

Final Report on the Tenth Kavli Futures Symposium:

Data Deluge from the Brain Activity Map

California Institute of Technology
January 17, 2013

“The great topmost sheet of the mass, that where hardly a light had twinkled or moved, becomes now a sparkling field of rhythmic flashing points with trains of traveling sparks hurrying hither and thither. The brain is waking and with it the mind is returning. It is as if the Milky Way entered upon some cosmic dance. Swiftly the head mass becomes an enchanted loom where millions of flashing shuttles weave a dissolving pattern, always a meaningful pattern though never an abiding one; a shifting harmony of subpatterns.” Sherrington, C.S. (1942).

Organizing Committee:

Miyoung Chun (Kavli Foundation)
Terrence Sejnowski (Salk Institute),

Participants:

Yaser Abu-Mostafa (Caltech)
Krastan Blagoev (NSF)
Miyoung Chun (Kavli)
Tom Dean (Google)
John Donoghue (Brown)
John Doyle (Caltech)
David Heckerman (Microsoft)
Tom Insel (NIMH)
Tony Lewis (Qualcomm)
Geoffrey Ling (DARPA)
Terry Sejnowski (Salk)
Fritz Sommer (UC Berkeley)
Larry Swanson (USC)
Clay Reid (Allen Institute for Brain Science)
Richard Weinberg (USC)
Rafael Yuste (Columbia)

Executive Summary:

The Tenth Kavli Futures Symposium on Data Deluge was a one day workshop to creatively explore the large-scale data management and analysis problems that need to be solved over the next decade as The Brain Activity Map (BAM) project matures. Large-scale brain recordings and our ability to manipulate the activity in neural circuits will generate immense amounts of data, requiring new analytical tools to discover how information is represented in brain activity patterns and disordered network activity can lead to brain dysfunction.

We anticipate that the advent of technology that allows a million neurons to be accessed simultaneously and continuously will produce not just a quantitative change in the way that we design experiments and analyze data, but a qualitative shift in the types of questions that can be asked and answered. A similar shift occurred in molecular biology when the human genome was sequenced, which shifted the focus from studies that examined one gene at a time to synoptic studies of the entire genome.

It will be important to plan ahead for what to do with the deluge of data when it arrives. Technical problems need to be solved for how to store, maintain and access the data. Equally important are issues regarding who has access to the data and under what conditions. Finally, we need to create an environment where many researchers can work together on the design of experiments, analysis of the data, and the development of new theoretical approaches to understanding brain function:

- BAM observatories should be set up to parallel those in astronomy.
- A data platform should be set up for shared public access based on technologies developed by Google, Microsoft and Amazon.
- A group of experts should be established to set the standards for database formats and annotation.
- Anatomical registration of the data is essential and a parallel Brain Anatomical Map project should be made part of BAM.
- Multiscale, nonlinear analysis of the high-dimensional brain activity and sensorimotor data will be a challenge for current machine learning techniques.
- Recordings from a million neurons will reveal many correlations between experience and activity, but manipulations of the neurons will be needed to test causal relationships.
- Better data will lead to better theories of brain function and ultimately to a better understanding of normal and abnormal brain states.

The million neuron march will arrive sooner than we imagine.

Introduction

Most of what we know about the properties of neurons has been derived from single unit recordings in experiments designed with a limited range of sensory stimuli and motor actions. There are exceptions, such as recordings from dozens of neurons simultaneously from the hippocampus of freely moving rodents. However, there are severe limitations on what can be learned by recording from randomly-sampled neurons in a single brain area.

We know, for example, that in sensory areas of the cortex there is trial-to-trial variability in the responses of single neurons, which are typically averaged over trials. But we can perceive a stimulus on a single trial perfectly well. This is because the activity hundreds of neurons together encode the features of sensory stimuli, and their responses are not independent of one another. We will need to record densely from many interacting neurons simultaneously to understand the neural code.

We also know that synaptic plasticity in the cortex depends on the relative timing of synaptic inputs and the postsynaptic action potential (spike time dependent plasticity). We can record the postsynaptic spikes, but do not know which synaptic inputs contributed to that spike. Dense recordings can provide data on the relative timing of spikes in many interacting neurons.

Simultaneous dense recordings from multiple cortical areas will be needed to discover how these areas are dynamically linked together during coordinated sensorimotor control and internal cognitive processing. We will need to record from the same neurons during the performance of many different behaviors.

The Brain Activity Map (BAM) project is a concerted multidisciplinary effort to develop and implement new tools to collect and process large scale brain activity and to change the activity patterns in neurons to test hypotheses for brain function. However, neither amassing immense amounts of neuroscience data nor large-scale simulations are likely to reveal the secrets of the brain without a new theoretical framework for exploring and interpreting the data, and creating a deeper understanding of the brain's most complex functions.

