Award #: 1931380

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

Grand Challenges

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

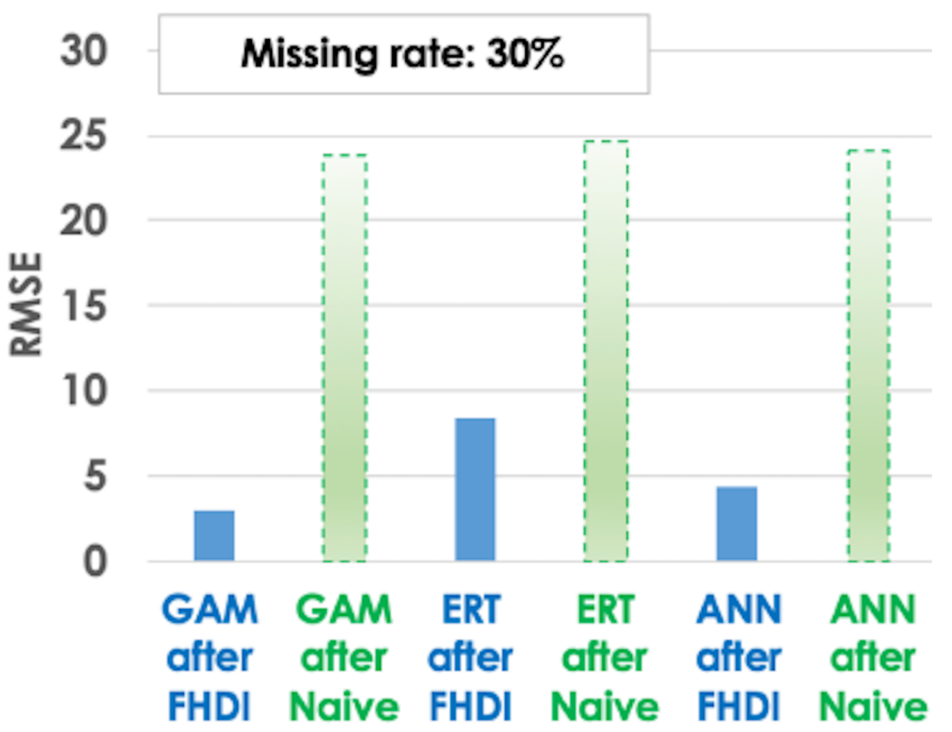


Fig. Positive impact of the proposed data curing method (FHDI) on statistical learning (SL) and ML predictions: Generalized additive model (GAM); Extremely randomized trees (ERT); Artificial neural network (ANN). Root mean square error (RMSE) is shown.

Research Objective

- Develop a new community-level data curing service running on NSF Cyberinfrastructure (e.g. XSEDE)
- No restriction of data size, type, high-dimensionality; No distributional assumptions or expert knowledge on data science
- Pursue a purely data-driven imputation by developing the **parallel fractional hot deck imputation** (P-FHDI)

➤ Assumption-Free, General Parallel Data Curing; **Only Observed Data** are Leveraged for Imputation (thus, "Hot-deck")

➤ Pursue Generality, Accuracy and Scalability in the Context of ML and SL

➤ Offer Information about ML/SL Predictions Using the Cured Data

Proposed Methods

Hybrid Parallelisms & Sure Independence Screening for P-FHDI's Core Steps

[Step 0] Sure Independence Screening (SIS)
Selectively Done for ***big-p*** (high-dimensional) Data

[Step 1] Parallel Imputation Cell Construction
Continuous → Discrete; Categorical → Unchanged

[Step 2] Imputation Cell's Joint Probability
Parallelized Modified EM Algorithm

[Step 3] Fractional Hot Deck Imputation
Parallelized Donor Selection and Imputation

[Step 4] Variance Estimation
Parallelized Jackknife Method

Results

Data: U(size, variables, missing rate)

