CSSI Elements: Development of Assumption-Free Parallel Data Curing Service for Robust Machine Learning and Statistical Predictions

PI: In-Ho Cho, Co-PI: Jae-Kwang Kim
Institutions: Iowa State University

Award #: 1931380

Grand Challenges

- Incomplete data is pandemic in broad science and engineering
- Lack of theory and software of missing data curing (called "imputation") for **large/big incomplete data**
- Naïve imputation may substantially hamper the accurate machine learning (ML) and statistical learning (SL)-based predictions

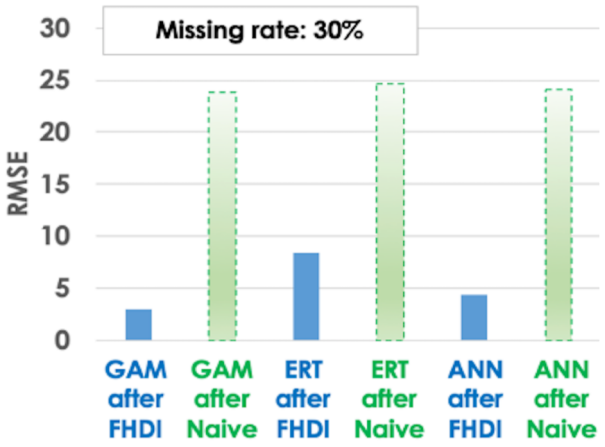

Fig. Positive impact of the proposed data curing method (FHDI) on statistical learning (SL) and ML predictions: Generalized additive model (GAM); Extremely randomized trees (ERT); Artificial neural network (ANN). Root mean square error (RMSE) is shown.

Research Objectives

- Develop a new community-level data curing service on NSF XSEDE and local academic HPC facilities
- No restriction of data size, type, high-dimensionality; No distributional assumptions or expert knowledge
- Pursue a purely data-driven imputation by developing the **parallel fractional hot deck imputation** (P-FHDI)

➤ Assumption-Free, General Data Curing | ➤ Pursue Generality, Accuracy and Scalability | ➤ Offer Information about ML/SL Using the Cured Data

Proposed Methods

- Hybrid Parallelisms & Sure Independence Screening (SIS) for P-FHDI's Core Steps
- Leverage **only observed data**, no artificial data generation; Selective SIS for big-*p* (high-dimensional) data
- Parallelized Cell Construction, Joint Cell Probability by Modified EM Algorithm, Imputation, and Variance Estimation
- Successfully Developing a Foundation of **Curing Large/Big Incomplete Data for Robust ML and SL**