ABI Development: Storage, Organization, and Machine-Learning -Based Analysis Platform for Multi-Scale 3D Biological Data

A Introduction

Goal: The goal of this project is to develop a general biological informatics platform for the storage, organization, and machine-learning-based analysis of large volume, high-resolution, multiscale 3D biological data.

Background: With rapid advances in 3D microscopy for biological imaging the amount and complexity of biological data are growing exponentially [22, 45, 63, 74, 101]. The resulting biological data sets are large in volume (tera bytes to peta bytes in data size), high resolution (sub micrometer), multiscale (spanning from cellular scale to whole organ scale), and in full 3D. Thus, simply storing and organizing the data for casual browsing is in itself a great challenge [18, 28]. Furthermore, the geometric structures in these specimens are highly complex and the data can contain various different sources of noise, thus making quantitative analysis very difficult. Due to these issues, manual methods or prescribed algorithms do not scale well or are not accurate enough in such cases, thus data-driven, machine learning-based algorithms are increasingly serving as a viable alternative [19, 37, 38, 53, 100, 102].

Gap: There are various web-based atlases of 3D biological data such as large volume brain atlases (e.g., Allen mouse brain connectivity atlas [82]), but full quantitative analysis does not often accompany such large data sources. General tools for biological image analysis do exist (e.g., Fiji [90]), however, they are not integrated with the data sources and most of the analysis steps are manual or based on prescribed algorithms that are not flexible enough.

Objectives: To address these gaps, this project will develop an informatics platform that has the following objectives. These objectives map to §D.1–D.5, respectively, in §D Research Plan.

1. **Web-based atlas:** Develop a multiscale, web-based data storage, organization, and browsing platform.
2. **Ground-truth annotation facility:** Develop an integrated manual and semi-automatic annotation interface for the platform, for easier ground truth generation.
3. **Machine-learning-based analysis facility:** Develop an integrated machine-learning-based segmentation and geometric reconstruction/analysis tool for the platform, using ground-truth from objective 2.
4. **Statistical analysis and visualization facility:** Develop an integrated statistical analysis and visualization tool for the platform, based on results from objective 3.
5. **Deployment on the cloud:** Deploy the platform on the cloud (see Table. 4 for definition and benefits) for increased accessibility and scalability (initially on a local cloud running OpenStack, eventually migrating to a commercial cloud).

Approach: This project will develop an integrated informatics platform that allows the user to store, organize, and visualize 3D biological data; and provide an easy-to-use interface for fast and accurate labeling of ground truth for machine-learning-based analysis. For increased accessibility, scalability, and long-term availability, the platform will be deployed on the cloud.

Intellectual merit: This project will involve the design and development of novel data storage, organization, and visualization methods for large volume 3D biological data, and a streamlined,

user-friendly interface for machine-learning-based analysis of the data. The end-result will be a biological informatics platform that can accelerate scientific discovery in biological sciences.

Broader impacts: The proposed informatics platform is expected to find broad use in the biological research community with needs for 3D data storage and analysis. The informatics platform will be made public, where the users can upload and analyze their own data. Graduate students and undergraduate students (through the REU mechanism) will be trained. Interactive exhibits for the general public will be organized at collaborating science museums. The project team will also organize exhibits and tutorials at scientific meetings and conferences.

B Background

Our project will develop a general informatics platform for the storage, organization, and machine-learning-based analysis of large volumes of multi-scale biological data. Here we will first review latest 3D microscopy techniques, and discuss how the high-volume, high-resolution, multi-scale nature of the data from these instruments serve as the main motivation for our proposed work. Next, we will survey existing 3D microscopy data sources, and discuss their shortcomings in at least one of the following aspects: multi-scale organization facility, online annotation for machine learning purposes, integrated machine learning pipeline, and cloud-based scalability/sustainability.

B.1 High-throughput, High-resolution, 3D Microscopy

Our biological informatics platform is strongly motivated by latest advances in high-resolution, high-volume 3D microscopy for biological imaging. Such methods can lead to the complete anatomical, physiological, and molecular mapping of entire organs or whole small animals. Examples of such methods include our Knife-Edge Scanning Microscopy (KESM, our own design) [63, 69, 70, 71, 72] (also see [57] based on the same principles), All-Optical Histology [101], Array Tomography [75], Serial-Block-Face Scanning Electron Microscopy (SBF-SEM) [22], Automatic Tape-Collecting Lathe Ultramicrotome (ATUM) [33, 34, 89], and Selective Plane Illumination Microscopy (SPIM: also see Light sheet fluorescence microscopy and combination of the technique with structured illumination) [35, 36, 44]. Table 1 shows a comparison chart of these microscopy technologies. Unlike Array Tomography and SBF-SEM, KESM can survey large volumes of biological tissue (whole small animal organs), and KESM is an order of magnitude faster than All-Optical Histology. SBF-SEM has the advantage of ultra high resolution, and Array Tomography is ideal for repeated imaging of a single specimen with multiple immuno stains and even with SEM. SPIM has broad applications in real-time imaging of dynamic processes in the biological specimen.

Table 1: **3D Microscopy Techniques.**

Method	Resol. (x & y)	Resol. (z)	Volume	Modality	Time
All-Optical Histology [101]	0.5 μm	1 μm	1 cm^3	Fluorescence	∼900 hours
KESM [63] (cf. [57])	0.3–0.6 μm	0.5–1 μm	1 cm^3	Bright field, Fluorescence*	∼100 hours
Array Tomography [74]	∼0.2 μm	0.05–0.2 μm	∼100^3 μm^3	Fluorescence, EM	N/A See SBF-SEM
SBF-SEM [22]	∼ 10 nm	∼ 30 nm	∼100^3 μm^3	EM	296.3 days†
ATUM [34]	5 nm	30 nm	$0.5 \times 2.5 \times 3$ mm^3	EM	1.5 years‡
SPIM [44]	0.25–0.5 μm	∼1 μm	$\sim 1.5^3$ mm^3	Fluorescence	90 seconds

*Experimental scans have been successful. † 200 μm cube at $10 \times 10 \times 50$ nm^3 resolution [22]. ‡ with selective scanning.

B.2 Biological Microscopy Data Collections and Data Dissemination Frameworks

The Allen Brain Atlas [2, 55, 59] contains detailed gene expression maps for ∼20,000 genes in the C57BL/6J mouse. A semi-automated procedure was used to conduct *in situ* hybridization and

data acquisition on 25 μm-thick sections (z-axis) of the mouse brain. The x-y-axis resolution of the images range from 0.95 μm to 8 μm. The Allen Brain Atlas is the first comprehensive gene expression map at the whole-brain level, and is currently accessed over 4 million times per month, with over 250 scientists browsing the data on a daily basis (see [2], the "community page").

BrainMaps.org [5, 78] is an internet-enabled, high-resolution brain map. The map contains over 10 million mega pixels (35 terabytes) of scanned data, at a resolution of 0.46 μm/pixel (in the x-y plane). The atlas provides an intuitive web-based interface for easy and band-width-efficient navigation, through the use of a series of subsampled (zoomed out) views of the data sets, similar to the Google Maps interface. Even though the x-y plane resolution is below 1 μm, the z-axis resolution is orders of magnitude lower (for example, one coronal brain set has 234 slides in it, corresponding to a sectional thickness of 25 μm).

Whole Brain Catalog (WBC) [107] is a 3D virtual environment for exploring multiple sources of brain data (including mouse brain data), e.g., Cell Centered Database (CCDB, see below), Neuroscience Information Framework (NIF), and the Allen Brain Atlas (see above). WBC has native support for registering to the Waxholm Space, a rodent standard atlas space [42]. Multiple functionalities including visualization, slicing, animations, and simulations are supported.

The Cell Centered Database (CCDB) [10, 60] houses high-resolution 3D light and electron microscopic reconstructions spanning the dimensional range from 5 nm^3 to 50 μm^3 produced at the National Center for Microscopy and Imaging Research (NCMIR) [80]. The current CCDB has over 80 tables containing descriptive data that models the entire process of reconstruction, from specimen preparation to segmentation and analysis.

EyeWire is a citizen science project (based on crowdsourcing [103, 104, 105]) hosted at Princeton University (previously at MIT) [94]. The platform serves high-resolution 3D electron microscopy data of the mouse retina, and is equipped with an interactive interface where users (over 82,000 volunteers) can segment and annotate neuronal cell boundaries in the data set. See [62] for an overview of EyeWire and other citizen science projects in biology.

CATMAID, Collaborative Annotation Toolkit for Massive Amounts of Image Data [87], is a system that is closest to our proposed system. CATMAID allows Google Map-like navigation, user data upload, user annotation, visualization, etc. (e.g., Open Connectome Project uses CATMAID for visualization and navigation [8]). The main difference between CATMAID and our proposed system is that CATMAID is a software suite that the users need to download and install on their system, unlike the web service we are planning to provide. Furthermore, CATMAID does not have 3D overlay visualization which is key to the exploration of 3D data sets, and does not have machine learning capability.

In summary, there are various biological 3D microscopy data collections available, but their informatics platforms are very specific to a small number of model systems, and user upload, use annotation, machine-learning-based analysis, and cloud-based scalability/availability are generally not supported. See Table 2 for a summary comparison.

B.3 Atlasing and Cell Description Standards

A rapid increase in web-based resources serving cellular morphology and atlas-scale data sets gave rise to the need for data representation standards. The Waxholm Space [42] is a standard atlasing space for rodents. The effort to build this standard space was motivated by multiple non-standard, yet widely used coordinate spaces such as the Allen Brain Atlas [2, 55] or Paxinos and Franklin [84]. For biological cells in general, CellML [20] has been developed (cf. Subcellular Anatomy Ontology [61]). CellML allows cellular level modeling of biological function. It is not domain specific, thus can be used to describe a wide variety of species and cell types. Cell func-

Table 2: **Microscopy Data Resources and Tools.**

Resource name	Web -based	User data upload	User annotation	Machine learning	Cloud enabled	Model system(s)
Allen Brain Atlas	✓					Brain
BrainMaps.org	✓					Brain
Whole Brain Catalog						Brain
Cell Centered DB	✓	✓	✓			Brain, Liver, Heart, etc.
EyeWire	✓		✓	✓*	✓†	Retina, Brain
CATMAID	✓	✓	✓			Any kind
FIJI		✓	✓	✓		Any kind
Proposed	✓	✓	✓	✓	✓	Brain, Plant, Lung, etc.