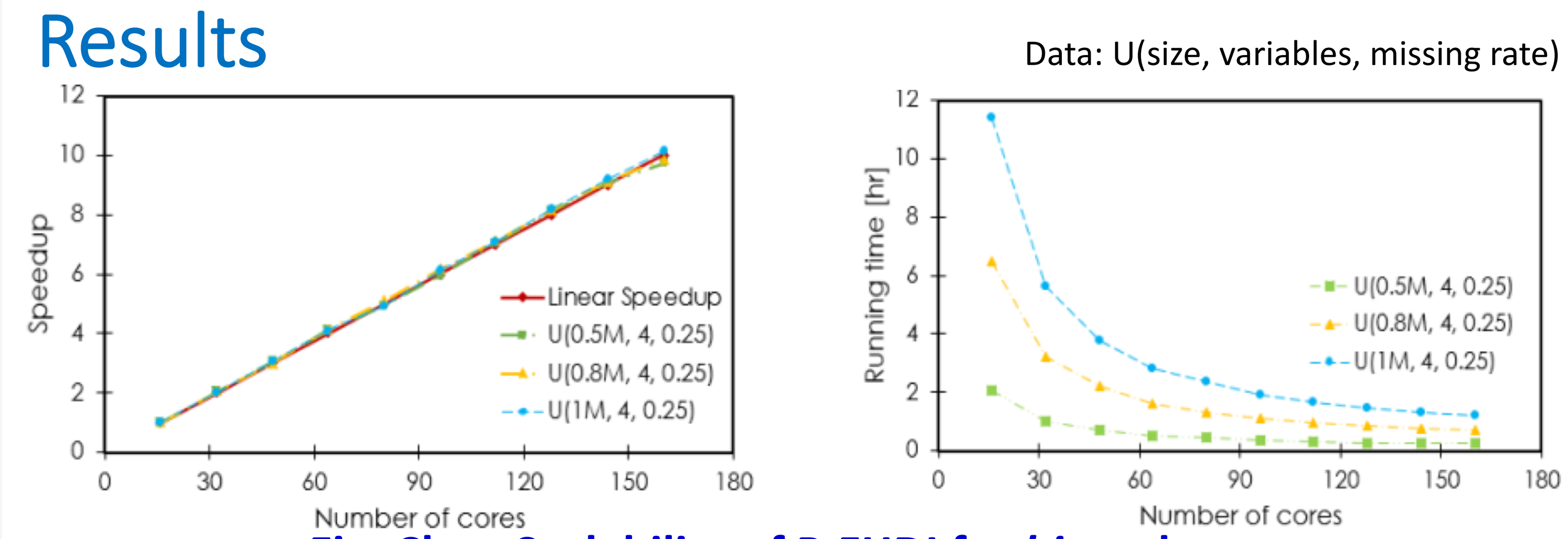


Fig. Clear Scalability of P-FHDI for *big-n* data

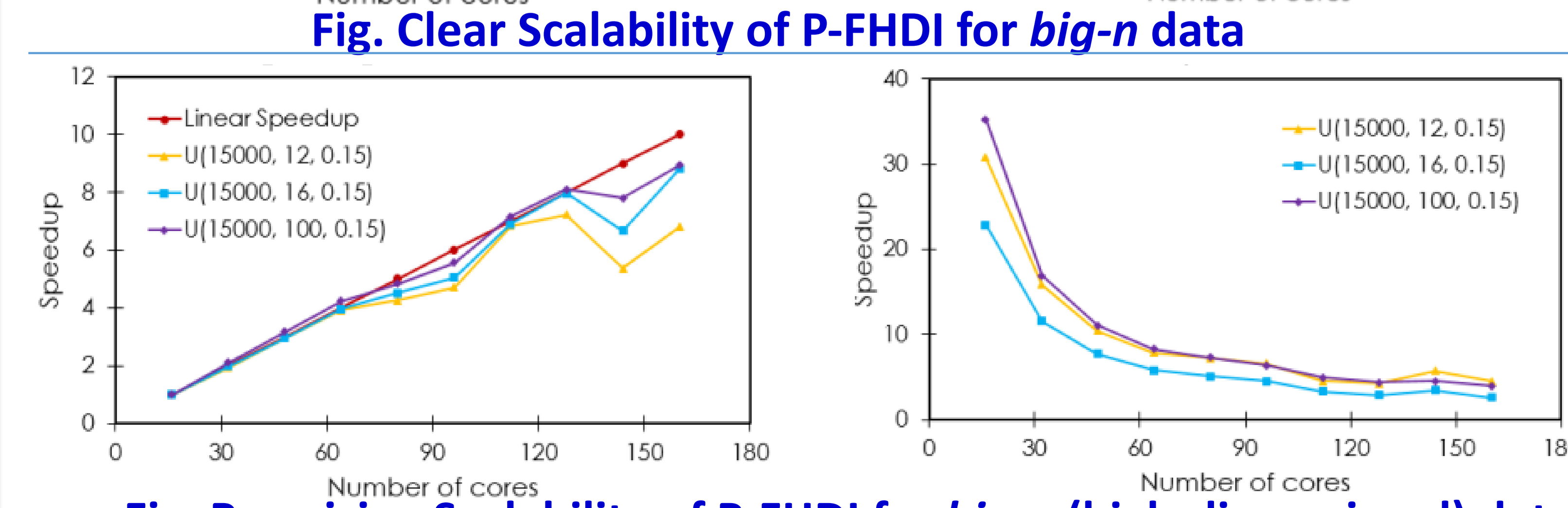


Fig. Promising Scalability of P-FHDI for *big-p* (high-dimensional) data

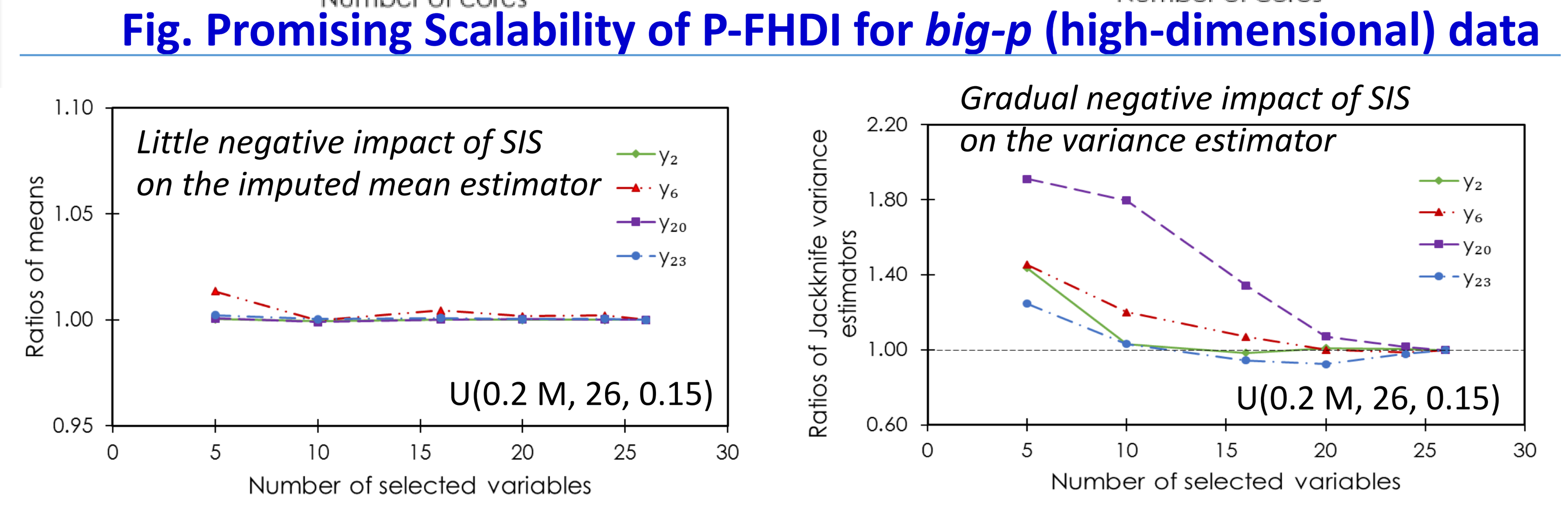
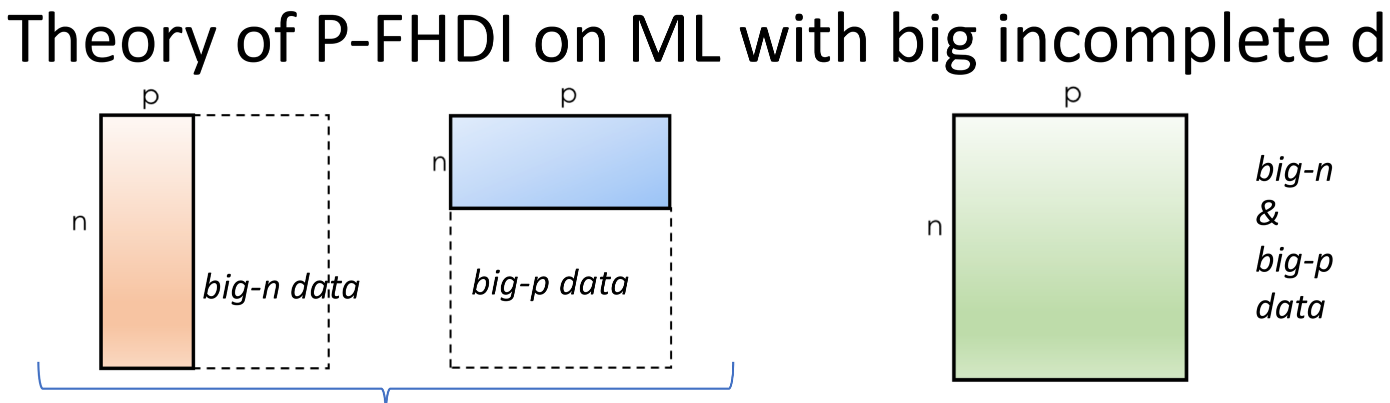
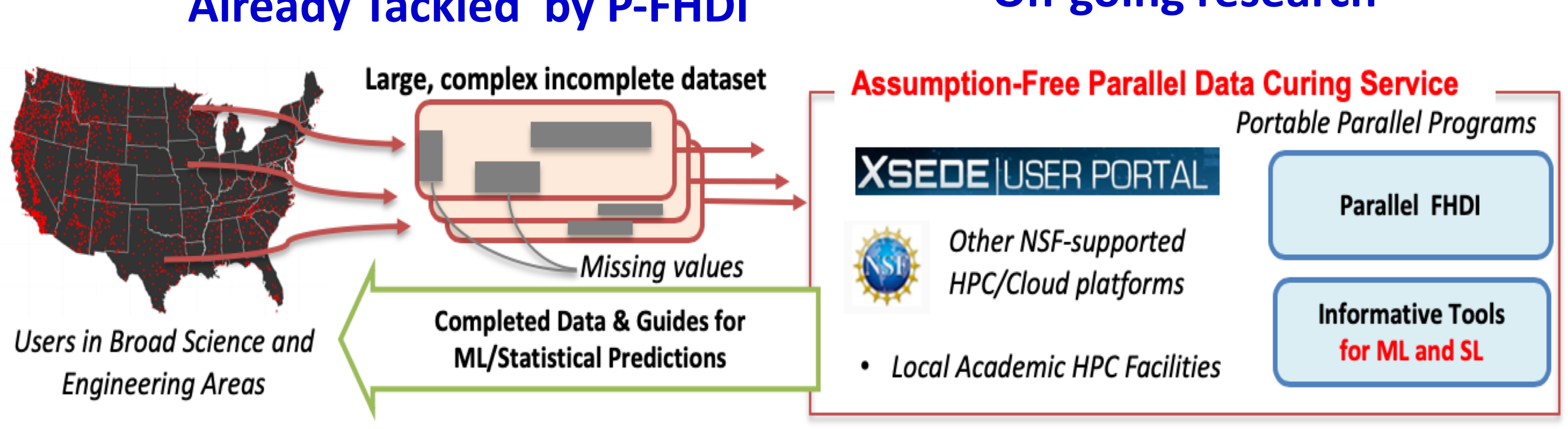


Fig. Impact of SIS on the P-FHDI's Mean and Variance estimator

Future Research Topics

- New Theory for Concurrently *big-n* & *big-p* data
- Deployment of the P-FHDI on *NSF XSEDE*
- Theory of P-FHDI on ML with big incomplete data





Conclusions

- For improving prediction accuracy of machine learning and statistical learning with large/big incomplete data, P-FHDI has been successfully developed
- Hybrid parallelisms and the sure independence screening (SIS) are key enabler of P-FHDI
- Current version P-FHDI can tackle *big-n* OR *big-p* data with the promising scalability and accuracy
- Developed P-FHDI program is available upon request to PIs (Note: serial version R Package ***FHDI*** readily available on *CRAN*)

Acknowledgement

- PIs work supported by NSF OAC-1931380 and NSF CBET-1605275
- HPC at Iowa State University supported by NSF MRI grant CNS-1229081 and CRI -1205413