EDWARD P. FITTS DEPARTMENT OF INDUSTRIAL AND SYSTEMS ENGINEERING

NC STATE UNIVERSITY

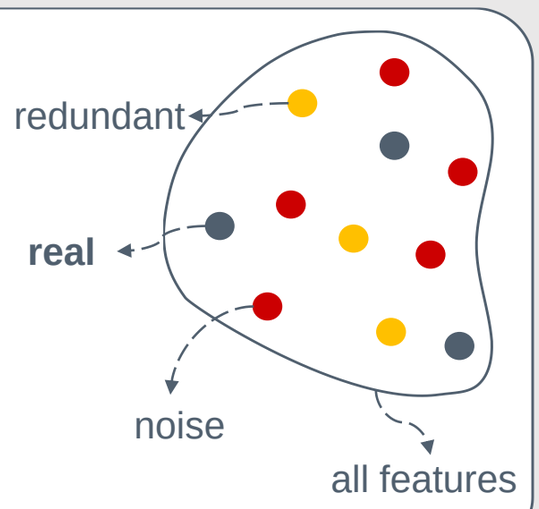
SIMULATION OPTIMIZATION BASED FEATURE SELECTION (SOFS)

Kimia Vahdat (kvahdat@ncsu.edu) and Sara Shashaani, Ph.D. (sshasha2@ncsu.edu)

Why Feature Selection?

Our goal is to find the true features to improve the prediction in:

- ✓ interpretability,
- ✓ accuracy, and
- ✓ efficiency.



Application- data analytic to predict unplanned hospitalizations

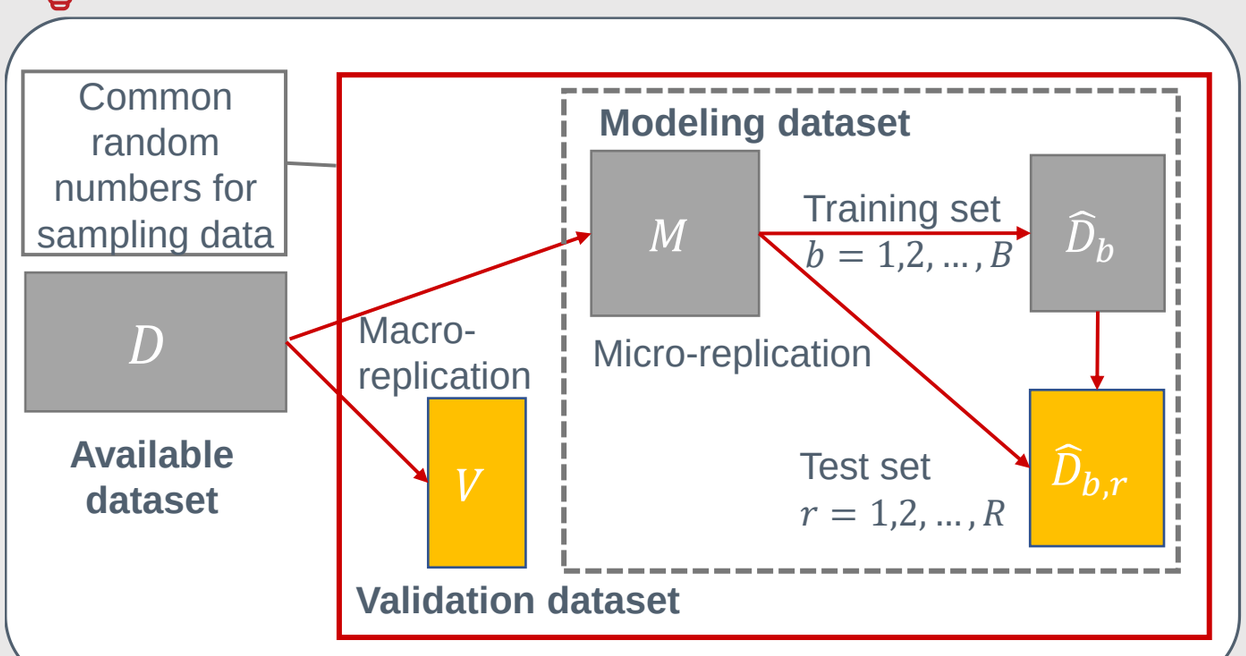
	Socio-demographics			Socio-economical	environmental			comorbidities			
Patient ID	Age	Gender	race	insurance	Income	County pollution	Average precipitation	Blood pressure	Diabetes	Obesity	Hospitalized in the next 30 days?
1	67	F	1	Yes	Average	Medium	225	Low	No	No	No
⋮							⋮				
n	75	M	2	Yes	High	Low	103	High	Yes	Yes	Yes

What we want to predict

Finding the most informative features in the big data applications is a challenge that can improve predictions and interpretability of the underlying systems.

