# Somali Chaterji: Research Statement

Assistant Professor, Purdue University, West Lafayette, IN 47907
CEO, KeyByte, https://keybyte.xyz
P: (765) 494-3652 [O]/(765) 714-8812 [C; preferred]; E: schaterji@purdue.edu
Blog: https://schaterji.io/blog/; W: https://schaterji.io

**Research Interests**: Internet of Things (IoT), Cloud and Edge Computing, Computational Genomics, and Data Analytics applied within these domains.

**Vision**: I direct the Innovatory for Cells and Neural Machines (ICAN) with research focused in two main areas.
**IoT, Edge, and Cloud Computing**: Given the proliferation of interconnected devices, my research aims to develop systems that streamline data processing and analytics across IoT, edge, and cloud infrastructures. The overarching goal is to design efficient, real-time solutions, especially in scenarios with limited resources and stringent latency requirements.
**Computational Genomics**: My work in this domain is centered on creating computational tools and frameworks for analyzing genomics data. This involves processing large datasets, designing algorithms for genetic sequence analysis, and detecting patterns critical for clinical and biomedical applications.

Both research areas share a common challenge: handling copious amounts of data and designing platforms that convert raw information into actionable insights. Resource constraints and latency bounds often intensify the challenge, but they also inspire innovative solutions. A comprehensive overview of my research can be found at: https://schaterji.io/research/

# 1 IoT, Edge, and Cloud Computing in Large-Scale Systems

In the rapidly evolving digital landscape, my research delves into the intricate nexus of *IoT*, *edge computing*, and *cloud computing*. The aim is to design systems characterized by resource efficiency and resilience. These systems find applications in myriad fields such as *digital agriculture*, *persistent surveillance*, *advanced manufacturing*, and *autonomous systems*. The pervasiveness of IoT technologies across varied sectors necessitates that our systems are not only robust from the outset ('resilient-by-design') but are also equipped with the adaptability to meet changing demands, ensuring resilience through responsive adaptations.

One significant shift reshaping this domain is the rise of *edge computing*. While edge computing optimizes bandwidth and reduces latency by situating computing resources proximate to data sources, it introduces concerns about its stability in comparison to the traditionally centralized cloud resources. A core undertaking in my lab is to assess the efficacy of data processing spanning client devices, edge devices, and centralized cloud servers. Noteworthy contributions from my lab in this domain encompass:

- Designing and developing a self-tuning NoSQL database optimized for IoT and genomics workloads [ACM Middleware-17, Usenix ATC-19, Usenix ATC-20].
- Forging a resilient IoT mesh network tailored for expansive agricultural settings [IEEE Internet of Things Journal-20].
- Comparative assessments of cutting-edge solutions from major industry players like Amazon, Google, and Microsoft [IEEE Cloud-20, featured in the Data Skeptic podcast - `https://schaterji.io/publications/2020/janus/podcast.html`].
- Innovating a leading video object detection system calibrated for real-time latency (circa 30 frames per second, FPS) on embedded devices [ACM Sensys-20, ACM EuroSys-22, CVPR-22, ACM TODAES-23].
- Advancements in optimizing serverless cloud computing workflows [Usenix ATC-21, Usenix OSDI-22, ACM Sigmetrics-22 (**best paper award**)].

In the vast expanse of IoT infrastructures, my quest is to weave together frameworks that strike a balance between ease of use and consistent reliability. The maturation of our flagship systems such as *LiteReconfig* [EuroSys-22], *SmartAdapt* [CVPR-22], and *Virtuoso* [ACM TODAES-23] stands as evidence of these efforts. In subsequent sections, I will delineate key features and innovations of LiteReconfig and Virtuoso, emphasizing their core functionalities and contributions.

**LiteReconfig - Adaptive Video Object Detection Framework:**
In the ever-evolving world of video object detection on mobile devices, there is an emerging trend toward adaptive vision systems. How can we mold a system that is fluid, responding seamlessly to the varied needs of the user? Whether it is the immersive world of an AR game or the meticulous scrutiny of surveillance footage, each scenario presents its unique demands. Sometimes, this means sifting through video data with a fine-toothed comb, emphasizing accuracy. Other times, it means being fleet-footed, dashing through processes for quick results. This is where adaptive vision systems have found their niche, as they promise both flexibility and performance tailored to individual needs of accuracy or latency. They are dynamic, intelligent, and have begun to set a new standard in the realm of video object detection.

