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 issue is everywhere in broad science and engineering
- Theories and methods of missing data curing (called "imputation") is **limited to small data**
- **Naïve imputation** may substantially hamper the accurate machine learning (ML) and statistical learning (SL)-based predictions (see Fig below)
- **Lack of theories and software** for large/big incomplete data curing

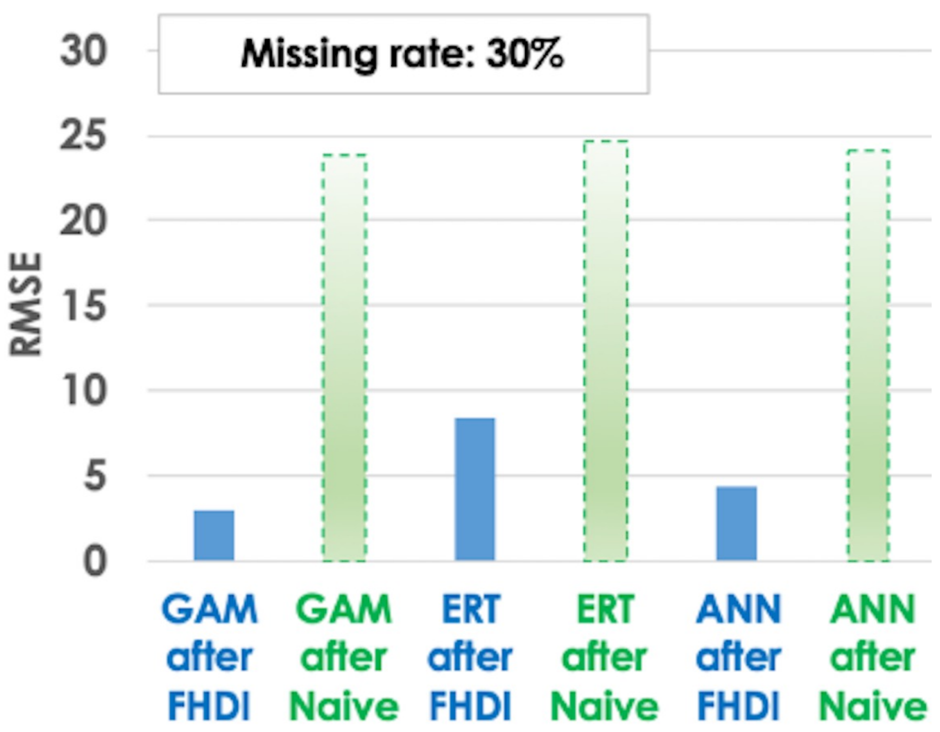


Fig. Positive impact of the proposed data curing method (FHDI) on statistical learning (SL) and ML predictions: Generalized additive model (GAM); Extremely randomized trees (ERT); Artificial neural network (ANN). Root mean square error (RMSE) is shown.

Research Objective

- Develop a new community-level **large data curing service** running on NSF Cyberinfrastructure (XSEDE) and local HPC
- **No restriction** of data sizes, types, high-dimensionality; No distributional assumptions or expert knowledge on data science required
- Pursue a purely data-driven imputation by developing the **ultra data-oriented parallel fractional hot deck imputation (UP-FHDI)**

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

➤ Provide **"Cured" large/big data set** for convenient subsequent ML and SL

Proposed Methods

Ultra Data-Oriented Parallel Fractional Hot Deck Imputation (**UP-FHDI**) [Ultra Data: Big-n & Big-p]

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

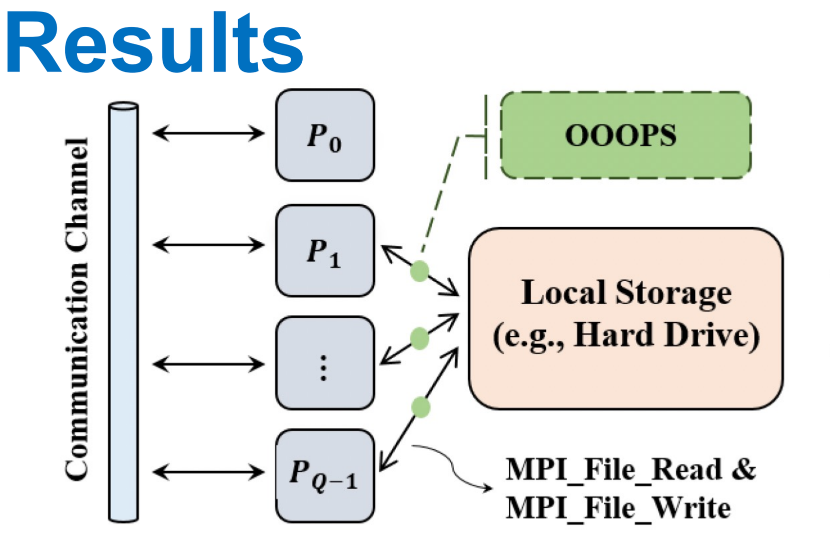
[Step 1] Parallel Imputation Cell Construction
Hybrid Data (Continuous & Categorical)

