Poster #59

IMPROVED FEATURE SELECTION WITH SIMULATION OPTIMIZATON

Sara Shashaani, Kimia Vahdat
Industrial and Systems Engineering
North Carolina State University

PRESENTER: Sara Shashaani

EDWARD P. FITTS DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

Why not include all features in the model?

- Overfitting
- Computationally expensive
- Less inference or interpretation power

Research Methodology

Given a learning model (linear regression, random forest, *etc.*) we look for the best subset of features

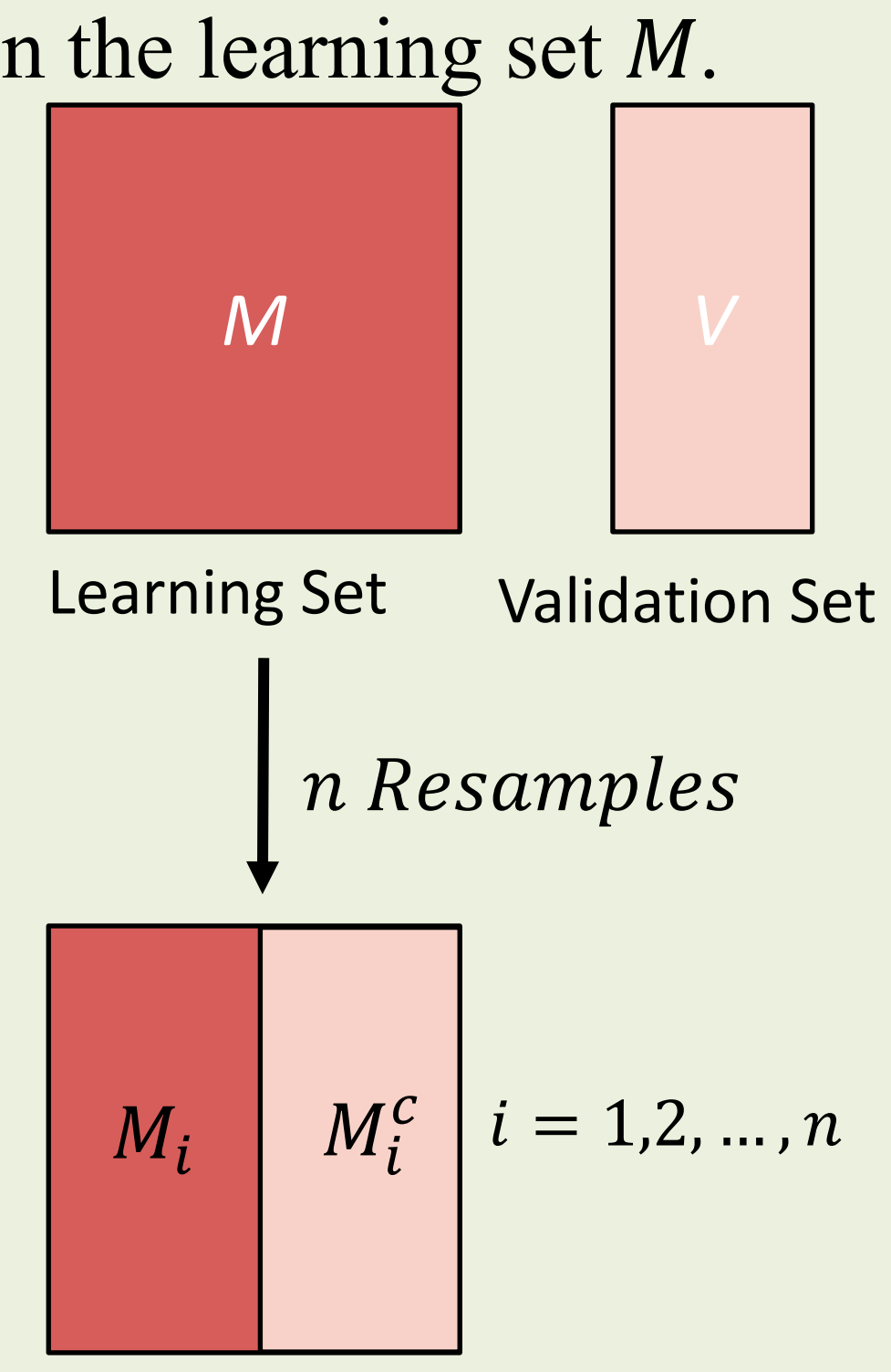
$$S^* = argmin_S \sum_{j \in V} (f_{M,A,S}(x_j) - y_j)^2$$

where $f_{M,A,S}(x_j)$ is the prediction model trained by the subset S of features of the learning set M with the learning algorithm A.