These new principles of brain function will have broad applications for treating mental health disorders and in solving many practical problems and engineering applications.

Organization of the workshop

The workshop was kept to the small number of key researchers and representatives of organizations who could fit around the table in the Milliken Board Room at Caltech. Companies that were represented included Google, Microsoft and Qualcomm. Agencies included NIH, NSF and DARPA. This group was able to cover a wide range of issues effectively and decisively.

The morning was devoted to laying the groundwork for the workshop. We started with an overview of the BAM project from Miyoung Chun and, via skype, John Donoghue and Rafa Yuste. Terry Sejnowski outlined the rationale for undertaking BAM from a computational perspective and the issues that arise when scaling up to a million simultaneously recorded neurons.

In the afternoon, we broke into smaller work groups that focused on four topics over two parallel sessions:

- 1. How should the database be organized and visualized?*
- 2. What platform is needed to manage the data?*
- 3. What tools are needed for analyzing the data?*
- 4. How will the data enlighten brain theory?*

How much data will be generated by BAM?

We envision BAM as a platform with three components: 1) Nanosensors and nanoactuators that can detect neural signals and in turn modify them; 2) A communications system for noninvasive two-way communication with the sensors and actuators; 3) A computational environment including data storage, analysis and visualization. In this workshop we focused on the third component. There are many types of sensors that can be devised, depending on the questions being asked, including the biochemical state of neurons as well as electrical activity. We will assume that these signals are continuous measurements that will be digitally sampled and communicated to a database.

The number of neurons that need to be interrogated to answer specific biological questions is an open theoretical question. We based our estimates for the size of the database on recordings from a million channels continuously: the numbers can be appropriately scaled for smaller or larger numbers of signals. For kilohertz sampling this would generate a gigabyte/sec, 4 terabytes/hour and 100

terabytes/day. These data can be compressed. Nonetheless, the data accumulate rapidly. Assuming a compression ratio of 10, one year of data would generate 3 petabytes/year. This is within the range of other major scientific instruments that generate big data.

The Large Hadron Collider (LHC) generates a terabyte/sec at the detector, but only keeps a gigabyte/sec for later analysis. This generates 10 petabytes/year. Towards the end of the decade, the 8 meter optical telescope Large Synoptic Survey Telescope (LSST) will begin its mission to survey over half the sky by taking 3 gigapixel images every 15 sec. This will produce over 10 petabytes/year, or 100 petabytes of archival data over 10 years.

How should the database be organized and visualized?

1) *Lessons from astronomy* (Chris Martin). The technology for BAM recording will be very expensive in the beginning. Thus, initially only a few brain observatories can be established, organized along the lines of astronomical observatories. Mechanisms need to be established for assigning “observation time” to researchers that propose the most interesting and potentially impactful experiments. Standardized formats should be developed as in astronomy (flexible image transport system FITS, virtual astronomic observatory us-vo.org). There are current neuroscience efforts towards a more standardized data description (INCF Task force for sharing electrophysiology data).

2) *All data should be shared with the community*. Data sharing will be an important component of the BAM initiative because: a) Data mining will benefit from involvement from many researchers, also potentially from various quantitative fields outside neuroscience, such as machine learning, statistics etc. b) Sharing the data can provide unparalleled new opportunities for education and teaching. c) Publishing the data is prerequisite for making published neuroscience results “reproducible”. d) Crowd sourcing has become an important resource in fields such as protein and RNA folding, where the complexity of the data requires human pattern recognition, and this is likely to be the case for brain data as well. Connectomics researchers are already developing the software needed for anatomical crowd sourcing.

3) *Save the raw recordings, not just the spikes.* The spike sorting problem for large data sets is not a solved problem for high-density recordings¹. Also, membrane potentials and local field potentials contain important additional clues about function. However, not all data are equally important and some aspects of the data do need to be made more accessible (Tom Dean). For example, the spike-sorted data might be an important first step in the analysis.

4) *Recordings need to be anchored to anatomy.* Without anchoring the recordings to specific neurons in specific parts of the brain their potential value will be much diminished. The mapping and consistent naming of the locations of the recording sites in the individual tissue is of extreme importance (Clay Reid). Since there is currently no consistent ontology for brain parts, the BAM project must include a strong anatomical component to work on these issues (Larry Swanson). There should be a parallel effort to BAM aimed at improving anatomical techniques, which could be called the Brain Anatomical Map project.

5) *Visualization of large data will be extremely important for the BAM initiative.* State-of-the-art techniques can be used for visualizing anatomical 3D structure (Richard Weinberg). However, for visualizing the high-dimensional spatiotemporal structure of brain activity new methods should be developed. At Microsoft a data wall is under development over the next 5 years that will allow researchers to interact directly with visualized data (Tom Dean).

