations, such as supercomputers. Recent projects have tried to combat this (for example, [14–18]), but each requires careful configuration. Container services [19,20] lower the barrier to entry by automatically launching machines in cloud environments, so hardware configurations are not required. But none of these solutions work with medium or big data. Moreover, they cater to technologists, not neurobiologists. TIBS will achieve repeatable analysis on arbitrarily big data by enabling neuroscientists to run an analysis in the cloud, without configuration.

6. Provenance Tracking. Provenance tracking is the process of tracking the history of all revisions to a research artifact. As data gets more diverse and distributed, tracking becomes more important to ensure repeatability and validity. *Current tools for provenance tracking [14,15,21–23] are not integrated with data storage, so that as pipelines get modified, researchers can lose track of which files were generated by which pipeline.* TIBS will automatically track data and workflows together.

7. Batch Effects. An experiment is reproducible if duplicating the design and analysis produces the same result, a much higher bar than data-analytic repeatability alone. The lack of reproducibility in science has been called a crisis [24,25]. It is partly caused by errors in reporting details, which can be addressed by tracking, but another cause is more troublesome. *Many large sources of experimental variability, such as sample handling, are easily detected but not biologically interesting.* This so-called “batch effect” [26] has been problematic in many disciplines, especially micro-arrays [27]. Neuroscience batch effects have drawn less attention, perhaps because pooling across datasets is common only in brain imaging [28]. TIBS algorithms will mitigate batch effects across spatial, temporal, and phylogenetic scales.

8. Scalable Data-Processing Algorithms. Many experimental modalities share a common sequence of data processing steps: geometric and chromatic aberration correction, registration to an atlas, object detection, graph inference and analysis, and modeling and testing. *A number of efforts have addressed each of these steps, although each has scalability and usability limitations*. Spark is a popular scalable machine learning library [29], but it requires cluster configuration and is inefficient in its consumption of computing resources [30]. FlashX [30] is an emerging option, but it lacks the desired functionality, only has R wrappers, and is poorly documented. TIBS will deploy reference scalable algorithms for each stage of analysis, with both R and Python interfaces, and extensive documentation.

9. Experimental Design. Experimental design is the art and practice of choosing data-collection and processing to optimize statistical efficiency, given a “budget” (say, of photons, scanner time, or money).

Now that more researchers can use a dataset to address a greater variety of questions, experimental design becomes more difficult and important, because we desire that the data are useful in a much wider variety of contexts. Existing approaches either require univariate data (e.g., ICC of a particular metric [[31]), or make strong parametric assumptions [32,33]. TIBS will include non-parametric experimental-design algorithms that allow researchers to optimize data collection and processing decisions for a wide variety of tasks.

10. Computational and Statistical Knowledge. *There is a skill gap.* Experimental neuroscience experts are often novices in computation and statistics, which may be another source of the reproducibility crisis [34]. Similarly, experts in computation and statistics typically lack detailed knowledge of the brain. TIBS will provide extensive education, including documentation, tutorials, workshops, hackathons, and summer courses, to train the next generation of data intensive brain scientists.

Once these challenges are met, researchers will be able to move seamlessly between models and data, regardless of its spatial, temporal, or phylogenetic scales, or the scale of the hardware or data resources.

Summary of Proposed Approach

Our project brings together a multidisciplinary team of experts to converge on a common problem. We have published together, supervised students together, and interact frequently, thanks to co-localizing at JHU and Harvard. We are also privileged to work with ~25 collaborators from around the country and the world to ensure that our work is useful in their daily laboratory practice. We will overcome the above challenges via four complementary tasks.

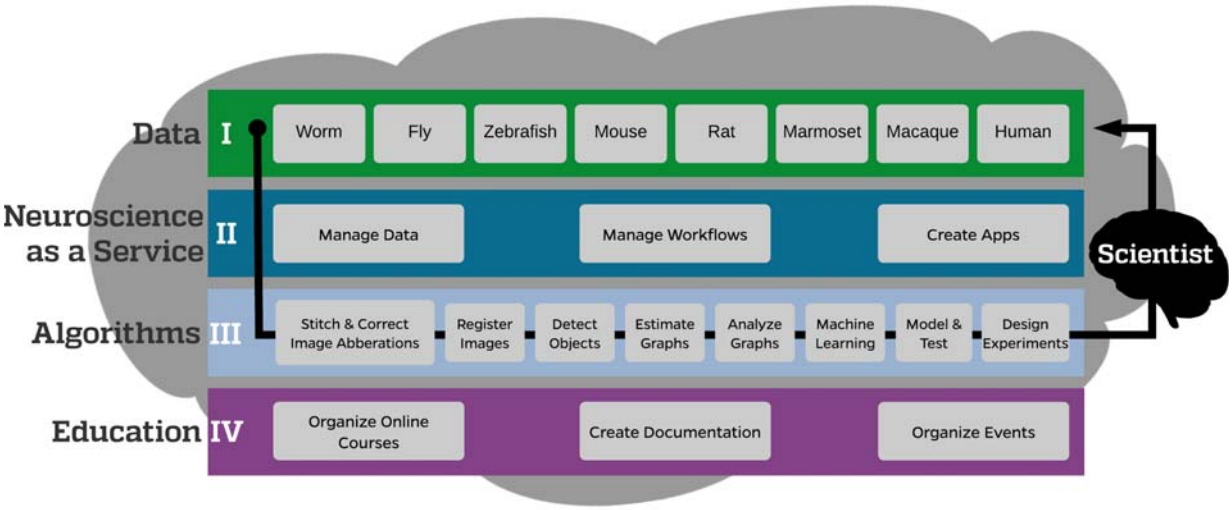

Figure 1: Schematic of four tasks and subtasks proposed to build The International Brain Station.

Task I: Data. We will take in, process, curate, and disseminate data across five modalities (activity, behavior, connectivity, development, epigenetics); eight species, from C. elegans to human; and spatiotemporal scales from nanometer to meter and milliseconds to minutes. All data will be stored in a common format with a standard API, enabling visualization, analysis, and comparison in the cloud. *Intake will include raw, derived, and metadata*. TIBS will enable uploading/downloading in many standard formats, so that anyone can use it. All open datasets will be provided with a unique digital object identifier (DOI) and user-specified license. This task addresses challenges (#1-3) of big, private, and disorganized data.

Task II: Neuroscience as a Service (NaaS). We will build a software ecosystem that lowers the hurdles to engaging with data, models, and analysis. **NaaS will provide a computational framework that will enable anyone to analyze brain science data of arbitrary scale and complexity.** NaaS provides a

complete scientific environment in the cloud that allows users to browse data, run models, and link to public data sources in a desktop and notebook abstraction implemented with, for example, Jupyter lab. In addition to a *data-management* system, we will deploy a *workflow-management* system that builds on the concepts of git versioning and git content storage, to enable users to copy or edit existing "digital experiments" that contain an entire pipeline, from data sources to visualizations. We will also provide a simple API to enable mobile-compliant *Apps*, which anyone can contribute. Apps customize or extend analysis and visualization to new data types and algorithms. The environment migrates seamlessly from a laptop to the cloud and runs identically in both. Our collaborators will provide feedback to ensure that NaaS accelerates their discovery. This task addresses challenges (#4-6) of private code, repeatable analysis, and provenance tracking.

Task III: Algorithms. We will develop *data-processing pipelines* from raw data to models. Each algorithm will be based on *foundational mathematics* and statistics, so it performs well across modalities and scales, while mitigating batch effects. We will develop enterprise-grade, scalable, and portable implementations of each algorithm, tools to guide *experimental design*s, and *qualitative and qualitative assessment* code to evaluate data quality before and after analysis. This task addresses challenges (#7-9) of batch effects, experimental design, and scalable data processing.

Task IV: Education. We will generate extensive educational content, allowing anyone to learn to use TIBS, regardless of their computational or neuroscience skills. We will create professional *documentation* and tutorials, massively *open online courses*, and *events,* such as workshops at top conferences, hackathons and trainings, and summer courses. This task addresses the knowledge gap (challenge #10).

Task I: Data

The data we include will determine the questions we can answer, and the tools that we build to answer them. Our faculty and collaborators have worked together to identify datasets upon which we will develop and test all of our functionality. These datasets span all the major neuroscience phylogeny, from the simplest to the most complex, from nanometer to millimeter spatial resolution, and from millisecond to minute temporal resolution. For each species and modality, we will in take, process, curate, and disseminate three different types of data: raw data, derived data, and metadata.

Raw data comes directly off the sensors or has been pre-processed in various ways. In preliminary work we have designed NeuroData [10] to store and manage raw data across modalities. The NeuroData repository is currently the largest and most diverse data repositories in neuroscience, hosting >20 different image datasets, and containing >50 teravoxels of image data, including electron microscopy [35–40], calcium imaging [4], array tomography [41–43], expansion microscopy [44], x-ray microtomography [45], CLARITY [46–48], gene expression maps [49], and magnetic resonance imaging (MRI) [50].

Derived data is the output of algorithms that convert the raw data into summary statistics, including volumetric annotations, skeletons and graphs. All derived data are stored in a neuro-ontological database called RAMON [51], which includes the compressed voxel list, in addition to a structured vocabulary of metadata fields (such as synapse, organelle) as well as unstructured key-value pairs. Once data are stored in this format, derived data can be visualized, overlaid on top of the image data, and filtered, sorted, or searched using a Web-interface (see Task II.3). NeuroData also stores derived datasets for a number of projects, including manually annotated saturated volumes containing neurotransmitter vesicles, mitochondria, spines, axons, and synapses from EM data [38], automatically detected cell bodies and blood vessels from X-ray microtomography [45], synapses [52], and anatomical regions of interest [46], and more [35,37,53]. NeuroData stores is over 500,000 annotation objects, including several teravoxels of annotations spanning eight publications and several species. In addition, NeuroData stores derived graphs from nearly 10,000 connectomes, spanning phylogenies and scales, mostly derived from diffusion and functional MRI. Each of these brain scans yields graphs at a variety of spatial scales; thus NeuroData stores >300,000 brain graphs. Metadata includes the information about the raw and derived data, including time stamps, user IDs, which software versions were used, etc. NeuroData already stores all metadata required for adding data to the spatial database.

The eight species that we will work with are: **1. *Caenorhabditis elegans*** (worm)**, 2. *Drosophila melanogaster*** (fly)**, 3. *Danio rerio*** (zebrafish)**, 4. *Mus musculus*** (mouse)**, 5. *Rattus norvegicus*** (rat)**,**

6. *Callithrix jacchus* (marmoset)**, 7. *Macaca mulatta*** (macaque)**, and 8. *Homo sapiens*** (human). Each species defines a subtask; for each species and modality, we will take in the raw data, derived data, and metadata into our cloud infrastructure. The specific scales for each modality and species will include:

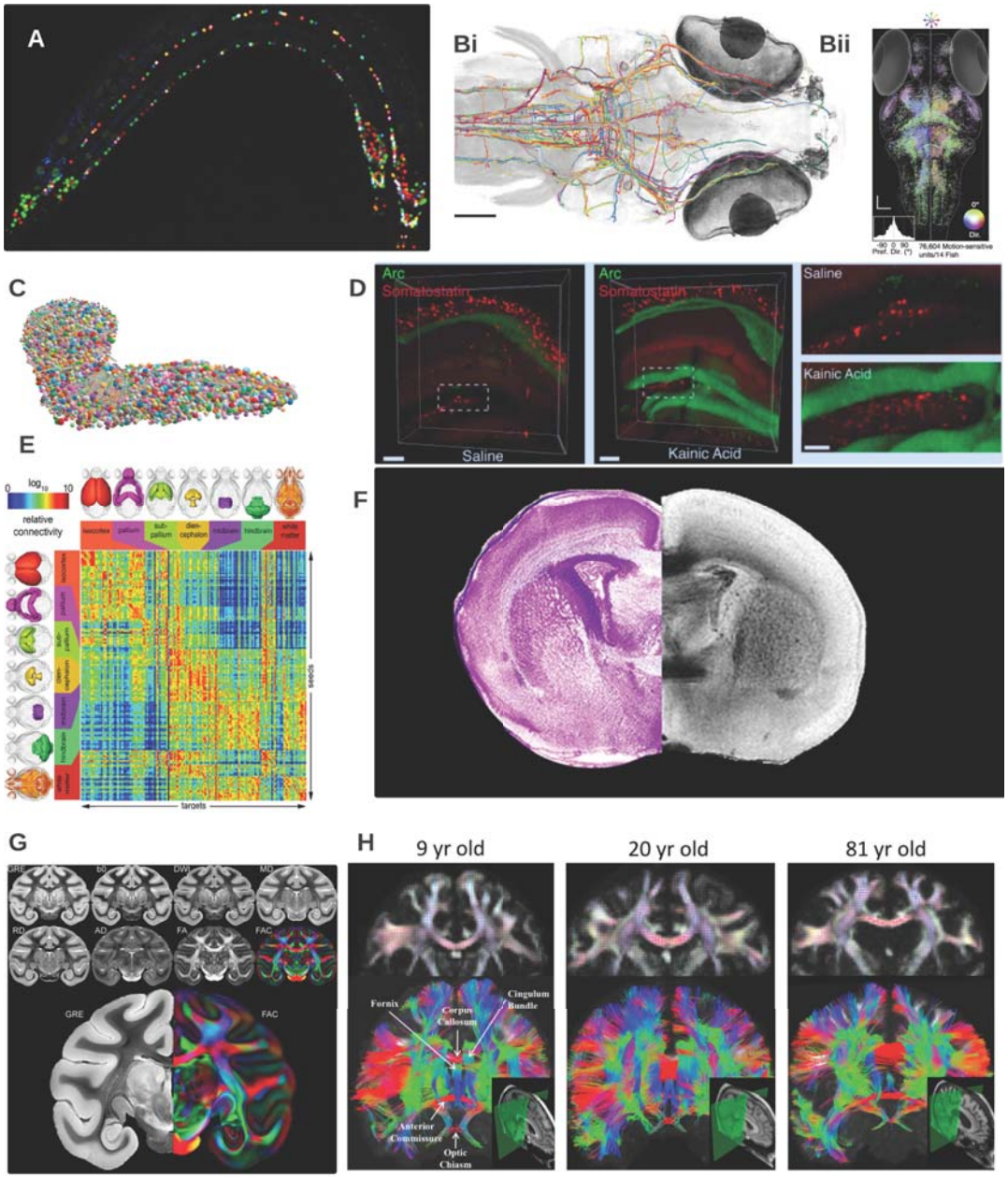

Figure 2: Images of raw and derived data across modalities and phylogenies. (A) Worm epigenetics (brainbow). (B) Zebrafish connectivity (i; EM data and traced axons) and activity (ii; 2-photon responses of all neurons). (C) Fly connectivity (EM and traced axons on all mushroom body cells). (D) Mouse epigenetics (CLARITY). (E) Mouse connectivity (derived from MRM). (F) Rat connectivity and development (histology and MRM). (G) Macaque and (H) human connectivity and development using MRM.

- *A for Activity:* In every species but human, microscale calcium imaging (whole brain for worm [54], fly [55,56], zebrafish [56–59], fast and deep scanning in mouse [60,61]), at rates from 10 to 300 Hz. *Frank* will also provide e-phys data for mouse (30 kHz), and both contributors who study humans will provide fMRI (1 mm^3 at 1 Hz).

- *B for Behavior*: Fly behavior includes epigenetic stimulation [40], and mouse is on a ball [62].
- *C for Connectivity*: Derived from nanoscale EM data for all species. For worm, fly, and fish, we will obtain multiple whole brains per species, for the others, we will obtain subvolumes. We will obtain larger volumes of X-ray μ-tomography (XCT) for a subset of those tissues [45]. All the mammals will also include milliscale MR microscopy (MRM) derived connectivity [63–65].
- *D for Development*: Worm data will include EM at each of the four larval stages plus an alternative stage, mouse includes gene-expression maps [66], rat includes histology and MRM, marmoset includes MRM acquired at each month across development, and human includes lifespan (rather than longitudinal) MRM from ages 5 to 80 [65].
- *E for Epigenetics*: Worm data will have a deterministic brainbow that uniquely labels all cell types, fly includes optogenetics for >2,000 single cell resolution cell types [67], zebrafish includes >150 single snapshot activity maps with antibody staining against phosphorylated ERK under different conditions [49], mammals will have CLARITY imaging using COLM [68–72], including whole brain for mouse, as well serial-section two-photon microscope with integrated vibratome [73], and expansion microscopy (XM) for both mouse and marmoset [44,74,75].

Table 2 (in Coordination Plan) lists for each species and modality who is responsible for providing the data (see Figure 2 for example figures). In addition to the above (not yet open access) data, for **subtask 9**, we will also take in many existing open-access datasets, including those from NeuroData, much of the Allen Institute for Brain Science's open data (with Collaborator Koch [76–84]), data from the Montreal Neurological Institute (with Collaborator Evans [85]), and the BigBrain histology dataset (with Collaborator Amunts [86]). Finally, for **subtask 10**, we will accept all data from Johns Hopkins Kavli Neuroscience Discovery Institute (KNDI), which includes over 45 neuroscience faculty across the different schools and campuses. Specifically, KDNI is purchasing a light sheet microscope, and we will be accepting and making public all data from that microscope, as well as developing all software for visualization and analysis of those data (see Task II and Task III).

Task II: Neuroscience as a Service

The goal of this task is to make neuroscientists agnostic to where the data and the code live, because all analyses can be applied across datasets and hardware configurations without change. Thus, once someone writes a workflow or pipeline to analyze some data, others could easily reproduce or extend it, without having to install dependencies, buy new hardware, or understand other eccentricities of the code base. The **Neuroscience as a Service** (NaaS) task will simplify integration of workflows and pipelines with data products. In particular, anybody would be able to upload an entire "digital experiment" that processes data and computes various summaries. Others would be able to run that experiment on their computers, modify it to point to their data, or use other algorithms. *All user-facing services will be designed with communications and digital media consultants to ensure maximally clear usability.*

Task II.1: Manage Data

As discussed in Task I, data includes raw, derived, and metadata. The current NeuroData infrastructure, while best in class, is inadequate to fully address the needs for the proposed work. The existing data is deployed on local institutional resources, so it cannot dynamically scale. For example, when a colleague obtains a new 40 terabyte (TB) dataset, we must request permission for shared resources, obtain hard drives, set up additional nodes, etc. Users also struggle to upload data, in particular derived data including volumes, meshes, graphs, skeletons, and more. Finally, the existing support for metadata currently only includes metadata essential for the intake process, such as the size of the data.