$*$: Convolutional Neural Networks used in preprocessing data; †: planned in the near future.

tions that can be described in CellML include metabolism, electrophysiology, signal transduction, cell division, immunology, muscle contraction, etc. For microscopy data, Open Microscopy Environment XML (OME-XML) provides facilities for tagging images with xyz dimensions, pixel type, metadata, structure annotation, and region of interest (ROI). See [58] for an overview of metadata for biological imaging, including a discussion of OME-XML. As for neuronal morphology, NeuroML [81] has become the de facto standard (XML). BrainML [6], on the other hand, provides an XML framework for the exchange of general neuroscience data at the whole brain scale.

B.4 Volume Visualizers and Image Analysis Applications

Once the region of interest (ROI) is identified from the source atlas, a fully interactive 3D visualization can be used to fully analyze the small ROI volumes. There are existing 3D visualization packages, both commercial and free (for noncommercial purposes), such as Amira [115] and MeVisLab [73]. Furthermore, open source tool kits such as VTK [92] are available for extensive extension, e.g., a volume viewer called Paraview, based on VTK. However, these visualization packages and tool kits do not natively support tracing and morphological analysis, and cannot handle or visualize extremely large data volumes (e.g., data that are greater than several GB).

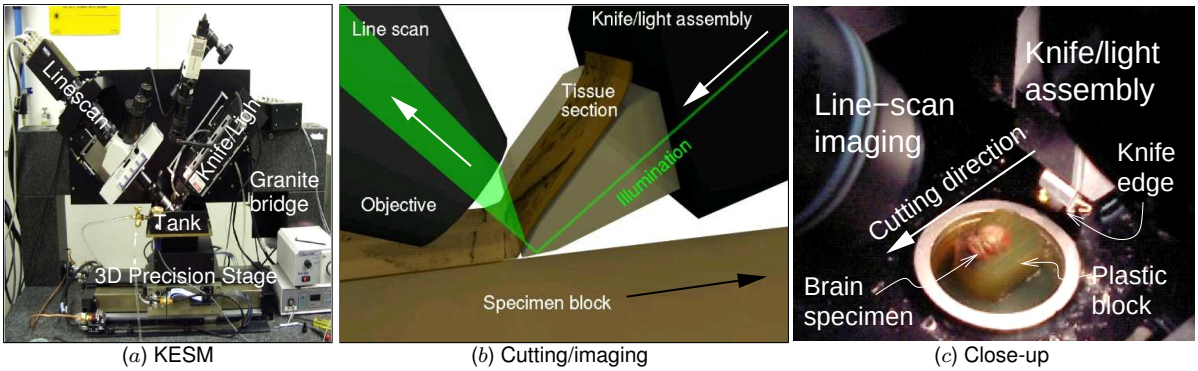
Applications for manual reconstruction e.g. [7, 27] and commercial software such as Neuro Lucida [77]; and commercial automated tracing packages such as Autoneuron [76] are available. However, availability of open source tracing and validation algorithms are very limited. One exception is the NIH-funded open source FARSIGHT toolkit [86]. However it is not a stand-alone application–it is a set of modules that can be combined to build a custom application. Another exception is Fiji [90], an open-source biological image analysis tool built on top of the popular scientific image manipulation tool ImageJ (see [91] for an overview and the tool's impact).

C Prior Work

In this section, we will summarize our prior works on imaging, informatics, and analysis.

C.1 Imaging with the Knife-Edge Scanning Microscope

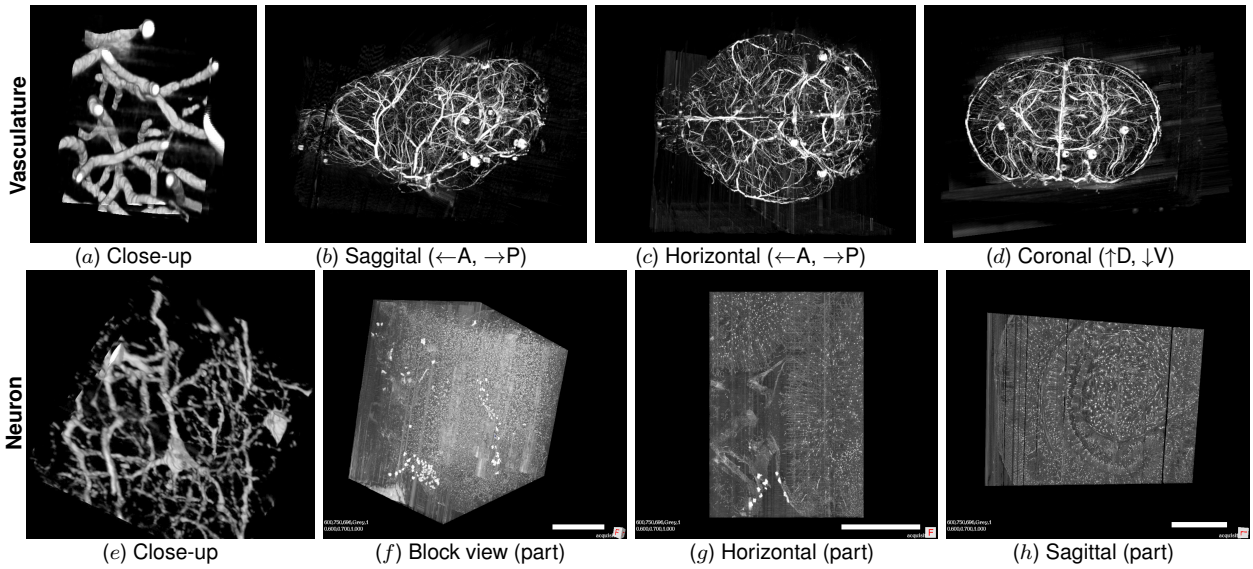
The Knife-Edge Scanning Microscope (KESM, US patent #6,744,572) [11, 51, 67, 69, 70, 72] has been designed at Texas A&M University (TAMU) in recent years with support from the National Science Foundation (MRI award #0079874; McCormick, PI), the Texas Higher Education Coordinating Board (ATP award #000512-0146-2001; Keyser, PI), and the National Institute of Neurological Disorders and Stroke (Award #1R01-NS54252; Choe, PI). The instrument, shown in Fig. 1*a*, is capable of scanning a complete mouse brain ($\sim$310 mm^3) at 300 nm sampling resolution within 100 hours when scanning in full production mode. The basic operation of KESM is shown in Fig. 1b. A white light source illuminates the rear of the diamond knife, and in turn illuminates the brain tissue at the leading edge of the diamond knife with a strip of intense illumination reflected



(a) KESM (b) Cutting/imaging (c) Close-up

Figure 1: **The Knife-Edge Scanning Microscope.** (a) Photo of the KESM instrument showing line-scan/microscope, knife/light assembly, granite bridge, and 3D precision stage. (b) Specimen undergoing sectioning by knife-edge scanner (thickness of section is not drawn to scale). (c) Close-up photo of the line-scan/microscope assembly and the knife/illumination.

from the beveled knife-edge. The microscope objective, aligned perpendicular to the top facet of the knife, images the transmitted light. A high-sensitivity line-scan camera repeatedly samples the newly cut thin section, imaging a region 20 µm along the length of the tissue ribbon and just beyond the knife-edge, prior to subsequent deformation of the tissue ribbon after imaging. See Fig. 2 for imaging results.



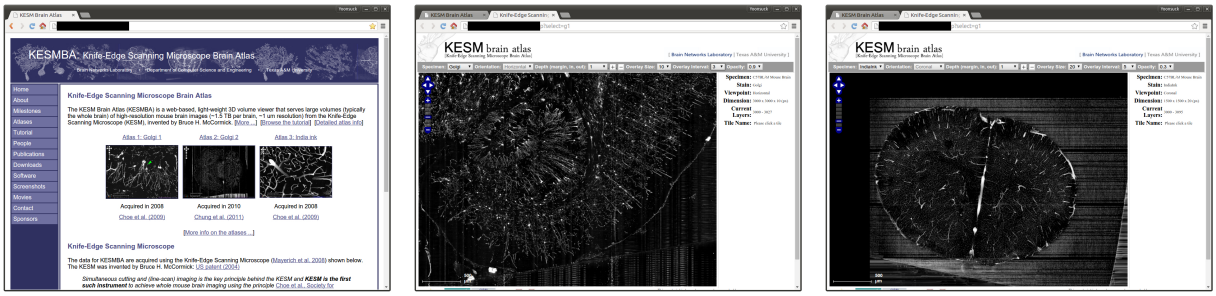
(a) Close-up (b) Saggital (←A, →P) (c) Horizontal (←A, →P) (d) Coronal (↑D, ↓V)

(e) Close-up (f) Block view (part) (g) Horizontal (part) (h) Sagittal (part)

Figure 2: **KESM Data.** Volume visualizations of KESM data stacks are shown for the vascular data set (top row, India ink stain) and the neuronal data set (bottom row, Golgi stain). (a) Close-up of the vascular data. Width $\sim$ 100 µm. (b–d) Three standard views of the whole mouse brain vasculature (subsampled from high-resolution data). Width $\sim$ 10mm. (e) Pyramidal cells from the visual cortex. Width $\sim$ 100 µm (f) A large sub-volume from the Golgi data set. Fine details are washed out. (scalebar = 1.44 mm) (g) A thin slab from (f) reveals intricate circuits (horizontal section). (scalebar = 1.44 mm) (h) A thin slab from (f) reveals intricate circuits (sagittal section). (scalebar = 1.44 mm)

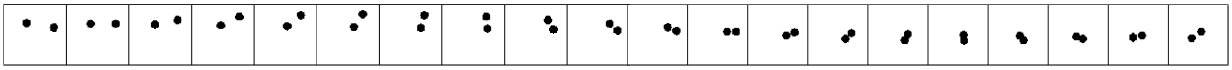
C.2 Rapid tracing of fibrous matter

In order to turn the raw data into a geometric description of the objects of interest (i.e., reconstruction), we are currently developing rapid tracing algorithms for fibrous matter such as neuronal

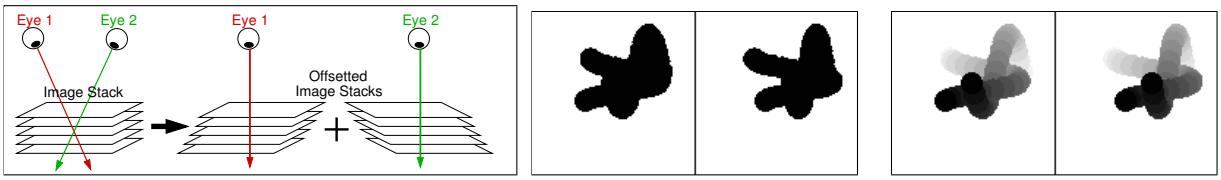


(a) KESMBA home page (b) Mouse brain (neurons) (c) Mouse brain (vasculature)

Figure 3: **KESM Brain Atlas.** Screenshots of the atlas running in the Google Chrome browser.



