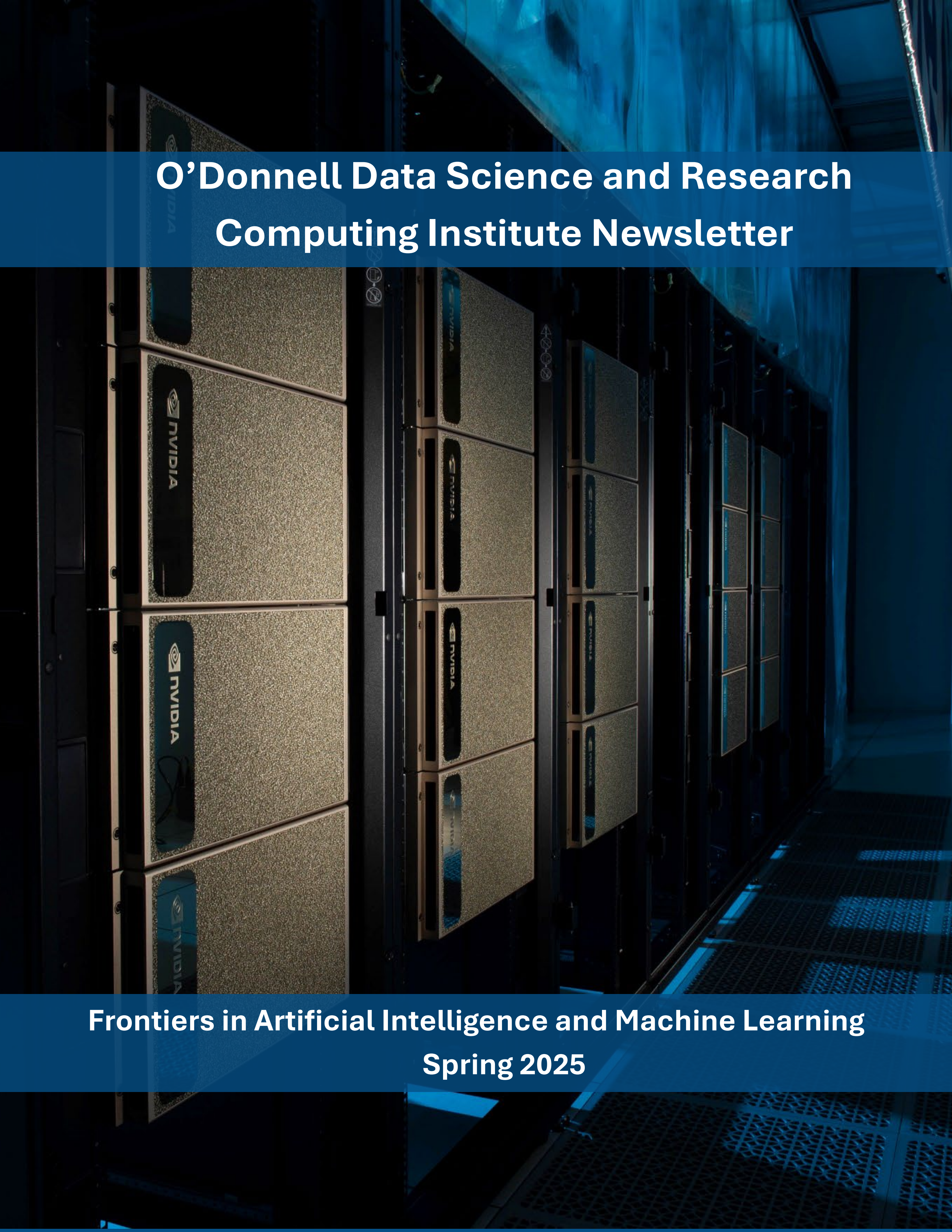


O’Donnell Data Science and Research Computing Institute Newsletter

Frontiers in Artificial Intelligence and Machine Learning

Spring 2025

A WORD FROM THE DIRECTOR

By: Dr. Neena Imam

Dear Colleagues,

I am happy to share the Spring 2025 O'Donnell Data Science and Research Computing Institute (ODSRCI) newsletter with you. This newsletter highlights the AI/ML research ongoing at SMU utilizing our HPC platforms. Our NVIDIA DGX SuperPOD and the ManeFrame III cluster are great resources for SMU researchers developing and training large-scale AI/ML models. I want to thank our faculty for contributing their big data computing research summaries for this newsletter.

Spring 2025 was a dynamic and very productive semester for the ODSRCI. We provided a series of impactful training sessions covering distributed python, graph machine learning, and privacy preserving machine learning. As part of the ODSRCI Seminar Series, we were honored to feature Dr. Seetharami Seelam, Distinguished Engineer at IBM, who delivered an insightful talk titled *Insights Gained from Delivering Two Generations of AI Supercomputing and Storage Solutions in IBM Cloud*. Slides for these presentations can be found here.

The ODSRCI recently announced the AREAD (**A**dvancing **R**esource-**E**fficient **AI** and Smart **D**ata Management) initiative. Please stay tuned for updates on this strategic initiative. I also want to congratulate the latest cohort of ODSRCI's graduate fellows: Yinglu Tan, Eli Laird, Marc de Vernon, Kang Liang, and Ding Lin. These five SMU graduate students have completed their ODSRCI fellowship, and we highlight their accomplishments in this newsletter. I expect to announce additional research grant opportunities soon and look forward to collaborating with SMU researchers.