Thus, we will port the NeuroData infrastructure first to the Amazon cloud, and then add flexibility to enable it to be run on other cloud service providers. This will include improved scalability, usability, and functionality for taking in both raw and derived data. We will work with Collaborator Evans to extend the LORIS metadata management system [8,87] to support our use cases, including a much wider variety of species and modalities, as well as object-level metadata (such as who generated each annotation).

In the resulting *Data Management System*, all input and output functions will have a well documented, simple API, will support common data formats, and will be easily extensible to new or custom formats. Data will also be compressed, with or without loss, as the contributor prefers. Separate services will be developed to address the different data types. All the activities of Task I depend on the functionality of this subtask. All datasets will be given a unique DOI via the JHU Library service, and granted one of several open access licenses. TIBS will become an official data repository for Nature Scientific Data, replacing NeuroData which currently serves that purpose.

Task II.2: Manage Workflows

Workflow management will enable users to publish "digital experiments", just as data management will enable users to publish datasets. A digital experiment is a set of data DOIs, algorithms, parameters, and results. For example, Alice might design a digital experiment that takes a collection of multi-modal MRI scans from a population of normal subjects, fully processes them, and then clusters them to identify groups, and reports on cluster stability. Once Alice publishes her digital experiment, Bob could make an exact replicate of it and run it on his own local hardware, or in the cloud, depending on his resources. If it performs as he expects, he can then extend it in a number of ways. For example, he could point it to a different dataset, or replace some of the algorithms with different algorithms. Bob could then publish his version of the experiment. All the provenance is automatically tracked by the system. So, if Bob's experiment only changed the last step of Alice's, when Bob runs his experiment, it would skip all the first steps, and just run the last. If Bob's experiment convinces Alice that his modification is an improvement, then Alice can merge her experiment with his. Chris might see that Alice and Bob have now converged, and replicate this version. But when Chris goes to run the experiment on his data, he might get a warning that his data are not of sufficiently high quality to trust the answers. So, Chris looks at the quality assessments that are output at each stage. He notices that the registration that worked for Alice's and Bob's data failed on his data. Now Chris can look for other algorithms that work on his data.

We have built a prototype digital experiment using the data services mentioned above. It takes a population of multi-modal MRI datasets and runs a subject-level analysis to estimate connectomes and plot 10 graph statistics, including degree distribution, and then a group-level analysis to compute the average connectome. The url http://scienceinthe.cloud currently hosts this digital experiment, so that anybody can run it. The container that includes all software dependencies and code is available from Docker hub, the source code including all the algorithms is on github, and the package is called "ndmg".

To make this story as simple as it sounds requires developing several components of the workflow management system. First, we will containerize many existing algorithms, so that each of them can be easily run on different software platforms. Second, we will design a "continuous integration system", built on top of services such as TravisCI. This system will track provenance and changes to data or code, so that when changes are made, all analyses can be run automatically. Upon completion of this subtask, generating digital experiments in the cloud will be standard practice for all TIBS users.

Task II.3: Create Apps

Apps provide user-facing tools for utilizing TIBS. Each app will be designed to address a particular "user story". User stories are a software-development practice designed to ensure rapid development of features to satisfy consumer needs [88], in the form: "As a <*role*> I want <*goal/desire*> so that <*benefit*>." We will define a specification to enable a wide variety of apps for different user stories. We will then build the basic apps to ensure all fundamental tasks are easy to perform. Below we provide a list of apps that we will deploy in this subtask. Each is designed to address at least one user story. This is merely a starting point; we expect that the community will develop apps beyond our current imagination.

1. *Intake* enables easy depositing of raw, processed and metadata into TIBS.
2. *Dashboards* directly interface with data-collection machines to provide live status updates.
3. *Console* controls access to data, products, and results.
4. *Discover* finds data via filtering against controlled vocabularies and unique identifiers, including quality-control metrics, and other data derivatives (rather than merely metadata).
5. *Leaderboards* compare performance metrics of appropriate algorithms on different datasets.
6. *Digital Experiments* provides scientific workflows from data discovery to scientific discovery.
7. *Explorers* enable navigation and data exploration including pan, zoom, filter, and select.
8. *Manual Annotation* enables users to label data for both human and machine consumption.
9. *Wrappers* connect algorithms to the infrastructure.

10. *Modeling Bots* automatically fit incoming data to an existing model, or fit existing data to a new model, given appropriate specifications of the data/model pair.

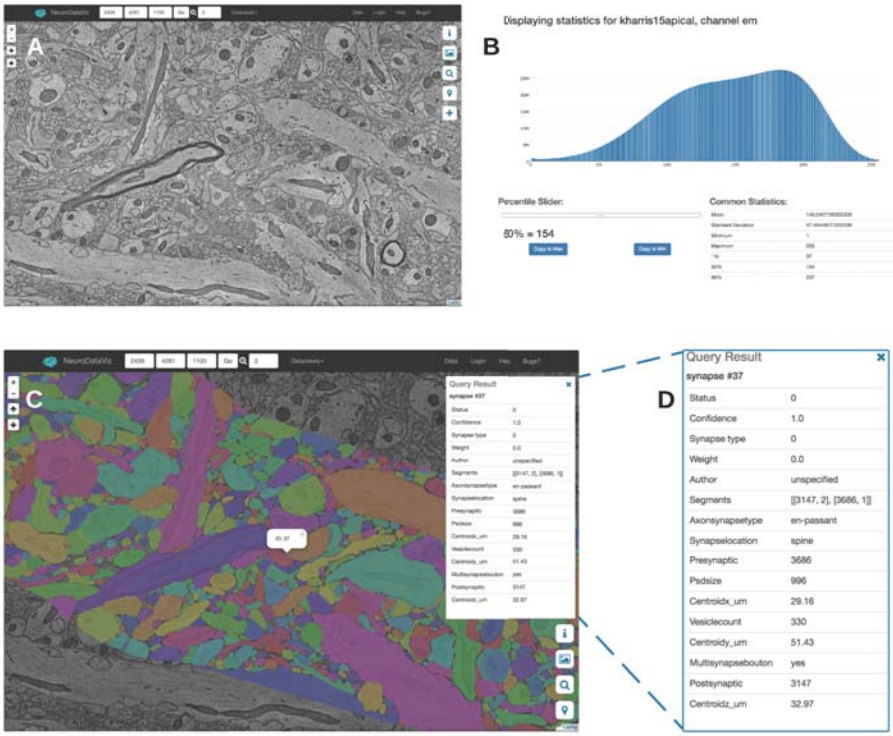

Figure 3: Several apps under development. (A) Mobile compliant visualization of 3D images; (B) Histogram computation; (C) Overlay of manual annotations; (D) Objects specific metadata for a selected synapse.

We have already built prototype versions of some of these apps—including intake (called ndingest), console (built into ndstore), digital experiments, for example using CLARITY data (inside ndreg), and explorers (including ndviz, Matrix Explorer, and Graph Explorer) (see Fig 3). All are open source with documentation, and available in GitHub from our NeuroData organization, under different repositories.

Task III: Algorithms

The goal of this task is to build mathematically principled, statistically justified, and computationally tractable algorithms for all stages of scientific discovery applicable to data across scales. Stages of analysis range from image filtering to modeling to experimental design. We will also build reference pipelines for each modality to enable comparisons across conditions, experiments, and scales. Below we emphasize a few important qualities for each step of the process that deserve special attention.

Mathematically principled methods are important for analyzing large, disparate, and multimodal data. Batch effects (sources of variance of no scientific interest, like which microscope was used) are more prevalent as data are combined and compared across experiments, institutions, and modalities. If the properties of brains differ across species because they were estimated using different techniques, the results are less interesting and less repeatable. To mitigate these effects, we can use the same algorithms across different datasets whenever appropriate. Developing mathematically principled methodologies will enable us to transcend existing discipline-specific boundaries, so algorithms can be applied in a wide variety of settings. Once we have processed disparate datasets using the same algorithms, we can compare outputs to harmonize results [28,89] (see Fig 4F). All of our algorithms will therefore be designed to be applicable across modalities and scales.

Quality assessment (QA) is important when running scalable algorithms on disparate data, because algorithms will run on datasets on which they were not explicitly tested. To address this, we will ensure that every algorithm outputs not just a final estimate, but also error bars or confidence measures. We will also provide qualitative assessments—visualizations of the results—as well as quantitative metrics computing performance along different dimensions. This will allow users to ascertain how well the algorithms are working, and whether they must make adjustments to achieve satisfactory results. In

preliminary work, we compiled the quality assessment tools from both the Harvard Diffusion Imaging QA Toolbox and CPAC [90] to build a comprehensive MRI QA package, which we have incorporated into our "ndmg" package (available on GitHub, including interactive QA tutorials). We will extend the functionality of this toolbox to be applied to all different scales and modalities of data.

Scalability is also necessary. We have used two complementary strategies to achieve it based on the properties of the data and algorithms. The intuition underlying both is that the complete data is too large to fit into memory for a single core, so the data must be distributed. Communication between hardware components can be very slow, so minimizing communication costs is key. For images, "embarrassingly parallel" operations are optimal: we built a *scalable distributed spatial database* (ndstore on GitHub) to enable efficient partitioning of petascale images into local slices or volumes [10], which can then be operated on each independently, and the results can be merged [91,92]. For graphs and matrices, we partition the vector/node *identifier* in memory whereas the entire graph/matrix is in storage. For such operations, we developed *semi-external memory* libraries: FlashGraph [30] and FlashMatrix [93]. For all algorithms we will use these approaches to ensure arbitrary scalability.

For Task III, we have defined 8 subtasks, one for each stage of analysis. For each, we will build upon our mathematically principled algorithms that already have demonstrated empirical utility to incorporate uncertainty estimates, and develop and disseminate scalable implementations, which then get wrapped by Apps to improve usability and generality across modalities and scales, and includes both R and Python wrappers for all functions including extensive documentation and interactive tutorials.

1. Stitch and Correct Image Aberrations*:* We have applied elastic stitching and registration [94] to a 100 terabyte (TB) EM dataset (Fig 4A). We have also been generalizing both 2D [91] and 3D [92] geometric correction algorithms (Fig 4B).

2. Register Images*:* Registration of images to reference templates is a problem that has seen much development [95,96]. In 1998 [97] we formalized diffeomorphometry within the field of computational anatomy, developing a mathematically principled framework for nonlinear registration based on the embedding of shape and form into an infinite dimensional metric space [97]. This has led to several algorithms for manipulating diffeomorphisms—e.g. ANTS and Large Deformation Diffeomorphic Metric Mapping (LDDMM)—that are widely available. Our focus on LDDMM [98] is for two reasons. First, it can be applied consistently to all submanifolds, including points (fiducials or cells), curves (such as cortical gyri) and full volumes [99–101]. Second, it provides a geodesic positioning system (GPS) [102]. Second, LDDMM and its derivatives are considered the state of the art [103]. Originally developed for the human 1 mm^3 MRI scale, it has been extended to high-field MRI on mice [104], the μm^3 scale of macaque histology [105], and mouse CLARITY data [106] (Fig 4C). Here, morphometry is indexed to a common template coordinate system by computing a smooth, invertible, differentiable correspondence between the template and the target [107,108]. LDDMM is currently deployed as a Software as a Service (SaaS) in MriCloud [109]. This is a crucial step for any cross-modality analyses.

3. Detect Objects: We have developed algorithms to detect large populations of neurons and infer their spiking from two-photon calcium imaging [110] and other imaging modalities. We showed [111] that the problem of inferring the spiking activity s from the fluorescence time series y of a *single* neuron can be posed as a Lasso style estimation problem $min_s||y - D\,s||^2 + \lambda||s||_1$, where D is a matrix that applies a convolution with a known decaying exponential to model the change in fluorescence resulting from a neural action potential. Then we formulated this problem as a structured matrix factorization problem, for which we proved our algorithm yields the globally optimal answer under suitable assumptions [112] (Fig 4D). A simpler special case is when there is no temporal activity, which can be applied to anatomical fluorescent data, including CLARITY and XM data, as well as histology, and even behavioral data. We have used other sparse methods to find cell bodies in XCT data [45].

4. Estimate Graphs*:* We have developed graph inference algorithms for many experimental modalities, including electron microscopy [51], calcium imaging [113], electrocorticographic data [114], functional [90,115,116] and effective [114,117] fMRI connectivity, and diffusion MRI [118,119]. These tools, for example, can show how brain connectivity changes with normal aging (Fig 4E). Many of these methods require multiple stages of processing. For example, we have developed reconstruction algorithms to

convert raw diffusion MRI data into orientation diffusion functions (ODFs) and algorithms for processing ODFs and estimating orientation using compressed sensing [120–123].

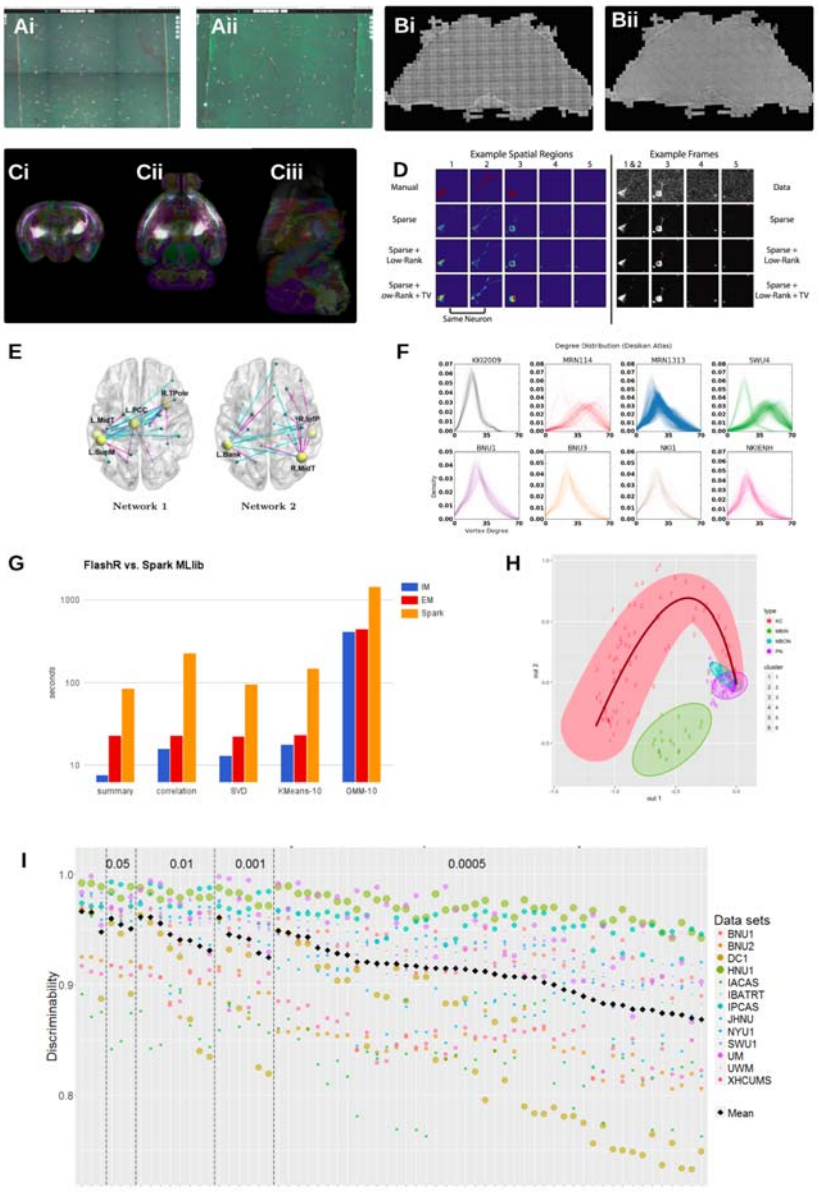

Figure 4: Preliminary results for all stages of analysis. (A) Before (i) and after (ii) geometric correction of array tomography data in the cloud [94]. (B) Before (i) and after (ii) chromatic correction on a 100 TB electron microscopy dataset [92]. (C) Three canonical axes showing a 1 TB whole mouse brain CLARITY image with the Allen Reference Atlas overlaid [46]. (D) Cell detection using NMF from two-photon calcium imaging [112]. (E) Network differences estimated from two populations using fMRI [160]. (F) Graph analysis across eight different datasets demonstrating batch effects [161]. (G) FlashMatrix scales 10x better than Spark for a variety of machine learning algorithms [93]. (H) Discovery of latent structure in larval Drosophila mushroom body EM derived connectome [40]. (I) Experimental design comparing 64 different pipelines on 12 different datasets to test which pipeline maximizes discriminability (related to reliability).

5. Analyze Graphs*:* We have written extensively on single network models, including testing [124,125], partitioning [126–129], and estimation [130]. We are developing a foundational statistical framework for classification of graphs [131–133], clustering of graphs [134,135], testing for differences between graphs [125], estimation of graphs with covariates [136], joint Bayesian modeling of graphs [89], and average [137] and median graphs [115], along with methods to estimate differences between graphs [138–141]. Finally, we developed a graph analytics library, FlashGraph, which enables anybody to process billion node and 100 billion edge graphs on a single commodity computer [30,128] (Fig 4F).

6. Machine Learning*:* We developed FlashX, which extends FlashGraph to FlashMatrix and FlashR. FlashMatrix provides a machine learning library that facilitates k-means, principal components analysis, non-negative matrix factorization (NMF), Gaussian mixture modeling, and more (Fig 4G). FlashR includes R wrappers for all of those functions, as well as optimized basic matrix operations, such as matrix multiply, variance, sum, etc. [30,128,142,143]. We also developed techniques for statistical shrinkage, which improves reproducibility in neuroimaging, for example, without discarding essential information

[144–147]. We will augment FlashMatrix with methods we developed for structured matrix factorization (including NMF and dictionary learning) [112], highly efficient variable selection and shrinkage [148], and deep learning [149]. Our prior work shows the utility of these methods for cell detection, for example [112].

7. Model and Test*:* The final stage of analysis that yields understanding is modeling and testing. It is a process of combining neurobiological insight with quantitative models. We have developed a large number of statistical methodologies for testing the various structured domains of an hypothesis, including testing in spatial domains [150], testing for local graph structure [124], robust tests for changes in brain shape [151], tests for anomalies [152], testing in manifold-valued data [153], and models relating behavioral maps and cell types [67]. Recently we have developed theory and tools to discover hierarchical structure in brain networks [126]. Applying these ideas to the larval Drosophila mushroom body connectome, we discovered a previously unknown latent quadratic relationship in the Kenyon cell connectivity (Fig 4H).