(a) Raw (synthetic) data slices



(b) Generating 3D views (c) Offsetted overlay (d) Dist.-atten./offs. overlay

Figure 4: **Generating 3D View Through Overlays.** Three-dimensional effects can be generated using a simple overlaying method. (a) A series of 20 image stacks taken from a synthetic volume data set is shown (left to right). (b) An illustration of how two stereo pairs can be generated using simple offsetting (shearing) and overlaying of the image stack. Such overlays can be easily produced in a web browser using Cascading Styling Sheets (CSS), from base images with transparencies (alpha channel). (c) Stereo-pair (for crossed viewing) of offsetted overlays are shown. The three dimensional effect is weak. (d) Stereo-pair (for crossed viewing) of offsetted overlays with distance attenuation is shown. Simply inserting semi-opaque layers with the background color between every data layer can achieve this effect. With this, the three-dimensional structure of the embedded object is clearly visible. See [26] for a report on our pilot results.

processes and neurovasculature [68], which is broadly classified as a vector tracing method [1, 9]. Our method has been used successfully on KESM data. See [64].

C.3 Web-based light-weight 3D microscopy image stack viewer

We have completed the design and development of a web-based 3D data stack viewer (called the Knife-Edge Scanning Microscope Brain Atlas), as a main product of our prior NSF grants (see §C.4 and Fig. 3). The atlas currently serves massive volume image stacks from KESM over the internet using standard web browsers without any add-on or plug-in. The basic idea is to use image overlays with distance attenuation and offsetting (shearing) to generate a stereo pair for a vivid 3D viewing experience (Fig. 4; stereo view feature is in the works). See [18, 26] for details.

C.4 Results from prior NSF support

The two recent NSF awards were strongly related, so we will discuss the two together. For both projects, the PI was Choe (PI of this project), and the Co-PIs were Abbott and Keyser (Co-PIs of this project). **(1) NSF award number, amount, and period of support**: i) #0905041, $114,024, 09/01/2009–08/31/2012. ii) #1208174, $200,868, 09/01/2012–08/31/2015. **(2) Title of the project**: i) CRCNS data sharing: Whole Mouse Brain Neuronal Morphology and Neurovasculature Browser. ii) CRCNS Data Sharing: Open Web Atlas for High-Resolution 3D Mouse Brain Data. **(3) Summary of results**. **(3-1) Intellectual merit:** The project developed a novel in-

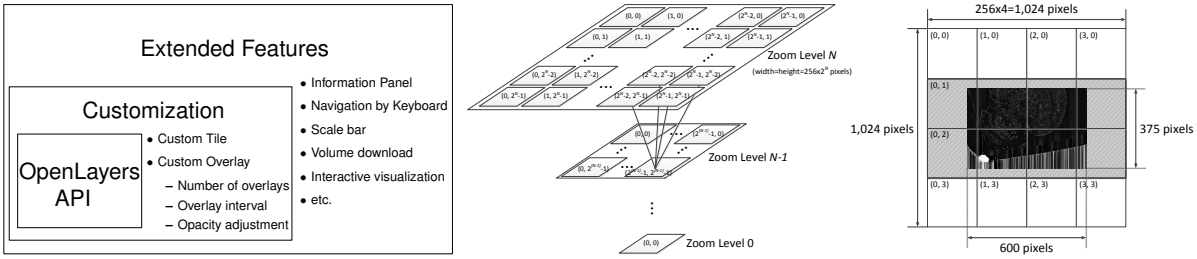
browser visualization method for large multiscale 3D microscopy data sets. The technique enabled real-time navigation and browsing of TBs of 3D microscopy data using only a standard desktop computer or even a smartphone/tablet. The high quality mouse brain data from the Knife-Edge Scanning Microscope we are serving on our data sharing platform is of high scientific value, and the tool allowed us to gain unprecedented insights into the organization of the brain at cellular level of detail. The resource is also freely available to the neuroscience research community and the general public, thus its scientific impact is potentially great. **(3-2) Broader impacts:** *Data and code dissemination:* Our main research product is the KESM Brain Atlas (two versions) that are currently serving data from four mouse brains (two Golgi-stained, one Nissl-stained, and one India-ink-stained). The web site is interactive and is publicly available (Fig. 3). Our atlasing code is also available on SourceForge. *Education:* As part of the two back-to-back grants, we trained 1 Ph.D. student, 9 M.S. students, and 7 undergraduate students (4: NSF REU, 3: other funding). See the publication list below. *Outreach:* Our data and technology were featured in San Francisco Exploratorium's exhibit titled "New Exhibition on Understanding, Influencing Brain Activity", which ran from 1/31/2015–3/1/2015 (Fig. 14b). *Tutorials, workshops, and exhibits:* To advertise our work and broaden the user base we ran a tutorial (International Joint Conference on Neural Networks, 2013, Dallas, TX), organized a workshop (Computational Neuroscience meeting, 2010, San Antonio, TX), and three exhibits (Society for Neuroscience, 2011 [Washington, DC], 2012 [New Orleans, LA], and 2014 [Washington, DC]). **(4) Publications**. 6 Conference proceedings (full papers): [14, 50, 53, 66, 100, 114]. 5 Abstracts: [12, 13, 16, 49, 98, 99]. 3 Journals: [15, 18]. 1 Ph.D. dissertation: [97]. 9 M.S. thesis: [17, 23, 47, 52, 79, 95, 96, 111, 113] (note: not all M.S. students were funded by the two NSF grants, although the topics covered were in line with the two projects). **(5) Evidence of research products and their availability**. *Online Brain Atlases:* Both brain atlases are currently hosted online, open to the general public (Fig. 3; access stats: Fig. 14c-d [4,727 visitors since 2012]). Both atlases (PNG and SVG version) are fully functional, serving whole mouse brain neuronal (Golgi stain and Nissl stain) and vascular (India ink) data. All servers are locally maintained at the PI's lab, with world-wide access. *Data shipping:* We are also shipping hard drives containing data to those who request the data. We've shipped data hard drives to: Johns Hopkins Univ. (Open Connectome project, Joshua Vogelstein); University of California, San Francisco (Nicolas Pannetier); Kettering Univ. (Jaerock Kwon); King Abdullah Univ. of Science and Technology (Markus Hadwiger); and Louisiana State Univ. (William Donahue).

D Research Plan

D.1 Web-based atlas

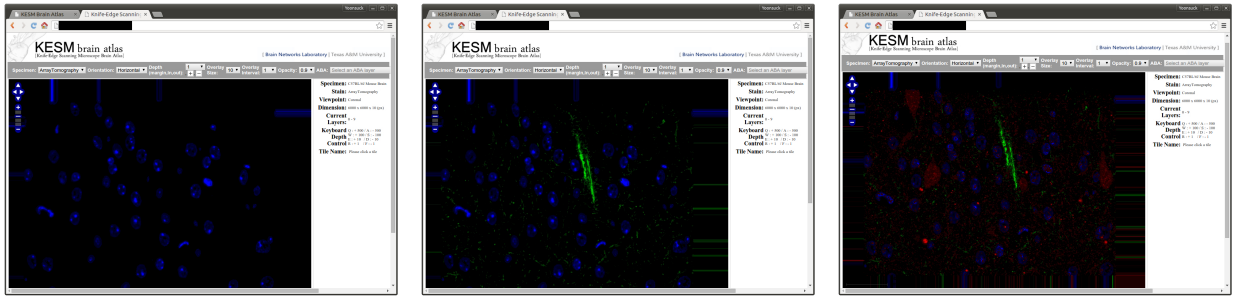
We will develop a multi-scale, web-based data storage, organization, and browsing platform, significantly extending our successful prior work on the Knife-Edge Scanning Microscope Brain Atlas (Fig. 3). We will migrate from our previous Google Maps-based system to an open-source mapping API called Open Layers [83]. The proposed platform's system architecture and data organization plan is shown in Fig. 5. The main concept for visualizing 3D data in a 2D mapping environment is shown in Fig. 4. We propose the following features for the web-based atlas: (1) Open Layers API, (2) image overlay-based 3D visualization, 3D stereo (Fig. 4) , and 3D multi-scale navigation, (3) multi-channel, multi-modal data overlay (see our pilot results on overlaying images with multiple molecular label from Array Tomography [Fig. 6]), (4) registration to standard atlasing space, when available (see Fig. 7 for our pilot results on registering KESM data to Allen reference atlas; for robust registration algorithms, see [88]) to enable correlative analysis (e.g., using Allen mouse brain gene expression atlas [55]) (5) Scalable Vector Graphics(SVG)-based storage for faster access and lower storage demand (preliminary studies showed 5X improvement [17]), (6)

user data upload feature, including automated multi-scale tiling and image processing pipeline, (7) meta data annotation facility, (8) user annotation, textual and ROI (also see §D.2), (9) machine-learning-based analysis (see §D.3), (10) statistical analysis and volume viewer (see §D.4), and (11) cloud-based scalability/availability/sustainability (see §D.5). The implementation will be in PHP, Javascript, and MySQL. Django will be used as a rapid prototyping framework for the web site. See Table 6 for a summary of software engineering practices to be used, technologies to be used, and policies.



