Massachusetts Institute of Technology

Big Data are Not Necessarily Good Data
What are Good Data Anyway?

Τα μεγάλα δεδομένα δεν είναι απαραίτητα καλά δεδομένα

FIPSE (ΦΨ)

Credits: Fabian Mohr, Weike Sun, Benben Jiang, Lee Rippon, Ibrahim Yousef, Yiting Tsai, Liang Cao, Sirish Shah

5th FIPSE conference, 2022

Bhushan Gopaluni
Richard Braatz

Department of Chemical Engineering
University of British Columbia
Massachusetts Institute of Technology

June 28, 2022, Crete, Greece

UBC THE UNIVERSITY OF BRITISH COLUMBIA

Misconception #1 Higher Dimensional Data are Better

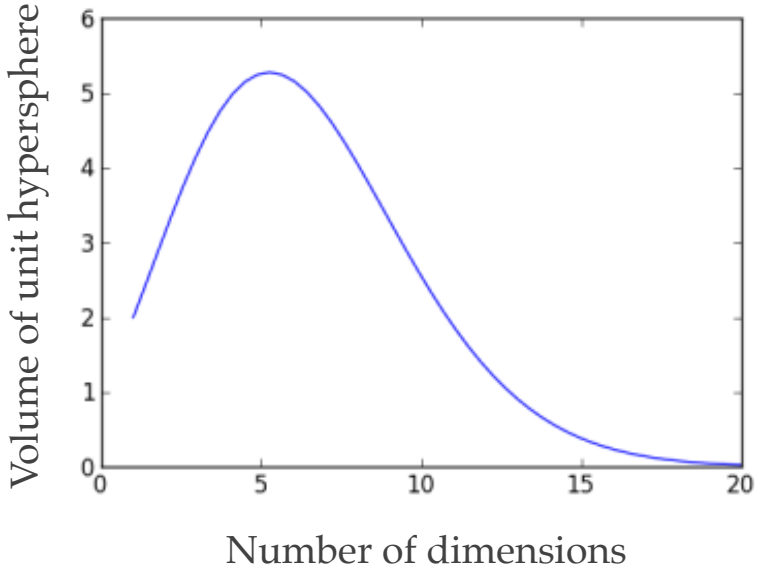
Adding more features NEVER hurts, in the worst case, they don't add new information

- **Common "wisdom"**

With the low cost of sensors and high storage capacity, conventional thinking is "why not measure as many variables as possible?"

- **The truth:** High-dimensional data can exhibit counter-intuitive phenomena

- The volume of a unit-radius sphere increases from dimensions 1 to 5 and then decreases to zero as dimension increases
- E.g., A trillion data points in a space of 100 dimensions cover only ~10^{-18}% of the input space



E.g., a biopharma application where better models were constructed when using fewer variables is in Comput Chem Eng, 80, 30-36, 2015.

Misconception #2 Volume Makes up for Quality

Machine learning algorithms with big data would make better decisions, irrespective of the quality of data

- **Common "wisdom"**

With more data, we should be able to obtain better models

- **The truth:** Big data in the process industries are characterized by noise, missing values, outliers, limited dynamics, etc.
- Even a simple least-squares algorithm will not provide good quality models if the data are not appropriately preprocessed and have “sufficient” excitation
- The question really is “Which subset of the data are useful for modelling?”

Misconception #3 - Machine Learning Algorithms are Superior

Machine learning algorithms with big data will outperform simple algorithms with small data