8. Design Experiments*:* As the father of modern statistics, Ronald Fisher put it "*To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*" The goal of experimental design is to optimize the quality/quantity trade-off for a scientific question (or set thereof) [154,155]. We developed a multivariate extension of intra-class correlation called I2C2 [32] using multilevel functional PCA [156], which we implemented to require little memory [157] and extended to network models [33]. These approaches, however, make parametric assumptions about the data (such as Gaussian noise), which can be inaccurate [158,159]. We are therefore developing a non-parametric approach to quantify reproducibility. Our new metric is called "discriminability", because it computes the probability that one can discriminate between measurements that are the "same" versus "different". We proved that this metric bounds the predictive accuracy for *any* inference task, even if the dependent variable is unavailable. Then we used it to compare 64 different pipelines, across nearly 20 repeated-measure fMRI datasets. The result (Fig 4I) demonstrates which pipelines yield maximally discriminable connectomes across all the datasets. Remarkably, the pipelines that work well on some datasets tend to work well on all. We used this metric to assess pre-collection decisions, such as repetition rate in fMRI, b-value in dMRI, and even comparing discriminability across fMRI and dMRI.

Task IV: Education

Task IV.1: Write Documentation

Every function we develop will include professional quality documentation, including an API with examples auto-generated using the same guidelines as the most popular numerical libraries, such as NumPy and SciPy. In addition, we will provide extensive interactive tutorials, leveraging our experience building both Shiny Apps (for R) and Jupyter Notebooks (for Python). Whenever possible, both documentation and tutorials will follow the guidelines of "Simple English Wikipedia", ideally restricting words to either "Basic English" or Voice Of America Special English. All content will be open source, and community contributions to documentations and additional tutorials will be encouraged and supported (for example, by using readme.io that includes a "suggest edits" link for every documentation entry).

Task IV.2: Provide Online Courses

JHU and its Bloomberg School of Public Health have an unparalleled record of developing massive open online courses (MOOCs). Programs in Data Science, Executive Data Science, Genomic Data Science, R Software Development and Neuroimaging Data Science have led to over *4 million enrollments*. Thus, the team has significant expertise in the development, deployment, and management of online courses. The Department of Biostatistics now boasts a recording studio along with a full-time production expert in video and audio recording and editing. In addition, we have a full-time support programmer, who has expertise in educational APIs and can support programming assignments in coursework. Finally, we have the half-time support of a MOOC management specialist (one of the few in existence), who assists with pedagogical best practices, university relations with corporate MOOC delivery partners, and community teaching assistant management. Most germane to this grant, our courses *"Statistical Analysis of fMRI*

Data" and *"Principles of fMRI I & II"*, which are hosted on the Coursera platform, have together had an enrollment of over 70,000 students since February 2014. These students come from 181 different countries, and many would not have had access to this type of material otherwise. In addition, the recently finished "Neurohacking in R" and courses in development form the foundation of our Neuroimaging Data Science program.

In addition, the group has active YouTube channels of educational content, as well as open and free or low-cost textbooks via the Leanpub platform. The team boasts several thousand followers on social media, which assists in advertising educational products and programs. **The result of these combined efforts is a major shift in free or ultra-low-cost educational opportunities for students and consequently for computational, statistical and data science literacy.**

We will develop MOOC programs that teach students how to interface and extend TIBS resources. These will teach the uses of TIBS open data, implementing NaaS on the TIBS platform, and understanding data processing and analysis. We will include written documentation in open textbooks. Each class will be 2 weeks to 1 month, featuring 6- to 15-minute lectures, auto-graded multiple-choice assessments and peer-graded projects, with rubrics from TIBS researchers.

These courses, programs, and open materials will share four coordinated missions. First, free or low-cost training will fill an educational void in multi-modal computational neuroscience. Second, highly produced training materials will dramatically increase adoption of the products developed from this grant. Third, the credentials from our program will be useful for students and granting agencies. We already see frequent use of our existing materials in federal grant applications. Fourth, the materials will be used for onboarding students, postdocs and junior faculty when they start to work with the group. This use of MOOCs is often overlooked, but we find that the faculty time it saves is worth the effort by itself.

Our priority list starts with introductory data intensive brain science (4 weeks), connectomics (4 weeks), regional analysis of brain imaging data (4 weeks), and an introduction to TIBS (2 weeks). Each course will be filmed in the Biostatistics studio. Dr. Caffo will assist with platform-specific deployment of the course and assessment development. Classes will run continuously, for free, with low-cost options for validated certifications. Every class will have a hands-on project that will be peer graded.

Task IV.3: Organize Events

Summer Bootcamp: With collaborators Littlewood and Kasthuri (Argonne National Laboratories; ANL) we will organize a 1-2 week summer bootcamp at Marine Biology Laboratory (which is run by ANL) on data intensive brain sciences. This will be more involved and deeper educationally than the MOOCs, including daily lab work. PI Vogelstein already teaches at Berkeley's summer course in mining and modeling of neuroscience data, and received excellent feedback. Students from around the world will be able to receive scholarships to attend these courses for free (ANL already has funding). The faculty presenting material will be faculty or collaborators of this grant, and the teaching assistants will be graduate students, postdocs, and research staff funded by this grant.

Conference Workshops, Symposia, and Tutorials: The faculty and collaborators in this proposal have a history of running successful workshops, symposia, and tutorials at leading conferences, including the Society for Neuroscience (SfN), Neural Information Processing Systems (NIPS; the leading machine learning conference), and Computer Vision and Pattern Recognition (CVPR, the premiere computer vision conference), among others. We will continue this tradition, adding more data-intensive brain science content, and applying for workshops, tutorials, or symposia at least at SfN and NIPS, annually.

Hackathons: Hackathons are hands-on meetings with targeted goals that all or some attendees work on collectively. We have run successful hackathons to (1) create necessary software, (2) process data, and (3) inspire newcomers to an area with project-based learning in a group environment. These include JHU DaSH and Medhacks, an official hackathon of Major League Hacking, attracting ~350 people to JHU annually to develop technological solutions to medical challenges. We will organize local hackathons to push forward data import and processing from large public repositories, software and API coding for basic plumbing of TIBS, and project-based learning from the completed TIBS project. The first two types will be local and initiated as needed. The latter will occur at the end of the grant, after the main computational, API, and data resources are created and organized. We will structure the larger event as three days at a hotel with appropriate facilities. We will create a collection of motivating projects and ask students to bring

their own. We will informally create groups to work on projects together and use faculty and graduate students as group leaders and technical assistance. Groups will present their work as the final event.

Broader Impacts

Our proposal is fundamentally about democratizing brain science along several fronts, including financial and educational, regardless of location, ethnicity, or disability. Collecting state of the art neuroscience data can be quite expensive. For example, certain experimental technologies now cost millions of dollars, in addition to the costs of running the experiments, causing scientific inequalities [162]. All the other stages of the scientific process, however, from designing the experiments to modeling the results, often merely require access, education, and ingenuity. Our proposal will therefore enable any neuroscientist, either professional and citizen, rich or poor, domestic or international, access and ability to visualize and analyze some of the most expensive neuroscientific data ever collected. The educational component will provide instruction for all citizens, from the naive to sophisticated, lacking computational or neuroscience backgrounds, to utilize the tools and understand the data.

Summer Program for Disadvantaged Students. Culturally disadvantaged and underprivileged individuals include under-represented minorities (URMs), women, and people of disability. People with hearing loss are the largest disability group in society. Nonetheless they must deal with reduced access to auditory information (a primary form of communication) as well as isolation, ignorance about their condition, and invisibility, similar to URMs and women. Several TIBS faculty have substantial experience in managing disadvantaged summer students. In 2016, TIBS faculty mentored Hispanic, female, and profoundly hearing impaired undergraduates, and four female (two URM) high-school students. We will create a summer internship program that will include students working and participating in laboratories and activities run by TIBS. All students will present a poster at the annual Summer Research Early Identification Program (SR-EIP) conference, and college students will be encouraged to present posters at national conferences. Students will include:

Undergraduates with hearing loss and high school students: We will implement a novel summer research program for college students with hearing loss who utilize spoken and listening language. We will work with a national organization–Alexander Graham Bell Association for the Deaf and Hard of Hearing–to recruit two college students with hearing loss. Four high school students will be recruited from the Baltimore area; they will be supported via programs at JHU such as the Center for Talented Youth (Vogelstein is an alum), Thread (he is a sponsor), and Engineering Innovation

Women and minority undergraduates: We will also recruit two URM college students who will be selected via SR-EIP or Morgan State University (MSU). Support for stipends, housing, travel, supervision and supplies is requested for these four undergraduate students. MSU is a historically black university in Baltimore, about 10 minutes from JHU. It ranks among the top 10 baccalaureate institutions for black PhD recipients in all STEM fields and third for doctorates in Engineering. NIH and NSF funding for undergraduate research training programs and partnerships with research institutions such as JHU have significantly contributed to these successes. The NIGMS RISE program, directed by Dr. Hohmann since 2003, has seen 63% of its 101 graduates enter graduate programs (37 into MS programs and 26 directly into PhD programs). Nine completed their PhDs, including one in Biomedical Engineering from JHU. This MSU program trains STEM undergraduates in basic research and writing skills, exposure to professional development workshops, responsible conduct training, and so forth, via symposia and weekly research seminars. This will create visibility for TIBS on the MSU campus and facilitate community building among our graduates and MSU undergraduates.

Extramural Graduate Student Training. All summer students will be supervised by two postdoctoral fellows and two graduate students. Those who work with high-school students will have been vetted beforehand. The mentors will participate in a day-long mentorship training workshop at MSU. In addition to their research activities in the TIBS labs, they will be important facilitators as either students or TAs in the new courses to be developed. These activities will accelerate the intellectual (and social) community of our graduate students.

Course and Curriculum Development. In Fall 2015, PI Vogelstein created the world's first undergraduate minor in Computational Medicine, and is the Director of Undergraduate Education for Computational Medicine. The minor includes a year-long introductory course co-taught by PI Vogelstein,

co-PI Miller, Faculty Ratnanather, and others. Its goal is to introduce students to methods of computational medicine through lectures and projects. Lectures present core principles and case studies of physiological and anatomical modeling of health and disease.

In Fall 2016, PI Vogelstein created a new course called “NeuroData Design”, in which students work in teams of 3-4 with an external advisor/client to provide use cases. The goal of the course is to develop Web-services that are immediately useful for the external collaborator, eliminating as many computational pain points as possible. In the process, the students learn best practices of data science, agile development, machine learning, and more. The class is held in the Center for Bioengineering Innovation and Design studio (see Facilities, Equipment, and Other Resources).

The faculty will build on these existing classes, and others, to augment the undergraduate and graduate research program, preparing our graduate students for academic careers, as well as leadership positions in industry and government. These new courses include “Data Intensive Brain Sciences,” which will cover many of the same contents as in the MOOC, but in a classroom environment; “Machine Learning Techniques for Neuronal Connectivity and Activity,” which will cover reconstructing connectomes from EM, tera-scale machine learning algorithms, and deep neural networks; and “Machine Vision,” which will cover anatomical reconstructions and automatic animal-behavior annotation.

Commercialization. Much of the technology that we propose to develop is commercializable, for neuroscience and related fields (such as pathology and radiology). At JHU, we have much experience in commercialization. CIS members have developed intellectual property, with many patents and at least four startup companies associated with image analysis, biometric, and brain-mapping technologies since 1998. PI Vogelstein has started two companies based on his research. One, d8alab (in partnership with Faculty Caffo), provides micro-consulting for big data analytics. The other, FlashX (in partnership with Co-PI Burns), aims to commercialize their academic development of FlashX, targeting brain science first, and then other disciplines. Other technologies we develop could lead to more commercialization opportunities. We will work with Johns Hopkins Technology Ventures (JHTV) to commercialize this technology, as appropriate. In particular, we will work with FastForward, an accelerator at JHTV that serves as a catalyst for the advancement and commercialization of an array of innovations that are derived at JHU. The goal of FastForward is to help early ventures increase the probability of realizing their potential and bring innovation and life-changing technologies to market.

Broader Dissemination. Here we propose to build software and take in data, but not to fund users around the world for unlimited computation. We are pursuing four strategies to make using our infrastructure feasible even for investigators with less money. First, we have a rich history, beginning with National Partnership for Advanced Computational Infrastructure, as one of the largest academic users of the Teragrid. Now through the high-performance computing infrastructure of XSEDE (www.xsede.org), we deliver about 30,000 mapped brains per year. We have an XSEDE grant for the computational anatomy gateway, nearly doubling our allocation of node hours to ~3 million. We deliver these resources through MRICloud, available to researchers across the world [109], where we have 10,000 registered users. Second, we have partnerships with commercial cloud service providers. As part of the MICrONS project from Intelligence Advanced Research Projects Activity, we are building a spatial database to store up to 10 petabytes in Amazon. We will leverage this relationship to obtain academic (or better) rates to use the services. Third, we generate all our technologies via open-source techniques, so that users can download our solutions and instantiate virtual machines on their local computational resources. Users will be able to run these virtual machines to reproduce exactly and extend cloud experiments. Fourth, we also work with the Department of Energy, National Science Foundation, and National Institutes of Health. Collaborator Littlewood is the head of a national laboratory committed to helping us deploy our tools on their supercomputers.

Timeline

Resource Sharing Plan includes a detailed timeline for all subtasks, which flows from data management and intake, followed by workflow management, then apps and algorithms, and finally education.

Prior NSF Support

(a) PI Vogelstein was funded by one NSF grant (BRAIN EAGER DBI-1451081, see co-PI Priebe below), and is now funded by NSF BRAIN E\GER 1649880 (see co-PI Burns) and NSF 1637376 (04/15/2016 — 04/14/2017, $97,950), which directly relates to the proposal. (b) This grant funded “A Scientific Planning Workshop for Coordinating Brain Research Around the Globe, Baltimore, Maryland, April 7-8, 2016”. (c) *Intellectual Merit:* At the workshop, the community defined the three grand challenges for global brain sciences and the unifying resource, The International Brain Station, proposed here. *Broader Impacts:* That conference led to additional workshops, including the “Open Data Ecosystem for Neuroscience”, the “Present and Future of the BRAIN Initiative”, and the upcoming “Coordinating Global Brain Projects”. (d) This workshop led to four manuscripts: on the grand challenges in global brain science [163], on The International Brain Station, requested for Neuron’s SfN special issue (V1), on a Global Brain Initiative (V2), and the first to demonstrate “science in the cloud” (SIC) (V3). (e) An example of SIC that TIBS would support is available at http://scienceinthe.cloud. (f) This proposal is not for renewed support.

(a) Co-PI Burns is the PI of one NSF grant, with Co-PIs Miller and Vogelstein: NSF 1649880 (1/1/2017--12/31/2018, $147,299). (b) This grant is entitled "Computational Infrastructure for Brain Research: EAGER: BrainLab CI: Collaborative, Community Experiments with Data-Quality Controls through Continuous Integration". (c) *Intellectual Merit:* Brainlab CI will prototype a cloud-based experimental-management system for reproducible science that provides workflows, visualization, and analysis. *Broader Impacts:* Brainlab CI will overcome major obstacles to data sharing. We envision a system that integrates thousands of publicly available data resources in MRI and neurophysiology and creates incentives for data sharing. (d) This new award has not generated publications. (e) Open-source releases will be available from GitHub. (f) This proposal is not for renewed support.

(a) Co-PI Priebe is not currently funded by NSF. A recent award related to this proposal is NSF BRAIN EAGER DBI-1451081 (9/1/2014-8/31/2016, $300K). (b) This project was titled “Discovery and characterization of neural circuitry from behavior, connectivity patterns and activity patterns”. (c) The project developed theory and methods for discovery and characterization from multi-modal connectomes. Like this proposal, it dealt with inference from connectome data, but as an initial development of statistical inference theory and methods, its scope was limited. *Intellectual Merit*: This project developed principled methodology for multi-modal connectome inference, with demonstrations on real data. *Broader impacts:* Funding for this project included material for graduate pattern recognition course at JHU; training of PhD students; and new theory and methods for general multi-modal inference. (d) Eight publications were produced: one conference paper, one accepted for publication [164], one under revision [165], three submitted and under review [153,166,167], and two in preparation. (e) This project used (and impacted the collection of) data from HHMI Janelia Research Campus. (f) This proposal is not for renewed support.

Co-PIs Miller and Engert do not have any results from prior NSF support to report.

Coordination Plan

Map and Tables show roles of **funded faculty** and *unfunded collaborators.* We will also fund three paid consultants: one for communications (to ensure clarity of documentation and tutorials), one for digital media (to ensure Web user interfaces meet the highest standards), and one for commercialization (to facilitate greater dissemination and sustainability) (budget line G3).

Collaboration Plan Map

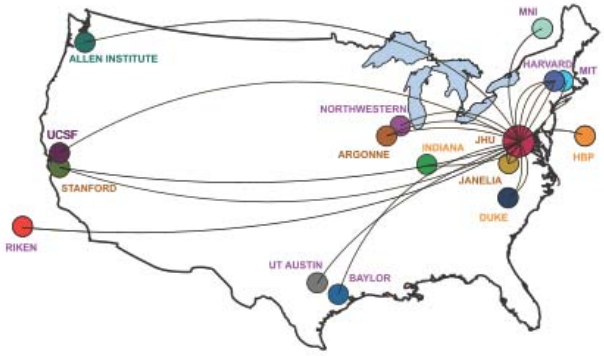

Table 1: **Funded faculty** organized by task (top row indicates lead).

I: Data	II: Neuroscience as a service	III: Algorithms	IV: Education
Vogelstein **Wang** **Perlman** **Engert** **Rosen**	**Burns** **Perlman** **Vogelstein** **Miller**	**Priebe** **Vogelstein** **Miller** **Vidal** **Caffo** **Lindquist** **Rosen**	**Ratnanather** **Vogelstein** **Caffo** **Lindquist**

Table 2: **Funded faculty** and *unfunded collaborators* subtasks for Task I, organized by species and modality.

	1. Worm	2. Fly	3. Fish	4. Mouse	5. Rat	6. Marmoset	7. Macaque	8. Human
Activity	*Samuel*	*Zlatic*	**Engert**	*Tolias* *Koch* *Frank*		**Wang**		**Rosen** *Evans*
Behavior	*Samuel*	*Zlatic*	**Engert**	*Tolias*		**Wang**		
Connectivity	*Samuel*	*Cardona*	**Engert**	*Lichtman* *Littlewood* *Kording* *Harris*	*Johnson* *Littlewood*	**Wang** *Johnson* *Littlewood*	*Johnson* *Littlewood*	**Rosen** *Sporns* *Evans* *Amunts* *Littlewood*
Development	*Samuel*			*Koch*	*Johnson*	**Wang**		**Rosen**
Epigenetics	*Samuel*	*Zlatic*	**Engert**	*Koch* *Boyden* *Deisseroth* *Spruston*	*Deisseroth*	**Wang** *Okano* *Deisseroth* *Boyden*	*Deisseroth*	*Koch* *Deisseroth*

Table 3: **Funded faculty** and *unfunded collaborators* subtasks for Task III (top row indicates lead).

