MISL: Multiple Imputation by Super Learning

Thomas Carpenito, M.A.
Justin Manjourides, Ph.D.
Northeastern University, Department of Health Sciences

Symposium on Data Science and Statistics

Motivation

- Missing data is ubiquitous in research
- Researchers must decide how to handle missing data (listwise and pairwise deletion, single imputation, *multiple imputation*, etc…)
    - Some examples include:
        - Predictive mean matching
        - Random sampling
        - Classification and regression trees
        - Mean/mode imputation
- Multiple imputation hinges on the assumption of correctly specifying an imputation model

What if any one imputation model does not satisfactorily capture the true underlying data distribution?

Motivation

- Multiple Imputation by Super Learning (MISL) is a missingness-agnostic multiple imputation mechanism
  - The algorithm consistently outperforms: mean imputation and Multivariate Imputation by Chained Equations (MICE - popular among researchers*) when data are:
    - Missing completely at random (MCAR)
    - Missing at random (MAR)
    - *Missing not at random (MNAR)*
  - Applications for: survey, cross-sectional, longitudinal, hierarchical data sets as well as “big data”

*Hayati Rezvan, P., Lee, K.J. & Simpson, J.A. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol* **15,** 30 (2015). https://doi.org/10.1186/s12874-015-0022-1

Overview

Multiple **I**mputation by **S**uper **L**earning

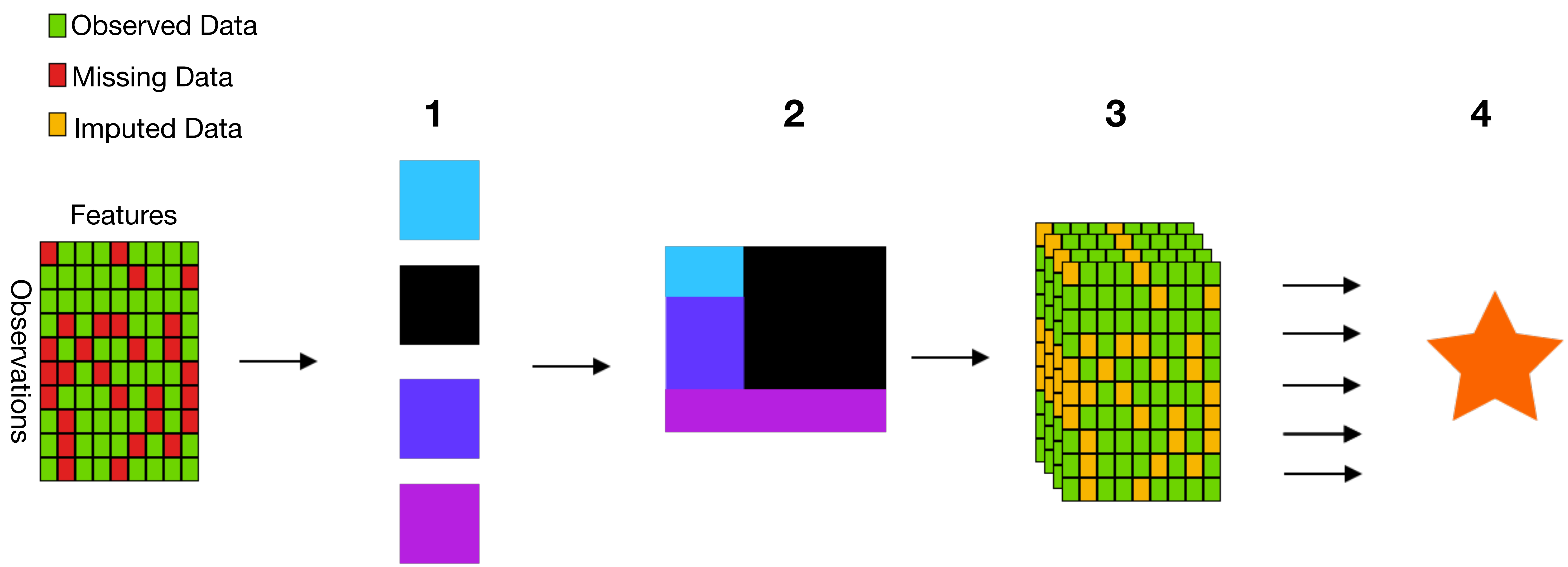
The generation of *m* distinct datasets allowing for uncertainty in the imputations