Estimate with its Sample Average Approximation

$$\hat{S}^* = argmin_S \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in M_i^c} (f_{M_i,A,S}(x_j) - y_j)^2$$

where Mi and M_i^c are resampled training and test sets within the learning set M.



Experiment

We compare the performance of Simulation Optimization based Feature Selection SOFS with *Genetic Algorithms* as the optimization method

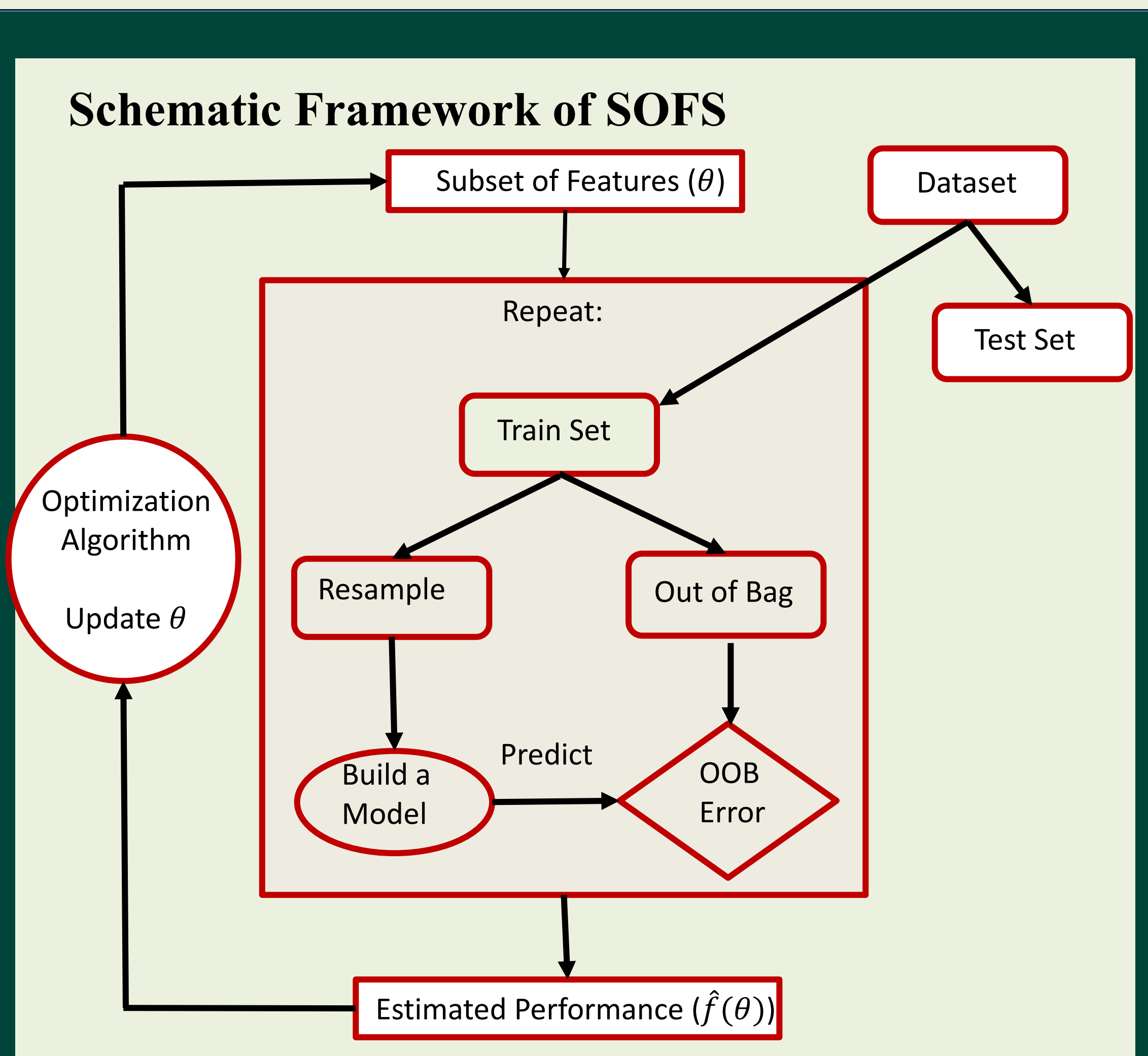
- in terms of mean absolute and squared error;
- with that of *Recursive Feature Elimination (RFE),* the commonly used greedy approach that looks for the best subset size

$$d^* = argmin_d \frac{1}{n} \sum_{i=1}^{n} \sum_{j \in M_i^c} (f_{M_i,A,d}(x_j) - y_j)^2 ;$$

- on a sample dataset from UCI repository with 55 features and 226 observations;
- on a two learning algorithms: linear regression (LM), and random forest (RF).

