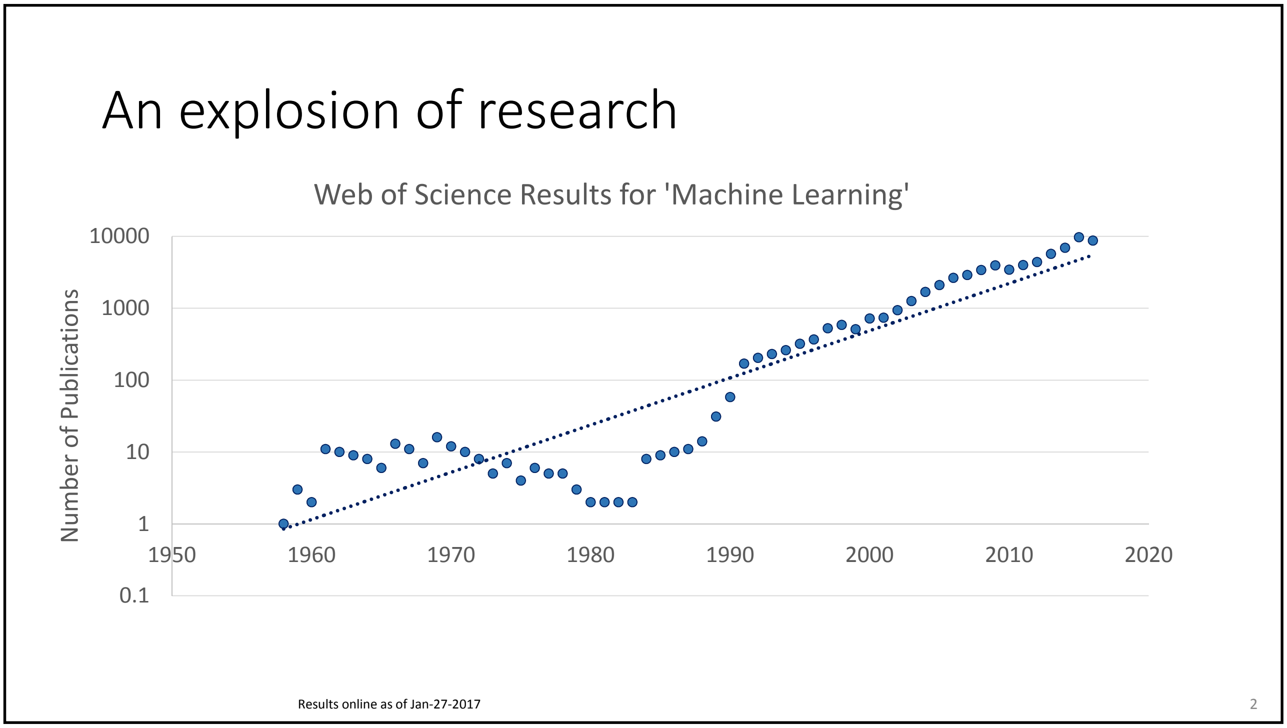
Workshop on Process Data Analytics and Machine Learning

Richard D. Braatz, S. Joe Qin, and Leo H. Chiang

Photo courtesy of Indeed Hiring Lab

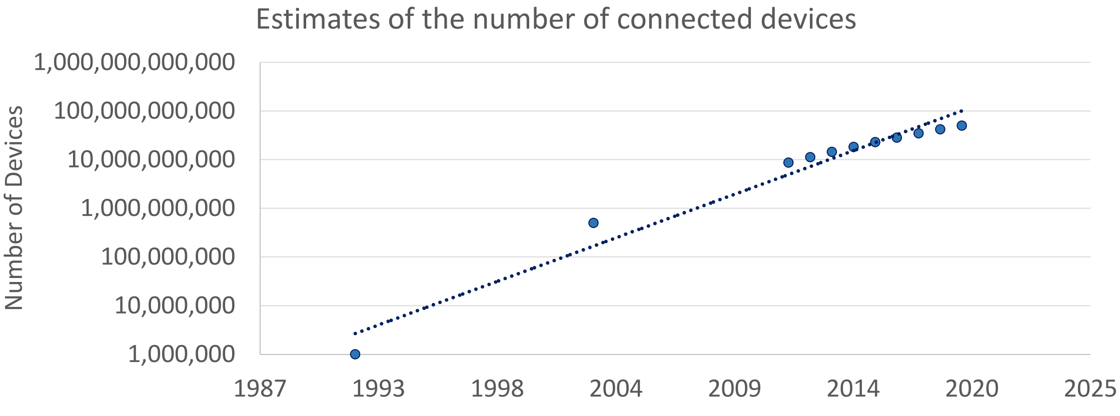


An explosion of research

Web of Science Results for 'Machine Learning'

Results online as of Jan-27-2017

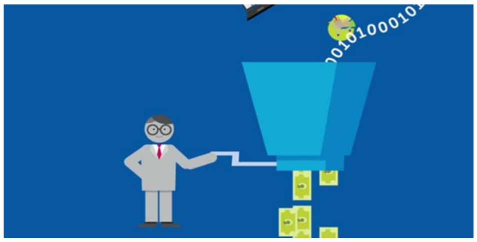
And an explosion of data

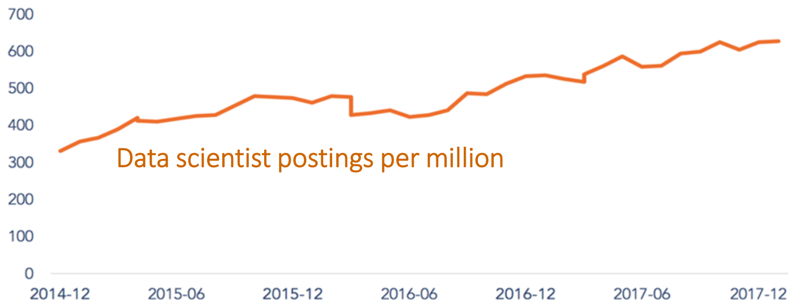


Data based on Cisco estimates

Demand for Data Analytics Expertise

- Companies are using data to streamline operations, improve reliability, optimize processes
- Enabled by huge increases in data and reductions in computer costs
- Resulting in high demand for people with expertise





