# Brain Cell Densities by Morphology across Brain Regions

Darius Irani, Ronan Perry, Hamilton Sawczuk, Bronte Wen

*Abstract*— **Recent advances in electron microscopy and machine learning classification technologies have created the possibility of a database of brain cell types and morphologies and their densities across brain regions at an unprecedented scale and resolution. Understanding brain cell morphologies and densities is critical to identifying differences between healthy and diseased at the nanoscale. Samples from twenty brains will be imaged - eight healthy brains and twelve with neurodegenerative diseases, four brains per disease. The images will be registered to 3D volumes, and using the 2D images and the 3D volumes, SVMs will be used to identify cell morphologies and densities. Additionally, the raw data, on the exabyte scale, will be publicly hosted along with processed data and a variety of tools for analysis. At the end of this project, a cutting-edge infrastructure will remain in place for further research.**

## I. INTRODUCTION

The human brain consists of an inordinate total number of cells. Brain cell types have been classified by their morphology, physiology, location, connectivity, and neurotransmitter type. Categorizing these brain cell types and their cell densities is necessary for understanding the differences between healthy and diseased brains. Neurodegenerative diseases such as Alzheimer's, amyotrophic lateral sclerosis (ALS), and Parkinson's, are well-known examples of differences in the brain due to neuronal cell death and deterioration. Furthermore, cell types and morphologies along with their cell densities could be useful in discovering evolutionary patterns and differences in intelligence between species. Advances in imaging and automated classification technologies bring forward the new possibility of a concise and unified database of brain cell types, morphologies, and density of unprecedented scale and resolution.

The development of the scanning electron microscopy (SEM) has given scientists the ability to view neuronal structures at the nanoscale. SEM is able to resolve structures at the sub-micron level; the resolution of a SEM is about 4 nm. SEMs have been used to image biological tissue and cells for analysis and provide extremely high resolution in comparison to previous technologies. Comprehensive efforts have been made to provide SEM images of cells, such as the Cell Image Library, a repository of images and movies of cells from a variety of organisms.

Processing these high-resolution images would be extremely time-consuming unless automated. Machine learning (ML) is especially useful for such tasks. Previously, Krishnan *et al.* were able to automate classification of sub-epithelial connective tissue cells using a support vector machine (SVM) based classifier. Lannin *et al.* used four ML algorithms to classify thousands of cells in a few minutes, greatly increasing the processing speed. Instead of classifying each image for each cell type, which is also likely to enable human error, ML can be used to increase the speed, accuracy, and consistency of the classifications.

## II. HARDWARE AND FACILITIES

### A. Hardware

- 750 ZEISS 505 MultiSEMs
- 3 IBM Sequoia Supercomputers
- 8000 Aberdeen Petaracks™

In order to fulfill our high throughput needs within our 5 year timeframe, we will acquire 750 ZEISS 505 MultiSEMs from Carl Zeiss Inc., each MultiSEM featuring 91 electron beams working in parallel in automated protocols (Carl Zeiss). Each MultiSEM equipped with a RMC Boeckeler ATUMtome tape collecting ultramicrotome allowing automated slicing of the resin-clad brain samples into sections ready for imaging. In order to process this amount of data, two IBM Sequoia Supercomputers will be used along with an additional one to allow for database search inquiries. With each cubic millimeter of data generated, 400 TB of storage are required (Titze & mGenoude, 2016).

Thus, to store all of the generated data, we will utilize 8000 Aberdeen Petaracks™. This storage system uses a ZFS filsystem that is discussed further in the Data Upload and Storage section.

### B. Facilities

Three facilities will be constructed to house the MultiSEMs, supercomputers, and Petaracks™. Imaging will take place in one building containing all of the MultiSEMs, data storage and database hosting in another, and computing, classification and user-querying in a third computing facility.