CONTENTS

Data-Intensive Computing

O'Donnell Institute Graduate Fellows Project Highlights

Resource-Efficient Artificial Intelligence Research

As Artificial Intelligence (AI) continues to expand into real-world applications, the need for resource-efficient models continues to be a pressing concern. Traditional deep learning models, particularly Transformer architectures, are notoriously resource-intensive, often requiring significant computational power and memory. This poses challenges for deploying AI in environments with limited resources, such as mobile devices, edge computing, and embedded systems. To address this, researchers are developing a myriad of model compression techniques that aim to reduce the size and complexity of models without sacrificing accuracy. These include pruning unnecessary parameters, reducing numerical precision (quantization), factorizing matrices, and distilling knowledge from large models into smaller, more efficient ones.

One of the most promising areas of innovation lies in rethinking how Transformers handle attention, a mechanism that allows models to weigh the importance of different inputs. Conventional attention scales poorly with input length, consuming ever-increasing resources. Emerging Transformer variants introduce creative solutions to this problem. Low-rank and kernel-based attention approximate or simplify the standard attention mechanism, drastically reducing computational demands while preserving performance. Sparse attention, whether based on fixed patterns or dynamically learned from data, prunes unnecessary connections to focus only on relevant information. Other models, like Synthesizers and memory-based Transformers, radically simplify or restructure the attention process, offering fast, scalable alternatives for long sequences. Still others, like Transformer-XL, use recurrence to break up long sequences and carry context forward efficiently, improving scalability for tasks like streaming and language modeling. **(cont.)**

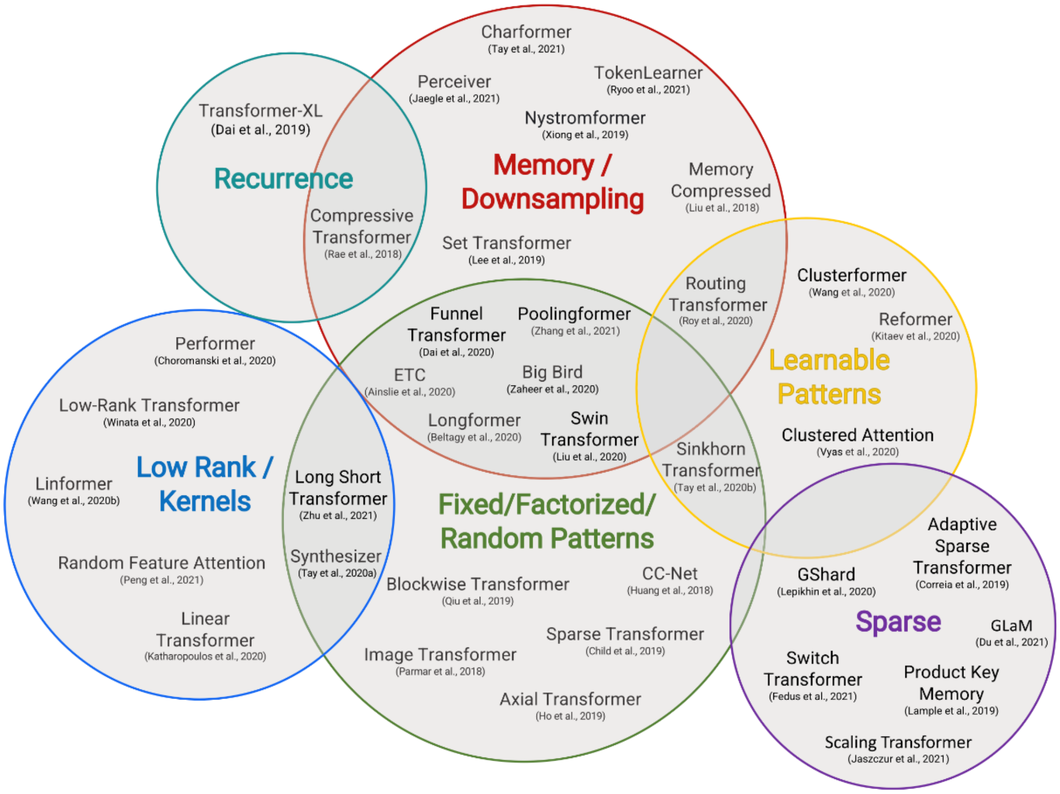


Figure 1: Taxonomy of Efficient Transformer Architectures (Y. Tay, M. Dehghani, D. Bahri, and D. Metzler *Efficient Transformers: A Survey ACM Computing Surveys*, 55(6), 2023)

These advances are more than just theoretical. Transformers are now being fine-tuned for high-impact applications in healthcare, finance, smart cities, agriculture, and energy systems, domains where efficient inference is critical. For example, medical devices, smart traffic systems, and remote environmental monitors often run on limited power and must deliver real-time results. Efficient models at the edge enable these large integrated systems to perform complex tasks like forecasting energy use or detecting fraud without relying on centralized infrastructure.

The future of AI hinges on making models both powerful and efficient. Systematic benchmarking of these attention variants will help determine not just their predictive capabilities, but also their performance in terms of speed, memory usage, and energy consumption. By aligning architectural innovation with practical constraints, researchers are paving the way for sustainable, accessible, and high-performance AI systems that can operate effectively even in constrained environments.