Figures courtesy of Opera Solutions and Indeed Hiring Lab

Some Workshop Goals

- Tools for solving problems based on data
- Ways to choose algorithms for a specific problem
- Numerous application examples
- Tips and tricks of the trade

Workshop Outline

1 - Introduction

1.1 Examples of typical data analytics applications

1.2 Unsupervised, supervised, and partially supervised learning

1.3 Least squares including sparse methods

1.4 Feature engineering

1.5 Kernel methods for nonlinear analytics

1.6 Neural networks and deep learning

Workshop Outline

2 - Latent Variable Methods and Application Case Studies

2.1 Principal component analysis

2.2 Partial least squares

2.3 Canonical correlation analysis

2.4 Dynamic principal component analysis and canonical variate analysis

2.5 Linear discriminant analysis and support vector machines

2.6 Process monitoring, diagnosis, and troubleshooting

Workshop Outline

3 - Industrial Experience and Tips, Interactive Discussions

3.1 Visualization

3.2 Outlier detection and data preprocessing

3.3 Method selection

3.4 How good is good enough? Industrial tips and tricks of the trade

3.5 Industrial case studies by guest speaker Dr. Ivan Castillo (Dow)

1. Introduction to Process Data Analytics and Machine Learning

Richard D. Braatz

Part 1 Outline

1 - Introduction

1.1 Examples of typical data analytics applications

1.2 Unsupervised, supervised, and partially supervised learning

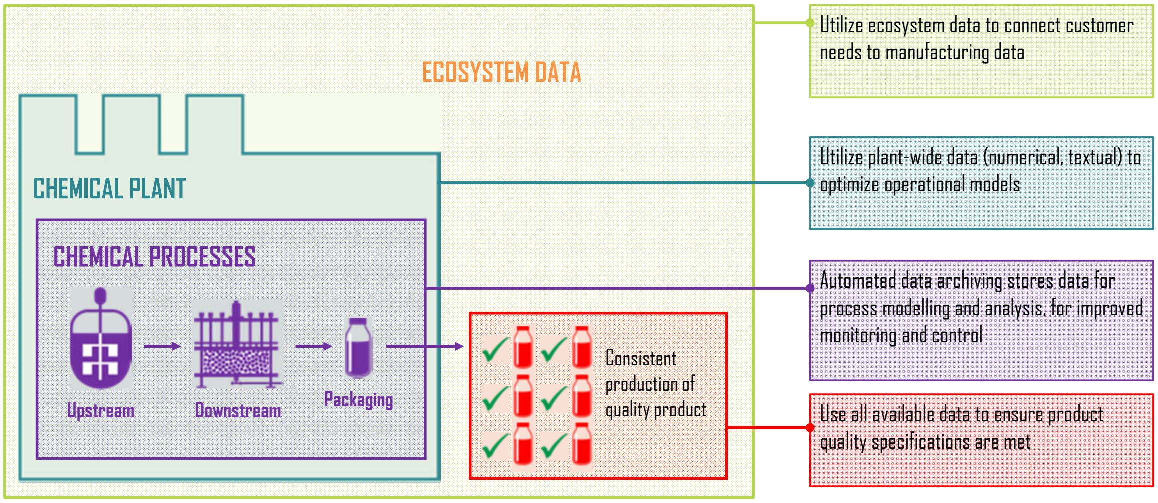
1.3 Least squares including sparse methods

1.4 Feature engineering

1.5 Kernel methods for nonlinear analytics

1.6 Neural networks and deep learning

Examples of Typical Data Analytics Applications



ECOSYSTEM DATA

CHEMICAL PLANT

CHEMICAL PROCESSES

Consistent production of quality product

Utilize ecosystem data to connect customer needs to manufacturing data

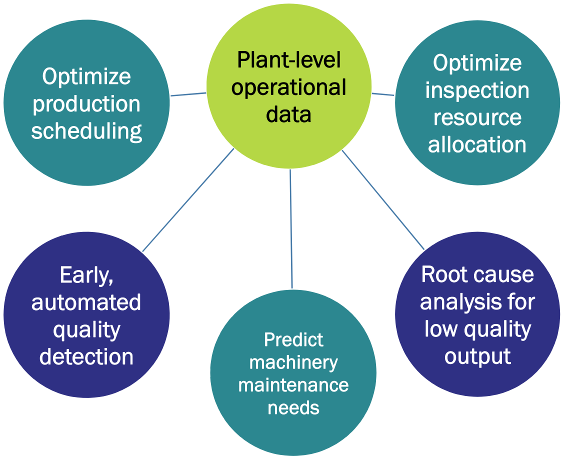
Utilize plant-wide data (numerical, textual) to optimize operational models

Automated data archiving stores data for process modelling and analysis, for improved monitoring and control

Use all available data to ensure product quality specifications are met

Objectives: Diagnostics/Prognostics, Continuous Improvement, and Optimal Decision Making

Examples of Typical Data Analytics Applications



Plant-level operational data

Optimize production scheduling

Optimize inspection resource allocation

Early, automated quality detection

Predict machinery maintenance needs

Root cause analysis for low quality output

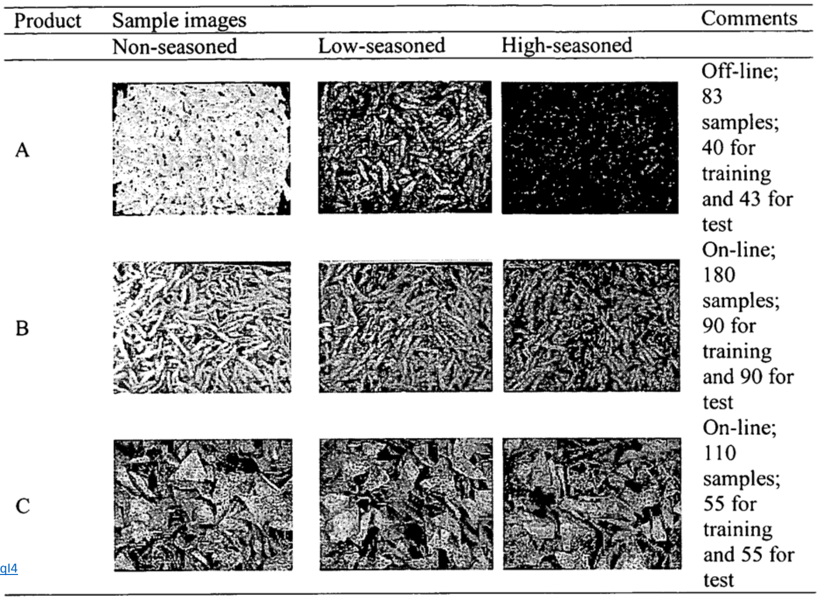
- Many companies have built the infrastructure to bring the data into one database for easy access
- Correlate plant data to off-line product quality specs
- Connect product quality data to the supply chain
- Troubleshoot problems in plant operation, e.g., causes of off-spec product (e.g. raw materials, operator error)
- Propose process or control design changes to reduce operational problems
- Optimize operations, e.g., selection of raw materials or mixtures from multiple suppliers
- Design predictive maintenance schedules
- Facilitate continuous improvement practices

Hasn’t data analytics been applied to chemical processes for decades?