When looking in a dataset with many features for the most informative ones, we can develop an **optimization** problem that estimates the predictive accuracy of a prediction model with any subset of features by mimicking a **simulation** of the system under consideration, for which we only have the available data, through resampled datasets.

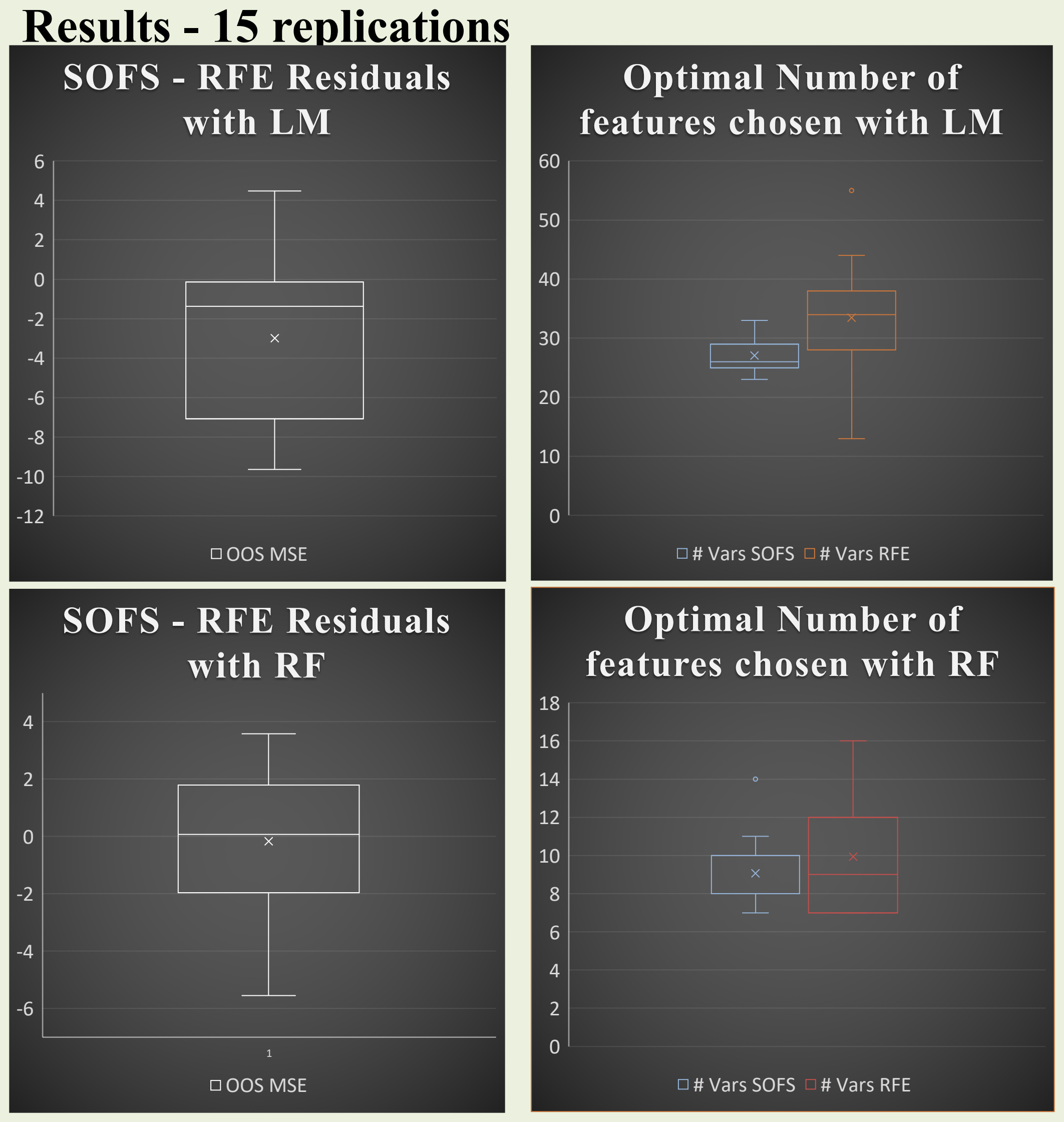
Schematic Framework of SOFS



Results - One replication

RFE vs. SOFS		# Feat	IS MAE	OS MAE	IS MSE	OS MSE	Time
LM	RFE	33	2.44	3.45	11.92	22.56	.
	GA	27	2.38	3.17	11.62	19.58	.
RF	RFE	10	2.87	3.49	15.15	22.15	0.30
	GA	9	2.79	3.44	14.35	21.98	1015.56

Results - 15 replications



Conclusion

- SOFS gains higher accuracy in predictions and more precision in number of features for both LM and RF.
- The optimization routine GA is only run for a limited budget so in RF it can stop before convergence. More efficient optimization routines are under study.

NC STATE UNIVERSITY