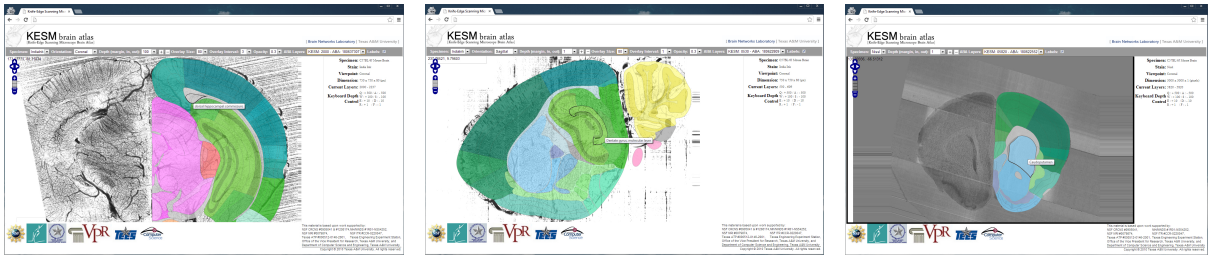
(a) System Architecture (b) Multiscale Tiling (c) Tile Indexing

Figure 5: **Proposed System with OpenLayers API.** (a) Proposed informatics platform's system architecture, using OpenLayers. (b) Multiscale tiling scheme and (c) tile indexing scheme. See [18] for details.



(a) DAPI (b) DAPI+GFP (c) DAPI+GFP+PV

Figure 6: **KESM Brain Atlas Array Tomography Data.** Screenshots of the atlas showing Array Tomography data (mouse cortex, data courtesy of Brad Busse, NIH/NICHD). Three different molecular labels are shown (DAPI, GFP, and PV). Each molecular label can be selectively turned on or off.

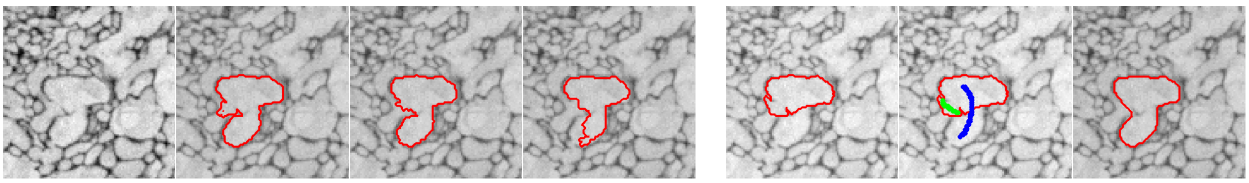


(a) Coronal (India ink) (b) Sagittal (India ink) (c) Coronal (Nissl)

Figure 7: **Registration of KESM Data to the Allen Reference Atlas.** (a-b) Mouse brain vasculature (India ink stain). (c) Mouse brain somata (Nissl stain). Pilot results [97] (Ph.D. work).

D.2 Ground-truth annotation facility

In order to allow users to utilize latest machine learning techniques for the analysis of their data, we will develop an integrated manual and semi-automatic annotation interface for our web-based informatics platform. Both binary and multi-class classification scenarios will be supported. This tool will make ground truth labeling much easier and intuitive.



(a) Multiple Choice **(b) Stroke-based Editing**

Figure 8: **Ground Truth Labeling.** Electron micrograph of zebrafish tectum. (a) Multiple choice. (1) Original image, (2)–(4) segmentation alternatives based on parameter variation. (b) Stroke-based editing. (1) initial segmentation, (2) user strokes (blue: connect, green: disconnect), and (3) edited result. [109]

For machine-learning-based analysis, generating ground truth data is a key requirement. Such ground-truth data can also be used in validating other automated methods of analysis [85, 106, 112]. Since we (and our users) will be dealing with a massive amount of data, complete ground truth data is not feasible, since it amounts to full manual analysis.

We propose to extend our interactive ground-truth generation method [109]. *(1) Multiple choice:* Instead of having users manually perform detailed segmentation and reconstruction to generate the ground truth, we will present a small number of parameterized candidate segmentation or reconstructions (multiple choice). This allows for rapid labeling of ground truth. *(2) Stroke-based edit:* Furthermore, we will use stroke-based correction using graph-cuts to quickly edit erroneous reconstructions. See [109] for details. Although our existing methods are for EM data, since KESM data (and other user data) are expected to be less complex than EM, we anticipate that the approach will generalize well. data sets. *(3) Annotation with 3D context:* Finally, it is important to note that labeling ground truth in 3D data can be difficult when the labeling is done based on viewing one 2D image in the image stack at a time. So, providing a local 3D context becomes important, and our overlay technique (Fig. 4) can naturally help provide this context. For example, Fig. 12a shows an example where an overlay is shown in the annotation display in our web-based platform [96]. Also see Fig. 9b for a similar ground-truth annotation interface (web-based, built on Django). Resulting annotations (in the form of region of interest [ROI]) will be stored in SQL tables and visualized on the web platform using overlays.

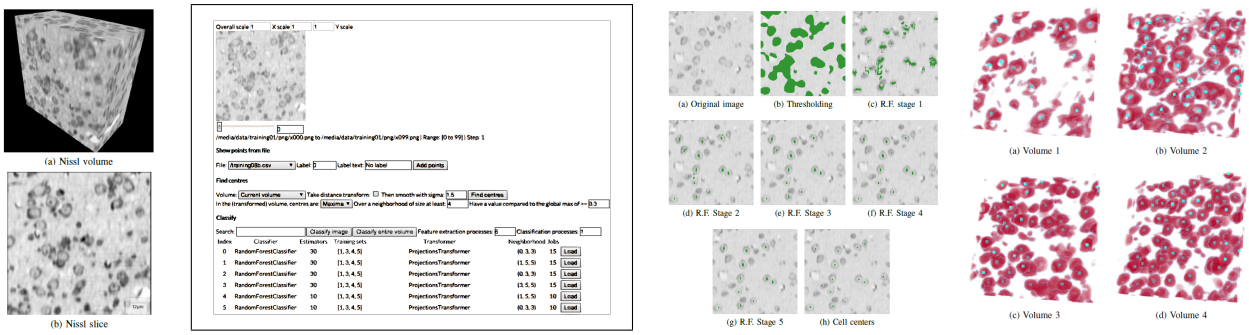
D.3 Machine-learning-based analysis facility

We will develop an integrated machine-learning-based segmentation and geometric reconstruction/analysis tool for our informatics platform. Ground-truth from objective 2 will be used extensively, both for testing and for production.

In biological imaging, machine learning plays an important role [56], and powerful algorithms emerged in the past few years such as deep neural networks that surpass human performance in many benchmarks (e.g., convolutional neural networks [19, 39, 54, 102]). Machine learning excels where prescribed algorithms are not general or flexible enough to deal with varying imaging conditions and noise modalities. Machine learning has been successfully used in various biological image analysis tasks: (1) region segmentation [40, 52, 53], (2) localization [29, 97, 100], (3) classification [25, 43], (4) noise removal or image enhancement [40, 46], and (5) feature extraction [24]. *Our machine-learning-based analysis facility will support all of the above.*

In this project, we will extend our prior work on the use of machine learning for biological image analysis, integrating the methods into our web-based informatics platform. Figs. 9 and 10 show our previous work, where we used machine learning to automatically locate the center of mass in mouse brain Nissl data.

For Fig. 9, we used a random forest classifier to identify the cell center. Note that for this



(a) Data (b) Ground truth annotation web interface (c) Random Forest (d) Results

Figure 9: **Ground Truth Annotation Interface and Machine Learning-Based Analysis.** Rat somatosensory cortex (Nissl-stained). See text for details. Pilot results reported in [53].

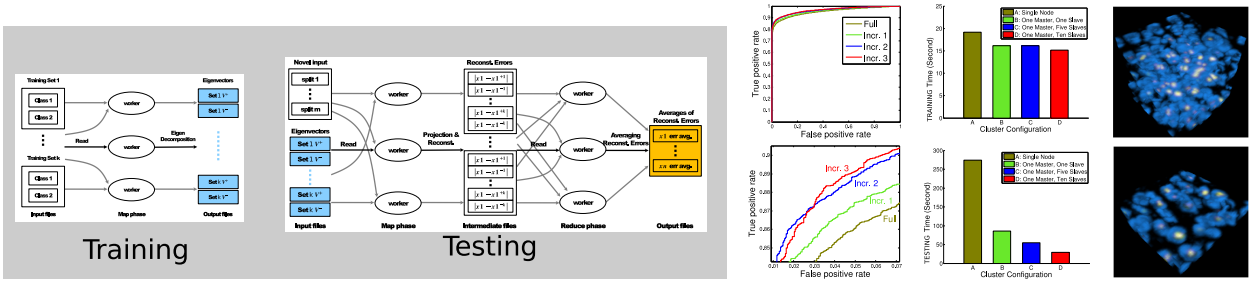
work, we developed a web-based ground truth annotation interface (Fig. 9b, using Django). The data, intermediate results from successive random forest stages, and results are shown in the figure. Precision varied from 0.802 to 0.954, recall from 0.890 to 0.936, and F1 measure from 0.864 to 0.934, showing promising performance [53]. We also experimented with scaling up machine learning methods to exploit cloud computing. Fig. 10 shows our machine learning approach that uses MapReduce, a simple framework for parallel processing developed by Google [21]. MapReduce greatly simplifies parallelization of large scale data analysis algorithms. Fig. 10a shows the Map and Reduce procedure for the training an testing phased in our novel learning algorithm based on Principal Component Analysis (PCA). The implementation was on Amazon Elastic Cloud (computing time generously donated by Amazon). The performance and speed of the experiments are shown in Fig. 10b (Receiver Operating Characteristics [ROC] curve shown), where the Area Under Curve (AUC) reached 0.9614, compared to 0.8228 in a comparable method [100]. Computing speed also scales well as the number of computing nodes are increased, more so for the testing phase than the training phase: this could be due to the identity mapping in the reduce stage in the training phase (Fig. 10c).

(1) We will first incorporate into the informatics platform our own existing methods described above. Then, gradually, (2) we will also include support for general machine learning packages such as Mahout (a machine learning package that runs on MapReduce) [3], WEKA (stand-alone machine learning package) [30], and Caffe (deep learning) [41]. (3) We will design and implement a batch processing configuration interface so that the users can easily set up the training and testing of large volumes of data. (4) Finally, we will implement a validation and editing facility so that the results from machine learning can be further cleaned up and curated.

