### Department of Biomedical Informatics

# Department of Biomedical Informatics at Emory University

Emory University, with its mission is to create, preserve, teach and apply knowledge in the service of humanity, is one of the world's leading research universities. It is recognized for its outstanding liberal arts[colleges, graduate and professional schools](http://www.emory.edu/home/academics/index.html), and one of the Southeast's[leading health care systems](http://www.emory.edu/home/about/healthcare.html). Emory is located on a[beautiful, tree-laced campus](http://www.emory.edu/home/_cms/OLD/campus/index.html) in Atlanta, Georgia's historic Druid Hills suburb where the more than 14,000 undergraduate and graduate students received an innovative and focused education. Emory consistently ranks amongst *U.S. News and World Report’s* top 25 Universities in the United States for its undergraduate and professional programs. It enjoys strong relationships with Georgia Institute of Technology through its shared Department of Biomedical Engineering.

# Space

As part of a $10M commitment from the university and the School of Medicine, the Department of Biomedical Informatics relocated to a newly renovated space on the 4th floor of the Woodruff Memorial Building in 2017. This space is adjacent to the Emory University Hospital, and this new location will offer students and trainees unprecedented access to senior researchers across campus in support of translational bioinformatics and clinical research bioinformatics research. The 7,745 square feet of usable space comprises offices for faculty and staff, open plan workstations for staff and two dedicated student workrooms. There is a dedicated 611 square foot multi-purpose classroom with seating for 16 individuals, a server room, an animal facility, a dedicated 7-seat hot-desking office for clinical faculty and visiting professors, a 255-square foot conference room with seating for 12 and several informal meeting spaces within the office suite to promote an environment centered around collaboration.

# Offices

Each faculty member has a private office of ~150 sq. ft. and adjacent workspaces for staff. All Emory University co-investigators included in this proposal have their own dedicated offices. Additional swing space is available for faculty of the proposed program. A departmental office houses the administrative assistants and department manager who handle grants management, teaching materials, manuscripts and grant applications.

# Communications

The Biomedical Informatics Department has multiple Logitech Group systems which allow researchers to host and join group video conferences simply by connecting a laptop to the Logitech hub managing the room’s conference camera and speakerphone. All conference rooms have dedicated video conferencing facilities.

# Computing Environment

**The HPC cluster** consists of 23 discrete nodes of multiple types:

●       Type A are 2U single CPU single GPU high-RAM large scratch-space systems (8 nodes)

○       AMD EPYC 7402P 24-core 2.80GHz CPU

○       256 GB DDR4 RAM

○       Nvidia A30 GPU (24 GB RAM, 3804 CUDA cores, 224 Tensor cores)

○       60 TB of NVMe-backed onboard scratch-space

●       Type B are 2U single CPU dual GPU high-RAM large scratch-space systems (7 nodes)

○       AMD EPYC 9254 24-core 2.90GHz CPU

○       384 GB DDR5 RAM

○       2 x Nvidia RTX 6000 ADA GPUs (48 GB RAM, 18716 CUDA cores, 568 Tensor cores)

○       60 TB of NVMe-backed onboard scratch-space

●       Type C are VMs with direct access to the Nvidia Quadro RTX 6000 GPUs and Infiniband network cards in their host systems (4 nodes)

○       AMD EPYC processor (virtual) 8-core 2GHz CPU

○       64 GB RAM

○       Nvidia Quadro RTX 6000 GPU (24 GB RAM, 4608 CUDA cores, 576 Tensor cores)

●       Type D are specialty systems that have unique configurations from the other nodes in the cluster, such as the Nvidia-DGX1 (4 nodes, each defined below)

○       2 x Intel Xeon E5-2698 v4 20-core 3.6GHz CPUs | 512 GB RAM | 8 x Nvidia Tesla V100 GPUs (32 GB RAM)

○       2 x Intel Xeon E5-2640 v4 10-core 3.40GHz CPUs | 1 TB RAM | 4 x Nvidia Tesla P100 GPUs (12 GB RAM)

○       2 x AMD EPYC 7742 64-core 2.25GHz CPUs | 1 TB RAM | 8 x Nvidia A100 GPUs (40 GB RAM)

○       2 x AMD EPYC 9354 32-core 3.25GHz CPUs | 1 TB RAM | 8 x Nvidia L40S ADA GPUs (48 GB RAM)

Most nodes have a QDDR Infiniband connection to a minimum of a 40G QDDR Infiniband switch, with many of the newer nodes having a connection to a higher-speed 100G Infiniband switch, for use as their data transfer network to communicate with our on-site 3.1 Petabyte dedicated storage infrastructure. They are also all joined by 10Gb ethernet connections to the rest of BMI’s internal network for use in other data transfers & receiving new computational jobs to run from our department’s researchers. The ethernet links connect through a set of high-performance switches with 48 x 10G ports and 6 x 40G ports each, while the Infiniband uplinks connect through multiple 40G & 100G Infiniband switches.