Understanding Execution Branches and what we mean by MBEKs in Video Object Detection on Mobile Devices: Diving deeper into the inner workings of these systems, we present two key components, as follows.
1. The *Multi-Branch Execution Kernel* (MBEK): Picture our MBEK as the control center of the adaptive vision system. Much like a multi-laned highway catering to different types of

vehicles and speeds, the MBEK houses multiple execution branches. Each branch processes video data with its specific balance of speed (latency) and quality (accuracy).
2. *Scheduler*: Acting as the traffic controller, our Scheduler determines which lane (or execution branch) the MBEK should take. It bases its decision on several factors: from the inherent features of the video to the urgency or precision the user requires. The Scheduler ensures that the system is always on the optimal path for the task at hand. For example, if a user prioritizes speed over quality, the scheduler might choose an execution branch that processes video quickly but with a moderate drop in accuracy.
While there has been great progress in developing models and systems lightweight enough to run on mobile devices, challenges persist. Historically, most approaches focused on static optimization, perfecting models for specific accuracy and efficiency standards. However, more adaptive models and systems have recently gained traction such as our own ApproxDet [SenSys 2020]. They are tailored to adjust dynamically based on video content, the device's available resources, and user objectives. Yet, a gap exists in current solutions:
1. Balancing Scheduler Efficiency with Quality Decisions: Previous systems tend to rely on lightweight video features, such as the video's height, width, or number of objects, to determine which execution branch to run. While these features are computationally efficient, they might not always make the best decisions in terms of accuracy. More detailed video features, like motion or appearance characteristics, could make better decisions but often require heavier computational resources.
2. Switching Overhead: If conditions change frequently, continuously switching between execution branches can incur computational overhead, reducing the system's efficiency.
To address these challenges, we designed *LiteReconfig*, which emerges as an innovative framework to address prevailing computational challenges. Its hallmark lies in a unique cost-benefit analyzer, which not only identifies the optimal features but also determines the most suitable execution branch during inference. Distinguishing LiteReconfig is its pioneering content-aware accuracy prediction model, which ensures each video frame is processed by its best-fit execution branch. LiteReconfig surpasses its peers, delivering markedly enhanced accuracy across diverse latency benchmarks. Remarkably, it realizes superior accuracy compared to existing systems under a broad range of latency objectives, all while ensuring up to 50 frames per second (fps) on the NVIDIA AGX Xavier board—a performance surpassing real-time latency requirements.
Our contributions can be summarized as follows:
1. Cost-Benefit Analyzer: We introduce a cost-benefit analyzer geared toward efficient online reconfiguration. This design optimizes the scheduler's efficiency and bolsters accuracy since it allows a larger portion of the latency budget to be allocated to the core object detection process.
2. Content-Aware Accuracy Prediction: Our framework boasts a content-aware accuracy prediction model for the execution branch. This ensures the scheduler's decision aligns with the video content. Intriguingly, this model leverages computationally-intensive features, seamlessly integrating with our overarching cost-benefit analysis.
3. Comprehensive Experimental Analysis: We have carried out a robust experimental evaluation on two distinct mobile GPU boards and juxtaposed our findings with previous methodologies. This deep dive highlighted two primary insights:

- The significance of accounting for interference from simultaneous applications.

- The imperative nature of judiciously selecting features for determining the optimal execution branch. This becomes even more pivotal when integrating content-aware features given their inherent computational weight. Noteworthy is the fact that LiteReconfig's full deployment meets even the strictest latency benchmarks, achieving 30 fps on the lesser-equipped TX2 board and an impressive 50 fps on the more robust AGX Xavier board.