- **Common "wisdom"**

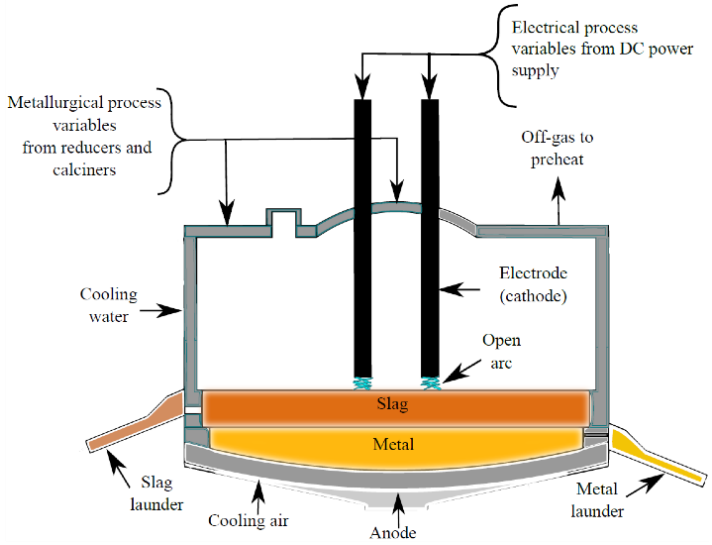
Classical modelling algorithms may not work, but "Big Data" machine learning algorithms will naturally provide superior performance

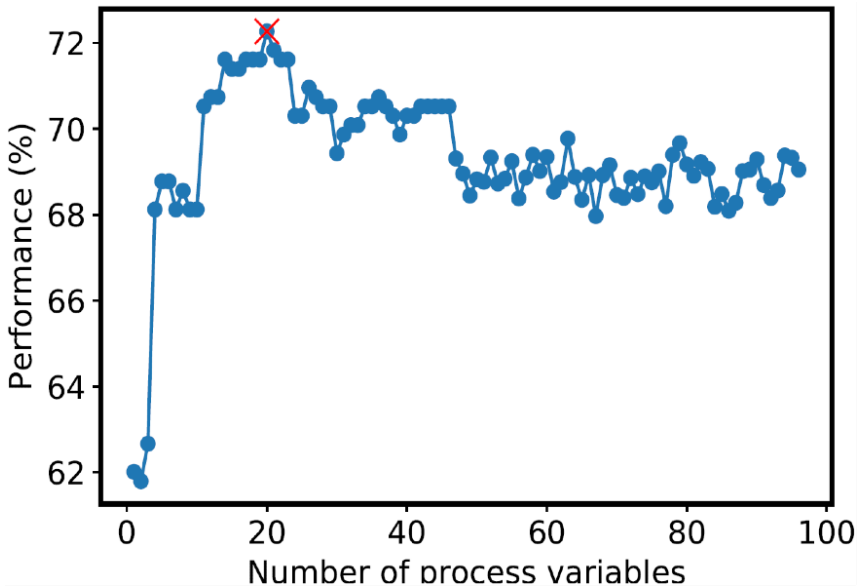
- **The truth:** The machine learning algorithms that require Big Data (e.g., convolutional neural networks) do not apply for those industrial processes for which Big Data are not available
- Often a model incorporating *a priori* process knowledge provides more accurate predictions than purely data-driven machine learning models
- There is no guarantee that machine learning algorithms will provide superior performance. In fact, several counterexamples have been reported*

* Sun, W. and Braatz, R.D., 2021. Smart process analytics for predictive modeling. *Computers & Chemical Engineering*, *144*, p.107134,
* Rippon et al Representation learning and predictive classification: Application with an electric arc furnace. *Computers & Chemical Engineering*, *150*, p.107304