The O'Donnell Institute's AREAD (**A**dvancing **R**esource-**E**fficient **A**I and Smart **D**ata Management) initiative targets the dual challenge of computational efficiency and intelligent data handling in scalable AI. AREAD supports research that advances techniques such as model compression, quantization, and lightweight model architectures to minimize energy and memory demands. It also promotes smart data strategies, such as adaptive storage, retrieval, and processing, to enhance overall system efficiency. With a focus on constrained, high-impact domains like IoT, autonomous systems, AR/VR, and digital twins, the AREAD aims to foster resource-aware solutions that align AI performance with sustainability and real-world deployment needs.

Privacy-Preserving Machine Learning Research

Machine Learning (ML) is a branch of artificial intelligence that enables systems to learn from data and improve their performance on a task without being explicitly programmed. Traditionally, ML models are trained in centralized environments, where all data is collected and stored in a single location. However, as data volumes grow, and privacy concerns become critical, centralized ML approaches face significant limitations. Federated Learning (FL) offers a solution to these challenges. FL is a decentralized ML approach designed to enable collaborative model training while preserving data privacy. Instead of sending personal data directly to a central server, each client (e.g., a smartphone, laptop, or data center) trains a local model on its own data and sends only the resulting model updates (e.g., weights) to the central server. The server then aggregates these updates to improve a global model. At a high level, FL involves two main components: distributed clients and a central server, as shown in Figure 1. The primary role of a client in FL is to train a local model on its own data and share the resulting model updates (rather than raw data) with the central server. The server acts as an aggregation hub: it receives the model updates from all clients, aggregates them, and then sends the updated global model back to the clients. In FL, aggregation methods are used to combine model updates from all clients. The most used aggregation method is Federated Averaging, which takes the average of the updates sent by the clients as the aggregated update. **(cont.)**

Through this iterative communication between clients and the server, the global model is improved. Once trained, the global model can be used for tasks, e.g., detecting a pandemic in healthcare.

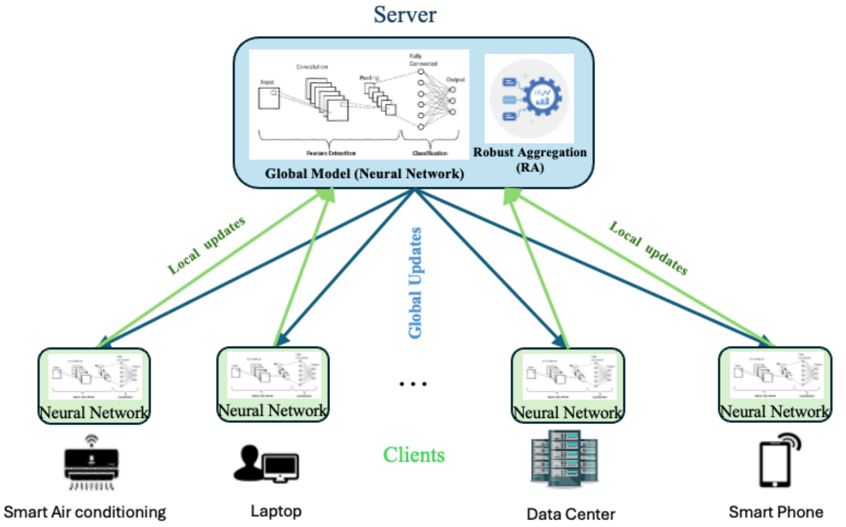


Figure 2: Federated Learning Architecture

Despite its many advantages, FL comes with emerging challenges. One concern is security. In practice, adversarial attackers may try to disrupt the entire system by sending incorrect model updates, so simply averaging everyone's updates doesn’t work anymore. At the SMU O'Donnell Institute, we're working on developing robust aggregation methods resilient to adversarial threats. Another challenge is heterogeneity. Some clients may prefer larger models, while others prefer smaller models due to resource constraints. This makes it difficult to combine updates effectively. We are actively researching heterogeneous model aggregation to address this. Additionally, clients often generate different types of data, e.g., images, videos, or tabular records. Handling this multimodal data is another area of research focus. Despite these challenges, FL is an exciting and rapidly growing field. These challenges provide new opportunities for innovation. Our goal is to improve FL by making it more robust, flexible, and applicable to real-world applications.

DATA-INTENSIVE COMPUTING

High-Throughput Data Management for Nanoscale Transport Simulations

Dr. Ali Beskok, George R. Brown Chair, Department of Mechanical Engineering

Dr. Ali Beskok’s research in micro- and nano-scale transport phenomena entails data-intensive molecular dynamics simulations using LAMMPS to model nanoscale fluid behavior, such as thin-film evaporation and water transport, across systems comprising hundreds of thousands of particles. Each simulation utilizes multi-GPU/CPU architectures and generates high-resolution output files, often exceeding 100 GB, stored in SQL database formats to enable low-latency retrieval and efficient data management. The research workflow includes remote post-processing, producing large volumes of log and auxiliary data, which are preserved to maintain reproducibility and support benchmarking. Experimental and simulation-driven projects, such as the group's graphene nano-channel studies, further contribute to significant **(cont.)**

storage demands, necessitating robust and efficient high-throughput storage infrastructure capable of sustaining I/O performance and hierarchical data organization at scale.