D.4 Statistical analysis and visualization facility

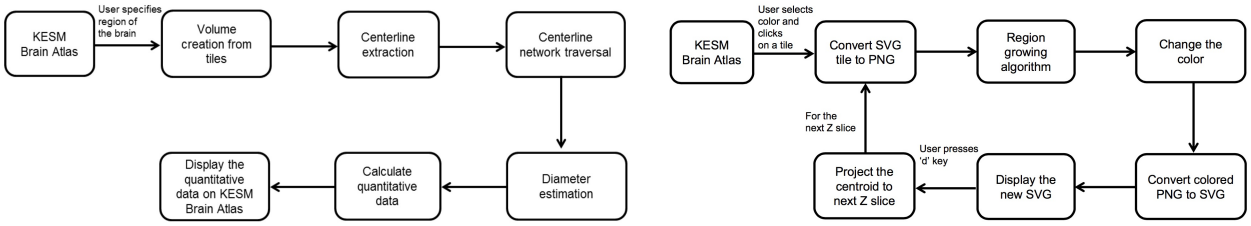
Once segmentation and other basic image analysis tasks are done, first, we need to reconstruct the 3D geometry of the objects in the data set. Statistical analysis and visualization can be done afterwards. We will develop (1) geometric reconstruction algorithms, (2) statistical analysis methods, and (3) in-browser visualization tool for the platform.

(1) Geometric reconstruction: Geometric reconstruction can be done both manually or automatically, depending on the data. In our prior work, we have developed various reconstruction algorithms [23, 31, 32, 95, 108, 111], and methods for fast manual reconstruction [96, 110]. Fig. 11 shows typical reconstruction pipelines for automatic and manual reconstruction. Fig. 12a and b show a reconstruction interface and results. We will incorporate our reconstruction algorithms into the proposed platform, and organize the results in SQL tables.



(a) Map-Reduce Procedure for Hadoop (b) Performance (c) Time (d) Results

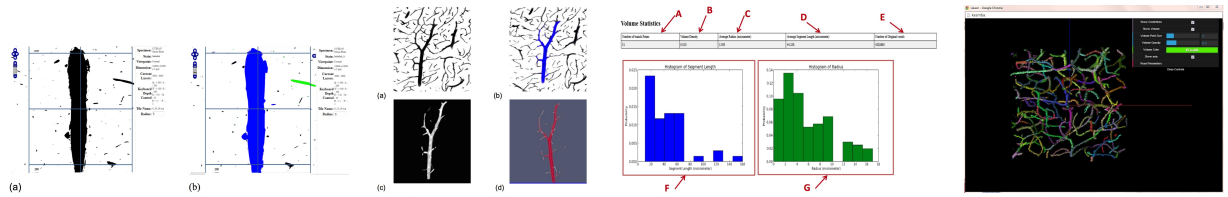
Figure 10: **Map-Reduce-Based Machine Learning for Cell Detection.** The machine learning task of cell center detection was executed on Amazon's EC2 cloud computing platform, running Hadoop, an open-source implementation of Google's Map-Reduce framework. Pilot results reported in [100].



(a) Automated tracing (b) Manual tracing

Figure 11: **Tracing Procedure.** Preliminary method reported in [52]. Note that automated tracing ends in one run, while manual tracing is iterative, requiring the user's repeated input. (MS from the PI's lab.)

(2) Statistical analysis: Once the geometries of the objects of interest are extracted, we can conduct statistical analysis. Fig. 12c shows the results of mouse brain vasculature statistics, shown in our prototype web interface [96] (also see Table. 3). The table and histograms show vascular segment length, diameter, number of branches, and other geometric properties. Such statistics can be used for the study of regional variations, structural organization principles, individual variations, etc. The key task here will be to identify biologically meaningful geometric measures, dependent on the specific data type (model organism and model system). Once these measures are identified, computing the statistics will be straight-forward. Standard statistical procedures will be used to guard against sampling bias. The results from statistical analysis will be visualized in the web interface, and internally stored in SQL tables. The user will also be given an option to download the data in standard data formats such as CSV, etc.



(a) Manual tracing interface (b) Results (c) Statistical summary (d) Volume viewer

Figure 12: **Semi-automated Tracing Interface, Statistical Analysis, and Volume Viewer.** Mouse brain vasculature (India ink-stained). (a) Semi-automated tracing interface where the user can view the local 3D context and click on an object of interest (marked Blue). (b) Original data (top left), tracing results of selected object (top right), in-browser visualization (bottom left), and external app visualization (bottom right). (c) Screenshot of geometry statistics in shown in the web interface. (d) In-browser interactive visualization tool using WebGL. Pilot results [96].

Table 3: **Morphological statistics (KESM vasculature data set) [65].**

Region	# Segments	Length (mm)	# Branches	Surface (mm^2)	Volume (mm^3)	Volume (% of total)
Neocortex	11459.7	758.5	9100.0	10.40	0.0140	1.4%
Cerebellum	34911.3	1676.4	19034.4	20.0	0.0252	2.5%
Spinal Cord	36791.7	1927.6	26449.1	22.2	0.0236	2.4%

(3) Interactive visualization: Since our web-based platform only shows one view point (or up to three orthogonal views at most), subtle aspects of the data may not be easy to explore. Once the region of interest is identified by the user, the user should be able to interactively view the local 3D data volume. For this, we propose to implement a in-browser 3D interactive visualization module using WebGL. Fig. 12d shows a prototype implementation of the module. Currently, the viewer only supports standard view point change and zoom-in/zoom-out. We will add more interactive features so that the user can pick out specific objects of interest and annotate or perform analysis.

D.5 Deployment on the cloud

To provide a scalable and highly available service to the biological research community, we propose to deploy our informatics platform on the cloud.

Cloud computing has come to prominence in the past few years [4]. The definition and benefits of cloud computing are summarized in Table 4. Cloud computing platforms such as Amazon Web Services (AWS) have enabled rapid development and deployment of web-based serviced at an affordable rate, while eliminating complications and risks surrounding IT infrastructure procurement, maintenance, and upgrade. We propose to deploy our informatics platform in two phases.

Table 4: **Cloud Computing: Definition and Benefits.**

What is Cloud Computing?: "on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing" – *Amazon*
1. Minimized infrastructure cost: No need to purchase, maintain, or upgrade hardware. 2. Scalability: Resources are dynamically increased or decreased based on demand. 3. Availability: Typical up time is 99.989% to 100% (e.g., Amazon Elastic Cloud) 4. Economy of scale: Low cost due to economy of scale, and dynamic, usage-based pricing.

(1) Phase I: Local cloud. Initially, we will set up a small-scale local cloud (see the budget request for a 5-rack system), running an open-source cloud computing framework called OpenStack [93]. This will allow us to figure out the exact configuration needed, and to estimate user demand (storage, computing, and data transmission) without incurring use-based charges of the commercial cloud.

(2) Phase II: Commercial cloud. Next, we will migrate our OpenStack-based cloud to commercial clouds such as Amazon Web Services (AWS). We already have experience with AWS [100], and we are currently discussing our needs and possible configurations with AWS solution architect. See budget item for monthly commercial cloud usage fee.

E User Community Engagement Plan

Immediate users of our proposed informatics platform are identified below in Table 5. Fig. 13 shows pilot results from on-going collaboration. The easy access to our multi-scale 3D microscopy data through standard web browsers will assist greatly in the dissemination of our data. The ability to upload and quantitatively analyze the users' own data volumes will help turn the raw data into information and knowledge, which in turn will provide deep insights into biological function.

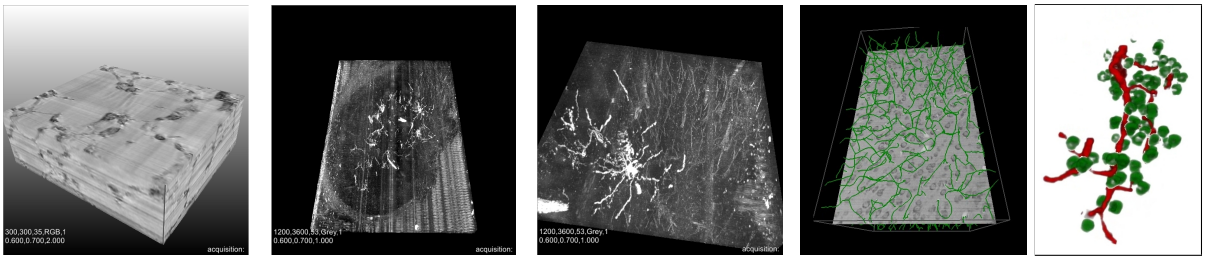
Since we anticipate our collaborators to send us large amounts of data, in most cases in raw image format, we will assist them with the following: (1) consultation on image processing pipeline, (2) multiscale tiling of the data set, (3) consultation on the use of machine learning for the analysis

Table 5: **Diverse User Community (Planned).**

Collaborator Name	Affiliation	Model organism/system	Collaboration Type
Wonbo Shim	Texas A&M (Plant Pathology)	Maize(root)/bacteria	New (see letter)
L. Rene Garcia	Texas A&M (Biology)	C. elegans (nervous system)	New (see letter)
Michael Smotherman	Texas A&M (Biology)	Bat (brain)	New (see letter)
Arum Han	Texas A&M (ECE)	Mouse (neuron cell culture)	New (see letter)
Hojun Song	Texas A&M (Entomology)	Locust (whole organism)	New (see letter)
Todd Huffman	3Scan – Startup company	Mouse, Rat (brain)	On-going (Fig. 14b)
David Edelman	Neurosciences Institute	Octopus (brain)	On-going (Fig. 13b,c)
Ching-Long Lin	University of Iowa (MIE)	Mouse (lung)	On-going (Fig. 13a)
David Mayerich	University of Houston (ECE)	Mouse (neurovasculature)	On-going (Fig. 13d,e)
Jaerock Kwon	Kettering University (ECE)	Mouse (neurovasculature)	On-going
Brad Busse	NIH/NICHD	Mouse (synapse)	On-going (Fig. 6)

* ECE: Electrical and Computer Engineering, MIE: Mechanical and Industrial Engineering
* NIH/NICHD: National Institutes of Health/National Institute of Child Health and Human Development

of their data (including types of ground-truth needed), (4) iterative user interface improvement and customization, (5) choice of statistical measures. We will develop step-by-step tutorials for our collaborators and users. We will also (6) identify potential users at exhibits, tutorials, and workshops we organize, and assist them in the same manner discussed above. (7) We will use Google Analytics (or similar) service to monitor the usage of our resource (see Fig. 14c-d for our current access stats).



