香港中文大學統計學系
Department of Statistics
THE CHINESE UNIVERSITY OF HONG KONG

# SEMINAR

## DEPARTMENT OF STATISTICS
## THE CHINESE UNIVERSITY OF HONG KONG

## Boosting Data Analytics with Synthetic Volume Expansion


**INVITED SPEAKER**

Xiaotong Shen
Professor
School of Statistics
University of Minnesota


**TIME**

January 11, 2024 (Thu) · 2:30 pm - 3:30 pm

**VENUE**

HYS G05 · Hui Yeung Shing Building G05 · CUHK


**ABSTRACT**

Synthetic data generation heralds a paradigm shift in data science, addressing the challenges of data scarcity and privacy and enabling unprecedented performance. As synthetic data gains prominence, questions arise regarding the accuracy of statistical methods compared to their application on raw data. Addressing this, we introduce the Synthetic Data Generation for Analytics framework, which applies statistical methods to high-fidelity synthetic data produced by advanced generative models like tabular diffusion models. These models, trained using raw data, are enriched with insights from relevant studies. A significant finding within this framework is the generational effect: the error of a statistical method initially decreases with the integration of synthetic data but may subsequently increase. This phenomenon, rooted in the complexities of replicating raw data distributions, introduces the "reflection point," an optimal threshold of synthetic data defined by specific error metrics. Through three case studies–sentiment analysis, predictive modeling, and inference of tabular data, we demonstrate the effectiveness of this framework. This work is joint with Y. Liu and R. Shen.