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

[Step 3] Fractional Hot Deck Imputation
Parallelized Donor Selection with KNN

[Step 4] Parallel Variance Estimation
Parallelized Jackknife & Parallel Linearization

Results



Adopted OOOPS (Huang and Liu, 2020)

Data: U(size, variables, missing rate)

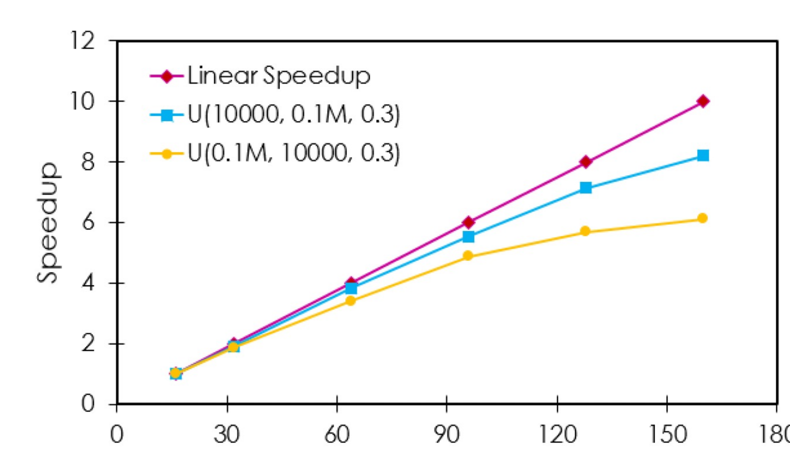


Fig. Scalability of UP-FHDI

SD	UP-FHDI	Naive
3	0.052	0.062
5	0.088	0.100
8	0.134	0.160

Fig. Average Standard Error

SD	UP-FHDI	Naive
3	0.077	0.081
5	0.079	0.088
8	0.080	0.092

Fig. RMSE

Synthetic Big-p Data (n=10,000, p=0.1 million, Missing=0.3) with increasing randomness (SD) in data

Test Data Sets

Dataset name	Number of Instances	Number of Variables	Discipline	Data Source
CT [Graf et al. 2011]	53500	380	Medicine	UCI
p53 [Lathrop 2010]	31159	5408	Genetics	UCI
Travel [Gao et al. 2021]	23772	50	Transportation	IEEE Dataport
Swarm [Abpeikar et al. 2020]	24016	2400	Biology	UCI

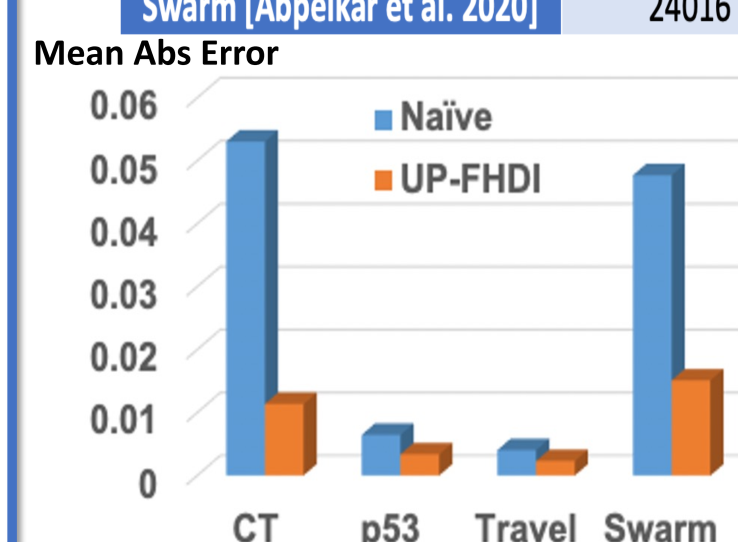


Fig. Accuracy of UP-FHDI

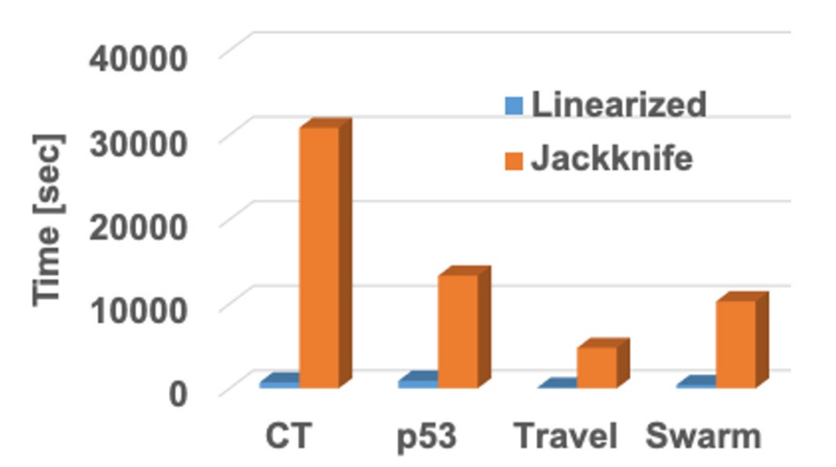


Fig. New Variance Est. Method's Efficiency

User-Friendly Service

- Deployment of the UP-FHDI on *NSF XSEDE*
- Graphical User Interface (GUI) for UP-FHDI
- Open data sets, examples, and manual videos