1. Stitch & Correct Image Abberations	2. Register Images	3. Detect Objects	4. Estimate Graphs	5. Analyze Graphs	6. Machine Learning	7. Model & Test	8. Design Experiments
Perlman *Kazhdan* *Saalfeld* **Vogelstein** **Burns**	**Miller** **Vogelstein**	**Vidal** **Vogelstein** *Kording*	**Vogelstein** **Priebe** **Burns** **Caffo** **Lindquist** **Vidal** *Venkataraman* *Kording* **Rosen**	**Priebe** **Burns** **Vogelstein** *Kording*	**Vidal** **Burns** **Vogelstein** **Caffo** **Lindquist** **Priebe**	**Priebe** **Vogelstein**	**Vogelstein** **Priebe** **Caffo** **Lindquist**

The funded team spans a great diversity of neuroscience scales and areas of expertise to converge on a common goal. For example, Burns and Perlman are both computer scientists, both have worked in industry, and both have worked almost exclusively on data intensive brain sciences since 2011. Priebe, Caffo, and Lindquist are all statisticians that focus on methodologies for brain sciences. Caffo and Lindquist have offices at JHU School of Public Health, a 20 minute bus ride from JHU KNDI and Center for Imaging Science (CIS), where the other JHU faculty have offices. Both Caffo and Lindquist plan to spend a sabbatical at the CIS/KNDI, a global leader in image analysis. Miller, Vidal, and Ratnanather are mathematicians who focus on mathematical models in brain sciences. Engert and Rosen are both

physicists by training at Harvard, and now collect state of the art neuroscience data. Vogelstein has formal training in biomedical engineering, neuroscience, and applied mathematics and statistics, so he is able to bridge between fields. The meeting and event schedule is listed in the color-coded table below. The field engineer will monitor “neurostars”, a community driven Q&A forum for brain sciences.

We have arranged 25 collaborating partners from around the US and abroad, spanning disciplines such as chemical and genetic engineering (Deisseroth and Boyden) and physics (Kording). Nearly all collaborators will provide data, to extend the modalities and scales upon which we operate. We will work with Amunts (from Human Brain Project) to link their informatics functionality with ours. Okano (from RIKEN) will provide genetic tools for marmosets, so that we can work with both Collaborators Deisseroth and Boyden with the most state of the art marmoset tools. Particularly important will be our collaboration with Huganir, who is the director of KNDI and will be providing all JHU KNDI data that we will be responsible for, ensuring that we are a critical node in all KNDI scientific processes.Our field engineer’s role will be to ensure that collaborators can get their data into the system, and understand how to use the tools. Moreover, at SfN, we will organize a 30’ x 10‘ booth in the non-profit section, as NeuroData has now for two years running. At the booth we will have a dedicated hackathon and support space for the duration of the conference, staffed by our software engineers (see budget line B and budget justification).

A large number of our collaborating partners are also applying for NeuroNex grants. We have begun coordinating synergistic activities with them. For example, with Collaborators Littlewood and Kasthuri, we are organizing a new week-long summer course at Marine Biology Laboratory (MBL). We are working with Collaborator Sporns to ensure that this network toolbox can build on our highly scalable network analytics library, FlashX. For Collaborators Deisseroth, Tolias, Lichtman, Harris, and Samuel, we are committed to ensuring any data collected and algorithms developed via their NeuroNex Hubs are integrated into TIBS. Collaborator Frank will be providing non-image data in the form of electrophysiology. We will work with Frank to integrate suitable algorithms for spike sorting from 1024 channel data.

Event	Frequency	Duration	Participants	Location	Budget Line
Task I subgroup	weekly	1 hr	Students, postdocs, relevant faculty	KNDI+telecon	N/A
Task II subgroup	weekly	1 hr	Students, postdocs, software engineers, relevant faculty	KNDI	N/A
Task III subgroup	weekly	1 hr	Students, relevant faculty	KNDI+telecon	N/A
Seminar Series	weekly	1 hr	Relevant faculty, trainees	KNDI	N/A
Blog post	monthly	N/A	Project manager	N/A	B
Online content	monthly	1 hr	Relevant faculty, educational staff, and consultants	JHSPH	G2
Site visits to collaborator	bi-monthly	2-3 days	Field engineer	See map	E
Hackathon	bi-monthly	1 day	Students, postdocs, software engineers, relevant faculty	KNDI	E
Scholar Visitor Program	bi-monthly	5 days	Collaborating trainee	KNDI	N/A
SfN symposium	annually	1/2 day	Relevant faculty & 350+ attendees anticipated	SfN	E
SfN booth	annually	5 days	Relevant faculty, interested parties, up to 30,000 visitors	SfN	E
NIPS workshop	annually	2 days	Relevant faculty & 50+ attendees anticipated	NIPS	E
Bootcamp	annually	10 days	Relevant faculty, TAs, & ~30 post college attendees	MBL	N/A
Stakeholders	monthly	1 hr	All faculty	KNDI+telecon	N/A
Collab' Coordination	quarterly	1 hr	Vogelstein, Miller, unfunded collaborators	KNDI+telecon	N/A
Retreat	annually	1 day	All parties	SfN+telecon	E
NeuroNex	annually	3 days	PI and Co-PIs	NSF	E

References

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

2. Forbes DA. So you want to be a professional astronomer! [Internet]. arXiv [astro-ph]. 2008. Available from: http://arxiv.org/abs/0805.2624

3. Burns R, Vogelstein JT, Szalay AS. From cosmos to connectomes: the evolution of data-intensive science. Neuron. 2014 Sep;83(6):1249–52.

4. Vladimirov N, Mu Y, Kawashima T, Bennett DV, Yang C-T, Looger LL, Keller PJ, Freeman J, Ahrens MB. Light-sheet functional imaging in fictively behaving zebrafish. Nat Methods. 2014 Sep;11(9):883–4.

5. Poldrack RA, Barch DM, Mitchell JP, Wager TD, Wagner AD, Devlin JT, Cumba C, Koyejo O, Milham MP. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 2013 Jul 8;7:12.

6. Crawford KL, Neu SC, Toga AW. The Image and Data Archive at the Laboratory of Neuro Imaging. Neuroimage. 2016 Jan 1;124(Pt B):1080–3.

7. Mennes M, Biswal BB, Castellanos XF, Milham MP. Making data sharing work: The FCP/INDI experience. Neuroimage [Internet]. 2012 Oct 30;null(null). Available from: http://dx.doi.org/10.1016/j.neuroimage.2012.10.064

8. Das S, Glatard T, MacIntyre LC, Madjar C, Rogers C, Rousseau M-E, Rioux P, MacFarlane D, Mohades Z, Gnanasekaran R, Makowski C, Kostopoulos P, Adalat R, Khalili-Mahani N, Niso G, Moreau JT, Evans AC. The MNI data-sharing and processing ecosystem. Neuroimage. 2016 Jan 1;124(Pt B):1188–95.

9. Teeters JL, Harris KD, Millman KJ, Olshausen BA, Sommer FT. Data sharing for computational neuroscience. Neuroinformatics. 2008 Jan;6(1):47–55.

10. Burns R, Lillaney K, Berger D, Deisseroth K, Kazhdan M, Szalay AS, Gray Roncal W, Manavalan P, Bock DD, Grosenick L, Lichtman JW, Vogelstein JT, Kleissas DM, Perlman E, Chung K, Kasthuri N, Reid RC, Vogelstein RJ. The Open Connectome Project Data Cluster: Scalable Analysis and Vision for High-Throughput Neuroscience. In: Scientific and Statistical Database Management [Internet]. 2013. Available from: http://arxiv.org/abs/1306.3543

11. Teeters JL, Godfrey K, Young R, Dang C, Friedsam C, Wark B, Asari H, Peron S, Li N, Peyrache A, Denisov G, Siegle JH, Olsen SR, Martin C, Chun M, Tripathy S, Blanche TJ, Harris K, Buzsáki G, Koch C, Meister M, Svoboda K, Sommer FT. Neurodata Without Borders: Creating a Common Data Format for Neurophysiology. Neuron. 2015 Nov 18;88(4):629–34.

12. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, Flandin G, Ghosh SS, Glatard T, Halchenko YO, Handwerker DA, Hanke M, Keator D, Li X, Michael Z, Maumet C, Nichols BN, Nichols TE, Pellman J, Poline J-B, Rokem A, Schaefer G, Sochat V, Triplett W, Turner JA, Varoquaux G, Poldrack RA. The brain imaging data structure, a

format for organizing and describing outputs of neuroimaging experiments. Sci Data. 2016 Jun 21;3:160044.

13. Gronenschild EHBM, Habets P, Jacobs HIL, Mengelers R, Rozendaal N, van Os J, Marcelis M. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One. 2012 Jun 1;7(6):e38234.

14. Chiu K, Chiu KP, editors. Galaxy Pipeline for Transcriptome Library Analysis. In: Next-Generation Sequencing and Sequence Data Analysis. BENTHAM SCIENCE PUBLISHERS; 2015. p. 128–46.

15. Rex DE, Ma JQ, Toga AW. The LONI Pipeline Processing Environment. Neuroimage. 2003 Jul;19(3):1033–48.

16. Ward B. The book of VMware: the complete guide to VMware workstation. 2002;

17. Merkel D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux J [Internet]. 2014 Mar;2014(239). Available from: http://dl.acm.org/citation.cfm?id=2600239.2600241

18. Rosenberg DM, Horn CC. Neurophysiological analytics for all! Free open-source software tools for documenting, analyzing, visualizing, and sharing using electronic notebooks. J Neurophysiol. 2016 Aug 1;116(2):252–62.

19. Lavoie-Courchesne S, Rioux P, Chouinard-Decorte F, Sherif T, Rousseau M-E, Das S, Adalat R, Doyon J, Craddock RC, Margulies DS, Chu C, Lyttelton O, Evans AC, Bellec P, Lavoie-Cour. Integration of a neuroimaging processing pipeline into a pan-canadian computing grid. Diversity . 2012 Feb;341(i):12032.

20. Bernstein D. Containers and Cloud: From LXC to Docker to Kubernetes. IEEE Cloud Computing. 2014;1(3):81–4.

21. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004 Nov 22;20(17):3045–54.

22. Callahan SP, Freire J, Santos E, Scheidegger CE, Silva CT, Vo HT. VisTrails: Visualization Meets Data Management. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: ACM; 2006. p. 745–7. (SIGMOD '06).

23. Mouallem P, Barreto R, Klasky S, Podhorszki N, Vouk M. Tracking Files in the Kepler Provenance Framework. In: Winslett M, editor. Scientific and Statistical Database Management. Springer Berlin Heidelberg; 2009. p. 273–82. (Lecture Notes in Computer Science).

24. Pashler H, Wagenmakers EJ. Editors' introduction to the special section on replicability in psychological science a crisis of confidence? Perspect Psychol Sci [Internet]. 2012; Available from: http://pps.sagepub.com/content/7/6/528.short

25. Baker M. Over half of psychology studies fail reproducibility test. Nature Online. 2015;

26. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733–9.

27. Parker HS, Leek JT. The practical effect of batch on genomic prediction. Stat Appl Genet Mol Biol. 2012;11(3):Article 10.

28. Mirzaalian H, Ning L, Savadjiev P, Pasternak O, Bouix S, Michailovich O, Grant G, Marx CE, Morey RA, Flashman LA, George MS, McAllister TW, Andaluz N, Shutter L, Coimbra R, Zafonte RD, Coleman MJ, Kubicki M, Westin CF, Stein MB, Shenton ME, Rathi Y. Inter-site and inter-scanner diffusion MRI data harmonization. Neuroimage. 2016 Jul 15;135:311–23.

29. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. In: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. USENIX Association; 2010. p. 10.

30. Zheng D, Mhembere D, Burns R, Vogelstein JT, Priebe CE, Szalay AS. FlashGraph: Processing Billion-Node Graphs on an Array of Commodity SSDs. In: File and Storage Technologies [Internet]. 2014. Available from: http://arxiv.org/abs/1408.0500

31. Buchanan CR, Pernet CR, Gorgolewski KJ, Storkey AJ, Bastin ME. Test–retest reliability of structural brain networks from diffusion MRI. Neuroimage [Internet]. 2013; Available from: http://www.sciencedirect.com/science/article/pii/S1053811913009907

32. Shou H, Eloyan A, Lee S, Zipunnikov V, Crainiceanu AN, Nebel NB, Caffo B, Lindquist MA, Crainiceanu CM. Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2). Cogn Affect Behav Neurosci. 2013 Dec;13(4):714–24.

33. Yue C, Chen S, Sair HI, Airan R, Caffo BS. Estimating a graphical intra-class correlation coefficient (GICC) using multivariate probit-linear mixed models. Comput Stat Data Anal. 2015 Sep;89:126–33.

34. Peng R. The reproducibility crisis in science: A statistical counterattack. Significance. 2015 Jun 1;12(3):30–2.

35. Bhatla N, Droste R, Sando SR, Huang A, Horvitz HR. Distinct Neural Circuits Control Rhythm Inhibition and Spitting by the Myogenic Pharynx of C. elegans. Curr Biol. 2015 Aug 17;25(16):2075–89.

36. Bock DD, Lee W-CA, Kerlin AM, Andermann ML, Wetzel AW, Yurgenson S, Soucy ER, Kim HS, Hood G, Reid RC. Network anatomy and in vivo physiology of visual cortical neurons. Nature. 2011 Mar 10;471(7337):177–82.

37. Harris KM, Spacek J, Bell ME, Parker PH, Lindsey LF, Baden AD, Vogelstein JT, Burns R. A resource from 3D electron microscopy of hippocampal neuropil for user training and tool development. Scientific Data. 2015 Aug;2:150046.

38. Kasthuri N, Hayworth KJ, Berger DR, Schalek RL, Conchello JA, Knowles-Barley S, Lee D, Vazquez-Reina, Kaynig-Fittkau V, Jones TR, Roberts M, Morgan JL, Tapia JC, Seung HS, Gray Roncal W, Vogelstein JT, Burns R, Sussman DL, Priebe CE, Pfister H, Lichtman JW. Saturated Reconstruction of a Small Volume of Neocortex. Cell. 2015 Jul;162:648–61.

39. Lee W-CA, Bonin V, Reed M, Graham BJ, Hood G, Glattfelder K, Reid RC. Anatomy and function of an excitatory network in the visual cortex. Nature. 2016 Apr 21;532(7599):370–4.

40. Ohyama T, Schneider-Mizell CM, Fetter RD, Aleman JV, Franconville R, Rivera-Alba M, Mensh BD, Branson KM, Simpson JH, Truman JW, Cardona A, Zlatic M. A multilevel multimodal circuit enhances action selection in Drosophila. Nature. 2015 Apr 30;520(7549):633–9.

41. Bloss EB, Cembrowski MS, Karsh B, Colonell J, Fetter RD, Spruston N. Structured Dendritic Inhibition Supports Branch-Selective Integration in CA1 Pyramidal Cells. Neuron. 2016 Mar 2;89(5):1016–30.

42. Collman F, Buchanan J, Phend KD, Micheva KD, Weinberg RJ, Smith SJ. Mapping synapses by conjugate light-electron array tomography. J Neurosci. 2015 Apr 8;35(14):5792–807.

43. Weiler NC, Collman F, Vogelstein JT, Burns R, Smith SJ. Molecular architecture of barrel column synapses following experience-dependent plasticity. Nature Scientific Data. 2014;

44. Chen F, Tillberg PW, Boyden ES. Optical imaging. Expansion microscopy. Science. 2015 Jan 30;347(6221):543–8.

45. Dyer EL, Roncal WG, Fernandes HL, Gürsoy D, De Andrade V, Vescovi R, Fezzaa K, Xiao X, Vogelstein JT, Jacobsen C, Körding KP, Kasthuri N. Quantifying mesoscale neuroanatomy using X-ray microtomography [Internet]. arXiv [q-bio.QM]. 2016. Available from: http://arxiv.org/abs/1604.03629

46. Kutten KS, Vogelstein JT, Charon N, Ye L, Deisseroth K, Miller MI. Deformably registering and annotating whole CLARITY brains to an atlas via masked LDDMM. In: SPIE Photonics Europe. International Society for Optics and Photonics; 2016. p. 989616–989616 – 9.

47. Tomer R, Ye L, Hsueh B, Deisseroth K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. Nat Protoc. 2014 Jul;9(7):1682–97.

48. Tomer R, Lovett-Barron M, Kauvar I, Andalman A, Burns VM, Sankaran S, Grosenick L, Broxton M, Yang S, Deisseroth K. SPED Light Sheet Microscopy: Fast Mapping of Biological System Structure and Function. Cell. 2015 Dec;163(7):1796–806.

49. Randlett O, Wee CL, Naumann EA, Nnaemeka O, Schoppik D, Fitzgerald JE, Portugues R, Lacoste AMB, Riegler C, Engert F, Schier AF. Whole-brain activity mapping onto a zebrafish brain atlas. Nat Methods. 2015 Nov;12(11):1039–46.

50. Grabner G, Janke AL, Budge MM, Smith D, Pruessner J, Collins DL. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. Med Image Comput Comput Assist Interv. 2006;9(Pt 2):58–66.

51. Gray Roncal W, Kleissas DM, Vogelstein JT, Manavalan P, Lillaney K, Pekala M, Burns R, Vogelstein RJ, Priebe CE, Chevillet MA, Hager GD. An automated images-to-graphs framework for high resolution connectomics. Front Neuroinform. 2015 Jan 13;9:20.

52. Gray Roncal W, Pekala M, Kaynig-Fittkau V, Kleissas DM, Vogelstein JT, Pfister H, Burns

R, Vogelstein RJ, Chevillet MA, Hager GD. VESICLE : Volumetric Evaluation of Synaptic Inferfaces using Computer vision at Large Scale. In: 26th British Machine Vision Conference (BMVC). 2015. p. 1–9.