## III. Data Collection

Twenty brains will be acquired - eight healthy brains and twelve with neurodegenerative diseases. Three neurodegenerative diseases will be selected, with four brains to study for each disease. Tissue samples from fifty regions in each brain will be collected and embedded in a resin. Each sample will be a 2mm x 2mm x 2mm cube in order to best capture a cortical microcolumn, a proposed fundamental unit of the cortex, estimated to be a cubic millimeter in size. Using the MultiSEM integrated RMC Boeckeler ATUMtome tape collecting ultramicrotome, these tissue samples will be automatically sliced into 2 mm x 2 mm x 50 nm thick sections at a rate of 1000 sections per machine per day (Carl Zeiss). These sections are further partitioned into tiles and are automatically collected and mounted on a roll of tape. Strips are cut from this roll and mounted on silicon wafers for analysis. First, a wafer will be imaged with a ZEISS light microscope to create an overview image. Then, the wafer will be transfered to the MultiSEM and the overview image will be used to navigate the sample. Efficient automation and the speed afforded by the 91 beams allows for a 1cm x 1cm x 50nm sized section to be imaged in under 3 hours (Carl Zeiss). The generated digital images are then ready for processing and the wafers will be stored for potential post-processing or additional studies.

## IV. Information Extraction

Following the initial data collection process, the generated digital images must be formed into a 3D representation of the region in a process known as ”registration.” The imaged tiles will be stitched together to form the 2D sections, which will in turn be vertically aligned to form the 3D volumes. This process and the necessary distortion and rotation corrections will be accomplished using the free and open source large-scale EM Aligner developed by HHMI/Janelia Research Campus and running on the supercomputers (Titze & mGenoude, 2016). With the 2D sections and the 3D volume, automated image analysis can commence. In order to identify cells and classify them based on their morphology, we utilize a support vector machine (SVM) machine learning pipeline that has seen prior success in cell classification (Sommer & Gerlich, 2013). Training data will come from the Allen Brain Atlas database which contains morphological and biological properties from individual cells present throughout the human brain (celltypes.brain-map.org). This database will determine the key cell morphological characteristics that define the basis for classification across cell type. Since SVM is a supervised machine learning method, it must first be trained. Before any images are generated form the MultiSEM process, our team will manually classify cells in a subset of images gathered from the Allen Brain Atlas. From this, the SVMs will generate parameters defining a trained cell classification model. As MultiSEM 2D sections and 3D volumes are generated, they will enter the machine learning pipeline run on one of the supercomputers. The images will first undergo pre-processing to remove any variations caused by the camera. Then, the objects of interest (cells) will be extracted from the background of the image. These cells will be complemented with a vector describing their quantitative features, automatically determined by the learned algorithm. The trained SVMs will then classify these cells based on their feature vectors to generate cell density counts across the various cell types and morphologies.

## V. Data Upload and Storage

Once an image has been processed, it will then be stored on one of the Aberdeen Petaracks™ along with the metadata. Data acquired and processed by NEURON will be shared publicly in a tiered system. Cell density counts in various cortial parcels, processed using our classification algorithm will be available through query on the

database. We plan to host an SQL database on the cloud to provide robust data-querying tools. The database will be hosted through a virtual server with MySQL server installed.

For the quantity of data being collected, cloud storage would be too costly and impractical. Instead, the raw image data will be held locally and appropriate requests for access will be granted. Aberdeen's Petaracks$^{TM}$ are petabyte scale storage that incorporates ZFS filesystems. ZFS filesystems allocate all data across devices into a shared storage pool. This storage infrastructure provides NEURON with (1) reliable, (2) flexible, and (3) scalable storage that will accommodate our requirements and those of our customers through the duration of the project. Raw image and processed data from each brain will be assigned to its own storage pool in the larger ZFS system.

(1) The ZFS infrastructure ensures data reliability through efficient error detection and checksum algorithms. Corrupt data blocks can also be identified and corrected through snapshots and self-healing. Preventing against data loss in such a high volume data storage system is especially important for keeping the raw image data. Data loss is prevented by maintaining a level of redundancy in the filesystem. In ZFS, data redundancy is maintained through a combination of mirroring and striping which protects data in the event of disk failure while maintaining performance.

(2) ZFS is flexible as device capacity does not need to be pre-allocated. Disk space is shared across each storage pool. The size of each disk grows automatically as data is added up to the allocated storage pool disk space. This is useful for storing raw image data, as it allows for simplified data transfer.