(a) mouse lung (b) octopus brain Supraesophageal mass(SEM) (c) octopus brain (SEM [zoomed in]) (d) Rat vasculature (e) Rat vasculature & nearby neurons

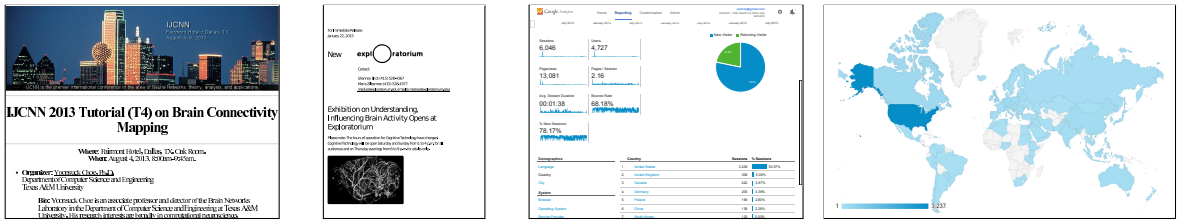
Figure 13: **Example Projects.** Pilot scans using the KESM are shown to highlight the broad applicability of the instrument to a wide range of biological research. (a) Alveoli in the mouse lung (210 µm-wide block). (b–c) Octopus brain (supraesophageal mass [SEM]). The block widths are 840 µm and 420 µm, respectively. (d–e) Rat brain (somatosensory cortex) vasculature and neighboring neurons (from Nissl data). In (d), dark gray spots are cell nuclei and the white discs are the cross sections of the vascular network (traced in green). In (e), traced vasculature (red) and neighboring neuronal cell bodies (green) are shown.

F Broader Impacts of the Proposed Work

(1) Education: We will train three graduate students so that they can become experts in biological informatics. We have trained undergraduate students in the past through the NSF Research Experience for Undergraduates (REU) program, and will continue to utilize this mechanism. The neuronal data from our whole mouse and other biological data are expected to serve as a rich resource for educational use. The multi-scale data scheme, together with custom annotations, will allow us to turn the web-based interface into a major educational resource for all levels of education (K-12 to graduate). We will set up a scaled-down version of the KESM Brain Atlas on a dedicated "for kids" page so that K-12 students and teachers can have a centralized access. The K-12 education portal will include (1) interactive tutorials highlighting the history of neuroscience and the key role played by neuroanatomy and (2) an Easter-egg hunt game using the web-based atlas, to allow students to search for key anatomical features they learned in the tutorial.

(2) Outreach: We will also organize exhibits and tutorials at biology conferences to broaden the user base and adoption of our informatics platform (PI Choe has extensive experience with organizing exhibits, workshops, and tutorials). Our work has also been featured in San Francisco's Exploratorium exhibit (2015 spring), and we will pursue similar opportunities for public outreach. See §C.4 and Fig. 14a-b for details.

(3) Data and code dissemination: Please see the *Plan for Preparation and Deployment* and *Software Development and Sustainability* in §G Management Plan below.



(a) Tutorial (b) Museum Exhibit (c) Access stats (d) World-wide access

Figure 14: **Broader Impact (Outreach).** Web page screenshots: (a) 2013 Int'l Joint Conference on Neural Networks Tutorial on connectivity mapping. (b) 2015 San Francisco Exploratorium Exhibit on the brain (featuring KESM India ink data at the bottom: cf. Fig. 2b; exhibit arranged by collaborator 3Scan, located in San Francisco). Google Analytics data on the KESM Brain Atlas: (c) access stats (February 2012 to present: 4,727 visitors, 13,081 page views) and (d) world-wide access plot showing broad world-wide usage. Some screenshots were edited (elements rearranged) to fit the page.

G Management Plan

(1) Plan for Preparation and Deployment: We have acquired a dedicated domain name to set up a web portal (see Fig. 3). First, the web-based informatics platform and all KESM data will be released to the public (year 1). Next, the source code for the web platform will also be made available online, with continued maintenance by the project team (years 2–3). The source code will be maintained as an open source project on GitHUB. We will use the SCRUM method (a form of Agile development) to facilitate rapid, accurate development. We will actively advertise this new resource through various channels: mailing lists, news briefs in scientific publications, personal contact, and tutorials/demos/exhibits at scientific meetings. Progress will be monitored using web site analytics services (e.g., Google Analytics: see Fig. 14c-d). We will also organize short courses and workshops to expand and support the user community. For the above, the PI (Choe) will depend on his extensive experience in the above activities (§C.4; Fig. 14a).

(2) Software Development and Sustainability: The PI/Co-PIs will design the main software architecture and do partial implementation, and three graduate students will assist in the design process and carry out the implementation. Co-PI Abbott (neuroscientist) and other collaborators identified in Table 5 will test the tools and provide continual feedback. In order to ensure continued support of the informatics platform developed through this project, we will design and implement a protocol for software development, documentation, and education. We will make extensive use of the collaborative software development platform GitHub (Texas A&M University has a GitHUB Enterprise license). All members of the development team (both internal and external) will be trained to abide by this protocol. We expect maintenance to require significantly less effort once the initial platform is implemented during the project period, and especially once the platform is migrated to a commercial cloud (see specific budget item for this). Migration to the cloud will obviate the need for expensive hardware upgrades, and data loss due to hardware failure. We will also seek continued funding from follow-up grants and from collaborations made through the web platform, e.g. the ABI Sustaining grant. Table 6 below summarizes software engineering practices,

Table 6: **Summary of Software Engineering Practices, Technology, and Policy.**

Devel/Tech/Policy	Choice of development practices/technology/policy
Code version control & sharing	GitHub Enterprise (Texas A&M has a site license)
Mapping API	OpenLayers
User authentication	Third party login (OAuth: Google, Twitter, Facebook, etc.)
Data standards	OME(Open Microscopy Env)-XML, NeuroML, BrainML, CellML, etc.
Cloud framework	OpenStack (development), Amazon Web Services (production)
Cloud computing	Spark+Hadoop on Amazon Elastic MapReduce
Machine learning	Mahout on Amazon Elastic MapReduce; WEKA
Website usage monitoring	Google Analytics, Amazon CloudBuddy Analytics
Software development approach	SCRUM (iterative development + prototyping) + Test-Driven Devel.
Rapid prototyping	Django (python-based web framework for rapid development) and Custom PHP, Javascript, and MySQL code
Code license	GNU public license (or similar): see Intellectual property, below
Data license	Creative Commons or similar (user uploaded data: elected by user)
Intellectual property	To be reviewed by Texas A&M technology commercialization office prior to public disclosure
Sustainability	NSF ABI Sustaining mechanism + commercial cloud = scalability + long-term availability + no infrastructure headaches

technologies to be used, and policies to be adopted by this project.

(3) Potential Risks and Risk Management: Since for most of the proposed works in §D we have done pilot studies, we do not expect major risks in terms of the conceptual and technical design. However, there are multiple risks typical for informatics platforms of the proposed scale: Hardware failure, security breach, data loss, incompatibility of legacy code during OS/API upgrade, etc. All of these potential risks are significantly lowered by moving the whole platform to a commercial cloud as we propose here. Furthermore, in the long run, the cost for sustaining the service is significantly lowered, both by avoiding expensive hardware upgrades and/or dedicated maintenance staff. Commercial clouds are not without inherent risks. For example, misconfiguration of the services can lead to security incidents or quick draining of the account balance. We will follow established guidelines regarding cloud security [48] and establish user quotas to avoid draining resources. Also, our initial implementation of the platform on a small-scale local cloud (see budget request for 5-rack server for this purpose) will help us accurately estimate the computing demand prior to deployment on the commercial cloud.

(4) Timeline and Responsibility: Table 7 below shows the timeline of the project and the responsible party for each task. Please see the budget justification for a more detailed breakdown of responsibilities of the PI/Co-PIs. The research team will hold a weekly meeting to discuss the progress and plan ahead. The research team and collaborators will have an online all-hands meeting at the end of each year (using Skype or Teamviewer, once per year), while keeping in constant contact over the year on an individual basis.

Table 7: **Tasks & Deliverables, Timeline, and Personnel**

Tasks and Deliverables	Year 1			Year 2			Year 3			Lead, Support
Task §D.1 Web-based atlas	▬	▬	▬	▬						Choe, Keyser
Task §D.2 Ground-truth annotation tool		▬	▬	▬	▬					Choe, Abbott
Task §D.3 Machine-learning-based analysis			▬	▬	▬	▬	▬			Choe, Keyser
Task §D.4 Statistical analysis and visualization						▬	▬	▬	▬	Keyser, Abbott
Task §D.5 Deployment on the cloud							▬	▬	▬	Choe, Keyser
Task §E User Community Engagement	▬			▬			▬	▬	▬	Abbott, Choe
Task §F Broader Impact	▬	▬		▬	▬		▬	▬	▬	Abbott, Choe

* Each year is divided into Summer, Fall, and Spring Semester.

REFERENCES CITED

[1] Al-Kofahi, K. A., Lasek, S., Szarowski, D. H., Pace, C. J., Nagy, G., Turner, J. N., and Roysam, B. (2002). Rapid automated three-dimensional tracing of neurons from confocal image stacks. *IEEE Transactions on Information Technology in Biomedicine*, 6:171–187.

[2] Allen Institute for Brain Science, Allen brain atlas. http://www.brain-map.org/.

[3] Anil, R., Owen, S., Dunning, T., and Friedman, E. (2010). *Mahout in Action*. Greenwich, CT: Manning Publications Co.

[4] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., et al. (2010). A view of cloud computing. *Communications of the ACM*, 53(4):50–58.

[5] BrainMaps, Brainmaps.org. `http://brainmaps.org`.

[6] BrainML. http://brainml.org.

[7] Brown, K. M., Donohue, D. E., D'Alessandro, G., and Ascoli, G. A. (2007). A cross-platform freeware tool for digital reconstruction of neuronal arborizations from image stacks. *Neuroinformatics*, 3:343–359.

[8] Burns, R., Lillaney, K., Berger, D. R., Grosenick, L., Deisseroth, K., Reid, R. C., Roncal, W. G., Manavalan, P., Bock, D. D., Kasthuri, N., et al. (2013). The open connectome project data cluster: scalable analysis and vision for high-throughput neuroscience. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, 27. ACM.

