Title: Efficient Knowledge Extraction and Visual Analytics of Big Data at Scale

Speaker: Soumya Dutta, full-time Scientist-2 in the Information Sciences group (CCS-3) at Los Alamos National Laboratory (LANL)

Date and Time: 3rd March 2022 (Thursday), 10:00 AM

Venue: Online

Talk abstract:

With the ever-increasing computing power, current applications, nowadays, produce data sets that can reach the order of petabytes and beyond. Knowledge extracted from such extreme-scale data promises unprecedented advancements in various scientific fronts, e.g., earth and space sciences, energy applications, chemistry, material sciences, fluid dynamics, just to name a few. However, finding meaningful and salient information efficiently and compactly from such big data and presenting them effectively and interactively is one of the fundamental problems in modern computer science research. Big data is often characterized by the 5 Vs, namely, Volume, Velocity, Variety, Veracity, and Value. Due to the excessive volume, big data is difficult to store, manage, curate, & analyze. Compared to the disk storage (I/O) speed, the data generation velocity has increased significantly and so keeping all the data is becoming prohibitive. The variety and complexity of the data are constantly challenging the existing analysis capabilities. Finally, the veracity and value of big data are of fundamental importance for it to be used reliably to make informed decisions and build trust.

My talk will focus on addressing these 5 Vs of big data while presenting novel strategies for big data analytics and visualization. I will discuss state-of-the-art data exploration methodologies that encompass the end-to-end exploration pipeline, starting right from the data generation time until when the data is being analyzed and visualized interactively to advance scientific discovery. I will present statistical and machine learning-based compact scientific data representations and models that are significantly smaller compared to the raw data and can be used as a proxy for the raw data to answer scientific questions efficiently. I will demonstrate the successful application of such model-based visual analytics techniques by showing applications from several scientific domains such as computational fluid dynamics, climate science, energy applications, etc. Finally, I will briefly discuss my future research plan involving statistical analytics of extreme-scale spatiotemporal data, uncertainty-aware machine learning techniques for visual analytics, and understanding the principles of deep learning for interpretability and explainability.

Bio:

Soumya Dutta is a full-time Scientist-2 in the Information Sciences group (CCS-3) at Los Alamos National Laboratory (LANL). Before accepting a scientist position, Dr. Dutta was a postdoctoral researcher in the Applied Computer Sciences group (CCS-7) at LANL from June 2018 - July 2019. Dr. Dutta obtained his Ph.D. degree in Computer Science and Engineering from the Ohio State University in May 2018. Prior to joining Ohio State, Dr. Dutta completed his B. Tech. in Electronics and Communication Engineering from the West Bengal University of Technology in 2009. His current research interests include Big Data Science & Visualization, Statistical Techniques for Big Data, In Situ Analysis, Machine Learning for Visual Computing, and HPC for Visualization. Dr. Dutta's research has won Best Paper Award at ISAV 2021 and Best Paper Honorable Mention Award at IEEE Visualization 2016. He was nominated for the Ohio State Presidential Fellowship in 2017 and was also recently selected for the Best Reviewer, Honorary Mention Award for the IEEE TVCG journal for the year 2021. He is a member of IEEE and ACM.