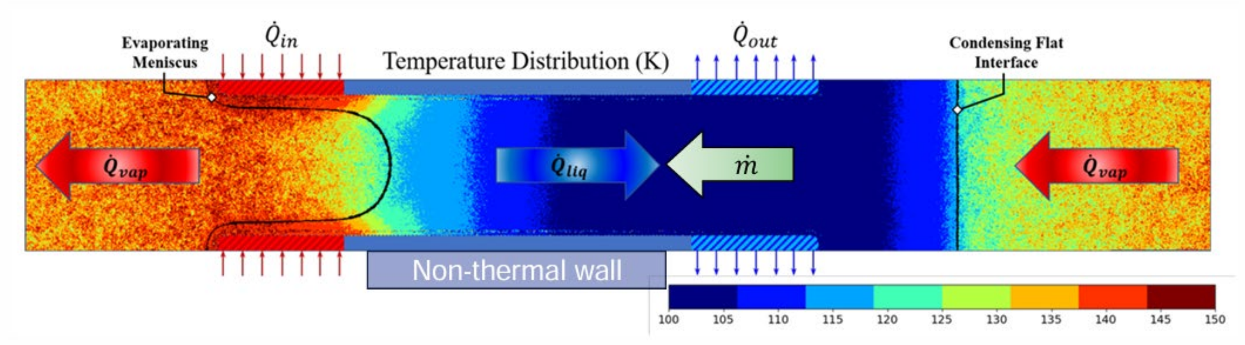


Figure 3: Temperature contours of thin film evaporation from a nano channel. (*Nanoscale meniscus dynamics in evaporating thin films: insights from molecular dynamics simulations*, M Ozsipahi, A Beskok, Langmuir 39 (50), 18499-18508)

Reconstructing Rice Domestication Through Ancient Environmental DNA and HPC-Enabled Genomic Analysis

Dr. David Meltzer, Henderson-Morrison Professor of Prehistory, Department of Anthropology

Figure 4: Aerial view of the lake in China where sediment cores were collected for ancient DNA analysis. The surrounding rice fields highlight the long-standing agricultural landscape of the region.

The Ancient environmental DNA project on SMU's HPC is led by SMU's Prof. David Meltzer in collaborations with Yucheng Wang of Cambridge University and the University of Copenhagen Center for GeoGenetics. In this project, we are primarily analyzing a very large ancient DNA dataset sequenced from ancient sediment samples from regions where rice was first domesticated some 10,000 years ago. The goal is to reconstruct the evolutionary and selection processes that led to the domestication of this important food crop, but also to gain a deeper understanding of the original genetic pool in order to identify functional and trait-related genes that have since been lost. So far, we have processed more than 20 billion sequencing reads against reference databases containing more than 100 million entries of reference genomes. From the data and results, we have preliminary profiles of the genetic processes of rice domestication, and identified more than 3,000 lost mutations (genotypes) on rice functional genes. This has enabled us to simulate the mutations in the rice genome, and predict their potential phenotypical effects by modelling the resulting protein structures. Several of these lost mutations are now being further tested in the lab. Facilitated by the SMU HPC, we now aim to extend the project using further optimized analysis methods and reference databases.

Interdisciplinary Data-Driven Research in Neuroscience, Genetics, and Political Science

Dr. Andrea Barreiro, Associate Professor, Department of Mathematics

Dr. Andrea Barreiro's research portfolio encompasses a diverse array of interdisciplinary projects, each leveraging large, complex datasets across neuroscience, computational genetics, and political science. In the realm of computational neuroscience, she investigates retronasal olfaction using multi-electrode electrophysiological recordings, fluid dynamics simulations, and computational modeling to dissect spatiotemporal activity in olfactory circuits, work that generates extensive raw and processed datasets. Complementing this, her development of FUSE, an open-source tool for long-term spike sorting, supports large-scale neural data analysis and contributes further to data-intensive processing pipelines. In computational genetics, Dr. Barreiro is updating analysis techniques originally developed for microarray data, to work for next-generation sequencing like single-cell RNA sequencing. Her political science research involves historical voter registration and electoral participation records, historical redistricting maps, and simulation-based assessments of gerrymandering, culminating in the construction of a temporal database and a suite of analytical tools. Collectively, these efforts depend on robust, scalable data storage and computing infrastructure to enable long-term retention, rapid retrieval, and high-throughput analysis of diverse and evolving datasets.