Example #1: Electric Arc Furnace

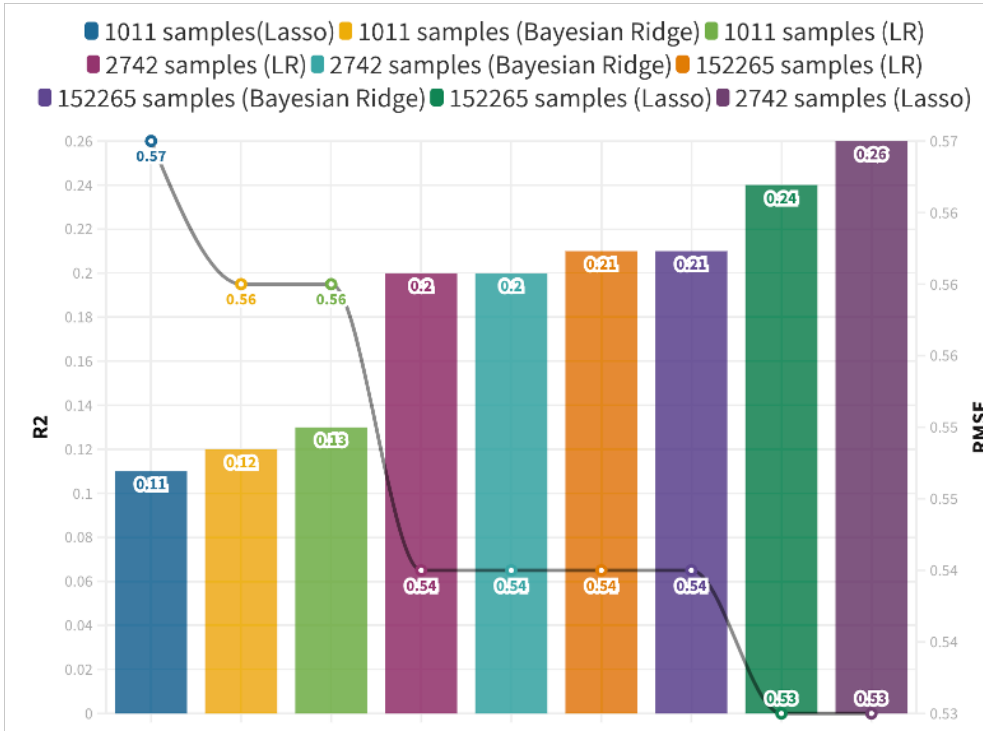
- **Data**: Collected from an entire year of operation, from the furnace, upstream and downstream processes
- 96 process variables (82 continuous & 14 categorical)
- Best performance achieved using 20 variables only
- **Observation**: Benefits of new features are outweighed by the curse of dimensionality