An algorithm that uses cross validation to generate predictions by combining a user-identified list of candidate models*

An imputation technique that iteratively generates *m* complete datasets with the use of ensemble learning

*van der Laan, Mark J.; Polley, Eric C.; and Hubbard, Alan E., "Super Learner" (July 2007). *U.C. Berkeley Division of Biostatistics Working Paper Series.* Working Paper 222. https://biostats.bepress.com/ucbbiostat/paper222

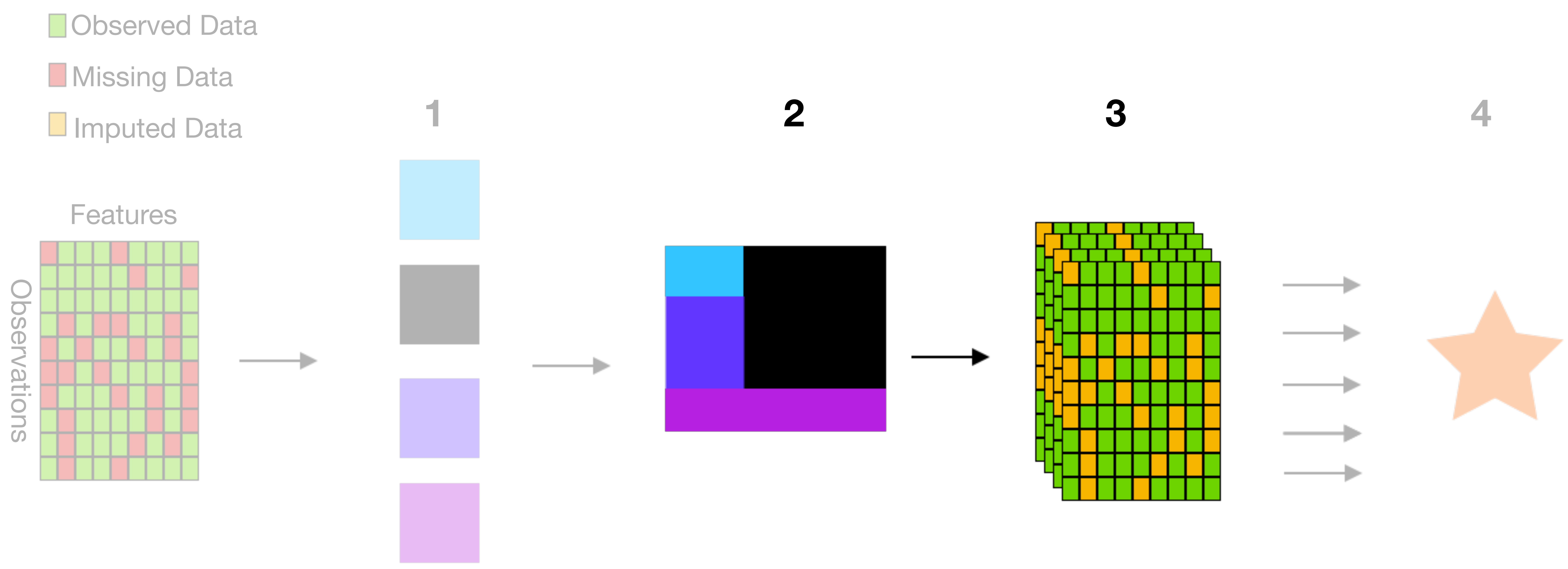
Overview



1. A selection of candidate algorithms are chosen for the super learner
2. The super learner uses cross validation to determine the column-specific combination of each algorithm for imputation
3. The MISL algorithm runs a set number of iterations and generates a number of complete datasets
4. The now full datasets can be analyzed (independently) and estimates can be combined using Rubin's Rules*

* Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Overview



1. A selection of candidate algorithms are chosen for the super learner
2. The super learner uses cross validation to determine the column-specific combination of each algorithm for imputation
3. The MISL algorithm runs a set number of iterations and generates a number of complete datasets
4. The now full datasets can be analyzed (independently) and estimates can be combined using Rubin's Rules*

* Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Between Steps 2 and 3

A. MISL selects the feature with the least amount of missing data (X_c) and imputes the mean/mode as placeholders for all other missing features

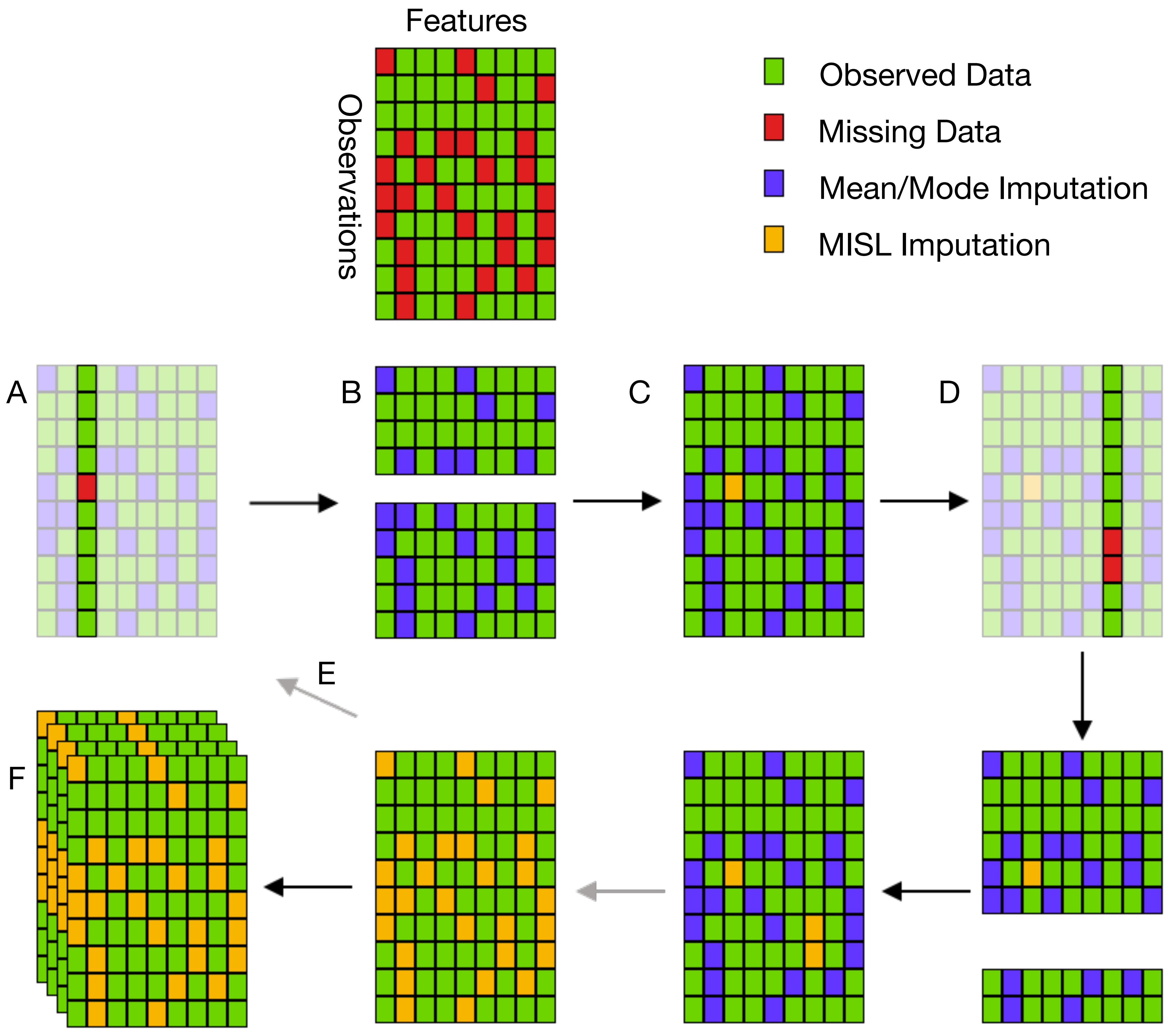
B. MISL Isolates observations for which a value for X_c exists