The general HPC cluster provides 644 cores of general-purpose computing power coupled with slightly over 8 TB of RAM. An additional 566,888 CUDAtm cores with 1,968 GB GDDR5 RAM provide GPU capabilities to make a very versatile and powerful HPC cluster.

A central set of file-server systems providing data via a combination of BeeGFS and NFS are connected by Infiniband & ethernet to the high performance switches and provide a total of 3.1 PB of long-term storage shared among all members of the BMI faculty and their researchers. In addition, a separate storage server provides an additional 80TB of space for use as scratch-space that’s backed by NVMe drives & shared across all nodes in the computational cluster via NFS. All 23 HPC nodes also have local storage ranging from 1.2TB to 60TB (depending on hardware configuration) for computational scratch space that doesn’t incur a network overhead. Software tools available on the cluster include MATLAB, Python 3.9 and 3.11 through the use of “Environment Modules” (with additional versions installable by end-users through tools like Conda), image processing tools from numerous sources and many other open-source tools and libraries. The base OS for the HPC cluster is Rocky Linux 8, and the cluster management software is Slurm.

**The cluster satisfies institutional HIPAA requirements to control storage and access for PHI data to provide cluster use with real patient data while still maintaining privacy of PHI records.**

**A virtual machine cluster** has been constructed from multiple high-core count nodes, both Xeon and Opteron, and managed by the open-source tool Ovirt. The virtualization method is KVM from Linux. There's substantial local storage on each virtualization node for each VM to have some local storage, as well as the ability to connect the VMs to the same NFS storage systems used by the BMI computational cluster. There are 4 virtualization nodes for general VM usage by the department with a total of 16 Xeon and 64 Opteron cores and 1 TB RAM overall, with a 10G connection available on each virtualization node. There are also some additional VM host systems which are dedicated to specific groups (Georgia CTSA, etc.) which provide them additional resources above those generally available to the department on the main VM cluster.

All computing equipment is housed in data centers with redundant cooling systems, redundant battery-backed UPS, and diesel generator backed-up power systems. The department has relationships with Amazon, Microsoft and Google to perform HIPAA-compliant storage and processing on AWS, Azure and GCP. The department also has a full-access, remote control network to facilitate rapid system repair and recovery using dedicated, out-of-band connectivity to each system.

# Imaging Resources:

The EICF AI Image Extraction and De-Identification Core (AI2EC) provided large-scale imaging extraction and de-identification services to researchers throughout Emory School of Medicine. AI2EC provides a robust pipeline that excels in extracting and safely de-identifying large volumes of imaging data in diverse modalities including MRI, CT, XR, NM, US, PET, and US. AI2EC can also provide support for ongoing regular extractions for research projects, clinical trials, or Centers. To date, AI2EC has extracted over 10 million images from the Radiology Picture Archiving and Communications System (PACS) system contributing to high-profile projects such as Emory Breast Imaging Dataset (EMBED), the world’s largest breast imaging dataset.

BMI jointly owns two slide-scanning microscopes with the Research Pathology Lab that are housed within the Emory Winship Cancer Institute:

1. A high-throughput Hamamatsu Nanozoomer 2.0-HT slide scanner is available for scanning large batches of slides in brightfield at magnifications up to 400X objective.

2. An Olympus VS120 fluorescence whole-slide scanner is available for brightfield and fluorescence specimens. This system is equipped with filters for both standard fluorescence as well as a quantum dot imaging and is capable of digitizing whole slides at up to 1000X magnification. Both systems are connected to our NAS storage to facilitate image analysis on the BMI computing resources.

**The Department of Biomedical Informatics Medical Informatics & AI Core**

The Medical Informatics & AI Core provides twenty-two distinct yet interrelated services to assist investigators across the Emory University community utilize data in research.  The services that will be provided are:

**CONSULTING: Planning, Early stages sample data pulling and exploration**

*SpringBoard:* A consulting service for investigators seeking insight and input on acquiring, storing, and using medical data for research. SpringBoard provides insight on self-management, level of effort and core service recommendations to ensure that investigators are appropriately resourced to complete their study with adequate support and statistical power. Consultations are available throughout an investigator’s entire research pipeline whenever data issues arise. This includes up to 2 hours of free assessment for faculty and staff.

*CohortCount:* Assessing patient population size at Emory Healthcare and affiliated facilities to determine if there is a sufficient patient population fulfilling the cohort criteria to conduct a study.