53. Takemura S-Y, Bharioke A, Lu Z, Nern A, Vitaladevuni S, Rivlin PK, Katz WT, Olbris DJ, Plaza SM, Winston P, Zhao T, Horne JA, Fetter RD, Takemura S, Blazek K, Chang L-A, Ogundeyi O, Saunders MA, Shapiro V, Sigmund C, Rubin GM, Scheffer LK, Meinertzhagen IA, Chklovskii DB. A visual motion detection circuit suggested by Drosophila connectomics. Nature. 2013 Aug 7;500(7461):175–81.

54. Venkatachalam V, Ji N, Wang X, Clark C. Pan-neuronal imaging in roaming Caenorhabditis elegans. Proceedings of the National Academy of Sciences [Internet]. 2016; Available from: http://www.pnas.org/content/113/8/E1082.short

55. Lemon WC, Pulver SR, Höckendorf B, McDole K, Branson K, Freeman J, Keller PJ. Whole-central nervous system functional imaging in larval Drosophila. Nat Commun. 2015 Aug 11;6:7924.

56. Ahrens MB, Keller PJ. Whole-brain functional imaging at cellular resolution using light-sheet microscopy. Nat Methods. 2013 Mar 18;10:413–20.

57. Dunn TW, Gebhardt C, Naumann EA, Riegler C, Ahrens MB, Engert F, Del Bene F. Neural Circuits Underlying Visually Evoked Escapes in Larval Zebrafish. Neuron. 2016 Feb 3;89(3):613–28.

58. Bianco IH, Engert F. Visuomotor transformations underlying hunting behavior in zebrafish. Curr Biol. 2015 Mar 30;25(7):831–46.

59. Ahrens MB, Li JM, Orger MB, Robson DN, Schier AF, Engert F, Portugues R. Brain-wide neuronal dynamics during motor adaptation in zebrafish. Nature. 2012 May 24;485(7399):471–7.

60. Froudarakis E, Berens P, Ecker AS, Cotton RJ, Sinz FH, Yatsenko D, Saggau P, Bethge M, Tolias AS. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. Nat Neurosci. 2014 Jun;17(6):851–7.

61. Cotton RJ, Froudarakis E, Storer P, Saggau P, Tolias AS. Three-dimensional mapping of microcircuit correlation structure. Front Neural Circuits. 2013 Oct 10;7:151.

62. Reimer J, Froudarakis E, Cadwell CR, Yatsenko D, Denfield GH, Tolias AS. Pupil fluctuations track fast switching of cortical states during quiet wakefulness. Neuron. 2014 Oct 22;84(2):355–62.

63. Badea A, Kane L, Anderson RJ, Qi Y, Foster M, Cofer GP, Medvitz N, Buckley AF, Badea AK, Wetsel WC, Colton CA. The fornix provides multiple biomarkers to characterize circuit disruption in a mouse model of Alzheimer's disease. Neuroimage [Internet]. 2016 Aug 10; Available from: http://dx.doi.org/10.1016/j.neuroimage.2016.08.014

64. Setsompop K, Kimmlingen R, Eberlein E, Witzel T, Cohen-Adad J, McNab JA, Keil B, Tisdall MD, Hoecht P, Dietz P, Cauley SF, Tountcheva V, Matschl V, Lenz VH, Heberlein K, Potthast A, Thein H, Van Horn J, Toga A, Schmitt F, Lehne D, Rosen BR, Wedeen V, Wald LL. Pushing the limits of in vivo diffusion MRI for the Human Connectome Project.

Neuroimage. 2013 Oct 15;80:220–33.

65. Fan Q, Witzel T, Nummenmaa A, Van Dijk KRA, Van Horn JD, Drews MK, Somerville LH, Sheridan MA, Santillana RM, Snyder J, Hedden T, Shaw EE, Hollinshead MO, Renvall V, Zanzonico R, Keil B, Cauley S, Polimeni JR, Tisdall D, Buckner RL, Wedeen VJ, Wald LL, Toga AW, Rosen BR. MGH–USC Human Connectome Project datasets with ultra-high b-value diffusion MRI. Neuroimage. 2016 Jan 1;124, Part B:1108–14.

66. Jones AR, Overly CC, Sunkin SM. The Allen Brain Atlas: 5 years and beyond. Nat Rev Neurosci. 2009 Nov;10(11):821–8.

67. Vogelstein JT, Park Y, Ohyama T, Kerr RA, Truman JW, Priebe CE, Zlatic M. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. Science. 2014 Apr 25;344(6182):386–92.

68. Chung K, Deisseroth K. CLARITY for mapping the nervous system. Nat Methods. 2013 Jun;10(6):508–13.

69. Tomer R, Ye L, Hsueh B, Deisseroth K. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. Nat Protoc. 2014 Jul;9(7):1682–97.

70. Lerner TN, Shilyansky C, Davidson TJ, Evans KE, Beier KT, Zalocusky KA, Crow AK, Malenka RC, Luo L, Tomer R, Deisseroth K. Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. Cell. 2015 Jul 30;162(3):635–47.

71. Sylwestrak EL, Rajasethupathy P, Wright MA, Jaffe A, Deisseroth K. Multiplexed Intact-Tissue Transcriptional Analysis at Cellular Resolution. Cell. 2016 Feb 11;164(4):792–804.

72. Ye L, Allen WE, Thompson KR, Tian Q, Hsueh B, Ramakrishnan C, Wang A-C, Jennings JH, Adhikari A, Halpern CH, Witten IB, Barth AL, Luo L, McNab JA, Deisseroth K. Wiring and Molecular Features of Prefrontal Ensembles Representing Distinct Experiences. Cell. 2016 Jun 16;165(7):1776–88.

73. Economo MN, Clack NG, Lavis LD, Gerfen CR, Svoboda K, Myers EW, Chandrashekar J. A platform for brain-wide imaging and reconstruction of individual neurons. Elife. 2016 Jan 20;5:e10566.

74. Tillberg PW, Chen F, Piatkevich KD, Zhao Y, Yu C-CJ, English BP, Gao L, Martorell A, Suk H-J, Yoshida F, DeGennaro EM, Roossien DH, Gong G, Seneviratne U, Tannenbaum SR, Desimone R, Cai D, Boyden ES. Protein-retention expansion microscopy of cells and tissues labeled using standard fluorescent proteins and antibodies. Nat Biotechnol. 2016 Sep;34(9):987–92.

75. Chen F, Wassie AT, Cote AJ, Sinha A, Alon S, Asano S, Daugharthy ER, Chang J-B, Marblestone A, Church GM, Raj A, Boyden ES. Nanoscale imaging of RNA with expansion microscopy. Nat Methods. 2016 Aug;13(8):679–84.

76. Li A, Gong H, Zhang B, Wang Q, Yan C, Wu J, Liu Q, Zeng S, Luo Q. Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. Science. 2010 Dec;330(6009):1404–8.

77. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, van de

Lagemaat LN, Smith KA, Ebbert A, Riley ZL, Abajian C, Beckmann CF, Bernard A, Bertagnolli D, Boe AF, Cartagena PM, Chakravarty MM, Chapin M, Chong J, Dalley RA, Daly BD, Dang C, Datta S, Dee N, Dolbeare TA, Faber V, Feng D, Fowler DR, Goldy J, Gregor BW, Haradon Z, Haynor DR, Hohmann JG, Horvath S, Howard RE, Jeromin A, Jochim JM, Kinnunen M, Lau C, Lazarz ET, Lee C, Lemon TA, Li L, Li Y, Morris JA, Overly CC, Parker PD, Parry SE, Reding M, Royall JJ, Schulkin J, Sequeira PA, Slaughterbeck CR, Smith SC, Sodt AJ, Sunkin SM, Swanson BE, Vawter MP, Williams D, Wohnoutka P, Zielke HR, Geschwind DH, Hof PR, Smith SM, Koch C, Grant SGN, Jones AR. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012 Sep 20;489(7416):391–9.

78. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ, Chen L, Chen L, Chen T-M, Chin MC, Chong J, Crook BE, Czaplinska A, Dang CN, Datta S, Dee NR, Desaki AL, Desta T, Diep E, Dolbeare TA, Donelan MJ, Dong H-W, Dougherty JG, Duncan BJ, Ebbert AJ, Eichele G, Estin LK, Faber C, Facer BA, Fields R, Fischer SR, Fliss TP, Frensley C, Gates SN, Glattfelder KJ, Halverson KR, Hart MR, Hohmann JG, Howell MP, Jeung DP, Johnson RA, Karr PT, Kawal R, Kidney JM, Knapik RH, Kuan CL, Lake JH, Laramee AR, Larsen KD, Lau C, Lemon TA, Liang AJ, Liu Y, Luong LT, Michaels J, Morgan JJ, Morgan RJ, Mortrud MT, Mosqueda NF, Ng LL, Ng R, Orta GJ, Overly CC, Pak TH, Parry SE, Pathak SD, Pearson OC, Puchalski RB, Riley ZL, Rockett HR, Rowland SA, Royall JJ, Ruiz MJ, Sarno NR, Schaffnit K, Shapovalova NV, Sivisay T, Slaughterbeck CR, Smith SC, Smith KA, Smith BI, Sodt AJ, Stewart NN, Stumpf K-R, Sunkin SM, Sutram M, Tam A, Teemer CD, Thaller C, Thompson CL, Varnam LR, Visel A, Whitlock RM, Wohnoutka PE, Wolkey CK, Wong VY, Wood M, Yaylaoglu MB, Young RC, Youngstrom BL, Yuan XF, Zhang B, Zwingman TA, Jones AR. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007 Jan 11;445(7124):168–76.

79. Thompson CL, Ng L, Menon V, Martinez S, Lee C-K, Glattfelder K, Sunkin SM, Henry A, Lau C, Dang C, Garcia-Lopez R, Martinez-Ferre A, Pombero A, Rubenstein JLR, Wakeman WB, Hohmann J, Dee N, Sodt AJ, Young R, Smith K, Nguyen T-N, Kidney J, Kuan L, Jeromin A, Kaykas A, Miller J, Page D, Orta G, Bernard A, Riley Z, Smith S, Wohnoutka P, Hawrylycz MJ, Puelles L, Jones AR. A high-resolution spatiotemporal atlas of gene expression of the developing mouse brain. Neuron. 2014 Jul 16;83(2):309–23.

80. Miller JA, Ding S-L, Sunkin SM, Smith KA, Ng L, Szafer A, Ebbert A, Riley ZL, Royall JJ, Aiona K, Arnold JM, Bennet C, Bertagnolli D, Brouner K, Butler S, Caldejon S, Carey A, Cuhaciyan C, Dalley RA, Dee N, Dolbeare TA, Facer BAC, Feng D, Fliss TP, Gee G, Goldy J, Gourley L, Gregor BW, Gu G, Howard RE, Jochim JM, Kuan CL, Lau C, Lee C-K, Lee F, Lemon TA, Lesnar P, McMurray B, Mastan N, Mosqueda N, Naluai-Cecchini T, Ngo N-K, Nyhus J, Oldre A, Olson E, Parente J, Parker PD, Parry SE, Stevens A, Pletikos M, Reding M, Roll K, Sandman D, Sarreal M, Shapouri S, Shapovalova NV, Shen EH, Sjoquist N, Slaughterbeck CR, Smith M, Sodt AJ, Williams D, Zöllei L, Fischl B, Gerstein MB, Geschwind DH, Glass IA, Hawrylycz MJ, Hevner RF, Huang H, Jones AR, Knowles JA, Levitt P, Phillips JW, Sestan N, Wohnoutka P, Dang C, Bernard A, Hohmann JG, Lein ES. Transcriptional landscape of the prenatal human brain. Nature. 2014 Apr 10;508(7495):199–206.

81. Bernard A, Lubbers LS, Tanis KQ, Luo R, Podtelezhnikov AA, Finney EM, McWhorter MME, Serikawa K, Lemon T, Morgan R, Copeland C, Smith K, Cullen V, Davis-Turak J, Lee C-K, Sunkin SM, Loboda AP, Levine DM, Stone DJ, Hawrylycz MJ, Roberts CJ, Jones

AR, Geschwind DH, Lein ES. Transcriptional architecture of the primate neocortex. Neuron. 2012 Mar 22;73(6):1083–99.

82. Neuroscience: An expression atlas of the developing macaque brain. Nat Methods. 2016 Aug 30;13(9):714–714.

83. Oh SW, Harris JA, Ng L, Winslow B, Cain N, Mihalas S, Wang Q, Lau C, Kuan L, Henry AM, Mortrud MT, Ouellette B, Nguyen TN, Sorensen SA, Slaughterbeck CR, Wakeman W, Li Y, Feng D, Ho A, Nicholas E, Hirokawa KE, Bohn P, Joines KM, Peng H, Hawrylycz MJ, Phillips JW, Hohmann JG, Wohnoutka P, Gerfen CR, Koch C, Bernard A, Dang C, Jones AR, Zeng H. A mesoscale connectome of the mouse brain. Nature. 2014 Apr 10;508(7495):207–14.

84. Ding S-L, Royall JJ, Sunkin SM, Ng L, Facer BAC, Lesnar P, Guillozet-Bongaarts A, McMurray B, Szafer A, Dolbeare TA, Stevens A, Tirrell L, Benner T, Caldejon S, Dalley RA, Dee N, Lau C, Nyhus J, Reding M, Riley ZL, Sandman D, Shen E, van der Kouwe A, Varjabedian A, Write M, Zollei L, Dang C, Knowles JA, Koch C, Phillips JW, Sestan N, Wohnoutka P, Zielke HR, Hohmann JG, Jones AR, Bernard A, Hawrylycz MJ, Hof PR, Fischl B, Lein ES. Comprehensive cellular-resolution atlas of the adult human brain. J Comp Neurol. 2016 Nov 1;524(16):3127–481.

85. Owens B. Montreal institute going “open” to accelerate science [Internet]. Science | AAAS. 2016 [cited 2016 Oct 6]. Available from: http://www.sciencemag.org/news/2016/01/montreal-institute-going-open-accel-erate-science

86. Amunts K, Lepage C, Borgeat L, Mohlberg H, Dickscheid T, Rousseau M-E, Bludau S, Bazin P-L, Lewis LB, Oros-Peusquens A-M, Shah NJ, Lippert T, Zilles K, Evans AC. BigBrain: An Ultrahigh-Resolution 3D Human Brain Model. Science. 2013 Jun 20;340(6139):1472–5.

87. Harlap J, Vins D, Evans AC, Turner JA. LORIS : a web-based data management system for multi-center studies. Neuroinformatics. 2012;5(January):1–11.

88. Cohn M. User Stories Applied: For Agile Software Development. Addison-Wesley Professional; 2004. 268 p.

89. Durante D, Dunson DB, Vogelstein JT. Nonparametric Bayes Modeling of Populations of Networks. arXiv [Internet]. 2014 Jun 30; Available from: http://arxiv.org/abs/1406.7851

90. Sikka S, Shehzad Z, Clark D, Khanuja R, Cheung B, Sebastian, Li Q, Lurie D, Vogelstein J, Tungaraza R, Craddock C, Watanabe A, Gorgolewski CF. C-PAC: CPAC Version 0.3.9 Alpha [Internet]. Zenodo; 2015. Available from: http://zenodo.org/record/16557

91. Kazhdan M, Surendran D, Hoppe H. Distributed Gradient-domain Processing of Planar and Spherical Images. ACM Trans Graph. 2010 Apr;29(2):14:1–14:11.

92. Kazhdan M, Gray Roncal W, Vogelstein JT, Vogelstein RJ, Kahzdan M, Burns R, Kasthuri N, Lichtman JW. Gradient-Domain Processing for Large EM Image Stacks. arXiv [Internet]. 2013; Available from: http://db.tt/yYfViFG5

93. Zheng D, Mhembere D, Vogelstein JT, Priebe CE, Burns R. FlashMatrix: Parallel, Scalable

Data Analysis with Generalized Matrix Operations using Commodity SSDs [Internet]. arXiv [cs.DC]. 2016. Available from: http://arxiv.org/abs/1604.06414

94. Saalfeld S, Fetter R, Cardona A, Tomancak P. Elastic volume reconstruction from series of ultra-thin microscopy sections. Nat Methods. 2012 Jul;9(7):717–20.

95. Klein A, Andersson J, Ardekani B a., Ashburner J, Avants B, Chiang M-C, Christensen GE, Collins DL, Gee J, Hellier P, Song JH, Jenkinson M, Lepage C, Rueckert D, Thompson PM, Vercauteren T, Woods RP, Mann JJ, Parsey RV. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage. 2009 Jul;46(3):786–802.

96. Klein A, Ghosh SS, Avants B, Yeo BTT, Fischl B, Ardekani B a., Gee JC, Mann JJ, Parsey RV. Evaluation of volume-based and surface-based brain image registration methods. Neuroimage [Internet]. 2010 Feb; Available from: http://dx.doi.org/10.1016/j.neuroimage.2010.01.091

97. Grenander U, Miller MI. Computational Anatomy: An Emerging Discipline. Quart Appl Math. 1998;56:617–94.

98. Beg MF, Miller MI, Trouvé A, Younes L. Computing Large Deformation Metric Mappings via Geodesic Flows of Diffeomorphisms. Int J Comput Vis. 2005;61(2):139–57.

99. Miller MI, Priebe CE, Qiu A, Fischl B, Kolasny A, Brown T, Park Y, Ratnanather JT, Busa E, Jovicich J, Yu P, Dickerson BC, Buckner RL. Collaborative computational anatomy: an MRI morphometry study of the human brain via diffeomorphic metric mapping. Hum Brain Mapp. 2009;30(7):2132–41.

100. Qiu A, Miller MI, Younes L, Glaunès J. Large Deformation Diffeomorphic Metric Curve Mapping. Int J Comput Vis. 2008 Dec;80(3):317–36.

101. Qiu A, Younes L, Miller MI. Principal component based diffeomorphic surface mapping. IEEE Trans Med Imaging. 2012 Feb;31(2):302–11.

102. Miller MI, Younes L, Trouvé A. Diffeomorphometry and geodesic positioning systems for human anatomy. Technology. 2014 Mar;2(1):36.

103. Ashburner J, Friston KJ. Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation. Neuroimage. 2011 Apr 1;55(3):954–67.

104. Aggarwal M, Manisha A, Wenzhen D, Zhipeng H, Neal R, Qi P, Ross CA, Miller MI, Susumu M, Jiangyang Z. Spatiotemporal mapping of brain atrophy in mouse models of Huntington's disease using longitudinal in vivo magnetic resonance imaging. Neuroimage. 2012;60(4):2086–95.