C. Super learner generates an ensemble and predicts X_c using remaining (mean/mode imputed) features

D. The cycle repeats for the next feature with the least amount of missing data using the newly imputed values

E. After all features have been imputed, the algorithm iterates a set number of times until convergence using the previous iterations imputations as placeholders

F. A complete dataset is generated and the algorithm continues *m*-1 more times



Simulation

Imputation with the following distribution using both small (100 observations) and large (1000 observations) datasets:

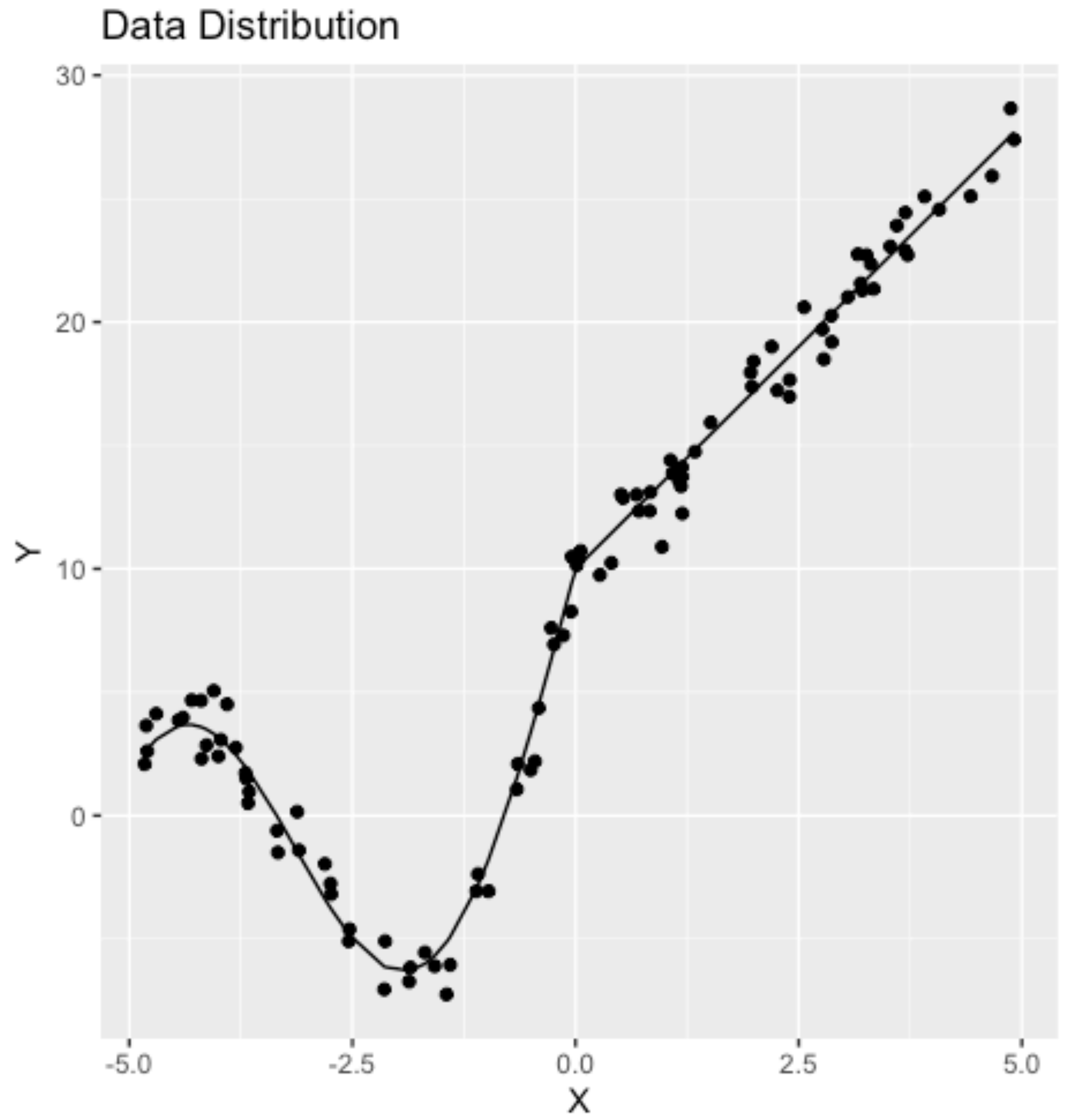
$$y = 10 + (10sin(x) \times I(x < 0)) + 3.6x + N(0,1)$$