**Virtuoso - The Triad of Accuracy, Energy Efficiency, and Latency:**
While many current solutions have ventured into optimizing computer vision tasks for mobile or embedded systems, they often overlook a holistic approach. Many fail to account for the crucial energy consumption during model inference. There is also an oversight in the synergy between latency, accuracy, and energy metrics. *Virtuoso* was conceived to fill these gaps. It aims to be a comprehensive solution that concurrently optimizes for accuracy, energy efficiency, and latency. At its core, Virtuoso houses a multi-branch execution kernel. This kernel is capable of operating at different levels across accuracy, **energy**, and latency. Complementing this is a nimble runtime scheduler that ensures the kernel aligns with user requirements. When benchmarked, Virtuoso's prowess is evident. Achieving up to 286 FPS on the NVIDIA Jetson AGX Xavier board, it is significantly faster than many existing models. Impressively, energy savings of up to 97.2% were noted compared to popular detectors like YOLO (v3). Furthermore, in a comprehensive comparison against 15 advanced or widely-used protocols, Virtuoso consistently exhibited superior performance. It led in terms of accuracy, surpassing models like FRCNN and YOLO by over 10%. With its holistic approach and impressive benchmarks, Virtuoso presents a promising future for object detection on embedded systems.

Both of our systems, LiteReconfig and Virtuoso, encapsulate a harmony of accuracy, latency, and energy efficiency, rendering them apt for a gamut of applications. I believe in contributing the source code and the data in almost all cases. For example, LiteReconfig got the *ACM-EuroSys "Results Reproduced"* badge, underscoring our lab's commitment to transparency and broad dissemination.

Final thoughts: In essence, as we look to the horizon, the future of video object detection on mobile devices is not just about faster models or higher accuracy. It is about building systems that understand, adapt, and evolve, ensuring that every frame, every pixel, and every user experience is nothing short of perfect.

# 2 Computational Genomics

My research endeavors to develop precision molecular therapeutics, extracting insights both from our physiological signals and the underlying genomic-epigenomic markers. While my efforts concentrate on the latter, collaborations are extensively pursued for the former. Notable contributions from my lab in this realm encompass:

- Exploiting machine learning, particularly in distributed settings, to design RNA-based molecular therapeutics. This is achieved by delving into genomic and epigenomic

data. [ACM BCB-15 (**best paper award**), Nature Scientific Reports-16, Theranostics-17, TCBB-18, Nature Scientific Reports-20].

- Engineering nanosensors to modulate cell biomechanics, paving the way for early and accurate disease detection. [Collaboration with UT Austin, UCLA, and Terasaki Institute of Biomedical Innovation - Biomaterials-18, Science Advances-20].

- Designing a repertoire of scalable and interpretative machine learning techniques, tailored for genomics datasets, with a particular focus on epigenomics and metagenomics data. By integrating biological insights, we not only enhance the scalability of our training but also heighten the clarity of our inferences. Our partnership with the MG-RAST team was instrumental in this journey, significantly boosting its processing capacity in a real-time operational environment. [IEEE/ACM Supercomputing-14, Nucleic Acids Research-15, Nature Scientific Reports-19, ACM KDD-22].

A recent endeavor led to the unveiling of *Kratos* [KDD-22] for single-cell RNA-sequencing (sc-RNA-seq) data processing. Conventional procedures involve a three-tiered workflow. Contrarily, Kratos employs an integrated biphasic process, enhancing the dependency among steps and producing finer clusters validated using marker genes. The upshot is an improved extraction of marker genes, facilitating superior cluster discrimination. Empirical benchmarks reveal Kratos' dominance, outperforming baselines such as Global Counterfactual Explanation (GCE) [ICML-20] and Adversarial Clustering Explanation (ACE) [ICML-21].

# 3 Future Directions: A Glimpse into Ongoing Initiatives

**Edge System Offloading:**

Building upon the advances in offloading computations, my focus is to enable more powerful edge systems by allowing them to seamlessly offload data and computation. Thus, my solution is enabling offload from the sensors to the edge servers to the cloud, depending on the applications, its latency requirements, and the network conditions. The goal here is multifaceted: ensuring reduced latency, optimizing computational efficiency, and catering to the bespoke needs of individual applications. At the core of this endeavor is a meticulous balancing act that weighs accuracy against latency, a challenge compounded by ever-evolving network conditions and the variable computational prowess of devices. Implementing data compression through variational autoencoders remains a pivotal strategy, more so when handling data-rich yet sparse 3D visual inputs like those sourced from LiDAR.