- Control charts, principal component analysis (PCA), and partial least squares (PLS) have been important tools in industry for decades
- New datasets present different attributes and challenges that are not addressed using classical techniques

Examples of Some Modern Datasets

- Example: images of snack foods (US Patent 7068817 B2)
- Real-time imaging used to control the amount of seasoning

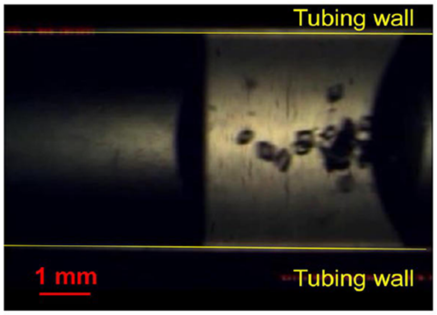


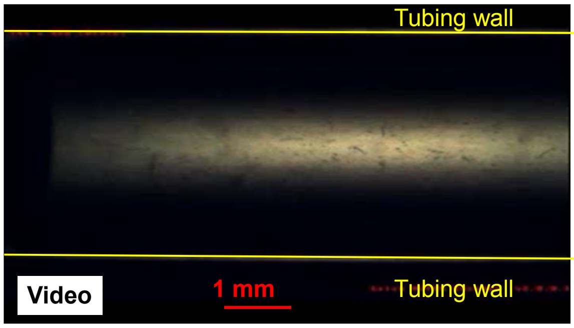
Product	Sample images			Comments
	Non-seasoned	Low-seasoned	High-seasoned	
A				Off-line; 83 samples; 40 for training and 43 for test
B				On-line; 180 samples; 90 for training and 90 for test
C				On-line; 110 samples; 55 for training and 55 for test

Images from https://www.youtube.com/watch?v=kvZfE3UeqI4
http://www.google.tl/patents/US7068817

Examples of Some Modern Datasets

- With cost < $100 for a color CCD camera, imaging is more widely used
- Example: grey-scale video with dimensions x, y, t
  (color adds an rgb dimension)





M. Jiang, Z. Zhu, E. Jimenez, C.D. Papageorgiou, J. Waetzig, A. Hardy, M. Langston, and R.D. Braatz. Continuous-flow tubular crystallization in slugs spontaneously induced by hydrodynamics. Crystal Growth & Design (2014) 14:851-860.

Examples of Some Modern Datasets

- Modern data include 1-way arrays (spectra),
  2-way arrays (2D particle size distributions), 3-way arrays (hyperspectral images), and 4-way arrays (color videos)
- Algorithms are available for handling these higher order data structures
- Unispectral and BGN Technologies Ltd. claim large cost reductions will occur by exploiting smart phone hardware

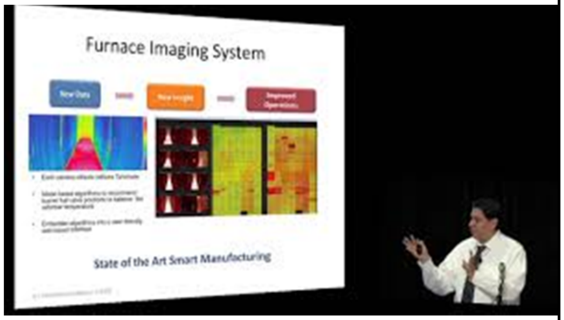
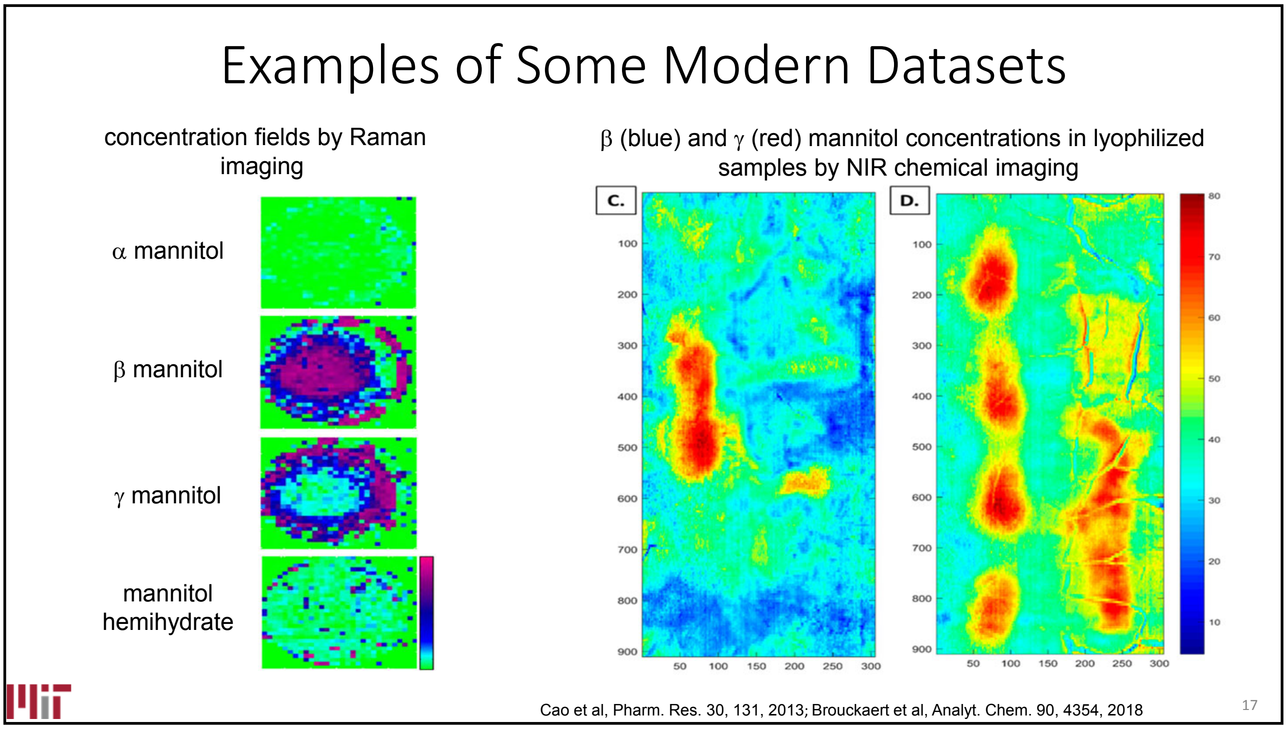