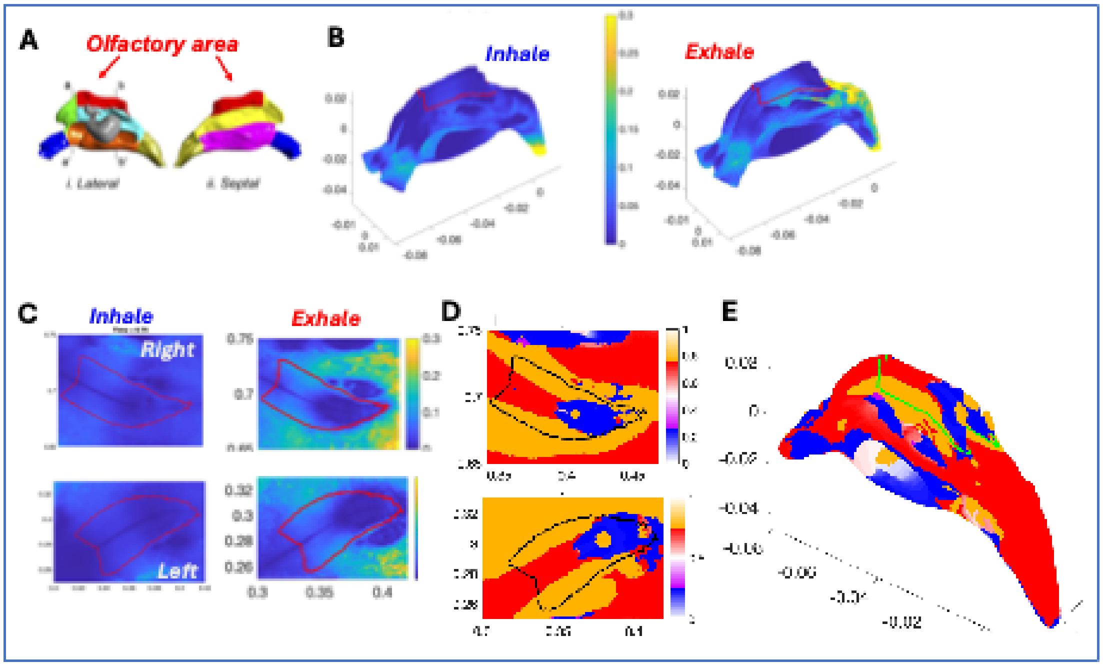


Figure 5: (A) Image of the human nasal cavity used in these simulations, with olfactory regions identified in red (adapted from from Shang et al. (2017) http://dx.doi.org/10.1016/j.euromechflu.2016.09.024.)(B) The magnitude of shear stress on the cavity at the peak of the inhale (left) vs. the peak of the exhale (right). The olfactory region is outlined in red. (C) By projecting the nasal surface into 2D, we can visualize right (top) and left (bottom) olfactory regions at once (outlined in red). (D) The phase preference map for the olfactory region: 0 ≤PPM ≤0.5 indicates ortho selective areas, 0.5≤PPM ≤1 indicates retro-selective areas. (E) The PPM in 3D, showing that the ortho-selective (blue) region lies to the back of the nasal cavity.

O’DONNELL INSTITUTE GRADUATE FELLOWS PROJECT HIGHLIGHTS

Algorithms for Reduced-Order Modeling and Data Assimilation with Applications to Digital Twins of Complex Flow System

Marc de Vernon, Mathematics | Advisor: Dr. Tom Hagstrom

We consider Reduced Order Models (ROMs) and domain decomposition for complex turbulent flow systems whose performance is dominated by propagation of coherent structures and shock waves. An example of these multi-scale, multi-physics simulations is the rotating detonation engine. This is a very complex system to simulate. A single instance of large eddy simulation in three dimensions with eighteen state variables and the necessary four million spatial degrees of freedom has a simulation time of five and a half million core hours, yielding just five milliseconds of simulation. Exploring all possible model parameters is too costly without ROMs.

Digital twins are a key focus across various fields and demonstrate the practical use of modern computing. A key challenge is learning how to mathematically couple ROMs at their interfaces while keeping numerical errors to a minimum. Our project offers guidance on implementing and minimizing these errors in regimes of increasing turbulence or chaos.

We introduce a low dimensional *proxy* problem allowing extensive high-fidelity simulations over a set of parameters that generate increasing chaos and apply this massive data set to study the propagation of error in the ROMs. This ‘perturbed Sine-Gordon’ equation exhibits a chaos cascade analogous to the full compressible Navier-Stokes equations.

SMU’s M3 High-Performance Computing (HPC) cluster under the O’Donnell Data Science and Research Computing Institute offers necessary capabilities to both generate the high-fidelity data at scale, and train and test a range of different ROMs, from fully intrusive to completely data driven. The insights gained from our proxy problem and high-fidelity simulations will inform best practices for ROM coupling strategies, enabling efficient, accurate, and robust digital twins.

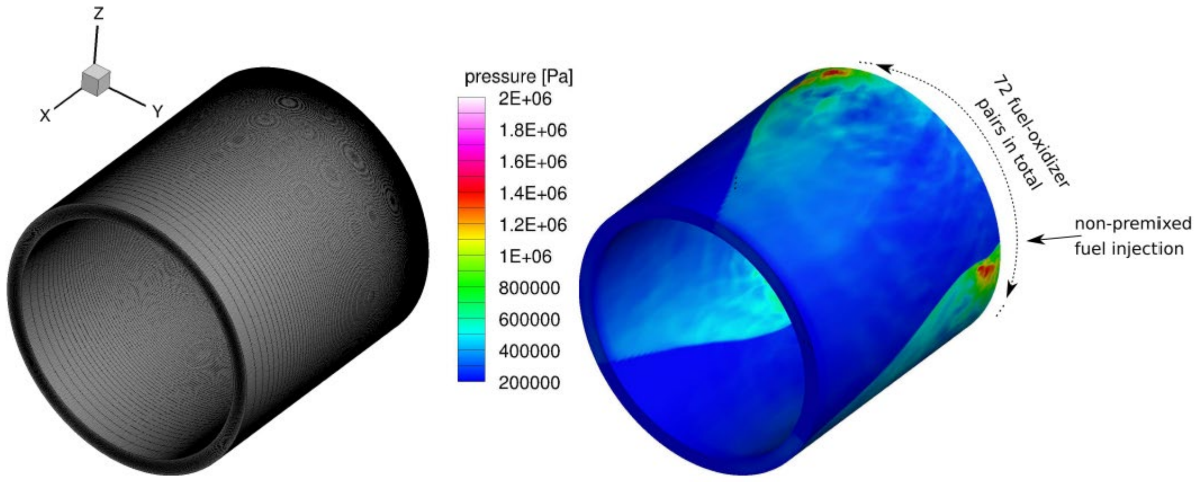


Figure 6: The computational domain and a single parameter simulation of the rotating detonation engine with three detonation waves.