Proportion of missing data by variable and pattern:

	Total Obs.	%X missing	%Y missing	%XY missing
MCAR	100/1000	30	40	13/12.4
MAR	100/1000	20/21	33/30	6/9.1
MNAR	100/1000	26/25	24/40	7/14.5

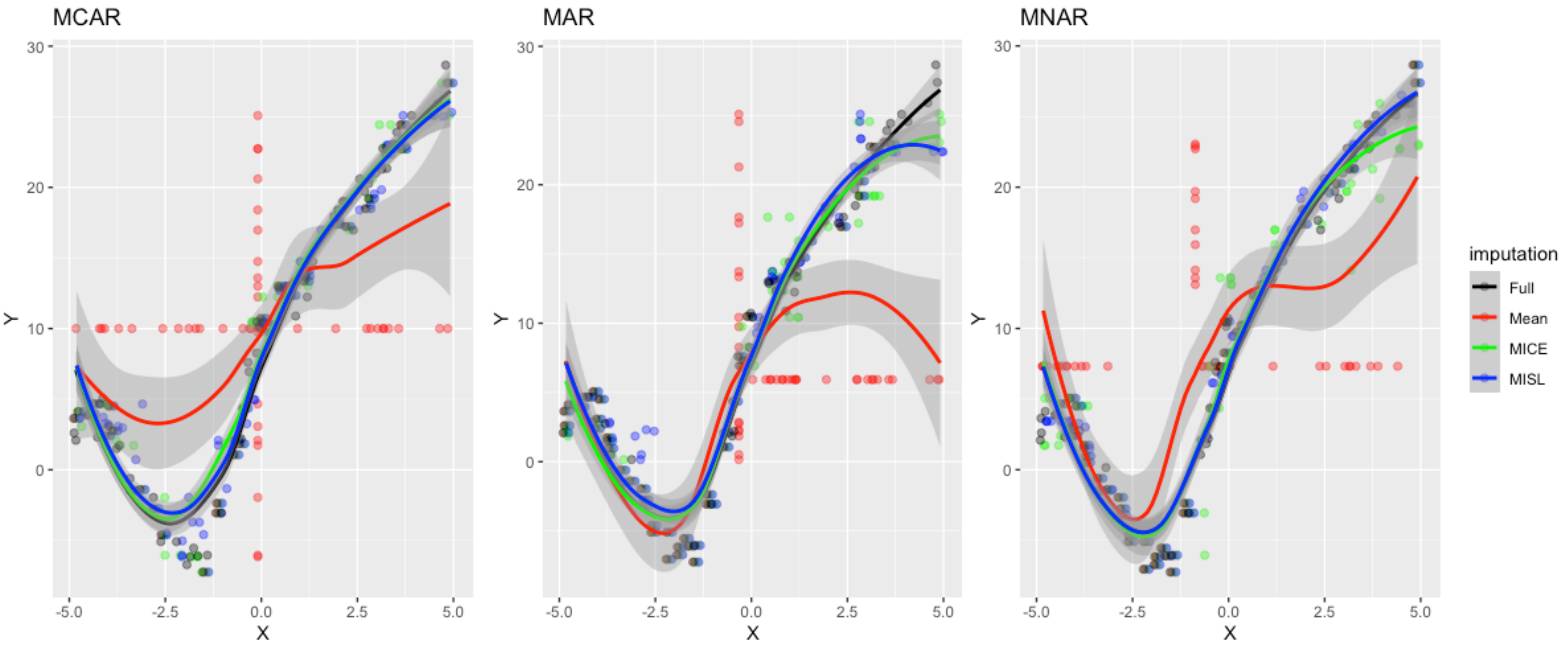
Candidate algorithms in MISL:

A. Generalized linear model
B. Gradient boosting
C. Random forest
D. Pruned regression tree
E. Mean imputation



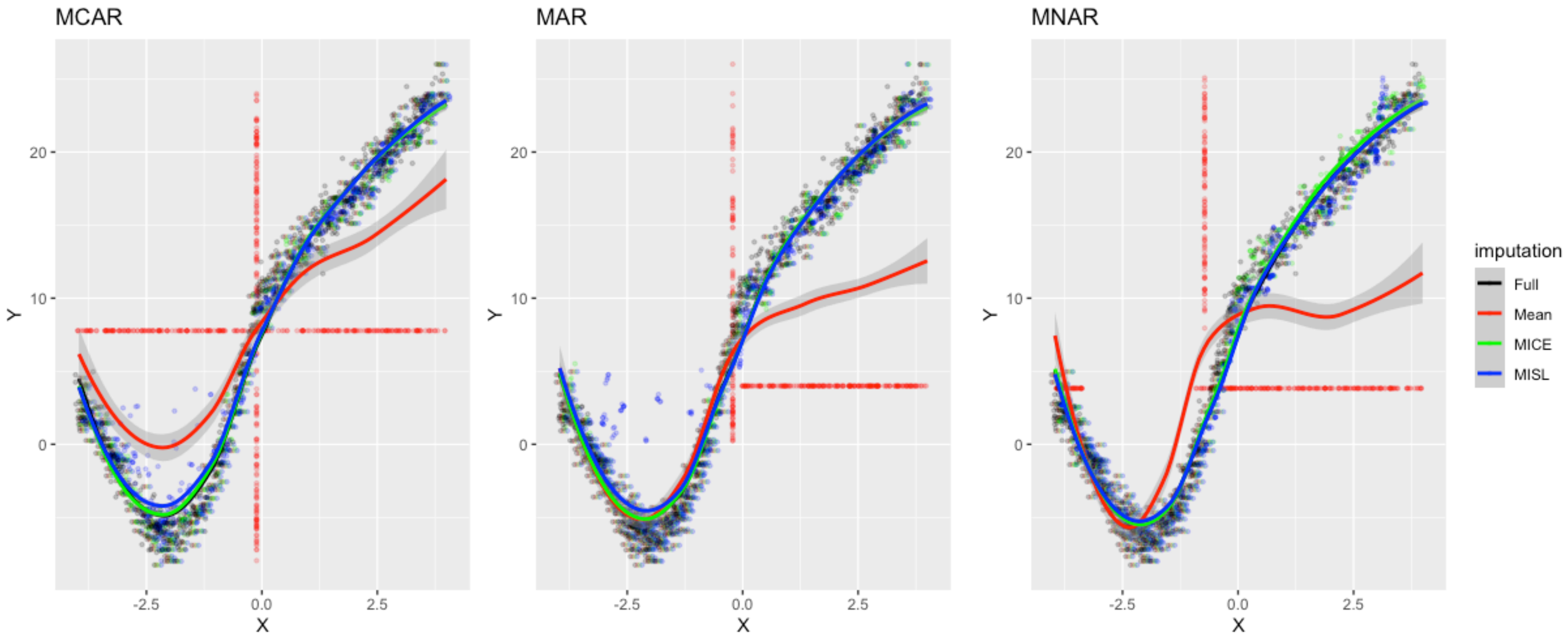
Observed (dots) and actual (line) data distribution for the “small” dataset