Image from https://www.youtube.com/watch?v=kvZfE3UeqI4



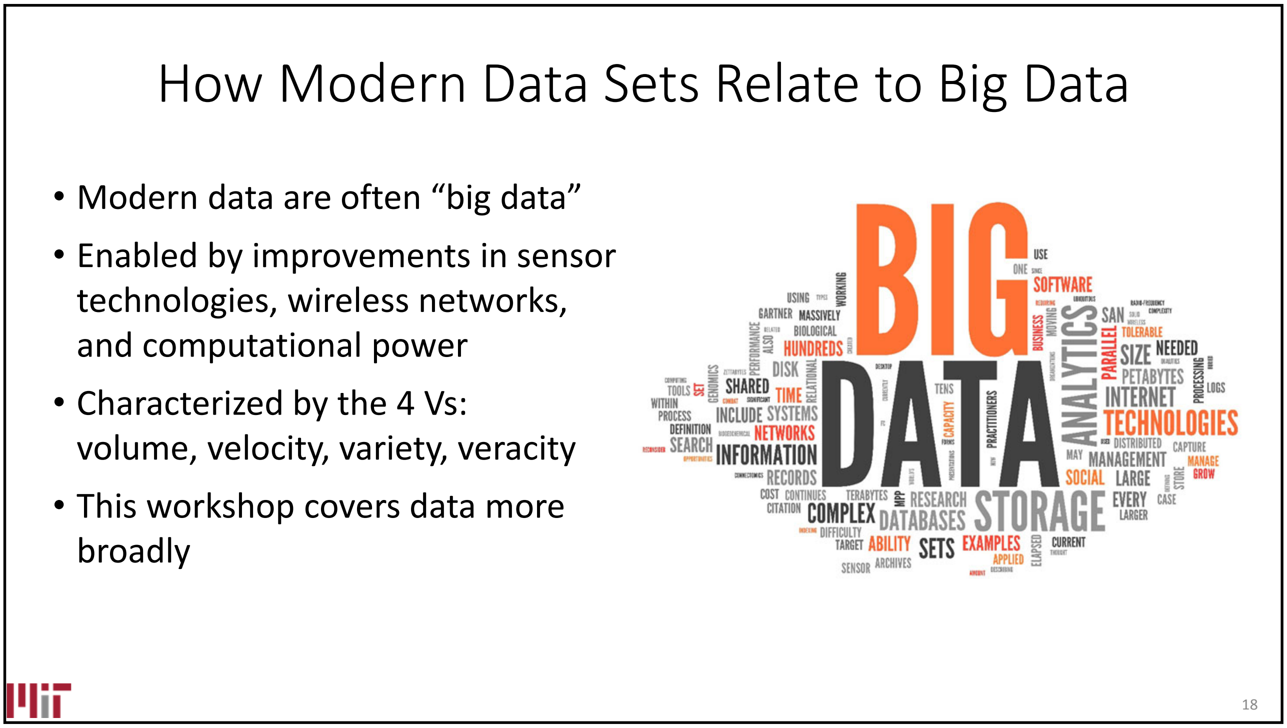
Examples of Some Modern Datasets

concentration fields by Raman imaging

β (blue) and γ (red) mannitol concentrations in lyophilized samples by NIR chemical imaging

Cao et al, Pharm. Res. 30, 131, 2013; Brouckaert et al, Analyt. Chem. 90, 4354, 2018

17

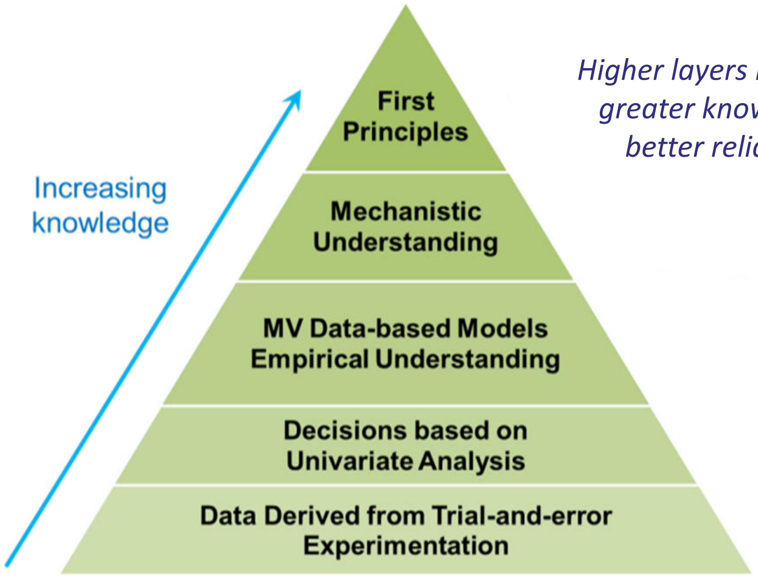


How Modern Data Sets Relate to Big Data

- Modern data are often “big data”
- Enabled by improvements in sensor technologies, wireless networks, and computational power
- Characterized by the 4 Vs: volume, velocity, variety, veracity
- This workshop covers data more broadly

18

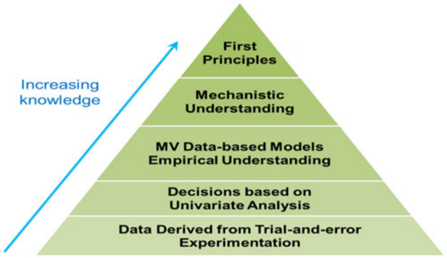
Which model type to use for a given application?