**3D Vision on Mobile Devices:**

As vehicles become smarter and more autonomous, the urgency of real-time environment perception soars. Indeed, 3D object detection—especially through LiDAR point clouds—provides a granular, near-photographic understanding of surroundings. However, this sophistication comes at a computational price. A case in point is the CenterPoint algorithm, which, while promising, lags on platforms like the NVIDIA Jetson AGX Xavier, clocking in at a prohibitive 400 ms per frame. Addressing this bottleneck, my team and I

are spearheading the development of a dynamic 3D object detection solution. At its heart lies the Multi-Branch Execution Framework (MEF)—an intricate web of 40 execution paths that deftly adapt to real-time conditions, such as scene complexity. The brilliance of MEF is its agility, ensuring that the chosen execution path always aligns with the real-time requisites of the task.

**Semi-Supervised Semantic Segmentation:**

The realm of image processing has always been a fertile ground for innovation. And among its myriad applications, semi-supervised semantic segmentation—especially in agritech for monitoring crop health—holds immense promise. Traditionally, the focus within this domain has remained tethered to pseudo-label consistency. However, my research is pushing the envelope by leveraging the power of latent feature embeddings, augmented further with the Sliced-Wasserstein Distance (SWD). This novel approach not only amplifies consistency but also streamlines the training process. Empirical benchmarks already place this method ahead of its contemporaries. As we look to the horizon, our aspirations are clear: to delve deeper into these latent embeddings, unlocking their potential to further refine and sharpen the boundaries that define semi-supervised segmentation.

**A Dual-Functional System for Single-Cell RNA Sequencing (scRNA-seq) Data:**

*Background on scRNA-seq Clustering:* Single-cell RNA sequencing (scRNA-seq) offers a granular perspective into individual cells by mapping gene expression profiles at a single-cell resolution, a stark contrast to traditional techniques that generalize over a cell population. As a result, scRNA-seq uncovers the diverse roles and states of cells that might be obscured in averaged data. From a computational standpoint, scRNA-seq clustering can be analogized as a problem of dimensionality reduction and classification. Two primary approaches are parametric and non-parametric clustering. Parametric methods operate under certain data distribution assumptions, often making them computationally efficient but potentially less adaptive to unexpected data structures. On the other hand, non-parametric methods operate without rigid distribution assumptions, allowing them to flexibly adapt to the data but often requiring more computational resources. Given the noisy and intricate nature of scRNA-seq data, having both tools in the toolkit is essential, as the best approach may vary based on the dataset's characteristics.

We are developing a system tailored for denoising and clustering scRNA-seq data. Under the non-parametric setup, it incorporates a denoising autoencoder framed around the Zero-Inflated Negative Binomial (ZINB) noise model, with the Leiden algorithm guiding initial cluster formation and the Dip test for cluster uniformity assessment. Conversely, the parametric approach integrates a graph autoencoder with spectral clustering directing the initial clustering. Benchmark tests place our system ahead of contemporaries like Monocle3 [Nature-19] and DipDeck [KDD-21]. Performance metrics reveal a marked improvement, with gains in Normalized Mutual Information (NMI) and the Adjusted Rand Index (ARI) being noteworthy. In trajectory inference accuracy, our system registers a 13.5% lead over scGAE [Nature-21]. Trajectory inference holds a unique place in the analysis of scRNA-seq data. It decodes the developmental pathways of cells, shedding light on cellular transitions and differentiation. These inferred trajectories offer deep insights into biological phenomena, from embryonic development to tissue regeneration and even disease evolution. For computational researchers, mapping these cellular time-

lines presents both a challenge and an opportunity, necessitating intricate algorithms to distill meaningful insights from the intricate maze of single-cell data.

Our system's main features include its dual-functionality in handling both parametric and non-parametric clustering for scRNA-seq datasets. Evaluation metrics such as K, ARI, and NMI confirm its performance. Notably, the non-parametric version provides consistent performance, even with varied hyperparameters. Using UMAP for visualizing latent embeddings, the system illustrates efficient clustering, exhibiting clear distinctions both within and between clusters. Additionally, the parametric version excels in cell trajectory inference, outpacing existing methods with a Kendall correlation coefficient (KCC) of 0.84.