[9] Can, A., Shen, H., Turner, J. N., Tanenbaum, H. L., and Roysam, B. (1999). Rapid automated tracing and feature extraction from retinal fundus images using direct exploratory algorithms. *IEEE Transactions on Information Technology in Biomedicine*, 3:125–138.

[10] Cell Centered Database. http://ncmir.ucsd.edu/CCDB/.

[11] Choe, Y., Abbott, L. C., Han, D., Huang, P.-S., Keyser, J., Kwon, J., Mayerich, D., Melek, Z., and McCormick, B. H. (2009). Knife-edge scanning microscopy: High-throughput imaging and analysis of massive volumes of biological microstructures. In Rao, A. R., and Cecchi, G., editors, *High-Throughput Image Reconstruction and Analysis: Intelligent Microscopy Applications*, 11–37. Boston, MA: Artech House.

[12] Choe, Y., Abbott, L. C., Miller, D. E., Han, D., Yang, H.-F., Chung, J. R., Sung, C., Mayerich, D., Kwon, J., Micheva, K., and Smith, S. J. (2010). Multiscale imaging, analysis, and integration of mouse brain networks. In *Neuroscience Meeting Planner, San Diego, CA: Society for Neuroscience*. Program No. 516.3. Online.

[13] Choe, Y., Abbott, L. C., Ponte, G., Keyser, J., Kwon, J., Mayerich, D., Miller, D., Han, D., Grimaldi, A. M., Fiorito, G., Edelman, D. B., and McKinstry, J. L. (2010). Charting out the octopus connectome at submicron resolution using the knife-edge scanning microscope. *BMC Neuroscience*, 11(Suppl 1):P136. Nineteenth Annual Computational Neuroscience Meeting: CNS*2010.

[14] Choe, Y., Mayerich, D., Kwon, J., Miller, D. E., Chung, J. R., Sung, C., Keyser, J., and Abbott, L. C. (2011). Knife-edge scanning microscopy for connectomics research. In *Proceedings of the International Joint Conference on Neural Networks*, 2258–2265. Piscataway, NJ: IEEE Press.

[15] Choe, Y., Mayerich, D., Kwon, J., Miller, D. E., Sung, C., Chung, J. R., Huffman, T., Keyser, J., and Abbott, L. C. (2011). Specimen preparation, imaging, and analysis protocols for knife-edge scanning microscopy. *Journal of Visualized Experiments*, 58:e3248. Doi: 10.3791/3248.

[16] Choe, Y., Sung, C., Choi, J., Srivastava, M., Priour, M. R., Mayerich, D., Keyser, J., and Abbott, L. C. (2014). Open web atlas for high-resolution 3d mouse brain data. In *Neuroscience Meeting Planner, Washington, DC: Society for Neuroscience*. Program No. 185.02. Online.

[17] Choi, J. (2013). *Knife-Edge Scanning Microscope Mouse Brain Atlas in Vector Graphics for Enhanced Performance*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[18] Chung, J. R., Sung, C., Mayerich, D., Kwon, J., Miller, D. E., Huffman, T., Abbott, L. C., Keyser, J., and Choe, Y. (2011). Multiscale exploration of mouse brain microstructures using the knife-edge scanning microscope brain atlas. *Frontiers in Neuroinformatics*, 5:29.

[19] Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Advances in neural information processing systems*, 2843–2851.

[20] Cuellar, A. A., Lloyd, C. M., Nielsen, P. F., Bullivant, D. P., Nickerson, D. P., and Hunter, P. J. (2003). An overview of cellml 1.1, a biological model description language. *Simulation*, 79(12):740–747.

[21] Dean, J., and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of ACM*, 51:107–113.

[22] Denk, W., and Horstmann, H. (2004). Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure. *PLoS Biology*, 19:e329.

[23] Dileepkumar, A. (2014). *Semi-Automated Reconstruction of Vascular Networks in Knife-Edge Scanning Microscope Mosue Brain Data*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[24] Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.

[25] Dreiseitl, S., Ohno-Machado, L., Kittler, H., Vinterbo, S., Billhardt, H., and Binder, M. (2001). A comparison of machine learning methods for the diagnosis of pigmented skin lesions. *Journal of biomedical informatics*, 34(1):28–36.

[26] Eng, D. C.-Y., and Choe, Y. (2008). Stereo pseudo 3D rendering for web-based display of scientific volumetric data. In *Proceedings of the IEEE/EG International Symposium on Volume Graphics*.

[27] Fiala, J. C. (2005). Reconstruct: A free editor for serial section microscopy. *Jounral of Microscopy*, 218:52–61.

[28] Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., Goldberg, D. H., Grafstein, B., Grethe, J. S., Gupta, A., et al. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, 6(3):149–160.

[29] Glory, E., and Murphy, R. F. (2007). Automated subcellular location determination and high-throughput microscopy. *Developmental cell*, 12(1):7–16.

[30] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[31] Han, D., Choi, H., Park, C., and Choe, Y. (2009). Fast and accurate retinal vasculature tracing and kernel-isomap-based feature selection. In *Proceedings of the International Joint Conference on Neural Networks*, 1075–1082. Piscataway, NJ: IEEE Press.

[32] Han, D., Keyser, J., and Choe, Y. (2009). A local maximum intensity projection tracing of vasculature in Knife-Edge Scanning Microscope volume data. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 1259–1262.

[33] Hayworth, K. J., Kasthuri, N., Schalek, R., and Lichtman, J. W. (2006). Automating the collection of ultrathin sections for large volume TEM reconstructions. *Microscopy and Microanalysis*, 12(Suppl. S02):86–87.

[34] Hayworth, K. J., Morgan, J. L., Schalek, R., Berger, D. R., Hildebrand, D. G., and Lichtman, J. W. (2014). Imaging atum ultrathin section libraries with wafermapper: a multi-scale approach to em reconstruction of neural circuits. *Frontiers in neural circuits*, 8.

[35] Huisken, J., and Stainier, D. Y. (2009). Selective plane illumination microscopy techniques in developmental biology. *Development*, 136(12):1963–1975.

[36] Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J., and Stelzer, E. H. (2004). Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, 305(5686):1007–1009.

[37] Jain, V., Bollmann, B., Richardson, M., Berger, D. R., Helmstaedter, M. N., Briggman, K. L., Denk, W., Bowden, J. B., Mendenhall, J. M., Abraham, W. C., Harris, K. M., Kasthuri, N., Hayworth, K. J., Schalek, R., Tapia, J. C., Lichtman, J. W., and Seung, H. S. (2010). Boundary learning by optimization with topological constraints. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2488–2495.

[38] Jain, V., Murray, J. F., Roth, F., Seung, H. S., Turaga, S., Briggman, K., Denk, W., and Helmstaedter, M. (2007). Using machine learning to automate volume reconstruction of neuronal shapes from nanoscale images. In *Society for Neuroscience Abstracts*. Washington, DC: Society for Neuroscience. Program No. 534.7. Online.

[39] Jain, V., Murray, J. F., Roth, F., Turaga, S., Zhigulin, V., Briggman, K. L., Helmstaedter, M. N., Denk, W., and Seung, H. S. (2007). Supervised learning of image restoration with convolutional networks. In *IEEE 11th International Conference on Computer Vision (ICCV 2007)*, 1–8.

[40] Jain, V., Seung, H. S., and Turaga, S. C. (2010). Machines that learn to segment images: a crucial technology for connectomics. *Current Opinion in Neurobiology*, 20:653–666.

[41] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, 675–678. ACM.

[42] Johnson, G. A., Badea, A., Brandenburg, J., Cofer, G., Fubara, B., Liu, S., and Nissanov, J. (2010). Waxholm space: An image-based reference for coordinating mouse brain research. *Neuroimage*, 53:365–372.

[43] Kamentsky, L., Jones, T. R., Fraser, A., Bray, M.-A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W., and Carpenter, A. E. (2011). Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. *Bioinformatics*, 27(8):1179–1180.

[44] Keller, P. J., Schmidt, A. D., Santella, A., Khairy, K., Bao, Z., Wittbrodt, J., and Stelzer, E. H. (2010). Fast, high-contrast imaging of animal development with scanned light sheet-based structured-illumination microscopy. *Nature methods*, 7(8):637–642.

[45] Keller, P. J., Schmidt, A. D., Wittbrodt, J., , and Stelzer, E. H. K. (2008). Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science*, 322:1065–1069.

[46] Kenig, T., Kam, Z., and Feuer, A. (2010). Blind image deconvolution using machine learning for three-dimensional microscopy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(12):2191–2204.

[47] Kim, D. (2011). *Automatic Seedpoint Selection and Tracing of Microstructures in the Knife-Edge Scanning Microscope Mouse Brain Data Set*. Master's thesis, Department of Computer Science, Texas A&M University, College Station, Texas.

[48] Krutz, R. L., and Vines, R. D. (2010). *Cloud security: A comprehensive guide to secure cloud computing*. John Wiley & Sons.

[49] Kwon, J., Lim, J., Lee, S., Mayerich, D., Keyser, J., Abbott, L. C., and Choe, Y. (2014). High-throughput and high-resolution 3d tissue scanner with internet connected 3d virtual microscope for large-scale automated histology. In *Neuroscience Meeting Planner, Washington, DC: Society for Neuroscience*. Program No. 185.15. Online.

[50] Kwon, J., Mayerich, D., and Choe, Y. (2011). Automated cropping and artifact removal for knife-edge scanning microscopy. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 1366–1369.

[51] Kwon, J., Mayerich, D., Choe, Y., and McCormick, B. H. (2008). Lateral sectioning for knife-edge scanning microscopy. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 1371–1374.

[52] Lal Das, S. (2014). *Cell Detection in Knife-Edge Scanning Microscopy Images of Nissl-stained Mouse and Rat Brain Samples using Random Forests*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[53] Lal Das, S., Keyser, J., and Choe, Y. (2015). Random-forest-based automated cell detection in knife-edge scanning microscope rat nissl data. In *Proceedings of the International Joint Conference on Neural Networks*. In press.

[54] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551.

[55] Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445:168–176.

[56] Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *Neuroimage*, 56(2):387–399.

[57] Li, A., Gong, H., Zhang, B., Wang, Q., Yan, C., Wu, J., Liu, Q., Zeng, S., and Luo, Q. (2010). Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science*, 330:1404–1408. *See the E-Letter commentary by Mayerich et al.*

[58] Linkert, M., Rueden, C. T., Allan, C., Burel, J.-M., Moore, W., Patterson, A., Loranger, B., Moore, J., Neves, C., MacDonald, D., et al. (2010). Metadata matters: access to image data in the real world. *The Journal of cell biology*, 189(5):777–782.