105. Selemon LD, Ceritoglu C, Ratnanather JT, Wang L, Harms MP, Aldridge K, Begović A, Csernansky JG, Miller MI, Rakic P. Distinct abnormalities of the primate prefrontal cortex caused by ionizing radiation in early or midgestation. J Comp Neurol. 2013 Apr 1;521(5):1040–53.

106. Kutten KS, Eacker SM, Dawson VL, Dawson TM, Tilak R, Miller MI. An image registration pipeline for analysis of transsynaptic tracing in mice. In: Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging [Internet]. 2016.

Available from: http://dx.doi.org/10.1117/12.2216233

107. Vaillant M, Marc V, Anqi Q, Joan G, Miller MI. Diffeomorphic metric surface mapping in subregion of the superior temporal gyrus. Neuroimage. 2007;34(3):1149–59.

108. Qiu A, Younes L, Miller MI, Csernansky JG. Parallel transport in diffeomorphisms distinguishes the time-dependent pattern of hippocampal surface deformation due to healthy aging and the dementia of the Alzheimer's type. Neuroimage. 2008 Mar 1;40(1):68–76.

109. Mori S, Susumu M, Dan W, Can C, Yue L, Anthony K, Vaillant MA, Faria AV, Kenichi O, Miller MI. MRICloud: Delivering High-Throughput MRI Neuroinformatics as Cloud-Based Software as a Service. Comput Sci Eng. 2016;18(5):21–35.

110. Stosiek C, Garaschuk O, Holthoff K, Konnerth A. In vivo two-photon calcium imaging of neuronal networks. Proc Natl Acad Sci U S A. 2003 Jun 10;100(12):7319–24.

111. Vogelstein JT, Packer AM, Machado TA, Sippy T, Babadi B, Yuste R, Paninski L. Fast nonnegative deconvolution for spike train inference from population calcium imaging. J Neurophysiol. 2010 Dec;104(6):3691–704.

112. Haeffele BD, Young ED, Vidal R. Structured low-rank matrix factorization. In: International Conference on Machine Learning [Internet]. jhu.pure.elsevier.com; 2014. Available from: https://jhu.pure.elsevier.com/en/publications/structured-low-rank-matrix-factorization-optimality-algorithm-and-4

113. Mishchenko Y, Vogelstein JT, Paninski L. A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. Ann Appl Stat. 2011;5(2B):129–1261.

114. Zhang T, Wu J, Li F, Caffo B, Boatman-Reich D. A Dynamic Directional Model for Effective Brain Connectivity using Electrocorticographic (ECoG) Time Series. J Am Stat Assoc. 2015 Jan 7;00–00.

115. Han F, Liu H, Caffo B. Sparse Median Graphs Estimation in a High Dimensional Semiparametric Model [Internet]. arXiv [stat.AP]. 2013. Available from: http://arxiv.org/abs/1310.3223

116. Qiu H, Han F, Liu H, Caffo B. Joint Estimation of Multiple Graphical Models from High Dimensional Time Series. J R Stat Soc Series B Stat Methodol. 2016 Mar 1;78(2):487–504.

117. Hedlin H, Boatman D, Caffo B. ESTIMATING TEMPORAL ASSOCIATIONS IN ELECTROCORTICOGRAPHIC (ECoG) TIME SERIES WITH FIRST ORDER PRUNING. 2010 [cited 2016 Oct 7]; Available from: http://biostats.bepress.com/jhubiostat/paper217/

118. Gray Roncal W, Koterba ZH, Mhembere D, Kleissas DM, Vogelstein JT, Burns R, Bowles AR, Donavos DK, Ryman S, Jung RE, Wu L, Calhoun VD, Vogelstein RJ. MIGRAINE: MRI Graph Reliability Analysis and Inference for Connectomics. Global Conference on Signal and Information Processing [Internet]. 2013; Available from: http://arxiv.org/abs/1312.4875v1

119. Gray Roncal W, Bogovic JA, Vogelstein JT, Landman BA, Prince JL, Vogelstein RJ. Magnetic resonance connectome automated pipeline: An overview. IEEE Pulse. 2010

Mar;3(2):42–8.

120. Goh A, Lenglet C, Thompson PM, Vidal R. A nonparametric Riemannian framework for processing high angular resolution diffusion images and its applications to ODF-based morphometry. Neuroimage. 2011 Jun 1;56(3):1181–201.

121. Goh A, Lenglet C, Thompson PM, Vidal R. Estimating orientation distribution functions with probability density constraints and spatial regularity. In: Medical Image Computing and Computer Assisted Interventions (MICCAI). 2009. p. 877–85.

122. Wolfers S, Schwab E, Vidal R. Nonnegative ODF estimation via optimal constraint selection. In: International Symposium on Biomedical Imaging (ISBI). 2014. p. 734–7.

123. Schwab E, Vidal R, Charon N. Spatial-Angular Sparse Coding for HARDI. In: Medical Image Computing and Computer Assisted Interventions (MICCAI) [Internet]. 2016 [cited 2016 Oct 7]. Available from: http://dx.doi.org/

124. Priebe CE, Coppersmith GA, Rukhin A. You say graph invariant, I say test statistic. Statistical Computing Statistical Graphics Newsletter. 2010;21(2):11–4.

125. Tang M, Athreya A, Sussman DL, Lyzinski V, Priebe CE. A semiparametric two-sample hypothesis testing problem for random dot product graphs. arXiv. 2014 Mar 8;44.

126. Lyzinski V, Tang M, Athreya A, Park Y, Priebe CE. Community Detection and Classification in Hierarchical Stochastic Blockmodels [Internet]. arXiv [stat.ML]. 2015. Available from: http://arxiv.org/abs/1503.02115

127. Fishkind DE, Sussman DL, Tang M, Vogelstein JT, Priebe CE. {Consistent Adjacency-Spectral Partitioning for the Stochastic Block Model. SIAM J Matrix Anal Appl. 2013 Jan;34(1):23–39.

128. Wang H, Zheng D, Burns R, Priebe CE. Active Community Detection in Massive Graphs. arXiv preprint [Internet]. 2014 Dec 30; Available from: http://arxiv.org/abs/1412.8576

129. Lyzinski V, Sussman DL, Tang M, Athreya A, Priebe CE. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. Electron J Stat. 2013 Oct 1;8(2):22.

130. Athreya A, Lyzinski V, Marchette DJ, Priebe CE, Sussman DL, Tang M. A limit theorem for scaled eigenvectors of random dot product graphs. arXiv. 2013 May 31;1305.7388.

131. Vogelstein JT, Gray Roncal W, Vogelstein RJ, Priebe CE. Graph Classification using Signal Subgraphs: Applications in Statistical Connectomics. IEEE Trans Pattern Anal Mach Intell. 2013;35(7):1539–51.

132. Vogelstein JT, Priebe CE. Shuffled Graph Classification: Theory and Connectome Applications. J Classification. 2015;32(1):3–20.

133. Vogelstein JT, Vogelstein RJ, Priebe CE. Are mental properties supervenient on brain properties? Sci Rep. 2009;1(100):11.

134. Koutra D, Vogelstein JT, Faloutsos C. DeltaCon: Measuring Connectivity Differences in Large Networks. In: SIAM International Conference on Data Mining [Internet]. 2013. Available from: http://knowledgecenter.siam.org/105SDM/1

135. Koutra D, Shah N, Vogelstein JT, Gallagher BJ, Faloutsos C. DeltaCon: A Principled Massive-Graph Similarity Function. ACM Trans Knowl Discov Data. 2015 May;

136. Binkiewicz N, Vogelstein JT, Rohe K. {Covariate Assisted Spectral Clustering}. arXiv [Internet]. 2014 Nov; Available from: http://arxiv.org/abs/1411.2158

137. Tang R, Ketcha M, Vogelstein JT, Priebe CE, Sussman DL. Law of Large Graphs [Internet]. arXiv [stat.ME]. 2016. Available from: http://arxiv.org/abs/1609.01672

138. Venkataraman A, Kubicki M, Golland P. From brain connectivity models to identifying foci of a neurological disorder. Med Image Comput Comput Assist Interv. 2012;15(Pt 1):715–22.

139. Venkataraman A, Kubicki M, Golland P. From connectivity models to region labels: identifying foci of a neurological disorder. IEEE Trans Med Imaging. 2013 Nov;32(11):2078–98.

140. Venkataraman A, Yang DY-J, Pelphrey KA, Duncan JS. Bayesian Community Detection in the Space of Group-Level Functional Differences. IEEE Trans Med Imaging. 2016 Aug;35(8):1866–82.

141. Sweet A, Andrew S, Archana V, Stufflebeam SM, Hesheng L, Naoro T, Joseph M, Polina G. Detecting Epileptic Regions Based on Global Brain Connectivity Patterns. In: Lecture Notes in Computer Science. 2013. p. 98–105.

142. Zheng D, Mhembere D, Lyzinski V, Vogelstein JT, Priebe CE, Burns R. Semi-External Memory Sparse Matrix Multiplication on Billion-node Graphs in a Multicore Architecture. 2016 Feb 9; Available from: http://arxiv.org/abs/1602.02864

143. Mhembere D, Zheng D, Vogelstein JT, Priebe CE, Burns R. NUMA-optimized In-memory and Semi-external-memory Parameterized Clustering [Internet]. arXiv [cs.DC]. 2016. Available from: http://arxiv.org/abs/1606.08905

144. Su S-C, Caffo B, Garrett-Mayer E, Bassett SS. Modified test statistics by inter-voxel variance shrinkage with an application to f MRI. Biostatistics. 2009 Apr;10(2):219–27.

145. Shou H, Eloyan A, Nebel MB, Mejia A, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA, Crainiceanu CM. Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI. Neuroimage. 2014 Nov 15;102 Pt 2:938–44.

146. Mejia AF, Nebel MB, Shou H, Crainiceanu CM, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA. Improving reliability of subject-level resting-state fMRI parcellation with shrinkage estimators. Neuroimage. 2015 May 15;112:14–29.

147. Mejia AF, Nebel MB, Eloyan A, Caffo B, Lindquist MA. PCA leverage: outlier detection for high-dimensional functional magnetic resonance imaging data [Internet]. arXiv [stat.ME]. 2015. Available from: http://arxiv.org/abs/1509.00882

148. You C, Robinson D, Vidal R. Scalable sparse subspace clustering by orthogonal matching pursuit. IEEE Conference on Computer Vision [Internet]. 2016; Available from: http://www.vision.jhu.edu/assets/CVPR16-SSCOMP.pdf

149. Haeffele BD, Vidal R. Global Optimality in Tensor Factorization, Deep Learning, and

Beyond. 2015 Jun 24; Available from: http://arxiv.org/abs/1506.07540

150. Ceyhan E, Priebe CE, Marchette DJ. A New Family of Random Graphs for Testing Spatial Segregation. Can J Stat. 2007;35(1):27–50.

151. Mohan NR, Priebe CE, Park Y, John M. Statistical Analysis of Hippocampus Shape Using a Modified Mann-Whitney-Wilcoxon Test. International Journal of Bio-Science and Bio-Technology. 2011;3(1):19–26.

152. Wang H, Tang M, Park Y, Priebe CE. Locality Statistics for Anomaly Detection in Time Series of Graphs. IEEE Trans Signal Process. 2013 Feb 1;62(3):703–17.

153. Trosset MW, Gao M, Priebe CE. On the power of likelihood ratio tests in dimension-restricted submodels. J Am Stat Assoc. 2016;https://arxiv.org/abs/1608.00032.

154. Priebe CE, Vogelstein JT, Bock DD. Optimizing the quantity/quality trade-off in connectome inference. Commun Stat Theory Methods. 2013;7.

155. Priebe CE, Sussman DL, Tang M, Vogelstein JT. Statistical inference on errorfully observed graphs. arXiv preprint [Internet]. 2013; Available from: http://arxiv.org/abs/1211.3601

156. Di C-Z, Crainiceanu CM, Caffo BS, Punjabi NM. MULTILEVEL FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS. Ann Appl Stat. 2009 Mar 1;3(1):458–88.

157. Zipunnikov V, Greven S, Shou H, Caffo B, Reich DS, Crainiceanu C. Longitudinal High-Dimensional Principal Components Analysis with Application to Diffusion Tensor Imaging of Multiple Sclerosis. Ann Appl Stat. 2014;8(4):2175–202.

158. Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A. 2016 Jul 12;113(28):7900–5.

159. Eklund A, Andersson M, Josephson C, Johannesson M, Knutsson H. Does parametric fMRI analysis with SPM yield valid results?-An empirical study of 1484 rest datasets. Neuroimage. 2012 Apr 10;61(3):565–78.

160. Venkataraman A, Yang DY-J, Pelphrey KA, Duncan JS. Bayesian Community Detection in the Space of Group-Level Functional Differences. IEEE Trans Med Imaging. 2016 Aug;35(8):1866–82.

161. Mhembere D, Gray Roncal W, Sussman DL, Priebe CE, Jung RE, Ryman S, Vogelstein RJ, Vogelstein JT, Burns R. Computing Scalable Multivariate Glocal Invariants of Large (Brain-) Graphs. In: Global Conference on Signal and Information Processing [Internet]. 2013. Available from: http://arxiv.org/abs/1312.4318

162. Marder E. The haves and the have nots. Elife. 2013 Nov 19;2:e01515.

163. Vogelstein JT, Amunts K, Andreou A, Angelaki D, Ascoli G, Bargmann C, Burns R, Cali C, Chance F, Chun M, Church G, Cline H, Coleman T, de La Rochefoucauld S, Denk W, Elgoyhen AB, Cummings RE, Evans A, Harris K, Hausser M, Hill S, Inverso S, Jackson C, Jain V, Kass R, Kasthuri B, Kording K, Koushika S, Krakauer J, Landis S, Layton J, Luo Q, Marblestone A, Markowitz D, McArthur J, Mensh B, Milham M, Mitra P, Neskovic P,

Nicolelis M, O'Brien R, Oliva A, Orban G, Peng H, Picchini-Schaffer A, Picciotto M, Poline J-B, Poo M-M, Pouget A, Raghavachari S, Roskams J, Sejnowski T, Sommer F, Spruston N, Swanson L, Toga A, Jacob Vogelstein R, Yuste R, Zador A, Huganir R, Miller M. Grand Challenges for Global Brain Sciences [Internet]. arXiv [q-bio.NC]. 2016. Available from: http://arxiv.org/abs/1608.06548

164. Adali S, Priebe CE. Fidelity-Commensurability Tradeoff in Joint Embedding of Disparate Dissimilarities [Internet]. arXiv [stat.ME]. 2013. Available from: http://arxiv.org/abs/1306.1977

165. Lyzinski V, Park Y, Priebe CE, Trosset MW. Fast Embedding for JOFC Using the Raw Stress Criterion [Internet]. arXiv [stat.ML]. 2015. Available from: http://arxiv.org/abs/1502.03391

166. Yoder J, Priebe CE. A model-based semi-supervised clustering methodology. J Am Stat Assoc. 2016;https://arxiv.org/abs/1412.4841.

167. Eichler K, Kumar AL, Park Y, Andrade I, Schneider-Mizell C, Saumweber T, Huser A, Bonnery D, Gerber B, Fetter RD, Truman JW, Priebe CE, Abbott LF, Thum A, Zlatic M, Cardona A. The complete wiring diagram of a high-order learning and memory center, the insect mushroom body. Nature. 2016;in review.

FACILITIES, EQUIPMENT AND OTHER RESOURCES

JOHNS HOPKINS UNIVERSITY

CENTER FOR IMAGING SCIENCE

PI Vogelstein, Co-PI Miller and Priebe, and Faculty Associates Vidal, and Ratnanather are all core members of the Center for Imaging Science (CIS), with their primary offices in the center. Co-PI Miller founded CIS in 1998, and continues to be the director.

Laboratory: CIS laboratory facilities on the 3rd floor of Clark Hall include 1,545 sq. ft. of graphics and computing laboratory and 3,582 sq. ft. of student laboratory/office space.

Clinical: Not Applicable

Animal: Not Applicable

Computer: CIS requires extensive computational resources for visualization and large storage management, and employs a full time information technologies specialist to ensure everything runs smoothly. The visualization workstations are used to develop volume-rendering algorithms. CIS has an open computer lab with 20 high-end, graphic workstations that are available and used by visiting faculty and scholars, and graduate and undergraduate students. In addition, CIS has 40 high-end, graphic workstations dedicated to research and located in the offices of faculty, graduate students, research technicians and research programmers, as well as three workstations used to support Center administration. The CIS internal infrastructure supports a gigabit network infrastructure to desktop; 800TB of data storage and 20 node development cluster. CIS also has an allocation of 3 million compute hours and a year of Extended Collaborative Support Services on XSEDE, an NSF supported compute cluster, and 2 million computer hours on MARCC (described below). The faculty and personnel on this proposal have been utilizing these resources in preliminary work for this proposal, as well as the funded CRCNS grant (CRCNS-1208044; co-PIs Vogelstein, Burns, et al.) and Big Data grant (BIGDATA-1251208; co-PIs Vogelstein, Burns, et al.), a CRCNS grant on computational infrastructures upon which the methods developed herein can operate at accelerated rates (NSF Proposal ID-1311505), and several others.

Office: CIS has 1,686 sq. ft. of offices and 651 sq. ft. of conference room and administrative areas. This space combined with the laboratory space provides a total of 7,500 sq. ft. It is a matter of policy that student space is not allocated to any individual investigator in CIS. Rather, we believe in the collegiality and intellectual cross-fertilization that takes place when students from different research laboratories occupy joint space. There is sufficient room for each investigator to comfortably support their lab.

Other Major Equipment: Not Applicable

Other Resources: The Center for Imaging Science is a multidisciplinary center comprising 10 core faculty, including 7 NIH-funded principal investigators, and 9 NSF-funded principal investigators, spanning the areas of biomedical engineering, applied mathematics and statistics and computer science. In addition to the sharing of high performance computing resources, CIS offers a rich intellectual environment, allowing for useful discussions among investigators, postdoctoral fellows and graduate students. This vibrant research environment has allowed for a number of fruitful collaborations among the different groups. In addition, CIS has strong relationships with the Kavli Neuroscience Discovery Institttue (KNDI), Institute for Computational Medicine (ICM), and the Center for Brain Imaging Science, which carry great presence on both the undergraduate and medical campuses. PI Vogelstein and co-PIs Priebe, Miller, and Vidal are all core faculty

in the Center for Imaging Science (CIS). Dr. Miller works hand in hand with Dr. Susumu Mori who directs the Center for Brain Imaging Science and the JHU High field scanning animal facilities, and Drs. Miller and Vogelstein are both members of the Steering committee for the KNDI. Dr. Miller is also co-Director of KNDI with Rick Huganir Chairman of Neuroscience.