# 4 Funding

Throughout my academic journey, I have secured consistent support from various prestigious programs at the national level. During my post-doctoral tenure, I was funded through the American Heart Association (AHA) Scientist Development Grant and the NIH Director's New Innovator Awards (DP2). My collaboration with Argonne National Labs was funded under the sponsorship of an **NIH R01 grant**, facilitating the creation of a self-tuning NoSQL database optimized for both on-premise and cloud infrastructures. Purdue University's WHIN (Wabash Heartland Innovation Network) Project, backed by a generous $20M fund from the **Lilly Endowment** (2018-24), supported my endeavors in digital agriculture analytics. Furthermore, my research in the Internet of Things and Edge computing domains is sustained by two **Army Research Lab** (ARL) contracts spanning from 2020-23 and 2020-25, respectively.

A pivotal aspect of my research landscape is my **NSF-CAREER project**, *Sirius*, more about the project is here: `https://schaterji.io/projects/sirius/`. This project addresses challenges presented by today's increasingly sensorized agricultural farms. With the proliferation of robust yet affordable sensors, there is a pressing need to process vast volumes of data efficiently. Sirius aims to revolutionize the Cyber-Physical Systems for digital agriculture by achieving on-device computation tailored to device heterogeneity, network conditions, and application demands. Moreover, it seeks to refine computations for intricate streaming analytics using inherently unreliable sensor networks, and to harness a continuum of sensors, edge devices, and the cloud to adapt computation, meeting user requirements. These innovations are poised to significantly advance sustainable agriculture, foster cross-disciplinary expertise, and democratize decision-making in our sensorized world. With backing from the NSF's Computer and Information Science and Engineering (CISE) Directorate, Sirius embodies my vision of empowering sensor nodes to be decision-making agents, driving real-time, rightsized, and decentralized solutions for sustainable agriculture and other IoT applications.

*Funding for Teaching Innovation:* A cornerstone of my mission extends beyond research; it emphasizes the importance of imparting advanced machine learning and data engineering knowledge to the next generation. I am ardently committed to curating a robust academic experience for both undergraduate and graduate students. Recognizing the rapid evolutions in the field of data science, I have been developing material for my

courses from the latest methodologies, from algorithms and databases to cloud computing, distributed systems, and the emerging domain of federated learning. Furthermore, the integration of ethics into these curricula ensures that students are equipped to navigate the complexities of the real world with integrity.

To substantiate this vision, we have secured two grants from NIFA. I have introduced innovative methods in content delivery, such as podcasting, to offer students an engaging learning mode. I also bring in industry professionals as guest speakers, providing students with insights into real-world applications of theoretical concepts. My educational podcasts can be accessed here: `https://schaterji.io/podcast.html`. Through these efforts, I aim to provide a comprehensive learning experience for all students.

*Industry Support:* My research on serverless cloud computing optimization has been funded by private companies, Amazon, Adobe Research, and Microsoft Research.

# 5 Research Group and Collaborations

I lead ICAN, comprising 20 dedicated researchers. Our team's makeup is diverse, with four staff members, which includes a talented female software engineer. The heart of our research efforts stems from our eight graduate students, two of whom have successfully cleared their preliminary exams and are set to graduate next year. Complementing this core team is a wonderful bunch of undergraduate researchers. Of these, 1 has graduated with his MS and is continuing with his PhD in my lab and 2 of the PhD students have passed their preliminary exams and are on track to graduate next year. I collaborate with faculty in ABE, ECE, and CS at Purdue; CS at Wisconsin Madison (Prof. Yin Li); bioengineering colleagues at Terasaki Institute of Biomedical Innovation (Prof. Ali Khademhosseini); MG-RAST team at Argonne National Lab (then headed by Dr. Folker Meyer); researchers at Microsoft Research (Dr. Sameh Elnikety); and Army Research Lab (Dr. Noah Weston, Dr. Venkat Dasari, Dr. Peng Wang). Here are some of our recent featured publications: `https://schaterji.io/publications/`.

I was also recently (July 2023) selected for the National Academy of Engineering (NAE) Frontiers of Engineering Symposium, held in Tokyo, jointly with the Engineering Academy of Japan. Through this symposium, I forged new collaborations (*e.g.,* Prof. Hironori Kasahara, Waseda University). Purdue's College of Engineering highlighted my selection in this news story: `https://schaterji.io/news/naejapan.html`