*GrantGen:* Providing researchers with tailored text for grant proposals, IRB protocols and project reports on Biomedical Informatics pipelines, including data acquisition, processing, infrastructure, and relevant statistical summaries.

*Project Management:* Providing comprehensive project management services, including planning, execution, monitoring, and reporting. This ensures projects are delivered on time and within scope. We provided services in planning and coordinating project activities, including core service activities, according to project-specific requirements and constraints, tasks, and timelines. Available for the project lifecycle from conceptualization to completion.

**DATA EXTRACTION: Large scale data extraction and exploration**

*DataDig (clinical data):* Facilitating the extraction of clinical data from electronic health records tailored to the needs of the project. This involves retrospective data extraction using various formats such as physiological data (time-series, images and video), text (e.g. clinical notes), flat files, and snapshots. In addition, DataDig provides services for multimodal data integration from various resources, data normalization, ontology encoding, de-identification, and cleaning.

*DataGrab (non-medical sources):* Addressing the need for multimodal data from devices or resources beyond standard medical databases, such as wearables, sensors, mobile phones, in addition to social media and public databases (e.g. for studying social determinants of health).

*Streaming Data (continuous data feed):* Providing continuous data feeds for real-time machine learning models. This service supports projects requiring up-to-the-minute data for dynamic model training and prediction.

*Synthetic Data and Digital Twins:* Generating synthetic datasets and creating digital twins resembling real data to support research on model development, data augmentation, and addressing challenges in imbalanced and insufficient data for building and training deep learning models. This service includes simulating real-world conditions and testing scenarios in an AI-ML friendly virtual environment. This service also extends to creating artificially generated datasets based on specified data distributions, suitable for development and testing of computational methods, and as an alternative to de-identification for data sharing.

**DATA ANALYSIS: Data analysis using machine learning, deep learning, natural language processing and large language models**

*DataHack:* BMI faculty and staff represent a valuable resource of state-of-the-art knowledge in the analysis of medical data, ranging from the application of signal processing to deep learning. The DataHack service runs feasibility studies and provides advice on which experts to work with and potential techniques to employ on data.

*Visual Analytics – Dashboards:* Developing visual analytics dashboards to aid researchers in interpreting complex datasets. These dashboards provide real-time insights and facilitate data-driven decision-making.

*ML-based Data Analysis and Processing:* Offering advanced data analysis and processing services using state-of-the-art model-based and data-driven techniques in machine learning and deep learning models.

*NLP-LLM Data Analysis:* Our team at BMI has substantial expertise in natural language processing (NLP) and large language models (LLMs). With the emerging demand for utilization and integration of NLP-LLM tools and services in biomedical informatics applications, we provide services including:

a)     LLM Servicese:

i)Prompting strategies including but are not limited to chain-of-thought prompting and in-context learning. This service also includes the exploration of prompting strategies and identifying optimal ones, and minimizing *hallucinations*. Both hard and soft prompting options will be available. In the near future, we also aim to provide trainable prompting (optimizing prompting automatically via training).

ii)Automated coding/annotation: This service will replicate human coding of text-based data to make the process more scalable.

iii)Information extraction and named entity recognition.

iv)Employing state-of-the-art information extraction methods, including supervised and unsupervised LLM-based approaches. Benchmark and optimize multiple named entity recognition algorithms (including rule-based and machine learning-based).

v)Fine-tuning/customizing LLMs.

vi)Fine-tune existing open-source LLMs with internal Emory Healthcare data. Fine-tuning strategies include domain-adaptive pretraining, source-adaptive pretraining and topic-specific pretraining.

vii)De-identification of textual health records using customized and high-precision de-identification methods. This includes physicians’ clinical notes, which are invaluable as they provide context by experts at the point-of-care; but in many cases include patient identifiers that are difficult to de-identify.

viii) Lexicon expansion: Most NLP of clinical notes still involve the creation of lexicons and knowledge bases. These are used to detect concepts. However, not all variations are typically encoded in a lexicon. We use large language models and semantic similarity-based strategies for automatically expanding existing lexicons with more EHR-specific variants.

ix)Customized language models: i) Train context-free language models from scratch (e.g., word or phrase level embeddings). ii) Further pretrain existing transformer-based models with EHR (or any other) data. iii) Fine-tuning existing open-source LLMs.

x)Retrieval Augmented Generation: Constrained text generation with retrieval engines at the back-end.

xi)Quantization of models: Task-oriented quantized models that can be deployed in low-resource environments.

xii)Feasibility consultation: Provide an assessment of the feasibility of conducting a specific NLP task involving LLMs or otherwise.