Building Adaptable and Efficient World Models for AI

Eli Laird, Computer Science | Advisor: Dr. Corey Clark

The future of embodied-AI depends on systems that can rapidly adapt to changing goals in complex environments while using minimal computational resources. Current model-based reinforcement learning approaches focus on constructing a "world model", an approximation of the environment, to roll-out future scenarios and develop the possible plans of action to accomplish their goals. These world models, however, fail to generalize to new environments, particularly the complex environments present in the real-world. In response to these limitations, this research proposes several improvements to the framework that improve adaptability, efficiency, and performance in complex environments.

To enable world models to adapt to different objectives, we propose a conditioning mechanism to the state-action pair encoder that provides a high-level "direction" bias to the world model. This mechanism allows the controller, or user, to provide high-level goals to the world model which constrain the world model's output to only the important paths to achieving the goal. Additionally, pretraining the world model with a latent conditioning vector enables the world model to switch between objectives by modifying the conditioning vector during runtime.

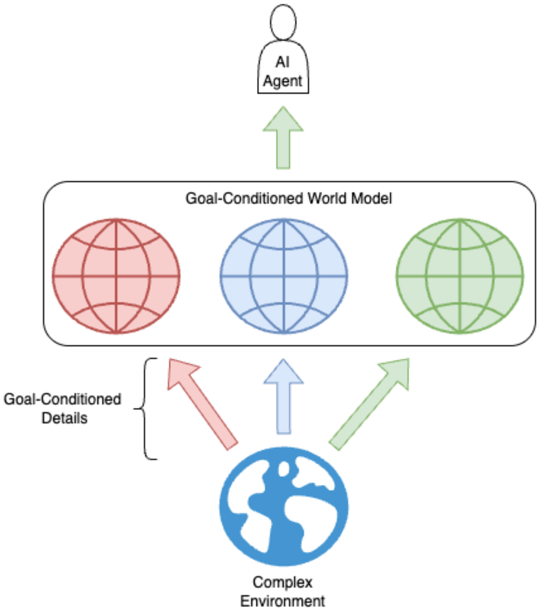


Figure 7: Goal-Conditions World Models

Traditional world models are trained using image reconstruction loss functions that tend to encode irrelevant information, such as precise pixel-level details that have little to no relation to the agent's primary goals. We propose a latent predictive loss function that removes the limitations of reconstruction losses, while importantly improving the efficiency of future scenario predictions. By creating more efficient and adaptable world models, this work enables AI deployment in previously impractical settings like mobile robotics and edge computing. Moreover, the objective-conditioning mechanism allows systems to rapidly adapt to changing goals without expensive retraining, making them more practical and responsive in dynamic real-world environments.

Noise2Fringe: A Self-Supervised Deep Learning Approach for Radar Interferometry Denoising

Kang Liang, Earth Sciences | Advisor: Dr. Zhong Lu

Synthetic Aperture Radar Interferometry (InSAR), as an active source geodetic technique, provides unique large scale earth surface deformation measurements with fine resolution, high precision, and is capable of operating in evening and cloudy weather conditions. However, the quality of the interferometric phase can be strongly affected by the noise introduced by decorrelation, a vexatious phenomenon mainly due to the change of ground target scattering characteristics. To remove the decorrelation noise, the recent supervised deep learning approaches outperform the classic filters in preserving the localized fringe patterns. Since real clean interferograms are not available, all supervised deep learning approaches are trained on synthetic datasets. Although the supervised models achieve start-of-the-art performance on synthetic datasets, fake fringes are observed when they are applied in complete incoherent interferograms. The reason is, for completely incoherent regions, it is impossible to reconstruct the clean interferograms while the synthetic datasets always have the corresponding clean interferograms that mislead the deep learning model.

To address the challenge of lacking real-world data for training, we propose a self-supervised training scheme, named Noise2Fringe, that trains denoising models with only real noisy interferograms. This self-supervised scheme is based on the observation that some pixels in interferograms, named Persistent Scatterers (PSs), only contain negligible decorrelation noises. That means, the observations on PSs can be treated as if they are from the “clean” interferograms. Therefore, the model can be trained based on the loss function that only considers PSs. Compared with the supervised training with synthetic datasets, Noise2Fringe doesn’t have the fake fringe issue as it is training on real-world noisy interferograms. The completed incoherent interferograms won’t be used for training as no PSs can be found in such interferograms. The performance of Noise2Fringe is tested with both synthetic interferograms and real interferograms. The test result shows that Noise2Fringe combines the advantages of previous methods. Compared with the classic transform-domain filters, Noise2Fringe preserves more fringe details. Compared with other supervised deep learning methods, Noise2Fringe does not generate fake fringes.