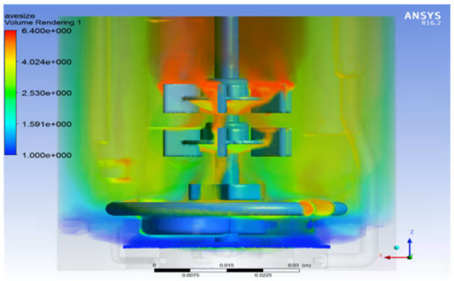

First Principles
Mechanistic Understanding
MV Data-based Models Empirical Understanding
Decisions based on Univariate Analysis
Data Derived from Trial-and-error Experimentation
Increasing knowledge

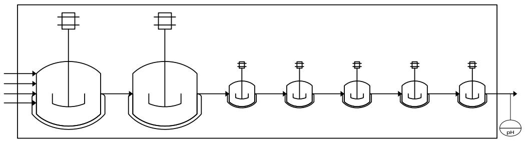
Higher layers in the pyramid correspond to greater knowledge, more informed and better reliability of decision making

The Best Model Complexity Depends on the Purpose

- Construct from first principles where possible
- Highest complexity models used for the intensification of process designs
- Lower complexity models for process control design and quality monitoring


First Principles
Mechanistic Understanding
MV Data-based Models Empirical Understanding
Decisions based on Univariate Analysis
Data Derived from Trial-and-error Experimentation
Increasing knowledge





Basic Principles of GMP, World Health Organization, May 2008
A.E. Lu et al., *IEEE Conf. on Control Appl.*, 1505-1515, 2015 A.E. Lu et al., *Proc. Am. Contr. Conf.*, 1741-1746, 2016

Questions?

Part 1 Outline

1 - Introduction

1.1 Examples of typical data analytics applications

1.2 Unsupervised, supervised, and partially supervised learning

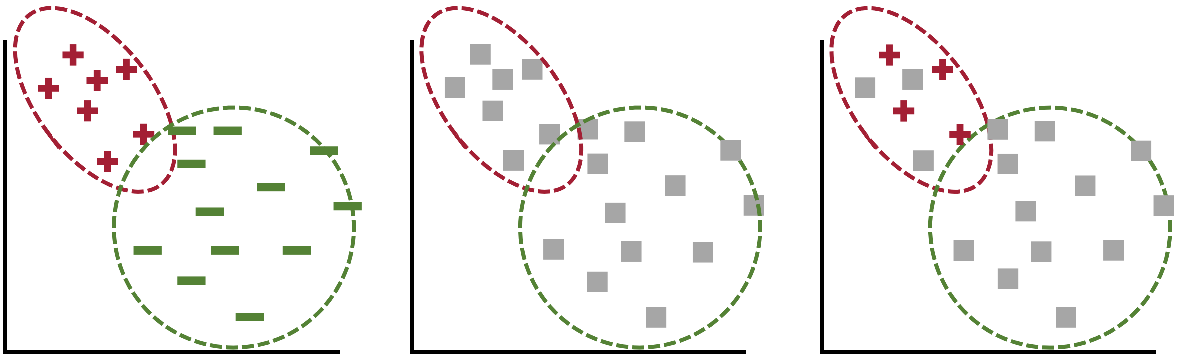
1.3 Least squares including sparse methods

1.4 Feature engineering

1.5 Kernel methods for nonlinear analytics

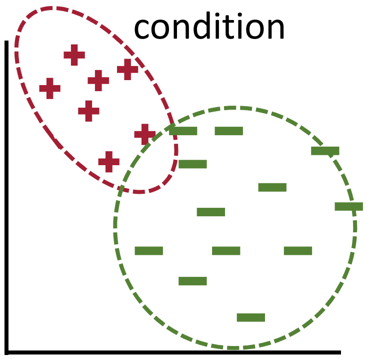
1.6 Neural networks and deep learning

Unsupervised, Supervised, and Partially Supervised Learning



Supervised Unsupervised **Partially supervised**

Unsupervised, Supervised, and Partially Supervised Learning



Known fault condition

Normal operating conditions

Supervised

Example Method: Linear Discriminant Analysis (LDA)

- Linear discriminant analysis is a type of generative classifier for binary problems
- Assumes that the data has class-specific means and a shared covariance matrix

$$Y \sim Binomial(\pi) \qquad \mathbf{X}|Y = c \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$$

Predict class + when:

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b > 0$$

where

$$\mathbf{w} = \Sigma^{-1}(\mu_+ - \mu_-)$$

$$b = \frac{1}{2}(\mu_+ - \mu_-)^{\mathrm{T}}\Sigma^{-1}(\mu_+ - \mu_-)$$

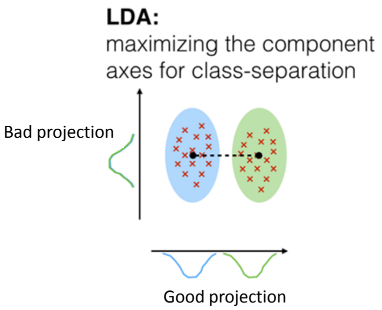
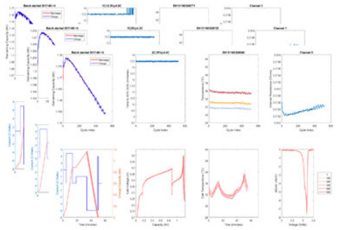


Image credit: http://sebastianraschka.com/Articles/2014_python_lda.html

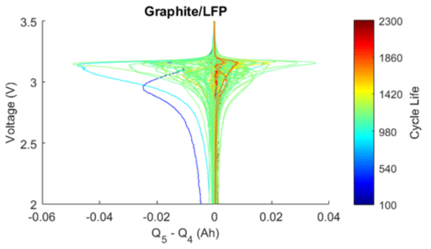
Example Application: Prediction and Classification of Battery Lifetime from High-throughput Cycling Data

- Predicted battery cycle life from data collected during the first 110 cycles
- Classified batteries into long and short lifetime based on data collected during only the first 5 cycles, before capacity degradation has occurred