Kavli Neuroscience Discovery Institute Hubs

Kavli Neuroscience Discovery Institute (KNDI) was founded in 2016, and is Co-Directed by Co-PI Miller. PI Vogelstein and Co-PI Burns are on the steering committee along with Co-PI Miller.

Laboratory: Not Applicable

Clinical: Not Applicable

Animal: Not Applicable

Computer: Not Applicable

Office: The Kavli Hub on the East Baltimore Campus (>2400 sq ft) includes a conference room with a live connection to the Homewood Campus Kavli Hub, to extend the casual "water cooler" effect across campuses, bridging the divide between neuroscience, neurodata, and neuroengineering. Additional adjacent space is specifically designed for students and postdoctoral fellows to gather. It will feature a small lounge area with grouped couches and chairs and smartboards to allow for informal conversations (also connecting the two hubs), small offices for studying or brainstorming, and a kitchen area. This space will have a video conferencing portal connected to the Homewood Campus Kavli Hub that will be live 24/7, virtually linking the two campuses at all times. This to be renovated space includes small networking rooms for groups of collaborators to brainstorm. The counterpart, the Kavli Hub on the Homewood campus, is a >3700 sq ft facility located on the 3rd floor of Clark Hall, which is contiguous with the Center for Imaging Science space. Similar to the East Baltimore Campus Kavli Hub, this space will include a 1000 ft^2 conference room with a 24/7 video link from the Kavli Hub from East Baltimore Campus Kavli Hub with built in smartboards and 10 Gbit ethernet connections. This space will double as a lounge area with grouped couches and chairs to allow for informal discussions and small networking rooms for groups of collaborators to brainstorm, as in the East Baltimore facility. The lounge/videoconference area is connected to a kitchen area, with tables and chairs for people to eat and discuss. Adjacent space includes 4 large offices for long time visitors (e.g., for sabbaticals), 3 smaller offices for shorter duration visitors, and 20 student workstations surrounded by floor-to-ceiling glass whiteboards interspersed with smart TVs.

Other Major Equipment: Not Applicable

Other Resources: Kavli NDI at Johns Hopkins brings together neuroscience, engineering, and data science – three traditional strengths at Johns Hopkins – with the ultimate goal of reaching a unified understanding of brain function. The members of Kavli Neuroscience Discovery Institute at Johns Hopkins are located across the University: in the Schools of Medicine and Public Health on the East Baltimore Campus and in the Schools of Engineering and Arts and Sciences on the Homewood Campus. These laboratories range from molecular biology and two photon imaging labs to human MRI imaging and computational labs to neuromorphic engineering labs.

Institute for Computational Medicine

PI Vogelstein, co-PI Miller and Faculty Associate Vidal, are also core faculty also in the Institute for Computational Medicine (ICM), have access to their space, personnel, and computing resources.

Laboratory: Not Applicable

Clinical: Not Applicable

Animal: Not Applicable

Computer: The Institute has a dedicated systems administrator who maintains the computational resources, including the server room of 1,000 sq. ft of space. This room is equipped with two 25 ton Liebert climate control systems and has a 10GbE backbone, providing connectivity to the JHU Homewood campus core, which is one hop away. The room has a dedicated Cisco 4500 series switch, which will provide 10GbE connectivity to the equipment housed there. All networking equipment is managed and monitored by Johns Hopkins Enterprise Networking, which provides 24/7 management of support to all Johns Hopkins networks. Dr. Sarma also has access to ICM's 256 dual quad-core node cluster, configured with 1 PB storage.

Office: The ICM occupies approximately 12,000 net square feet of space on the 2nd and 3rd floors of Hackerman Hall. Facilities include 12 faculty offices, six student labs adjacent to faculty mentor offices, three dedicated conference rooms, and an administrative suite. It is a matter of policy that student space is not allocated to any individual investigator in the ICM. Rather, we believe in the collegiality and intellectual cross-fertilization that takes place when students from different research laboratories occupy joint space. There is sufficient room for each investigator to comfortably support their lab.

Three ICM conference rooms are available for class meetings in Hackerman Hall, each with seating capacity up to ~15. The Dean's Conference Room in Hackerman Hall can seat up to ~60 students and can be scheduled for classes and seminars. The Hackerman Hall Basement Auditorium has staggered seating and dual screen/projector systems for larger audiences, seating up to 125 people.

Equipment: ICM operates a dedicated machine room on the 2nd floor of Hackerman Hall that are *freely available to all ICM core faculty and their trainees*: This machine room is designed for projects involving HIPAA-controlled, individual-level data access. The room is 1,000 net square feet, with space for 22 standard 42" x 24" racks in 3 aisles. A dedicated 300kVA transformer supplies power. Two 25-ton computer room air-conditioning units provide cooling, with facilities and space available for a third. This equipment is managed and monitored by Johns Hopkins University Plant Operations 24/7 and is configured to generate audible, visual, and cell phone alarms in cases of emergency (e.g., loss of chilled water or air conditioning failures leading to increased room temperature). These alarms are sent to the ICM Systems Administrator and JHU Plant Operations. Hackerman Hall has a 10GbE backbone, providing 10GbE connectivity to the JHU Homewood campus core, which is one hop away. The machine room has a dedicated Cisco 4500 series switch, providing 10GbE connectivity to the equipment housed there. All networking equipment is managed and monitored by Johns Hopkins Enterprise Networking, which provides 24/7 management of support to all Johns Hopkins networks. Uninterruptible power supplies are installed in each rack as necessary. The University provides all power and machine room maintenance. This machine room has carefully controlled card key access, recorded by the Johns Hopkins University Office of Security Management. A camera is used to maintain a visual and time-stamped record of entry and egress to the room at all times.
Resources in the room comprise:

- One 250-node IBM iDataPlex compute cluster. The compute nodes each have 2 Intel Xeon E5472 (Seaborg chipset) 3.00GHz Quad Core processors, for a total of 8 processors per node. 16 nodes have 64 GBs of memory. The remaining compute nodes each have 32 GB memory. Each compute node communicates with the SAN via a 4X DDR InfiniBand switched fabric. GPFS, IBM's high-performance parallel file management solution, provides file system access.
- 1 Petabyte storage area network with fiber channel connection to the cluster
- One IBM 3584 Ultrascalable Tape Library with 8 LTO4 drives compressed storage capacity 800 TBs. Tivoli Storage Manager 6.3 is used to manage backup, archive, and restore for all systems.

Other Resources: The ICM was chartered in 2005 and is an Institute in both the Whiting School of Engineering (WSE) and the School of Medicine (SOM). The ICM comprises 20 faculty members from the Departments of Neurosurgery, Biomedical Engineering, Applied Mathematics and Statistics, Emergency Medicine, Mechanical Engineering, and Computer Science. In addition, the Institute employs five research faculty and eight research staff members. The ICM reports administratively to the Whiting School of Engineering. The ICM has a full-time systems administrator, a full-time administrative manager, and three finance and administrative staff.

Center for Bioengineering Innovation and Design

PI Vogelstein uses the Center for Bioengineering Innovation and Design (CBID) facilities for his NeuroData Design course. In its first year the course enrollment has been capped at 12 enrolled students and 1 TA. Reception has been quite positive, and enrollment is expected to grow as space facilitates.

Laboratory: Not Applicable

Clinical: Not Applicable

Animal: Not Applicable

Computer: Readily available departmental software includes Solidworks, Matlab, SimBiosys and LabView. A computational laboratory for all JHU students is rich in finite element, CAD and animation software.

Office: Our school has recently constructed a 5,000 square foot design studio at our undergraduate campus that is open 24 hours a day for team use, including a conference room. A second design studio is available for DT students at the medical school campus ("The Carnegie Center for Surgical Innovation"). This studio will serve as a home base for the clinical and translational immersion programs and has tables with seating for 24 students, as well as basic hand tools and instrumentation equipment. This studio is frequently used for project committee meetings since it does not require clinicians to travel off site. Both of these studios were primarily built for teaching BME Design Teams and our students have priority for their use.

Equipment: The design studio at our undergraduate campus contains a large workshop area, a wet-lab, a fume hood, a machine shop, sewing machines, six 3D printers, a 3D scanner, mobile instrumentation carts, microscopes, equipment for injection molding, soldering stations, and an electronics shop. Students also have access to an additional 5,000 square foot machine shop ("The WSE Advanced Manufacturing Center") for large or complex machining.

Other Resources: CBID is a joint effort of the Whiting School of Engineering and the School of Medicine at Johns Hopkins University. Our focus is translational engineering focusing on health care technology. No other program combines clinical immersion, travel to developing countries for global health immersion, state-of-the-art facilities, and world-renowned lecturers to inspire students to deliver real innovation. CBID receives institutional support and commitment from the BME department, the Whiting School of Engineering, and the JHU School of Medicine in order to educate the 120 Design Team (DT) students in the Fall semester and 190 DT students in the spring semester. The DT program receives personnel support from the BME department in the form of 2.5 FTE instructors, 14 MSE TAs that mentor one DT each, one FTE for course administration and logistics, and the CBID organization supports additional mentors for smaller roles in the instruction of our students.

Johns Hopkins University, Bloomberg School of Public Health

Laboratory: Not Applicable

Clinical: Not Applicable

Animal: Not Applicable

Computer: The computing resources in the Department of Biostatistics consists of a joint high performance computing (HPC) cluster, dedicated department web server and file sharing server, faculty research servers, and personal desktop/laptop for each faculty and staff. As of September 2015 the HPC cluster has 67 nodes with a total of 2496 cores and about 19 TB DDR-SDRAM in production. The total mass storage is about 1.9PB. Home directories are backed up to a Lustre storage file system. This shared resource provides the needs for compute-intensive and data intensive research and teaching.

The dedicated department web server provides a platform for users to provide teaching resources and present their research work on the internet. The file sharing server allows effective collaboration. Some faculties have their own research servers to run special projects. Each faculty, staff, and student has either a desktop or a laptop with external monitor(s) in the office. They access the server(s) via secure shell (SSH) protocol. Except for standard operating systems, MS Office, and adobe acrobat, major statistical and mathematical computing packages are available, including R, SAS, Stata, Matlab, and Mathematica.

Office: All investigators on this project currently have office space in the Johns Hopkins Bloomberg School of Public Health within the Department of Biostatistics.

Other Resources: The Department of Biostatistics in the Bloomberg School of Public Health is led by Dr. Karen Bandeen-Roche. Established in 1917, this is the oldest autonomous department of its kind in the world, and is among the most productive Departments of Biostatistics in research and in training of masters and doctoral students. The Department aims to promote effective statistical reasoning and applications in health research. It currently includes 37 full-time faculty, 40 part-time faculty, 50 doctoral candidates, 13 master degree and 12 joint MHS students, 12 postdoctoral fellows and 13 staff. Sixty percent of the department's Ph.D. graduates over the last 5 years have gone on to tenure-track academic positions, including at the Universities of Minnesota, Pennsylvania, and Washington, Stanford and Johns Hopkins. Faculty members spend roughly half their research time on developing statistical methods and on applications of substantive importance. Methodological research is conducted on a broad array of topics, including foundations of inference, clinical trials, longitudinal data analysis, latent variable modeling, spatial statistics, nonparametric smoothing methods for very large data streams, and statistical genetics and genomics. In addition to aging, the department has major applications in basic science, environmental health, epidemiology, health services research, ophthalmology, psychiatry, neurology, pediatrics, and oncology. The Department of Biostatistics offers educational programs leading to the Ph.D., Sc.M, and M.H.S. degrees and more than 65 graduate courses in various learning formats. Approximately half of the courses are designed for students outside of biostatistics; the other half are for students in biostatistics or related fields. In addition to course work, the department supports weekly seminars and "working groups" in which students and faculty interested in a particular topic meet biweekly for an informal seminar or discussion. Currently, working groups are active in aging, causal inference, environmental epidemiology, genomics, medical imaging and spectra, and longitudinal/multivariate data analysis. Students learn the application of statistics by collaborating with faculty in research on health or in brief consultation through The Johns Hopkins Biostatistics Center, the Department's unit devoted to consultation. The result is an active, engaging intellectual environment.

Data Science Lab: The JHU Biostatistics Data Science Lab is a center in the Department of Biostatistics comprised of faculty Roger Peng, Jeff Leek and Brian Caffo. The group boasts of a dedicated office for recording with professional recording equipment (HD cameras), audio equipment (boom, lapel and multichannel microphones), green screens, lighting, editing software (Adobe Create Suite) and computational resources (HD monitors and computing towers). The group also has a Wacom monitor with an active digitizer and screencasting software (Camtasia) for recording speech over handwriting.

Additional Computing Resources

NeuroData Mini-Cluster Dr. Burns and Dr. Vogelstein have established a NeuroData mini-cluster for development and rapid prototyping. In aggregate, the cluster consists of 11 machines with over 200 cores, 3.5TB memory, 100TB of traditional storage, 10TB of SSD storage. All of the hardware is connected to the Johns Hopkins Research Network with 10 Gigabit Ethernet. These machines are all actively used by members of both PI Vogelstein and co-PI Burns' labs.

Maryland Advanced Research Computing Center (MARCC): Using designated funds from the State of Maryland, Johns Hopkins, in partnership with the University of Maryland, has recently opened a new HPC facility called Maryland Advanced Research Computing Center: MARCC. MARCC is located at the East Baltimore Bayview Campus of Johns Hopkins University and its resources are available to all program faculty and trainees. The system initiated operation on July 1, 2015. Computational and storage resources will be available for major research efforts and to students in the context of coursework. The system already has an operational 100G connection to both the JHU core and to Internet2. Multiple members of the Burns, Vogelstein, and Priebe labs are regular users of MARCC, each lab has between 50,000 and 250,000 compute hours per quarter.

Bluecrab is the main cluster at MARCC with 21,000+ cores and a combined theoretical performance of 1.1 PFLOPs. The compute nodes are a combination of Intel Haswell, Ivy Bridge and Broadwell processors and Nvidia K80 GPUs. It also features two types of storage: 2 PB Lustre (Terascala) and 18 PB ZFS (under Linux). The standard compute nodes have dual Intel Xeon E5-2680v3 processors (12 cores at 2.5GHz) and 128 GB DDR4 memory. The GPU nodes are Dell PowerEdge R730 servers with dual Intel Xeon E5-2680v3 processors (12 cores at 2.5Ghz), 128GB DDR4 memory and two Nvidia K80 GPUs. The large memory nodes are Dell PowerEdge R920 servers with quad Intel Ivy Bridge Xeon E7-8857v2, (12 cores at 3.0GHz) with 1 TB RAM.

Networking: Johns Hopkins Research Network (HORNET) The Johns Hopkins Research Network (HORNET) is the collaborative network platform for MARCC. The HORNET architecture is built around a high-speed, 100-Gigabit full mesh core. Research systems can take full advantage of the MARCC research facility and efficiently collaborate with other John Hopkins teams through their ability to connect to HORNET at speeds of 10-Gigabit and higher. High speed connectivity is not limited to Johns Hopkins as HORNET also provides 100-Gigabit, Internet2 connectivity to other remote research institutions. The NSF has provided funds as an STCI grant (OCI-1137045) to help JHU to build this high speed data connection that will be utilized throughout this proposal.

Harvard University, Department of Molecular and Cellular Biology

Laboratory: The Engert laboratory consists of ~3000 square feet of combined laboratory and office space. Three large rooms contain six two-photon microscopes and house two more imaging/electrophysiology set-ups (250 square feet each). Furthermore, the Lichtman lab houses a state of the art core facility for scanning electron microscopy.

Clinical: Not Applicable

Animal: Zebrafish are housed in the zebrafish facility (2500 sq. ft.; installed by Marine Biotech Inc.) in the BioLabs basement. The facility has 6000 aquaria for keeping families of up to 30 fish and an additional 4000 aquaria for keeping individual adult fish. More than 2000 different strains of wild-type and mutant fish can be kept. A separate quarantine facility (300sq.ft.; 500 tanks) is located in the BioLabs basement. The PI

supervises the housing of the animals under the direction of the Associate Director of Animal Care of Harvard's Office of Animal Resources (OAR). OAR veterinary staff perform routine rounds through the zebrafish facility.

Computer: Members of the labs have PC or MAC computers for word processing, data management and internet access. All of these are connected to an Ethernet network that allows access to the other computers in the Institute. For image analysis and sequence analysis, we have access to high performance systems and computer clusters in the Center for Biological Imaging, Center for Brain Science, Dana Farber Cancer Institute and Broad Institute.

Office: The PI has an office of about 200 sq.ft. and the fellows and graduate students a larger one of 450 sq.ft.

Other Major Equipment: All the necessary equipment and facilities for molecular biology experiments and zebrafish embryology are available in the Engert lab, including cold and warm rooms, 24-well PCR machines, multi-well and repeating pipette devices, agarose and acrylamide gel apparatuses with power pacs, 3 dissecting microscopes with digital cameras and fluorescence capability, water baths, orbital shakers, rockers, three balances, microwave oven, 6 microinjection set ups, refrigerators, and -20 and -80 freezers and 2 chemical hoods. The core equipment in the laboratory consists of eight custom dual channel two-photon microscope, including high speed scanning mirrors, custom photon counting, and sophisticated motion control for automated operation and fast 3D scanning. The microscopes are controlled by custom software written in C# (Microsoft). Calcium images are scanned at 50 Hz while a high power microscope objective (Olympus) is swept across the depth axis in order to scan a volume of 200 µm x 50 µm x 50 µm at 1 Hz. A behavioral camera simultaneously collects images for tail tracking to allow synchronization of calcium activity and behavioral events. Additionally, we have four set-ups that allow combined *in-vivo* fluorescence and patch-clamp recordings and six custom built microscopes for the monitoring and online analysis of restrained and unrestrained fish behavior. A large and well equipped machine-shop is located in the basement of the neighboring Center for Brain Science.

Other Resources: *Environment:* The Biological Laboratories at Harvard University house two different departments – the Department of Molecular and Cellular Biology and the Department of Organismic and Evolutionary Biology. Furthermore they are in very close proximity to the Departments of Physics and Chemistry as well as the School of Engineering and Applied Sciences, and the recently inaugurated Center for Brain Science. Over the last ten years a vibrant and productive exchange of ideas and technologies has developed between all of these neighboring laboratories and through the intermingling of these different disciplines a whole series of successful collaborations has come into being. Such an environment is clearly the ideal place for scientific projects that rely on the development of new technologies and techniques since these often require the expertise of scientists in different fields. A wide spectrum of expertise is presented in various disciplines. In neurobiology: Drs. Sanes, Dowling, Zhang, Ölveczky, Dulac, Hoekstra, Kunes, Murthy, Samuels. In quantitative analysis and systems biology: Ramanathan, Cluzel, Needleman, Nelson, Berg, Mahadevan. In imaging: Cluzel, Needleman, Zhuang, Xie. In bioinformatics: Regev, Liu. In zebrafish biology: Zon, Dowling.