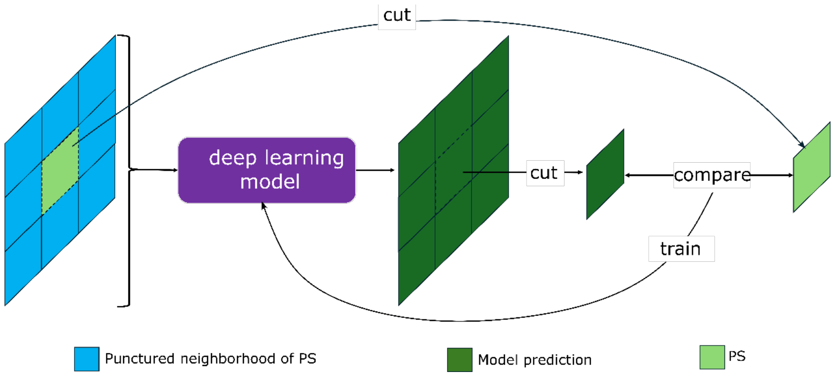


Figure 8: Schematic diagram of the unsupervised Noise2Fringe training scheme. This method incorporates two key modifications compared to the supervised training approach. First, the model takes a noisy interferogram as input, with one persistent scatterer (PS) pixel intentionally excluded. Second, the training objective is to minimize the difference between the model's output and the corresponding clean interferogram, specifically at the excluded PS pixel.

Quantum Reinforcement Learning in Power System Operations

Ding Lin, Electrical and Computer Engineering | Advisor: Dr. Jianhui Wang

This project presents a novel quantum-classical reinforcement learning framework for solving the Multi-Objective Transmission Switching (MO-TS) problem in power systems, which involves optimizing network topology to balance economic and reliability objectives. Due to the nonlinearity and computational complexity inherent in MO-TS, especially with increasing system scale, traditional solution methods often struggle with efficiency and scalability. To address these challenges, we develop a Two-Stage Quantum Soft Actor-Critic algorithm that harnesses the potential advantages of quantum computing for high-dimensional decision-making.

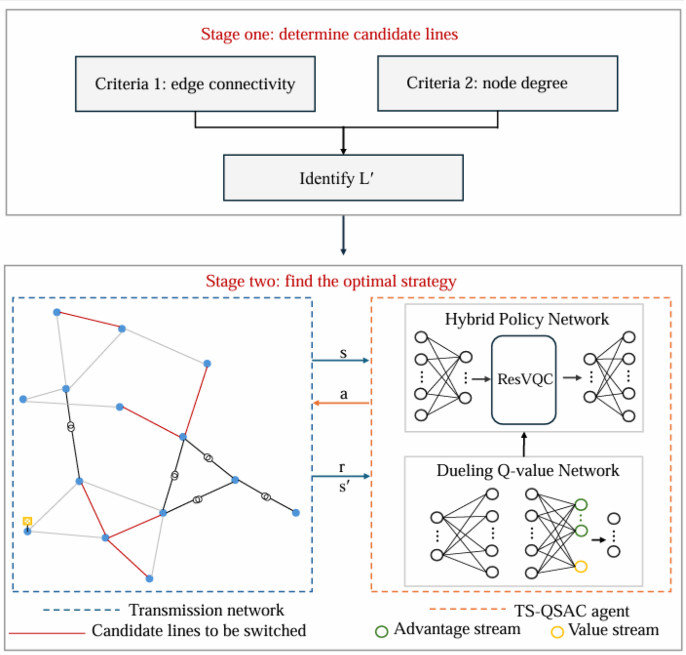


Figure 9: The Proposed Framework

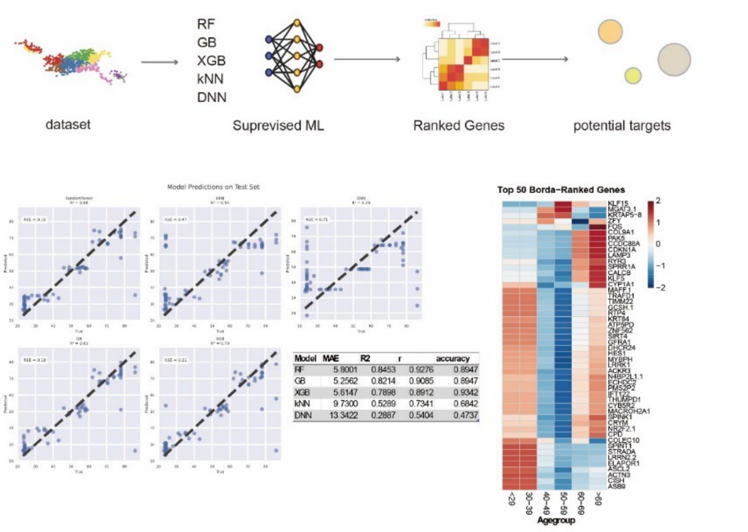
As shown in Figure 9, the first stage employs a graph-theoretical heuristic to identify candidate transmission lines for switching, which can effectively reduce the problem's dimensionality. The second stage introduces a hybrid quantum-classical reinforcement learning architecture, where a novel CNN-ResVQC policy network is proposed. To enable efficient quantum processing, this architecture compresses high-dimensional system states into a low-dimensional latent space using a classical autoencoder. It also incorporates residual variational quantum circuits with data re-uploading to counteract the vanishing gradient problem and enhance model trainability. Additionally, we introduce a learnable measurement-based approach, replacing fixed Pauli measurements with trainable Hermitian operators that can model a richer set of observables. This approach expands the expressive power of the quantum circuit and allows for better adaptation to the complex reward structures typical in MO-TS.

Extensive experiments on IEEE 14-, 57-, and 118-bus systems validate the effectiveness of the proposed method, and demonstrate its faster convergence, improved training stability, and significant model compression (using only 1% of parameters compared to classical models). Moreover, the method exhibits robust performance under common quantum noise channels, highlighting its potential for deployment on near-term quantum devices.