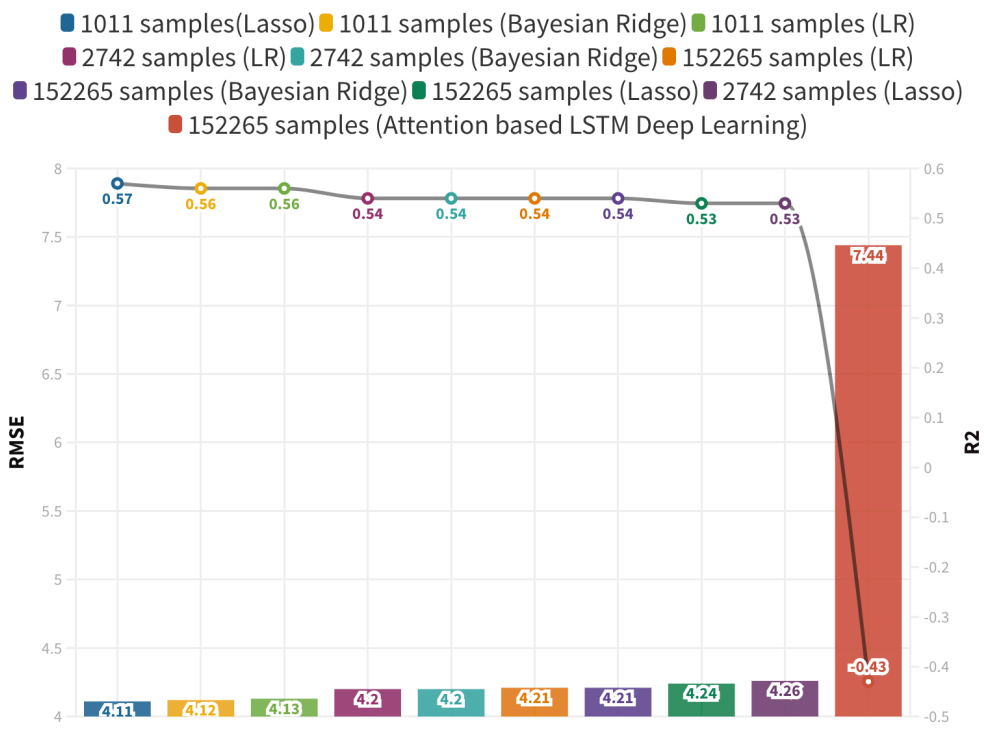




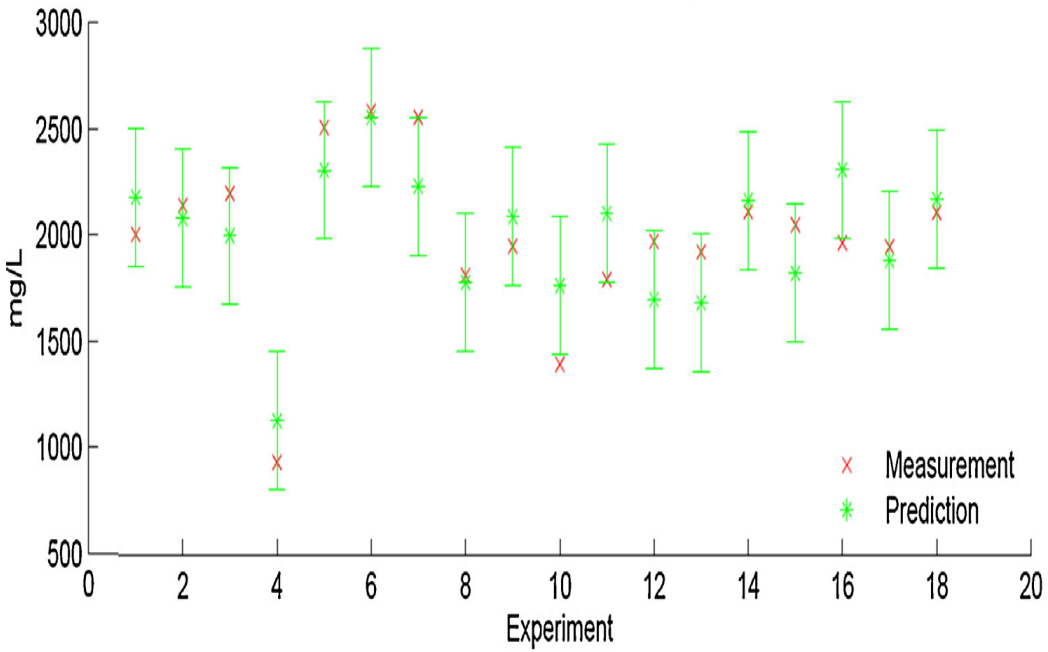
Example #2 Fluid Catalytic Cracking in a Refinery

- **Data**: Three years of operational data from an FCC in a refinery, 10 min sampling time, 10 process variables, 1 quality variable, 152265 samples
- Removing data with low variance resulted in only 2742 samples (1.8% of the original dataset), and using change point detection to identify informative data resulted in only 1011 samples (0.66% of the dataset)





Example #3 Biopharmaceutical Monoclonal Antibody Manufacturing CQA Prediction at Biogen



		Variance of the Prediction Using ...		
Unit Operation	**Output Variable**	**PCR**	**PLS**	**ENwMC**
Bioreactor	G0 product quality	0.146 (4)	0.148 (1)	**0.087 (3)**
	Final titer	0.281 (4)	0.287 (2)	**0.178 (3)**
	DNA	0.209 (4)	**0.201 (1)**	0.223 (4)
	HCP	0.258 (6)	0.210 (2)	**0.150 (6)**
Protein A Column	DNA	0.151 (4)	0.143 (1)	**0.095 (4)**
	HCP	0.268 (6)	0.202 (3)	**0.080 (4)**
	Total impurity	0.286 (4)	0.256 (1)	**0.164 (5)**
	HMW	0.117 (6)	0.092 (1)	**0.045 (4)**
Cation Exchange Column	HCP	0.226 (9)	0.132 (2)	**0.083 (4)**
	Total impurity	0.323 (5)	0.348 (2)	**0.226 (2)**
	HMW	0.058 (3)	0.063 (1)	**0.010 (3)**
Anion Exchange Column	HCP	0.189 (7)	0.140 (2)	**0.048 (3)**
	Total impurity	0.228 (4)	0.227 (3)	**0.115 (4)**
	HMW	0.067 (9)	0.050 (4)	**0.007 (2)**