[59] MacKenzie-Graham, A., Jones, E. S., Shattuck, D. W., Dinov, I. D., Bota, M., and Toga, A. W. (2003). The informatics of a C57BL/6J mouse brain atlas. *Neuroinformatics*, 1:397–410.

[60] Martone, M. E., Gupta, A., Wong, M., Qian, X., Sosinsky, G., Ludscher, B., and Ellisman, M. H. (2002). A cell-centered database for electron tomographic data. *J. Struct. Biol.*, 138:145–155.

[61] Martone, M. E., Tran, J., Wong, W. W., Sargis, J., Fong, L., Larson, S., Lamont, S. P., Gupta, A., and Ellisman, M. H. (2008). The Cell Centered Database project: An update on building community resources for managing and sharing 3d imaging data. *Journal of Structural Biology*, 161:220–231.

[62] Marx, V. (2013). Neuroscience waves to the crowd. *Nature methods*, 10(11):1069–1074.

[63] Mayerich, D., Abbott, L. C., and McCormick, B. H. (2008). Knife-edge scanning microscopy for imaging and reconstruction of three-dimensional anatomical structures of the mouse brain. *Journal of Microscopy*, 231:134–143.

[64] Mayerich, D., and Keyser, J. (2008). Filament tracking and encoding for complex biological networks. In *Proceedings of ACM Symposium on Solid and Physical Modeling*, 353–358.

[65] Mayerich, D., Kwon, J., Choe, Y., Abbott, L., and Keyser, J. (2008). Constructing high-resolution microvascular models. In *Proceedings of the 3rd International Workshop on Microscopic Image Analysis with Applications in Biology (MIAAB 2008)*. Online.

[66] Mayerich, D., Kwon, J., Panchal, A., Keyser, J., and Choe, Y. (2011). Fast cell detection in high-throughput imagery using gpu-accelerated machine learning. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 719–723.

[67] Mayerich, D., McCormick, B. H., and Keyser, J. (2007). Noise and artifact removal in knife-edge scanning microscopy. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*, 556–559.

[68] Mayerich, D. M., Melek, Z., and Keyser, J. (2007). Fast filament tracking using graphics hardware. Technical Report TAMU-CS-TR-2007-11-3, Department of Computer Science, Texas A&M University.

[69] McCormick, B. H. (2003). The knife-edge scanning microscope. Technical report, Department of Computer Science, Texas A&M University. http://research.cs.tamu.edu/bnl/.

[70] McCormick, B. H., System and method for imaging an object. USPTO patent #US 6,744,572 (for Knife-Edge Scanning; 13 claims).

[71] McCormick, B. H., Abbott, L. C., Mayerich, D. M., , Keyser, J., Kwon, J., Melek, Z., and Choe, Y. (2006). Full-scale submicron neuroanatomy of the mouse brain. In *Society for Neuroscience Abstracts*. Washington, DC: Society for Neuroscience. Program No. 694.5. Online.

[72] McCormick, B. H., and Mayerich, D. M. (2004). Three-dimensional imaging using Knife-Edge Scanning Microscope. *Microscopy and Microanalysis*, 10 (Suppl. 2):1466–1467.

[73] MeVis Medical Solutions AG, and Fraunhofer MEVIS, Mevislab. Http://www.mevislab.de.

[74] Micheva, K., and Smith, S. J. (2007). Array tomography: A new tool for imaging the molecular architecture and ultrastructure of neural circuits. *Neuron*, 55:25–36.

[75] Micheva, K., and Smith, S. J. (2007). Array tomography: A new tool for imaging the molecular architecture and ultrastructure of neural circuits. *Neuron*, 55:25–36.

[76] MicroBrightField, Inc., Autoneuron. http://www.mbfbioscience.com/autoneuron.

[77] MicroBrightField, Inc., Neurolucida. http://www.mbfbioscience.com/neurolucida.

[78] Mikula, S., Trotts, I., Stone, J. M., and Jones, E. G. (2007). Internet-enabled high-resolution brain mapping and virtual microscopy. *Neuroimage*, 35:9–15.

[79] Miller, D. E. (2014). *A Combined Skeleton Model*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[80] National Center for Microscopy and Imaging Research. http://www.ncmir.ucsd.edu/.

[81] NeuroML. http://www.neuroml.org/.

[82] Oh, S. W., Harris, J. A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A. M., et al. (2014). A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214.

[83] Open Source Geospatial Foundation, Openlayers version 3. `http://openlayers.org/`.

[84] Paxinos, G., and Franklin, K. B. J. (2001). *The Mouse Brain in Stereotaxic Coordinates*. San Diego, CA: Academic Press. Deluxe second edition. (with CD-ROM).

[85] Pham, D. L., Xu, C., and Prince, J. L. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2:315–337.

[86] Roysam, B. et al., FARSIGHT. Http://www.farsight-toolkit.org.

[87] Saalfeld, S., Cardona, A., Hartenstein, V., and Tomančák, P. (2009). Catmaid: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics*, 25(15):1984–1986.

[88] Schaefer, S., McPhail, T., and Warren, J. (2006). Image deformation using moving least squares. *ACM Transactions on Graphics*, 25:533–540.

[89] Schalek, R., Wilson, A., Lichtman, J., Josh, M., Kasthuri, N., Berger, D., Seung, S., Anger, P., Hayworth, K., and Aderhold, D. (2012). Atum-based sem for high-speed large-volume biological reconstructions. *Microscopy and Microanalysis*, 18(S2):572–573.

[90] Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7):676–682.

[91] Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671–675.

[92] Schroeder, W., and Martin, K. (2006). *Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. Clifton Park, NY: Kitware. Fourth edition.

[93] Sefraoui, O., Aissaoui, M., and Eleuldj, M. (2012). Openstack: toward an open-source solution for cloud computing. *International Journal of Computer Applications*, 55(3):38–42.

[94] Seung, S., and Burnes, L., Eyewire. `http://eyewire.org/`.

[95] Singhal, A. (2015). *Skeletonization-Based Automated Tracing and Reconstruction of Neurovascular Networks in Knife-Edge Scanning Microscope Mouse Brain India Ink Data*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[96] Srivastava, M. (2015). *Knife-Edge Scanning Microscope Brain Atlas Interface for Tracing and Analysis of Vasculature Data*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[97] Sung, C. (2013). *Exploration, Registration, and Analysis of High-Throughput 3D Microscopy Data from the Knife-Edge Scanning Microscope*. PhD thesis, Department of Computer Science and Engineering, Texas A&M University.

[98] Sung, C., Chung, J. R., Mayerich, D., Kwon, J., Miller, D. E., Huffman, T., Keyser, J., Abbott, L. C., and Choe, Y. (2011). Knife-edge scanning microscope brain atlas: A submicrometer-resolution web-based mouse brain atlas. In *Neuroscience Meeting Planner, Washington, DC: Society for Neuroscience*. Program No. 328.05. Online.

[99] Sung, C., Mayerich, D., Kwon, J., Miller, D. E., Abbott, L. C., Keyser, J., Huffman, T., and Choe, Y. (2012). Web-based knife-edge scanning microscope brain atlas in vector-graphics for enhanced performance. In *Neuroscience Meeting Planner, New Orleans, LA: Society for Neuroscience*. Program No. 328.05. Online.

[100] Sung, C., Woo, J., Goodman, M., Huffman, T., and Choe, Y. (2013). Scalable, incremental learning with MapReduce parallelization for cell detection in high-resolution 3D microscopy data. In *Proceedings of the International Joint Conference on Neural Networks*, 434–440.

[101] Tsai, P. S., Friedman, B., Ifarraguerri, A. I., Thompson, B. D., Lev-Ram, V., Schaffer, C. B., Xiong, Q., Tsien, R. Y., Squier, J. A., and Kleinfeld, D. (2003). All-optical histology using ultrashort laser pulses. *Neuron*, 39:27–41.

[102] Turaga, S. C., Murray, J. F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., and Seung, H. S. (2010). Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Computation*, 22(2):511–538.

[103] von Ahn, L. (2006). Games with a purpose. *IEEE Computer*, 39:96–96.

[104] von Ahn, L., Kedia, M., and Blum, M. (2006). Peekaboom: A game for locating objects in images. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHO 2006)*, 55–64.

[105] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321:1465–1468.

[106] Warfield, S. J., Zou, K. H., and Wells, W. M. (2002). Validation of image segmentation and expert quality with expectation-minimization algorithm. In *Lecture Notes In Computer Science, Vol. 2488; Proceedings of the 5th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 298–306.

[107] Whole Brain Catalog. http://wholebraincatalog.org/.

[108] Yang, H.-F., and Choe, Y. (2010). Electron microscopy image segmentation with estimated symmetric three-dimensional shape prior. In *Proceedings of the 6th International Symposium on Visual Computing*.

[109] Yang, H.-F., and Choe, Y. (2011). Ground truth estimation by maximizing topological agreements in electron microscopy data. In *Proceedings of the 7th International Symposium on Visual Computing (LNCS 6938)*, 371–380.

[110] Yang, H.-F., and Choe, Y. (2011). An interactive editing framework for electron microscopy image segmentation. In *Proceedings of the 7th International Symposium on Visual Computing (LNCS 6938)*, 400–409.

[111] Yang, W. (2014). *Automated neurovascular tracing and analysis of the Knife-Edge Scanning Microscope India ink data set*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[112] Yoo, T., Ackerman, N. J., and Vannier, M. (2000). Toward a common validation methodology for segmentation and registration algorithms. In *Lecture Notes In Computer Science, Vol. 1935; Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Intervention*, 422–431. London: Springer.

[113] Zhang, W. (2014). *Real-time Image Error Detection in Knife-Edge Scanning Microscope*. Master's thesis, Department of Computer Science and Engineering, Texas A&M University.

[114] Zhang, W., Yoo, J., Keyser, J., Abbott, L. C., and Choe, Y. (2015). Real-time detection of imaging errors in the knife-edge scanning microscope through change detection. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*. Accepted.

[115] Zuse Institute Berlin (ZIB) and Mercury Computer Systems, Berlin, Amira: Advanced 3D visualization and volume modeling. http://www.amiravis.com.