MGH/HST Athinoula A. Martinos Center for Biomedical Imaging

Laboratory: The imaging facilities of the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital are located on the Hospital's Research Campus in the Charlestown Navy Yard. The Martinos Center currently occupies ~85,000 sq.ft. of space and comprises basic and clinical research laboratories, as well as educational areas and administrative offices. The imaging laboratory comprises 8 large-bore MRI systems, 5 small-bore MRI systems and other imaging facilities including MEG, EEG, TMS, Optics, MicroPET and MicroPET-SPECT-CT imaging facilities. In particular, the Connectome Scanner is based on a Siemens Skyra 3T with the 300mT/m SR=200T/m/s "connectome" gradients. The

system comes with 64 RF channels and home-built 32- and 64-channel brain arrays available. The platform also contains visual (rear projection) and auditory stimulation setups, as well as a triggering interface.

Clinical: Not Applicable

Animal: Not Applicable

Computer: The Center's IT infrastructure consists of over 300 Linux workstations and 150 Windows and Macintosh desktops in offices and labs owned by individual research groups. There is a server farm with over 25 Linux servers that handles central storage, email, web, print, specialized processing and other shared services. The overall storage capacity of the center, including disks in local workstations and central storage, exceeds 2 petabytes. For high-performance image reconstruction, the center is equipped with a custom-designed ScaleMP vSMP computer equipped with sixteen 8-core Xeon E5472 and 1TB shared RAM. In April 2013, the Center added a Dell PowerEdge R910 server with four 10-core Intel Xeon E7-4850 processors and 1TB of quad ranked, DDR3 RAM. The center also has three Dell PowerEdge R910 with four 8-core Intel Xeon X7560 processors and 256GB of RAM.

Office: Dr. Rosen has an office of about 200 sq.ft. and the fellows and graduate students a larger one of 450 sq.ft.

Other Major Equipment: Not Applicable

Additional Computing Resources: The Center has a 126-node computing cluster for batch analysis jobs. Each node consists of two Quad Core Xeon E5472 3.0 GHz CPUs with 32GB of RAM, which together equal a total of 1024 compute cores available for batch jobs. Each node is connected by both a 1 GBit/s Ethernet link and a 20 GBit/s DDR Infiniband backplane. The Infiniband connection is used by parallelized jobs using MPI (message passing interface) to utilize multiple cores. The IT facilities are supported by a small IT staff comprising one full-time PhD-level manager, who directs two full-time system administrators and a part-time support technician. The Center also has three full-time programmers who support in-house-developed software for data analysis and management. Available commercial software includes AVS (Advanced Visual Systems, Waltham, MA), MATLAB (The MathWorks, Natick, MA) and MEDx (Sensor Systems, Sterling, VA) for general-purpose computation and simulation and image analysis; XWIN-NMR (Bruker BioSpin), Origin (OriginLab Corp., Northampton, MA), and Nuts (Acorn NMR, Livermore, CA) for analysis of NMR spectra; and the Siemens IDEA development environment for pulse sequences and image reconstruction software (Siemens, Erlangen, Germany). A substantial level of internal software development for image and data analysis is ongoing, using LAMP, C, C++, Java, FORTRAN, Ruby, Python, Perl and TCL/TK.

Data Management Plan

All data resources associated with this project are digital. There are no physical samples and no new neuroscience imaging data will be collected. This proposal will create and manage several distinct classes of data:

- neuroscience data contributed by the open science community;
- software and documentation developed by the project team;
- operational data, such as logging and usage statistics;
- derived or enhanced data products, metadata, and models; and
- publications and Web pages.

All data sets will be archival; they are intended for long-term preservation and subject to the data management policies described below.

The project will be conducted within the new Kavli Neuroscience Discovery Institute (KNDI) at Johns Hopkins University. As a new institute founded in 2016, KNDI is actively developing and codifying its best practices for the management of data. KNDI has inherited from multiple organizations: the NeuroData (http://neurodata.io) project, which provides community access to neuroscience data and is a Nature Data repository meeting Nature's standard for preservation and access; the Institute for Data Intensive Science and Engineering (http://idies.jhu.edu), which manages the Sloan Digital Sky Survey (http://sdss.org) and provides public access to more than 11 PB from 50 open-science observational and numerical simulation datasets; and the Center for Imaging Science (http://cis.jhu.edu), which provides metadata and computing for analysis of MRI images in the Cloud (http://mricloud.org). KNDI has developed data management policies from the best practices of these organizations that are also suited to our mission of providing publicly accessible data in the cloud.

Data Sharing: The software developed by this project will be freely and continuously available on the collaborators' and project's GitHub repositories. Data will be public, linked to the project's Web sites, and available for download and analysis through the project's Web services. All resources will be publicly available and distributed by default. Users of the system will have the option to keep data private or embargo contributed data for a period of time while awaiting publication. We will not store any private health information; all human data from Rosen will be de-identified using best practices prior to our receiving them. No login or registration will be required to access either code or data resources.

Licenses: All data sets will be released under the Open Data Commons Attribution (ODC-By) license. The license allows for free use of all data products, including copying, sharing, and redistributing data, and deriving new, customized data products. Our specific license terms require that all subsequent data use attributes of the original publication that contributed the data, as well as the project that hosts the data. All software and course material will be released under the Apache 2.0 license. This license is permissive; it places few restrictions on the use of software. Derived software products or software systems that integrate our code or course materials do not need to be open-source.

Documentation and Metadata: Source code documentation will be collocated and released with software, governed by the same open source license, and hosted on public repositories. Documentation for data sets are treated similarly. They will be managed in version control public repositories and will be available as links that are collocated with the Web-services that provide access to data.

Laboratory Information Management System (LIMS): Metadata for stored datasets will be collected and stored in a LIMS system that describes the source, protocol, instrument, location of resources, tools used to process data, and links derived data to original sources. KNDI is currently evaluating different neuroscience LIMS software for deployment in this project, including LORIS, COINS, and XNAT.

Digital Object Identifiers (DOIs): All published data products will have DOIs associated with them that are persistent and citable. All source code, tools and repositories will be managed in GitHub and DOIs will be

issued by Zenodo for those repositories. Datasets will be published to the cloud and owned and managed by the project. DOIs for datasets will be issued by the Johns Hopkins Sheridan Libraries. This is standard practice for the NeuroData team.

Logging and Reporting Data: All access to Web services, site visits, and data downloads will be logged and recorded. This includes requests forwarded to NSF's advanced computing resources. This practice starts with the Sloan Digital Sky Survey. Based on our logging infrastructure, we will be able to analyze the data usage patterns and contributions. The goal is to inform the evaluation and reporting process with data-driven metrics at an arbitrarily fine granularity.

Security, Privacy, and Embargo: The privacy of data sets is implemented with a combination of user authentication in our project management system, user configurable access controls on a per data set basis,and secure Web-services (https://). Optionally, users may use an external service (e.g. Google or Facebook) for authentication using OAuth or OpenID. The combination allows users to maintain control of their data even after placing it in the system, and self-manage access to their data. They may set individual data sets, images, and annotation projects as public or private. They can also determine when to make data public, when to release data from embargo, etc. Derived data products may be kept private by registered users. Private data sets will be encrypted end-to-end for all Web-services and visualization tools.

Ownership, Copyright, Intellectual Property: Our project encourages scientists and communities that "open source" their data in exchange for storage and analysis services. Open-source data becomes a community resource, unrestricted for non-commercial use. Data providers retain copyright privileges and reserve licensing and approval rights for commercial uses of data. Users of the data reserve rights to their algorithms and analysis techniques.

Storage and Backup During the Project: The project will store data principally in the Amazon Cloud. Amazon provides guarantees about the reliability of data in the service level agreements, e.g. some data services implement triply redundant storage. Data on local systems and other computer systems are considered active scratch and working data, which are copied from persistent images on the cloud, and then stored back to the cloud.

Long-Term Archival and Preservation: At the end of the project, select data stored in the cloud will be ingested back in the KNDIs local storage cluster. This addresses the long-term accessibility of data after project completion when cloud resources can no longer be afforded. We have partnered with Google to store all image data on an ongoing basis. Development of new data-intensive clusters (GrayWulf 2006 and Data-Scope 2012) has provisioned a small fraction of resources to maintain the entirety of our previously collected data. Our preservation ethic includes preserving the function and semantics of the data long after project operation completes. We will do so by defining archival packages that include algorithms and methods as well as data. We recognize sustainable preservation as one of the most challenging aspects of developing data resources and note our commitment and experience in defining strategies to fund and maintain resource sharing beyond project lifetimes.

Personnel: Co-PI Randal Burns will serve as the data management officer and will ensure that all collaborators and subcontractors comply with our data management policy. He will also contribute yearly reports on data management, which PI Vogelstein will include in year end reports. Senior personnel Eric Perlman will aid Burns and would serve as a replacement were Burns to become unavailable.

Postdoctoral Researcher Mentoring Plan

During this project, a postdoc will be employed at Johns Hopkins University and two at Harvard University as a subrecipient. This Postdoctoral Researcher Mentoring Plan recites our uniform mentoring guidelines for all postdocs.

Recruitment and Orientation. Postdocs are recruited through an open recruiting process. Postdoc candidates will be interviewed by one or more investigators. At interview time, mutual expectations will be established for a) the amount of independence the Postdoctoral Researcher will have, b) interaction with other team members, c) productivity including the importance of scientific publications, d) work habits and laboratory safety, and e) documentation of research methodologies and experimental details so that the work can be continued by other researchers in the future.

Professional Responsibility. After joining the team, postdocs will be expected to immediately take and pass any required courses on research conduct and research ethics. Postdocs will also be encouraged to join a professional society if they are not already members as students.

Research Mentoring. Postdocs will be included in the intellectual leadership of the research. They will be encouraged to develop their own unique lines of research and to produce first-author publications. Postdocs also serve the very important role of integrating their work at their laboratories with TIBS. This includes, for example, ingesting all relevant data and utilizing/testing TIBS functionality designed for their user stories, as well as contributing algorithms developed in-house for integration with TIBS, and providing frequent feedback on usability and comprehensiveness. They will also be included in discussions with industrial partners, thus providing them with exposure to the practical and intellectual shaping of the research agenda. Finally, postdocs will be instructed and included in technology transfer activities including applicable confidentiality requirements and preparation of invention disclosure applications. In addition, at Harvard (Engert) all mentees are encouraged to attend group meetings not only of the Engert lab but also those of the neighboring labs of Schier, Lichtman, Olveczky and de Bivort, where they can interact with fellow scientists working on disciplines as diverse as stem cell biology, embryonic development and systems neuroscience. At Harvard (Rosen), mentees are expected to attend all Human Lifespan group meetings and the weekly BrainMap seminar, where trainees will hear talks by their fellow scientists, as well as invited speakers outside the institute. At JHU, mentees are expected to attend the Kavli Neuroscience Discovery Institue/Center for Imaging Science seminar series and the Wang laboratory seminar series at the School of Medicine.

Grant Writing and Meetings. Postdocs will also be involved in the development of research proposals. This will include identification of key research questions, definition of objectives, description of approach and rationale, and construction of a work plan, timeline, and budget. Mentees at Harvard University will receive intensive and focused training in writing R01 style grants: They will attend Grant Writing 101. They are also encouraged to apply for K-style training grants. Workshops and lecture series tailored for training grant preparation are also available to all mentees. All mentees are encouraged and financially supported to attend at least two meetings each year to keep up with the latest research in the field and to set up future collaborations to bring outside expertise into their projects. Postdocs will also assist in developing curricula, including local courses, summer courses, and online courses, as well as running workshops adjoining the major conferences.

Mentoring Students and Teaching. All members of the research team are active in teaching, and teach courses in topic areas that would be relevant to any postdoc employed on this project. Postdocs will be encouraged to deliver lectures in these courses and they may, if appropriate, develop their own courses. Postdocs will also be intimately involved with the graduate students and undergraduates on the project and will thus gain valuable mentoring experience.

Career Counseling. Each postdoc will be mentored with the skills, knowledge, and experience needed to excel in his/her chosen career path. This guidance will be provided by all investigators of this proposal; indeed the sharing of postdocs will allow the postdoc to create a larger career mentoring network than would otherwise be afforded.

Resource Sharing Plan

Anticipated range of uses of the proposed neurotechnologies for research and education in neuroscience and other fields

The TIBS Neuroscience as a Service (NaaS) platform creates a complete computational environment in which users can apply state of the art tools to world-class data sets from a thin client-laptop or mobile device. The framework adds value for a broad range of users. This includes experimental neurobiologists that contribute data and computational neuroscientists who bring models to confront data. It also includes amateur neuroscientists, educators, and students who can observe and engage in scientific discovery with nothing more than a network connection. End-to-end experiments in TIBS connect data from processing and analysis pipelines all the way through to visualizations and publications. This allows anyone to enter the system and see exactly how a result was achieved, and provides a baseline for extending that result. In the classroom, it allows the learner to replicate previous discoveries, providing an authentic science experience.

The overarching goal is to provide all users with a low barrier to entry, and activities that build expertise incrementally. We will support educational users with Massively Open Online Courses that use TIBS experiments as the learning environment. For non-neuroscientists, amateurs, and global users, we will create well-documented experiments in which they can innovate on specific aspects without a comprehensive understanding of the whole process. For example, a statistician could provide a new machine learning algorithm to detect connectome changes across time without knowledge of imaging protocols.

Coordination with related available resources and infrastructure, and potential for integration with such resources.

TIBS will leverage existing research investments by integrating existing computational and data resources into its NaaS environment. Data sources include NeuroData (http://neurodata.io), which provides access to more than 50 multi-modal data sets that span from the nanoscale (electron microscopy) to millimeter (MRI). We will also link to data in the Human Brain Project (http://humanbrainproject.eu) and the Human Connectome Project (http://humanconnectomeproject.org). Data at these sites represent a substantial investment (over $100M USD) by NSF, NIH, DARPA, and the European Union. TIBS provides global access to world class data without having to take ownership of the data or pay for its maintenance.

TIBS links with advanced computing resources at the NSF to contribute computation to the open science community. The TIBS workflow system will launch processing of MRI images through MRICloud, and ingest and manage the results in its versioning and provenance framework. MRICloud will contribute 2.8M core hours of processing on XSEDE resources to the public (registered users) through its *Computational Anatomy Gateway* (NSF ASC140026). This relationship enhances MRICloud by making the results of its computations persistent, public, repeatable, and discoverable. The Computational Anatomy Gateway provides one positive example of TIBS sustainability models: TIBS will provide free access to open-science users of public data, and training and education applications.

Strategy for disseminating the neurotechnology resource rapidly to users that are most likely to benefit from its development.

TIBS reduces barriers to using developed technology by choosing development and deployment strategies to ensure that: (1) experiments are usable immediately, (2) tools and data run everywhere, and (3) no licensing restrictions prevent reuse. All software development, both infrastructure and experiments, will be done in an agile fashion and freely available on GitHub. Agile means that software is tested and pushed multiple times a day—only operational software makes it into the master branch. Landmark versions of experiments are tagged, e.g. those linked to a publication, but collaborators can engage at any point in the development process. The TIBS NaaS platform runs all experiments within Docker containers that

encapsulate all software dependencies. This ensures that experiments run uniformly across platforms—from a scientist's desktop to a Jupyter notebook in the cloud. TIBS adopts permissive licenses for both data and software (see Data Management Plan). This allows for the arbitrary reuse of software products without the restriction that derived software is open source. It also allows for the unrestricted reuse of data subject to citation of the data generator and data provider, with the Open Data Commons *Attribution* license.

Plans for outreach and community input, and Anticipated implementation timetable and strategy for evaluation and management over the course of the award period.

The concept of TIBS was developed in an NSF-funded workshop attended by 60+ leading neuroscientists from around the world. From this group, we will draw an advisory board that will meet twice a year, once virtually, and once in person at the Society for Neuroscience conference. The board will represent different user groups, educators, experimental biologists, computational neuroscientists, international users, and philanthropists (including our 25 collaborators). The goal of the board will be to provide feedback on how TIBS is serving the neuroscience community from each of these diverse perspectives.

Our outreach efforts span the purely virtual to physical, and provide a diversity of activities, from casual to immersive, that allow the learner/scientist to engage at their level of interest and grow their expertise and interest incrementally. The Broader Impacts statement details these activities, from a YouTube channel of specific activities at the most casual level, to Massive Online Open Courses, all the way to a summer school program at the most immersive.

Timeline

For Task I, we begin py porting existing data to the cloud in Year 1 (Subtask 9). In Year 2 we post data from our funded collaborators (Subtask 3,6,8), because we have postdoctoral fellows working directly with that data. In Year 3, we will post all existing data from our unfunded collaborators (Subtask 1,2,4,5,7), once we have extensive experience working with the postdocs. In all years we will be accepting all data from JHU KNDI (Subtask 10). Years 4 and 5 we will post data as it becomes available, from both named participants in this proposal, and beyond. For Task II, we begin by deploying the data management system in the cloud in Years 1 and 2 (Subtask 1), which is required for Task I. Year 2 begins our development of a workflow management system, which we complete in Year 3 (Subtask 2). Once both of those are in place, we focus on building Apps for the remaining 3 years (Subtask 3). Task III consists of the development of eight algorithms for different stages of the scientific process. Each year we work on two different stages, the first year dedicated to developing mathematically principled methods, the second devoted to incorporated quality assessment and scalable implementations. All code from Task II and III will be developed open source. Finally, for Task IV we begin generating documentation and tutorials for the existing tools (Subtask 1). In Year 2 we start augmenting our workshops and other events with new TIBS material (Subtask 2). Finally, in Year 3 we begin creating content our online courses, filming in Year 4, and refining in Year 5 (Subtask 3).

Task	I: Data										II: NaaS			III: Algorithms								IV: Education		
Subtask	1	2	3	4	5	6	7	8	9	10	1	2	3	1	2	3	4	5	6	7	8	1	2	3
Year 1									X	X	X			X	X							X		
Year 2			X			X		X		X	X	X		X	X	X	X					X		X
Year 3	X	X		X	X		X			X		X	X			X	X	X	X				X	X
Year 4										X			X					X	X	X	X		X	X
Year 5										X			X							X	X		X	X