Imputations (small dataset)



Imputed values for a **single dataset** of the MICE/MISL algorithm using the small dataset

Imputations (large dataset)



Imputed values for a **single dataset** of the MICE/MISL algorithm using the large dataset

Squared Prediction Error

Small (100 observation) dataset

Method	MCAR_X	MCAR_Y	MAR_X	MAR_Y	MNAR_X	MNAR_Y
Mean	1.58(4.31)	38.54(78.92)	1.37(4.07)	42.96(105.89)	1.33(4.26)	28.87(70.46)
MICE	0.22(1.52)	1.39(3.72)	0.47(2.7)	2.2(7.41)	0.16(0.56)	3.31(9.52)
MISL	0.28(1.61)	0.93(2.13)	0.19(0.83)	1.56(5.67)	0.03(0.14)	0.73(2.64)

Large (1000 observation) dataset

Method	MCAR_X	MCAR_Y	MAR_X	MAR_Y	MNAR_X	MNAR_Y
Mean	1.11(3)	33.52(64.48)	0.71(2.7)	44.09(96.46)	1.08(3.68)	37.05(89.28)
MICE	0.18(1.18)	0.68(1.95)	0.23(1.47)	0.93(5.35)	0.02(0.09)	0.68(2.25)
MISL	0.09(0.62)	0.52(1.58)	0.14(0.82)	0.34(1.23)	0.03(0.1)	0.47(1.38)

Mean(sd) of the squared prediction error of each missingness mechanism for each imputation technique

Euclidean Distance

Small (100 observation) dataset

Method	MCAR	MAR	MNAR
Mean	3.61(5.24)	3.37(5.77)	3.07(4.59)
MICE	0.67(1.08)	0.75(1.46)	1.02(1.56)
MISL	0.62(0.91)	0.57(1.2)	0.44(0.75)

Large (1000 observation) dataset

Method	MCAR	MAR	MNAR
Mean	3.4(4.81)	3.27(5.85)	3.12(5.33)
MICE	0.48(0.8)	0.41(1)	0.4(0.74)
MISL	0.4(0.67)	0.31(0.62)	0.35(0.61)

Mean(sd) of the euclidean distance of each missingness mechanism for each imputation technique

MISL Iterations

Iteration	% GLM	% GB	% RF	% PRT	% Mean
1	3	37	0	60	0
2	4	41	0	54	1
3	2	5	33	59	0
4	2	0	59	38	1
5	2	36	15	47	1

Weights assigned to each candidate algorithm (GLM = generalized linear model, GB = gradient boosting, RF = random forest, PRT = pruned regression tree) by the super learner by dataset (m) and iteration for the Y variable when data are MAR

Conclusions

- MISL generates predictions that are:
  - Sensical/logical: as they graphically appear like the underlying data distribution
  - Accurate:
    - The average squared prediction error is as good (if not better) than both MICE and mean simulated imputations
    - The average euclidian distance between observed and actual data points is smaller amongst MISL when compared to MICE and MEAN simulated imputations
- MISL is respectful of the following assumptions:
  - The imputation is only as good as the models supplied to the super learner
  - The underlying missingness mechanism can be appropriately explained

Future Directions

- Support for more user customizations (including: custom learners, guidance on choosing algorithms for the super learner, automate Rubin's rules, etc…)
- Incorporating current prediction models into MISL
  - ex: MICE, decision tree classifiers, and voting methods
- Create an R package
- Further research regarding theoretical properties of MISL and implications of MNAR

Contact

Email: t.carpenito@northeastern.edu

Twitter: @CarpenitoThomas

Thank you!