Fast, Furious, and Interpretable for your Hardest Machine Learning Workloads in the era of the Internet-of-Small-Things (IoST)

Somali Chaterji, Purdue University; https://schaterji.io

I will talk to you about the work in my lab, an applied machine learning lab --- the Innovatory for Cells and Neural Machines (ICAN), and our wondrous discoveries. ICAN innovates at the nexus of computer vision and mobile systems [Thrust 1], on the one hand, and at the interface of machine learning and genomics [Thrust 2], on the other.

Vignette 1: Bioengineering and genomics generate datasets that are high-dimensional and noisy, and with the promise of enhancing healthspan. Single-cell sequencing datasets enable the high-throughput interrogation of genome biology including gene expression, histone modifications, and methylation events to make sense of the complex data patterns. ICAN's goal is to disentangle these complex and rich datasets to draw insights and contribute to increasing the healthspan.

Vignette 2: Computer vision applications generate high rate of data streams and often need low latency analytics on the stream. Driven by this, my research is targeted at downsizing computation to fit within mobile and embedded devices, while satisfying client SLOs (service level objectives), focusing on computer vision applications and drones.

Vignette 3: Cloud computing applications should rightsize servers to map to the continually changing machine learning (and other) workloads and we have the best algorithm to profile the serverless DAGs and invoke the best-fit VMs to optimize performance by ($) cost.

Solution 1: **Kratos** *[KDD 2022]* jointly optimizes dimensionality reduction and clustering for the most interpretable cell clusters validated by the panels of marker genes that define these clusters.

Solution 2: **LiteReconfig** [*EuroSys 2022*] and SmartAdapt [*CVPR 2022*] perform principled approximation in streaming video analytics so that they can run on mobile or embedded devices and keep up with the video rate. They perform the approximation in a cost-benefit and video content-aware manner. We have also created ApproxLive, a dashboard that makes our innovation available to end users. Finally, we have an energy-aware variant of Litereconfig, Virtuoso, which also lets you define your energy SLA (in addition to the latency SLA and is more useful for energy-constrained IoST).

Solution 3: **WiseFuse** [*Sigmetrics 2022* ***best paper award***] and **Orion** [*OSDI 2022*], perform end-to-end optimization of serverless DAG workflows, driven by our analysis of real serverless cloud computing workloads from Microsoft Azure. Concretely, our work introduces two optimizations: horizontal colocation or *bundling* of parallel invocations of a function and vertical *fusion* of in-series functions, while rightsizing the VMs hosting these functions.

Bio: Somali Chaterji (pronounced shoh-MAH-lee CHA-ter-jee) is an Assistant Professor in Agricultural and Biological Engineering (ABE) and in the Elmore Family School of Electrical and Computer Engineering (ECE, by courtesy). Dr. Chaterji is also the founder and CEO of a blazing fast cloud computing company, **KeyByte** (from 2021). Dr. Chaterji leads ICAN, the **Innovatory of Cells and Neural Machines**, at Purdue. She did her PhD in Biomedical Engineering (BME) from Purdue, winning the Chorafas Dissertation Award and the Future Faculty Fellowship, followed by a postdoc in UT Austin and then a Visiting Assistant Professor stint in Purdue Computer Science. Her work has appeared at the top applied machine learning, computer systems, and computational genomics conferences and journals. She has received funding from NIH (R01), DOD (ARL), NSF (CISE-CNS, CAREER), NIFA, as well as private industries like Amazon, Microsoft Azure, and Adobe Research. Dr. Chaterji also goes by her childhood alias Sonata Chaterji, Sonata from Beethoven's Moonlight Sonata that her mom loves. Read more about her from this recent Purdue's CoE blog. She serves on the Technical Program Committees of conferences in IoT, computer systems, machine learning, and computational genomics like Usenix ATC, EuroSys, HotStorage, Middleware, ICML, and NeurIPS. She won the **NSF-CAREER award** from CISE (CNS) in January 2022, which is shaping up its cyber nook here: https://schaterji.io/projects/sirius.html

More: https://schaterji.io/extended-bio.html