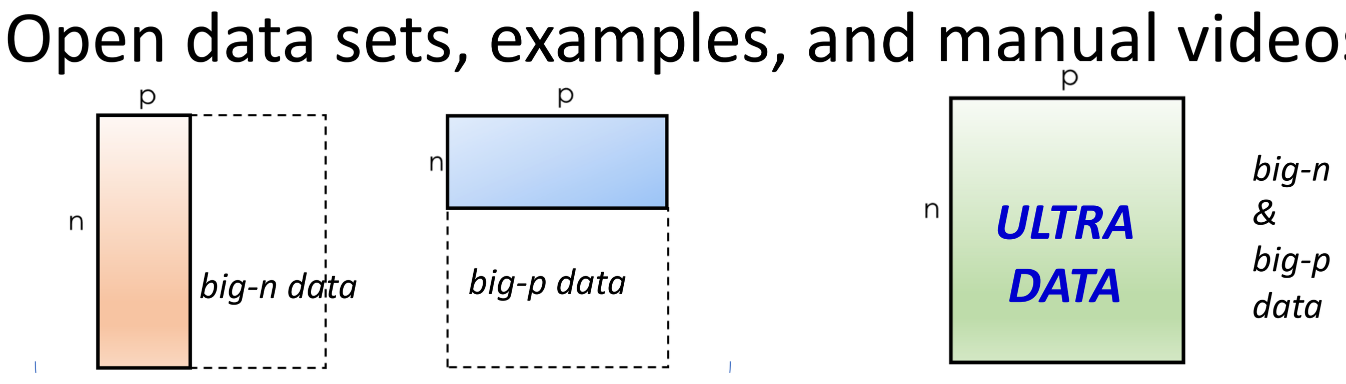
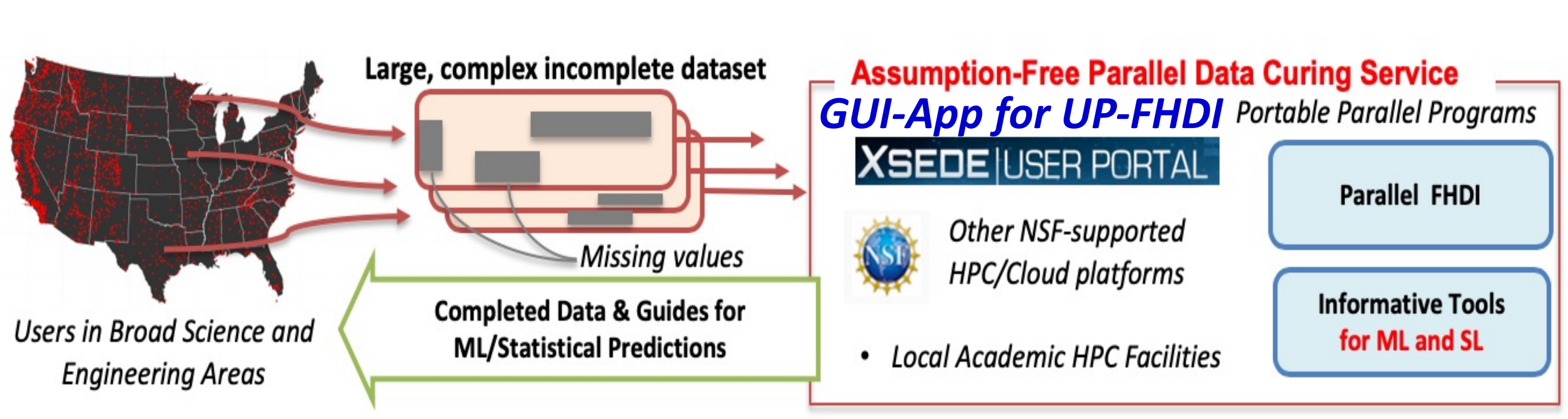


Fig. UP-FHDI Can Tackle Three Data Types



Conclusions

- **UP-FHDI** has been developed for improving prediction accuracy of ML and SL with Ultra incomplete data (**up to millions of instances and 10,000 variables**)
- The program is deployable on NSF XSEDE and local HPC
- Serial version ***R Package FHDI*** available on *CRAN*

References of the PIs

- Yang et al., 2022, *IEEE TKDE (under 2nd review)*
- Yang et al., 2020, *IEEE TKDE*
- Song et al., 2019, *IEEE TKDE*
- Im et al., 2018, *The R Journal*

Acknowledgement

- PIs work supported by NSF OAC-1931380
- HPC at Iowa State University: *Condo* supported by NSF MRI grant CNS-1229081 and CRI-1205413; *Nova* by MRI-1726447 and 2018594