What platform is needed to manage the data?

1) *BAM data should be shared and publicly accessible.* This will require a change in community norms, which now favor ownership of the data by the laboratory, even when the recordings were supported by public funding. The rules for who has access and conditions for publication should be developed by the community and be made clear to all.

2) *The problems associated with storing, maintaining and accessing large data sets have essentially been solved* by information based companies and there was interest expressed from Google, Microsoft and Qualcomm to help out (Tom Dean, David Heckerman, and Tony Lewis). The general rule is that you store the data once and don't move it – Co-locate performance computing architecture with data storage. Analyses and modeling of the data can be performed on the

¹ S. Leski, K.H. Pettersen, B. Tunstall, G.T. Einevoll, J. Gigg, and D.K. Wojcik: Inverse Current Source Density method in two dimensions: Inferring neural activation from multielectrode recordings, *Neuroinformatics* 9, 401-425 (2011)

high-performance servers. Software can be designed to control the analysis and display the data through web browsers. It will only be possible to download small subsets of the data for search and extraction computed locally (a lesson from genomics). Building the BAM data platform will rely on the help of companies experienced in big data.

3) *Setting the standards for storing and describing the shared data on the platform will be crucial* (David Heckerman): This should include:

- Anatomical structure and connectivity of the individual brain and relative positions of recording sites.
- Standardized method to store and transmit recorded activity data (Fritz Sommer - INCF task force)
- Description of the experimental procedure, ideally in a high-level programming language for robots (Tony Lewis). Is there a high level language for describing physical human actions required to perform the experiments?
- Detailed specification of the behavioral paradigm of the animal (e.g., stimuli, tracking data of skeletal posture need to be part of the shared data set, CRCNS.org). Means to quantify behavior for various paradigms will be necessary.

4) *BAM data should be made widely visible, accessible, and useful to as many communities, researchers and students as possible*. Impressive examples from genomics suggest that such open access can enable groundbreaking results from “non-experts” (Tom Insel).

- Provide visualization tools for viewing, screening and examining the data. The need for new tools to visualize and manipulate brain activity data should not be underestimated (see above). For example, an interactive multi-monitor 3D display of superimposed anatomical structure, nomenclature, location, function and activity could be developed (Richard Weinberg).
- The community can be engaged by formulating challenge questions that are easy to understand but hard/laborious to solve (For example, the Neural Prediction Challenge in CRCNS.ORG) (Fritz Sommer).
- Involve gaming experts in order to find ways to engage crowd sourcing for solving problems arising with exploiting and understanding the BAM data.
- Allow theories with experimental predictions to be posted and discussed on the platform.

5) *Collaborative organization of BAM research.* As with other major scientific instruments, teams of researchers will need to collaborate to carry out the design, data collection, data analysis and modeling. One of the a general problems that needs to be solved for how to help nonexperimental researchers to use the database and make predictions that can be used to guide further experiments (Clay Reid). We also foresee the need to design collaborative filter mechanisms for:

- Assigning observation time to the best proposals from researchers (based on the experience in astronomy).
- Assigning computation time (if limited) to the best proposals for how to model or analyze the data.
- Agreeing on rules to fairly assign credits to researchers involved in creating results (co-authorship versus acknowledgement for sharing data or methods, etc.).

What tools are needed for analyzing the data?

1) *Compression of the data should be pinned down* (Yaser Abu-Mostafa). This depends on what needs to be preserved. For example, delta compression can be used for compressing analog recordings². Spike sorting and other preprocessing steps should be carried out early so these parts of the database can be quickly and efficiently accessed during the experiment to help guide subsequent steps.

2) *There is a wide range of mathematical tools in use for analyzing recordings from neurons*, such as information theory, Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Support Vector Machines (SVM). The challenge, however, will be to scale these up from a few hundred neurons to a million neurons. The servers that will handle these large datasets will be distributed across many cores and the algorithms should be adapted to run in parallel across hundreds of thousands of cores, which is a major challenge in computer science.

3) *The low hanging fruit will be identifying the “eigenfunctions” of brain activity in different parts of the brain* (David Heckerman). The dynamical state of the brain, visualized as a trajectory in the million dimensional state of neural activity, should be matched with the dynamical state of the world, with which the brain is interacting. These high-dimensional states of the brain can be projected down to a lower number of dimensions by PCA and ICA for visualization, but this only

² A compression ratio of 800 was achieved with recordings from the Utah array [IEEE JOURNAL OF SOLID-STATE CIRCUITS, 44, 995-1005 (2009)].

captures a small fraction of the total variance in the recordings. It may be possible to use dimensionality reduction techniques from machine learning on high-dimensional nonlinear manifolds to extract more complex patterns from the data.