Exploring Machine Learning Techniques for Discovering Aging Biomarkers in Human Skeletal Muscle

Yinglu Tang, Biology | Advisor: Dr. Zhihao Wu

Aging-related changes in gene expression within skeletal muscle significantly contribute to the progressive decline in muscle mass and strength observed in older adults—a key factor driving physical frailty. However, aging is inherently a nonlinear biological process, characterized by complex, time-dependent molecular alterations, with notable shifts occurring around ages 44 and 60. Previous studies on muscle aging have often assumed a linear model, which may reduce prediction accuracy and overlook critical transition. To better capture these dynamic changes, this study leverages nonlinear Machine Learning (ML) models for more accurate and biologically informed predictions.



Model	MAE	R2	r	accuracy
RF	5.8001	0.8453	0.9276	0.8947
GB	5.2562	0.8214	0.9085	0.8947
XGB	5.6147	0.7898	0.8912	0.9342
kNN	9.7300	0.5289	0.7341	0.6842
DNN	13.3422	0.2887	0.5404	0.4737

Figure 10: workflow and performance of age predicting models.

As illustrated in Figure 10, the analytical workflow begins with gene expression array data from aging skeletal muscle. To reduce potential bias, batch effects were removed by stratifying the samples into three age groups: <44, 44–60, and >60 years. The cleaned data were then processed through supervised ML models trained to predict aging-related traits. These models not only assess prediction performance but also rank genes based on their contribution to accurate age prediction.

To account for the nonlinear nature of aging, several advanced ML models were evaluated, including ensemble methods such as Random Forest (RF), Gradient Boosting (GB), and XGBoost (XGB), as well as k-Nearest Neighbors (kNN) and Deep Neural Networks (DNN). Among these, RF, GB, and XGB demonstrated superior performance, achieving the highest R^2 values (~0.86) and the lowest Mean Absolute Error (~5–6 years), confirming their effectiveness in modeling nonlinear aging trajectories.

The heatmap of top-ranked genes further supports the validity of the selected biomarkers, revealing clear age-related expression patterns. Notably, several top genes are involved in protein synthesis and proteostasis—processes frequently disrupted in muscle aging. For example, *RPS4X.1*, *EEF1G.1*, and *RPL7.1* are key components of the translation machinery, ensuring efficient protein production, while *UBA1* and *LONP2* play crucial roles in protein degradation and quality control. The inclusion of these biologically relevant genes highlights the strength of this integrative ML approach in identifying robust biomarkers tied to core mechanisms of skeletal muscle aging, paving the way for future therapeutic strategies.

SPRING 2025 EVENTS

WORKSHOPS

January 23: Distributed Python

This workshop was designed for researchers seeking to optimize their workflows for large-scale data processing and parallel computing. Participants explored the fundamentals of distributed computing with Python, learning how to leverage Dask for efficient task parallelism and Ray for scalable machine learning and reinforcement learning applications. Through hands-on exercises, attendees gained practical experience in deploying these tools to handle massive datasets, optimize computational resources, and streamline workflows.

February 6: Graph Machine Learning Fundamentals

This workshop was designed to equip researchers with the skills to apply Graph Neural Networks (GNNs) to complex datasets. Participants learned the fundamentals of graph-based data representations, followed by practical training on using PyTorch Geometric (PyG) for building, training, and evaluating GNN models.

February 11 & 18: Fundamentals of Deep Learning Parts 1 and 2

This workshop, designed by NVIDIA Deep Learning Institute (DLI), covered foundational techniques for learning deep learning models, including CNNs (Convolutional Neural Networks), data augmentation, pretrained models, and natural language processing, with practical implementation in Python and PyTorch. The workshop objectives were to learn fundamental techniques and tools for training deep learning models, gain practical experience with common deep learning model architectures, and build confidence to tackle projects using modern deep learning frameworks. After successfully completing the assessment and attending both sessions, participants received an NVIDIA DLI certificate to validate their skills and enhance their career opportunities.

March 10: Privacy-Preserving Machine Learning

This workshop introduced the basics of federated learning and included a hands-on session with NVIDIA FLARE (an open-source federated learning framework developed by NVIDIA). It featured an image classification demo using federated learning on SMU’s NVIDIA SuperPOD (a high-performance computing platform for large-scale data analysis). The workshop also explored current research challenges and future directions in federated learning, highlighting key opportunities in the field.

SEMINAR SERIES

March 4: Insights Gained from Delivering Two Generations of AI Supercomputing and Storage Solutions in IBM Cloud | Dr. Seetharami Seelam, Distinguished Engineer at IBM Research

AI Supercomputers in public clouds serve as crucial components in the swift and cost-effective creation and deployment of cutting-edge AI models. This heightened demand for potent cloud-native AI supercomputers stems from the increasing prevalence of generative AI and foundational models. In these systems, numerous GPUs collaborate to facilitate model training, optimization, and serve countless concurrent applications without disruption. To ensure optimal performance, reliability, and adaptability for various AI workloads, a comprehensive solution integrating hardware, software, and holistic telemetry is essential. This solution enables the efficient and high-performance execution of multiple AI workload types while maintaining resilience. In this talk, Dr. Seelam discussed two generations of Vela cloud-native AI systems in IBM Cloud, which form the backbone of IBM's AI endeavors. He explored the scaling, performance, and high availability challenges confronted during their development and operation.