(3) ZFS is scalable as there are currently no limits to the number of storage pools or files that can be contained within the structure. When a new device is added to a storage pool in ZFS, all other devices in the pool have immediate access to the data. ZFS also provides controls to change the permissions for which devices can access different classifications of data. The scalability of this filesystem makes using a ZFS filesystem attractive because additional data racks can be purchased and added to the filesystem seamlessly to expand the shared disk space.

The Aberdeen Petaracks provide the infrastructure to store large volumes of image data efficiently and safely. Processed data, with cell density counts and cell classification found on the database will be hosted to allow other research scientists to query data of interest and then request the full dataset corresponding with the queried results. Datasets will be shipped to scientists on harddrives similar to The Connectome Project's Connectome-in-a-box program. A visualization toolkit will be provided on the database to allow for 3D and embedded 2D visualizations and cell identification among other analytical tools.

## VI. Cost Analysis

The $1 billion allotted to infrastructure will be go towards acquisition of a warehouse and the necessary modifications to accommodate 750 SEM microscopes, 8,000 Petaracks$^{TM}$, 3 supercomputers, and lab/office space. These modifications will include provisions for high electricity and water usage, as well as safety and security. This money will also cover setup of microscopes, petaracks, and supercomputers.

The $3 billion in imaging costs are due to the cost of purchasing the 750 SEM microscopes and prepping the brain samples for imaging.

The $2.4 billion associated with data storage primarily covers the cost of purchasing the 8,000 petaracks needed to store all raw and process data.

The $750 million of computing costs allow for the purchase of 3 supercomuters used to handle user inquiries and data processing.

The $750 million reserved for labor costs will cover the salaries of the 2,000 employees. Man power will consist of scientist, engineers, and support staff.

The $1.5 billion utilities budget will cover energy and water consumption, which will be especially high over the initial three years of image processing.

TABLE I

COST BREAKDOWN

| | |
|---|---|
| Infrastructure | $1.0 bil. |
| Imaging | $3.0 bil. |
| Data Storage | $2.4 bil. |
| Computing | $0.75 bil. |
| Labor | $0.75 bil. |
| Utilities | $1.5 bil. |
| Total | $9.4 bil. |

## VII. DISCUSSION

At the conclusion of the five-year timeline, NEURON expects to have a unmatched database of brain tissue 3D volumes at nanoscale scanning electron microscope resolution. Key cell density will be extracted as the images are acquired, but the database will be maintained as a source of free data for future research work in the international scientific community. With this data easily available, we hope to spur further research that leverages our high resolution data and enable novel discoveries. With such a vast collection and free SQL querying tools, we hope to enable research that otherwise would have been impossible due to the cost of obtaining the data. Our project is an ambitious one, requiring huge imaging throughput and a massive amount of long term storage. We anticipate a variety of challenges to emerge as this proposal is put into action, from running these machines and facilities over the course of five years to implementing the data pipeline without error. However, the project is fully possible given the large grant budget available and will yield unprecedented scientific advances in fields like connectomics.

## REFERENCES


[1] M.M.R. Krishnan, M. Pal, S.K. Bomminayuni, C. Chakraborty, R.R. Paul, J. Chatterjee, et al. Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosisan SVM based approach. Comput. Biol. Med., 39 (2009), pp. 1096-1104

[2] Lannin, T. B., Thege, F. I. and Kirby, B. J. (2016), Comparison and optimization of machine learning methods for automated classification of circulating tumor cells. Cytometry, 89: 922931. doi:10.1002/cyto.a.22993

[3] Titze, B. and Genoud, C. (2016), Volume scanning electron microscopy for imaging biological ultrastructure. Biol. Cell, 108: 307323. doi:10.1111/boc.201600024

[4] Carl Zeiss, Inc., MultiSEM 505/506. MultiSEM - The World's Fastest Scanning Electron Microscope

[5] celltypes.brain-map.org

[6] Sommer C, Gerlich DW (2013). Machine learning in cell biologyteaching computers to recognize phenotypes. J Cell Sci 126, 55295539.