Due to the uncertainty in the data, we address feature selection for any machine learning algorithm of choice, as a stochastic optimization. The resulting feature subsets are more robust to the changes in the data and lead to better predictions in simulated and real datasets.

How does SOFS solve it?



Datasets:

1 CMS dataset with binary response (**zero-inflated** with 9% non-zero), **19k** instances and **380** features.

Prediction of unplanned admission due to heart failure, in the following month.

For patients >65 years old, who have:

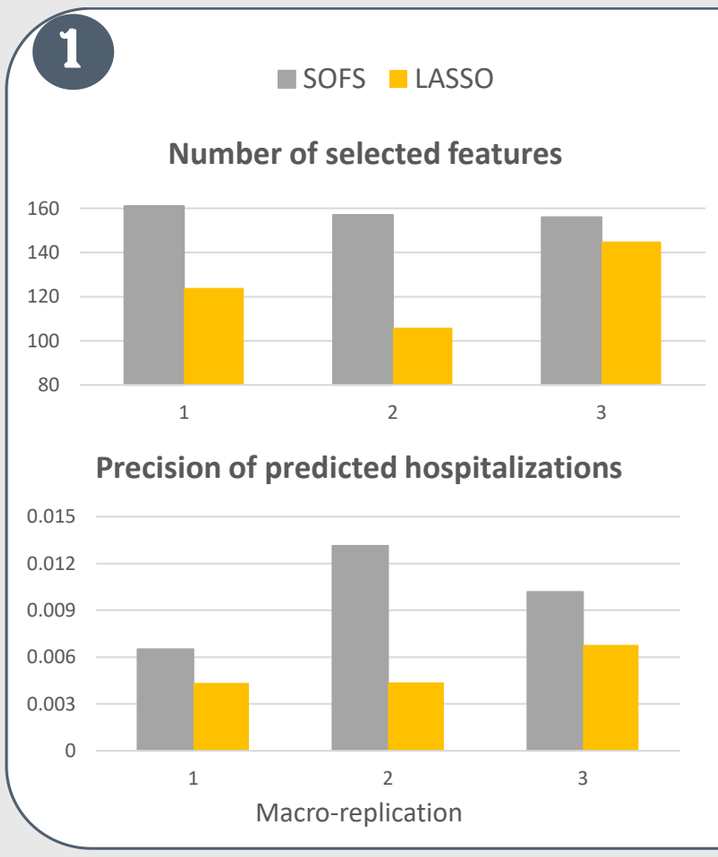
- no cancer history,
- more than 10 comorbidities,
- other minor conditions.

2 Simulated dataset with continuous response, 300 instances and 220 features; where real variables$\sim Gamma(2,2)$.

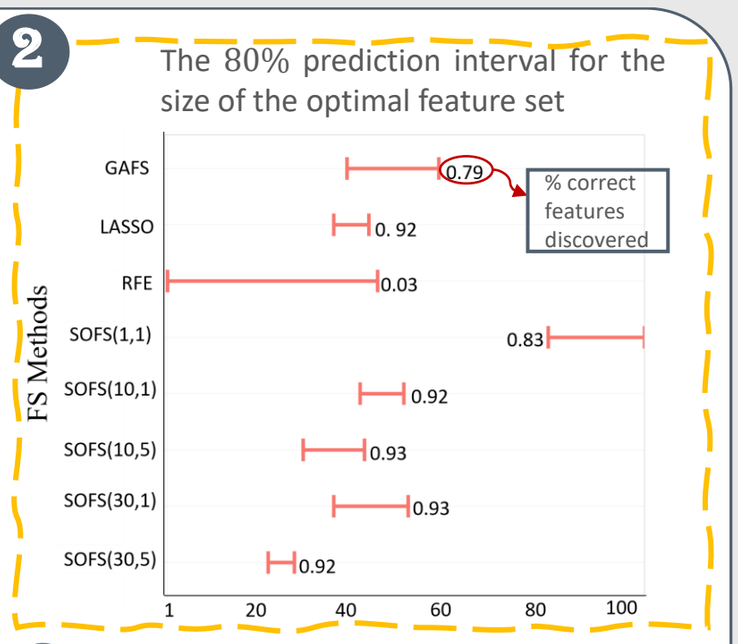
Feature type	Count
Real	12
Redundant	100
Noise	108

How does SOFS perform?

1



2



✓ Better performance in: Accuracy and Sharpness

✗ Very time consuming

Conclusion

- Increased reliability in the results (more accuracy and less variability)
- Flexibility with any learning algorithm and response type
- Improved performance in identifying true features
- Surpassed all evaluated benchmarks

Future Work

- Employing adaptive sampling to reduce computation time
- Incorporating data sampling correlation in estimation
- More comparison on selected subsets
- Estimating the estimator's bias

Thanks To Seth Guikema (University of Michigan) for initial discussions, ISE faculty, Drs. Swann, Mayorga, and Ivy for their comments, and Edward P. Fitts Graduate Fellowship for funding support.