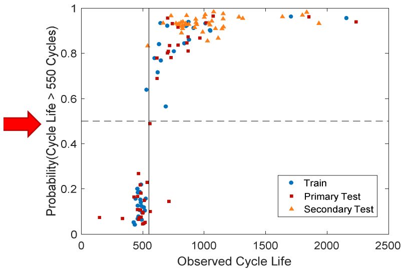
Experimental Cycling Data



Feature Engineering & Elastic Net

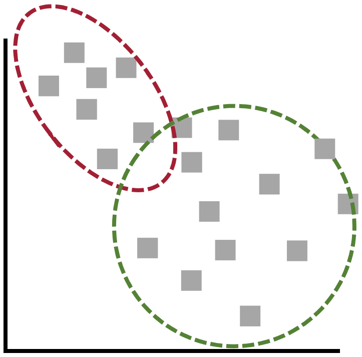


Classification Modeling



Severson et al., Data-driven prediction of battery cycle life before capacity degradation, *Nature Energy*, 4, 383-391, 2019

Unsupervised, Supervised, and Partially Supervised Learning



Unsupervised

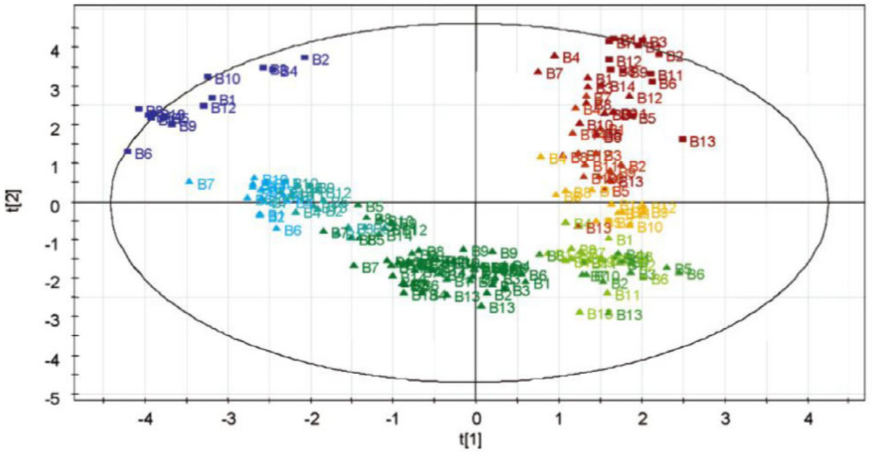
Years of historical data

Some abnormal operations must have occurred, but uncharacterized

6

Example Application: Scatter Plot of Bioreactor Data Projected onto Two Principal Components

Data are within 95% confidence ellipse, but clearly show different behavior



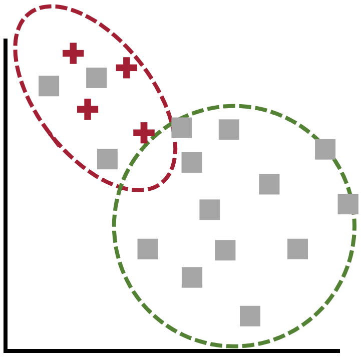
Kirdar et al., *Biotechnology Progress*, 23(1), 61-67, 2007

Unsupervised, Supervised, and Partially Supervised Learning

Years of historical data

Some past data are associated with faults (+), e.g., stuck valves

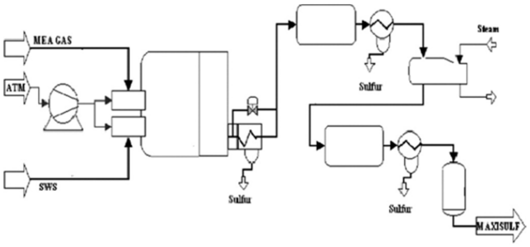
Probably some other abnormal operations occurred, but are uncharacterized

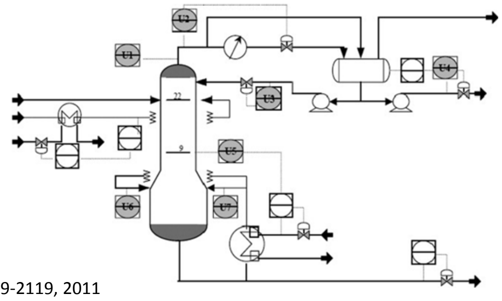


Partially supervised

Example Application: Scatter Plot of Bioreactor Data Projected onto Two Principal Components

- Can have more input process data than output quality data
- Applied to a sulfur recover unit and a debutanizer column
- Soft sensors had similar prediction error when 50% of the output values were unknown