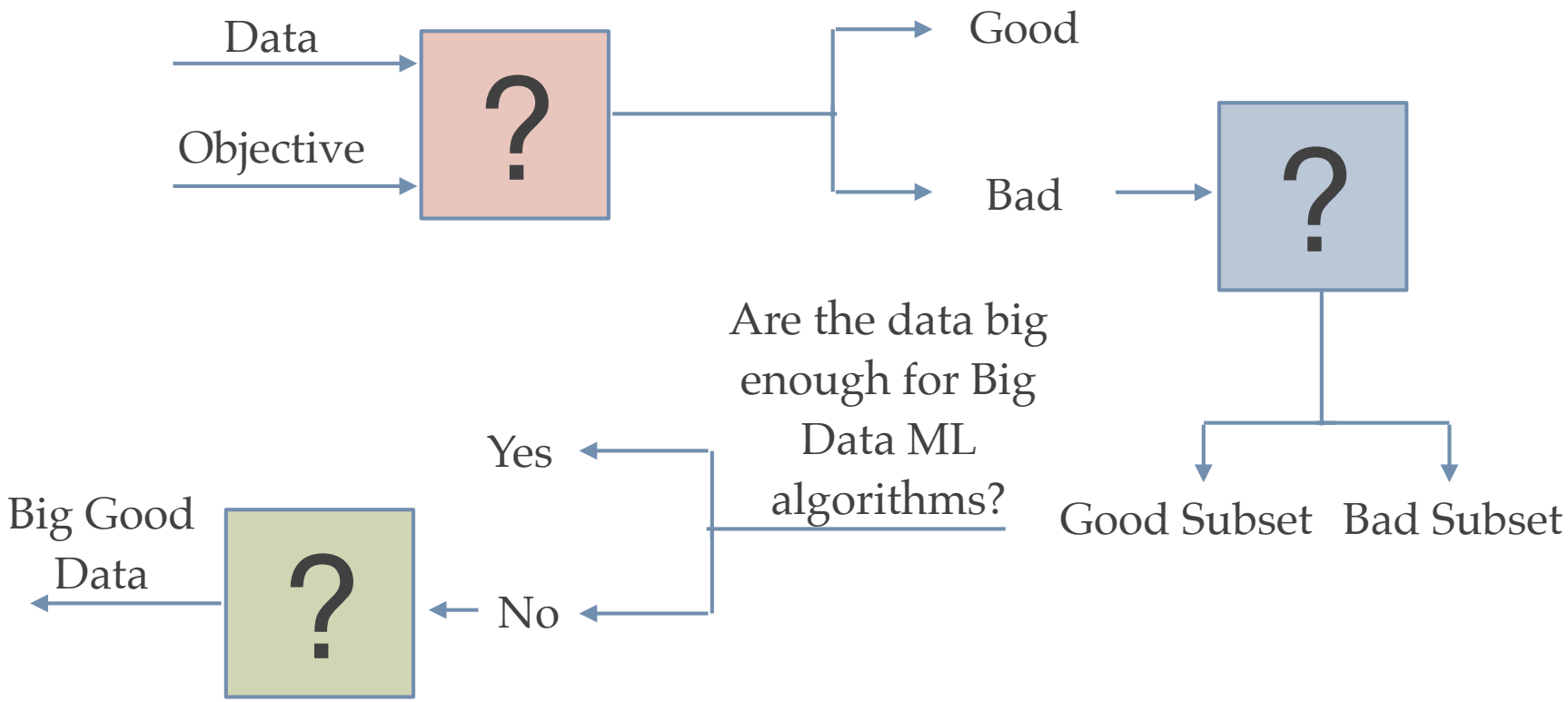
K Severson et al., in *Multivariate Analysis in the Pharmaceutical Industry*, Elsevier, Chapter 12, 2018

Using only 3 to 6 variables gave much better predictions than using all variables

Open Problem - What are Good Data?

Good Data are data that are useful for a specific purpose

- How do we formulate a mathematical problem to determine if a given set of data are good for an objective?



Questions to Ponder

For a given objective, how do we characterize and quantify the usefulness of Big Data?

What are Good Data, and how do we generate Big Data that are also Good Data?

How can we fully exploit the power of modern machine learning tools such as deep and deep reinforcement learning in the process industries?