b)     NLP Services:

i)Regular expression preparation: Customized for detecting complex lexical patterns in clinical notes.

ii)Fuzzy matching: Inexact matching with thresholding for detecting concepts even when they are misspelled or expressed in some non-standard form.

iii)Cohort discovery from text: Develop strategies for creating cohorts from clinical notes when ICD codes do not cover the population of interest. Particularly useful for detecting rare cohorts. Methods employed for cohort creation involve rule-based NLP (e.g., fuzzy matching) and supervised classification.

iv)Supervised classification: Employ state-of-the-art supervised classification methods. Benchmark and optimize multiple supervised classification algorithms (including traditional and transformer-based) on the same data to identify the best strategy. Can be useful for a variety of tasks including cohort discovery.

v)End-to-end pipelines: Provide solutions (design and implementation) for end-to-end processing of clinical notes involving multiple NLP and machine learning modules.

**SOFTWARE/MODEL DEVELOPMENT:**

*Application Development (frontend and backend):* Developing mobile and web-based applications and backend systems necessary for research activities. This involves both software development and model creation to support data analysis and processing. This applications will support data collection, analysis, and real-time monitoring. This includes creating user-friendly interfaces and ensuring data security.

*Machine Learning Algorithm Development:* Creating and refining machine learning algorithms tailored to specific research needs. This includes standardized training, validation and testing of state-of-the-art machine learning and deep learning models.

*ML/NLP/LLM Computation Optimization:* machine and deep learning models trained on very large datasets require significant computational resources. In cloud-based systems, the computational load directly translates to the cost of cloud credits. In on-premises processing, this relationship is indirect, but in the long-term equates or even exceeds cloud-based computations (factoring for expert human resources, maintenance, repair, replacement, electricity, cooling systems, backup mechanisms, etc.). Therefore, with a team of skillful staff and faculty with formal training and extensive experience in computer science and AI, we provide services for computational optimization of these models.

*ML/NLP/LLM model tuning and transfer learning:* transfer learning, model tuning/adaptation are key techniques in big data, where it is computationally unaffordable/infeasible to retrain models from scratch. We have significant transfer learning and model tuning experiences in our team and will provide this as a core service.

**SOFTWARE/MODEL DEPLOYMENT:**

*Application/Software Hosting:* Providing hosting and deployment services for applications and software on the BMI HPC cluster.

*Infrastructure Support:* Offering comprehensive support for computation, model development and deployment infrastructure, including maintenance and integration with IT services to facilitate smooth operation and data flow on funded projects.

*Software/Model Validation*: We are planning to provide model validation as a service, by conducting rigorous validation of software and models to ensure they meet the best and standard procedures of training, validation and testing on unseen data. This includes testing for accuracy, reliability, and compliance with regulatory requirements. We will provide this as a service to Emory and external researchers seeking regulatory clearances such as FDA approvals. FDA has a Medical Device Development Tools (MDDT) program, which aims to streamline the development, evaluation, and innovation of medical devices by qualifying tools/data that sponsors can use to support regulatory submissions. Entities with access to large and standardized datasets or with expertise in model development and validation, may apply to contribute to the MDDT program. Qualified entities serve as third-parties in evaluating algorithms and software for companies seeking FDA approval. Leveraging our team’s unique experience in organizing standardized biomedical informatics-focused data challenges (including the PhysioNet Challenge and related hackathons), and building on Emory’s data warehouse, our vision for the MIAI core is to join the MDDT program, enabling us to provide software and model validation as a service.

*ML Model Deployment:* Facilitating the deployment of machine learning models into operational environments and dashboards.

*NLP/LLM Model Deployment:* We will provide a similar service for NLPs and LLMs. While many aspects of NLP/LLM model deployment are similar to ML model deployment, there are also unique challenges, requirements and current opportunities, which will make this an appealing service for the core customers.

*Data Management:* We will provide the systematic organization, storage, and maintenance of identified, limited, and deidentified clinical and associated research data to ensure its accuracy, accessibility, and security using on-premises and cloud resources. Key aspects include:

Grant Writing: We assist PIs in designing and implementing Data Management plans for their grants.

Data Management and Storage: Gathering data from various sources and storing it in a AI-ML ready structured formats in databases and data warehouses, and unstructured data in data lakes.

Data Security: Protecting data from unauthorized access and breaches through encryption, access controls, and other security measures.

Data Governance: Establishing policies and procedures for data management to ensure compliance with regulations and standards such as existing IRB approvals and data sharing agreements.

Data Lifecycle Management: Managing data throughout its lifecycle, from creation and storage to archiving and deletion.

Advise and support data upload and submission to public data repositories based on Data Management Plans.

For more information, please contact hyelyon.lee@emory.edu.