*Ge and Song, *AIChE Journal*, 57(8), 2109-2119, 2011

Part 1 Outline

1 - Introduction

1.1 Examples of typical data analytics applications

1.2 Unsupervised, supervised, and partially supervised learning

1.3 Least squares including sparse methods

1.4 Feature engineering

1.5 Kernel methods for nonlinear analytics

1.6 Neural networks and deep learning

Linear Regression

Models that are linear functions of the parameters

$$y = \sum_{i=1}^{n} \theta_i a_i + e = a\theta + e = \tilde{y} + e$$

calibration parameters

n sensor signals

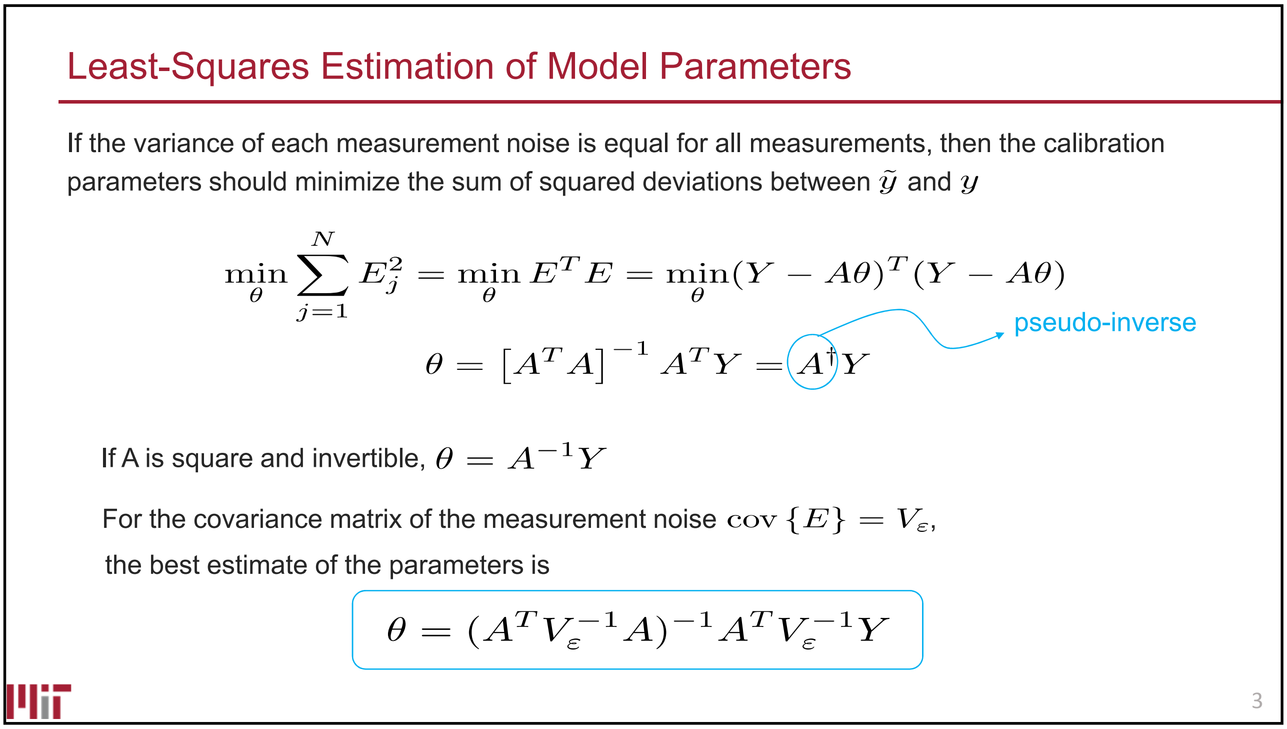
measurement noise

(zero mean, independent)

a: row vector $\qquad$ θ: column vector

N samples ⟶ stack the variables into vectors and matrices:

$$Y = \begin{bmatrix} y^1 \\ \vdots \\ y^N \end{bmatrix} = \begin{bmatrix} a^1 \\ \vdots \\ a^N \end{bmatrix} \theta + \begin{bmatrix} e^1 \\ \vdots \\ e^N \end{bmatrix} = A\theta + E$$



Least-Squares Estimation of Model Parameters

If the variance of each measurement noise is equal for all measurements, then the calibration parameters should minimize the sum of squared deviations between $\tilde{y}$ and y

$$\min_{\theta} \sum_{j=1}^{N} E_j^2 = \min_{\theta} E^T E = \min_{\theta} (Y - A\theta)^T (Y - A\theta)$$

$$\theta = \left[A^T A\right]^{-1} A^T Y = A^{\dagger} Y$$

pseudo-inverse

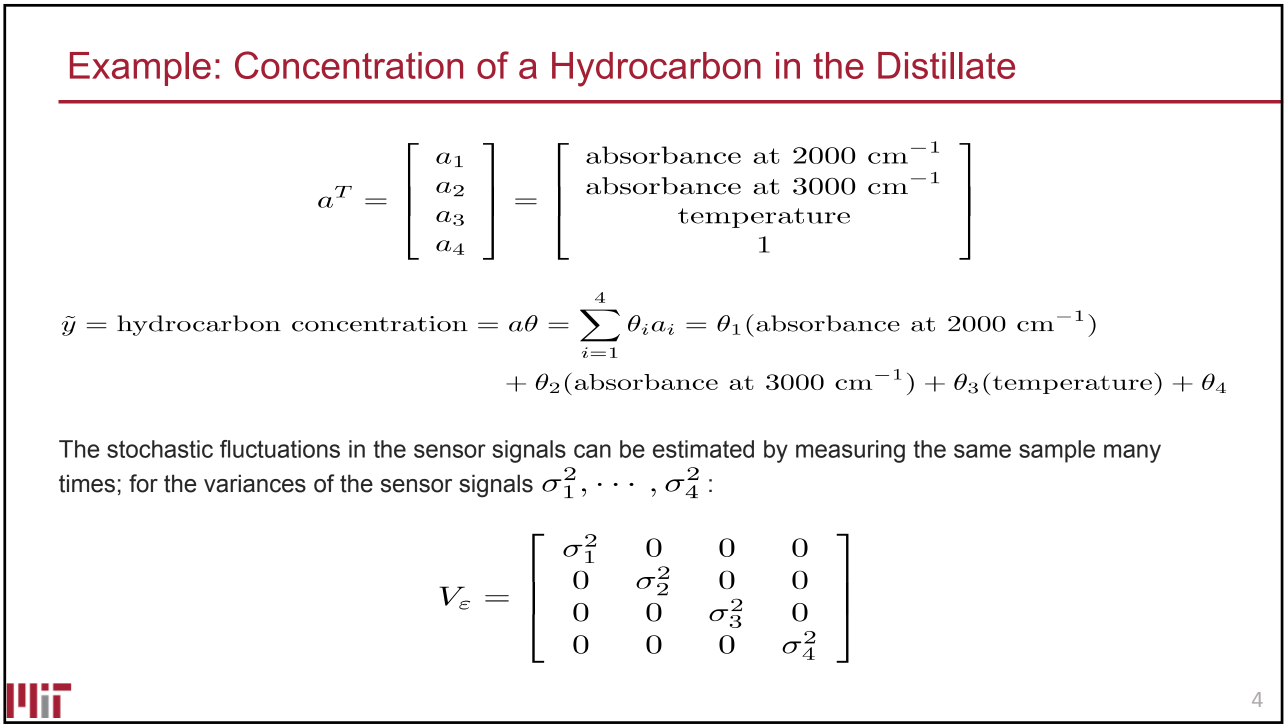
If A is square and invertible, $\theta = A^{-1} Y$

For the covariance matrix of the measurement noise $\operatorname{cov}\{E\} = V_{\varepsilon}$,

the best estimate of the parameters is

$$\boxed{\theta = (A^T V_{\varepsilon}^{-1} A)^{-1} A^T V_{\varepsilon}^{-1} Y}$$