4) *A wide range of natural behaviors should be studied* (Fritz Sommer). Only a limited number of stimuli can be tested when studying a single neuron for a few hours. The ability to record continuously for hours and days from a million neurons will open up new ways to design experiments that make better use of richer sensory stimuli and behavioral conditions. The observatories should be equipped with virtual reality systems and motion capture so that freely-moving animals can be tracked and their sensory state accurately controlled. The Fly-O-Rama used by Michael Dickinson demonstrates the insights that can be achieved by observing flies behaving in more natural conditions. The data to be analyzed will include the continuous sensory stimuli and motor state of the animal, as well as the locations of the neurons and other anatomical features.

5) *Brain activity is dynamically changing on time scales from milliseconds to years, over eight orders of magnitude.* The longer time scale for BAM recordings will make it possible to observe nonstationary aspects of behavior, particularly the changes that occur in brain activity during learning. We know that even the simplest forms of adaptation affect processing in many parts of the brain, which can be tracked by BAM to reveal the elusive “engram”. This will require multiscale analysis of brain activity.

How will the data enlighten brain theory?

1) *What we can measure constrains what we can understand.* Current theories of brain function based on recordings from single neurons are limited in scope. Population codes have been deduced by stitching together recordings from many neurons recorded separately. Observing population codes as they unfold in real time should reveal much richer and dynamically shifting relationships between the neurons and the ongoing processing occurring in brain circuits. Bill Newsome has been recording multiunit activity from hundreds of neurons in area MT of the visual cortex in response to sensory stimuli for two different tasks. The population dynamics of the neural ensemble carries a far richer representation of the stimuli and task than the properties of neurons at any single recording site.

2) *The reason we need to collect so much data is that we understand so little about brain function,* but as our understanding increases, the data problem

should become easier (John Doyle). A similar evolution occurred in our understanding of shear flow turbulence, which unfolded over several decades as measurement techniques improved. Simulations became much better but ultimately new mathematics was required to achieve a deeper understanding. But without the data and the simulations the path to the new math would have been much more difficult if not impossible.

3) *The most exciting questions will be the new ones that it will be possible to ask once the BAM observatory is operational.* For example, the responses of neurons in the visual cortex have latencies that range from 20 ms to 100 ms, but we have the impression of a single moment when a stimulus flashes. Is there a deeper neural correlate among a large population of neurons that is more closely associated with our subjective impression of time? BAM may bring us closer to answering ultimate questions, such as the nature of consciousness. Key to achieving this deeper understanding of brain function will be the ability to manipulate as well as record brain activity. This will allow us to test theories, as well as to devise new technologies for rapid, high bandwidth man-machine interaction and to treat mental disorders.

Data Deluge from the Brain Activity Map

California Institute of Technology
January 17, 2013

Schedule:

10:00 am – Welcome - Miyoung Chun

– Introduction to BAM – John Donoghue and Rafa Yuste (via skype)

11:00 am – Problem statement and charge to the group - Terry Sejnowski

12:00 pm – *Working Lunch*

1:00 pm – Work Groups: Session 1

1) *Database and Visualization*

2) *Data Mining and Modeling*

2:15 pm – *Break*

2:30 pm – Group Reports

3:00 pm – Work Groups: Session 2

3) *Data Platform*

4) *Brain Architecture*

4:15 pm – *Break*

4:30 pm – Group Reports

5:00 pm – *Dinner (Athenium)*

Work Groups:

1) *Database and Visualization:*

Fritz Sommer (UC Berkeley) - Chair

Clay Reid (Allen Institute for Brain Science)

Larry Swanson (USC)

Tom Dean (Google)

Richard Weinberg (USC)

Geoffrey Ling (DARPA)

Krstan Blagoev (NSF)

Miyoung Chun (Kavli)

2) *Data Mining and Modeling:*

Terry Sejnowski (Salk) - Chair
Yaser Abu-Mostafa (Caltech)
John Doyle (Caltech)
David Heckerman (Microsoft)
Tony Lewis (Qualcomm)
Miyoung Chun (Kavli)

3) *Data Platform:*

Fritz Sommer (UC Berkeley) - Chair
Tony Lewis (Qualcomm)
Terry Sejnowski (Salk)
Tom Insel (NIMH)
David Heckerman (Microsoft)
Miyoung Chun (Kavli)
Richard Weinberg (USC)

4) *Brain Theory:*

John Doyle (Caltech) – Chair
Clay Reid (Allen Institute for Brain Science)
Larry Swanson (USC)
Tom Dean (Google)
Geoffrey Ling (DARPA)
Krastan Blagoev (NSF)
Miyoung Chun (Kavli)