3



Example: Concentration of a Hydrocarbon in the Distillate

$$a^T = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} \text{absorbance at 2000 cm}^{-1} \\ \text{absorbance at 3000 cm}^{-1} \\ \text{temperature} \\ 1 \end{bmatrix}$$

$$\tilde{y} = \text{hydrocarbon concentration} = a\theta = \sum_{i=1}^{4} \theta_i a_i = \theta_1(\text{absorbance at 2000 cm}^{-1}) + \theta_2(\text{absorbance at 3000 cm}^{-1}) + \theta_3(\text{temperature}) + \theta_4$$

The stochastic fluctuations in the sensor signals can be estimated by measuring the same sample many times; for the variances of the sensor signals $\sigma_1^2, \cdots, \sigma_4^2$:

$$V_{\varepsilon} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

4

Example: Concentration of a Hydrocarbon in the Distillate

Given the known hydrocarbon concentrations of the 10 samples

$$Y = \begin{bmatrix} y^1 \\ \vdots \\ y^{10} \end{bmatrix} \qquad A = \begin{bmatrix} a_1^1 & a_2^1 & a_3^1 & a_4^1 \\ a_1^2 & a_2^2 & a_3^2 & a_4^2 \\ \vdots & \vdots & \vdots & \vdots \\ a_1^{10} & a_2^{10} & a_3^{10} & a_4^{10} \end{bmatrix}$$

known hydrocarbon concentrations

sensor signals

$$\Rightarrow \theta = (A^T V_\varepsilon^{-1} A)^{-1} A^T V_\varepsilon^{-1} Y$$

(weighted least squares)

This method produces good calibration models when the number of calibration parameters is low, but performs poorly when the number of calibration parameters is high, due to high correlations between the sensor signals

Ridge Regression

Simplest method for producing accurate models for highly correlated data

$$\min_\theta \sum_{j=1}^{N} E_j^2 + \alpha \sum_{i=1}^{n} \theta_i^2 = \min_\theta E^T E + \alpha \theta^T \theta$$

$$= \boxed{\min_\theta (Y - A\theta)^T (Y - A\theta) + \underline{\alpha} \theta^T \theta}$$

a positive small number

$$\Rightarrow \theta = \left[A^T A + \alpha I\right]^{-1} A^T Y$$

More sophisticated methods covered in Part 2

Sparse Models

Models with higher predictive capability are often obtained by assuming that the model is sparse, that is, the most elements of the vector θ are equal to zero.

Lasso (least absolute shrinkage and selection operator):

$$\min_{\theta} \sum_{j=1}^{N} E_j^2 + \alpha \sum_{i=1}^{n} |\theta_i| = \min_{\theta} E^T E + \alpha \sum_{i=1}^{n} |\theta_i|$$

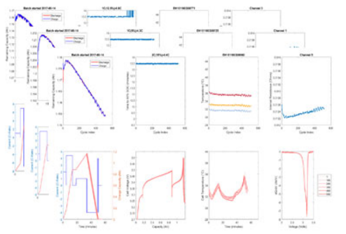
$$= \boxed{\min_{\theta} (Y - A\theta)^T (Y - A\theta) + \alpha \sum_{i=1}^{n} |\theta_i}$$

The selection of nonzero elements for lasso can be sensitive to small perturbations in the data
Elastic net combines ridge regression and lasso to generate sparse models with higher robustness

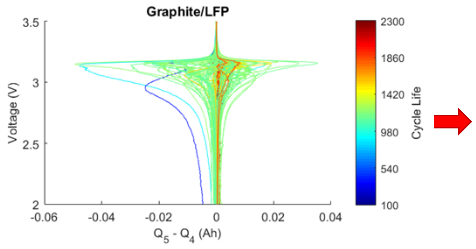
Example Application: Prediction and Classification of Battery Lifetime from High-throughput Cycling Data

- Predicted battery cycle life from data collected during the first 110 cycles
- Classified batteries into long and short lifetime based on data collected during only the first 5 cycles, before capacity degradation has occurred
- Elastic net identifies one feature that gives ~90% of the prediction accuracy

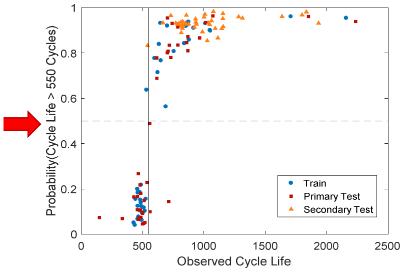
Experimental Cycling Data



Feature Engineering & **Elastic Net**



Classification Modeling

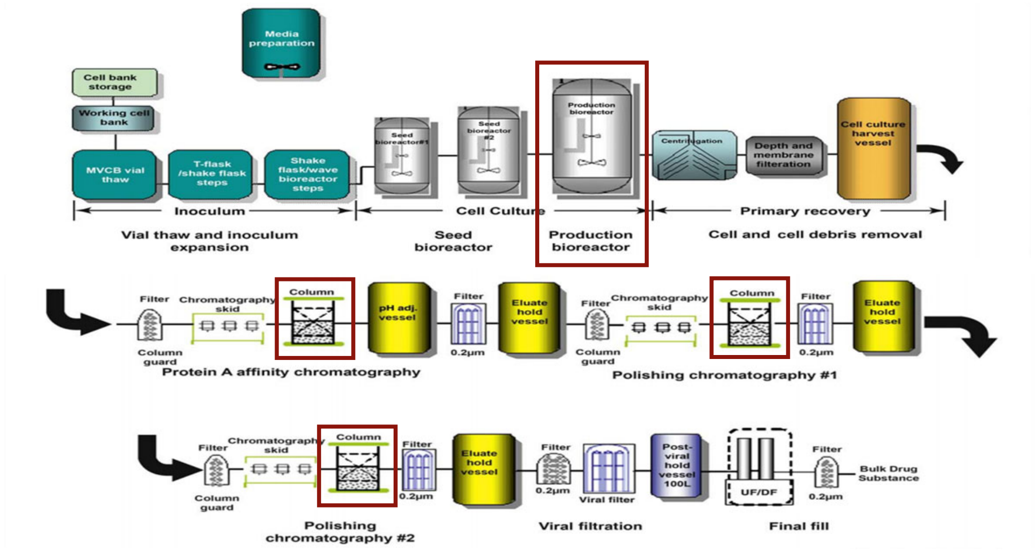


Severson et al., Data-driven prediction of battery cycle life before capacity degradation, *Nature Energy*, 4, 383-391, 2019

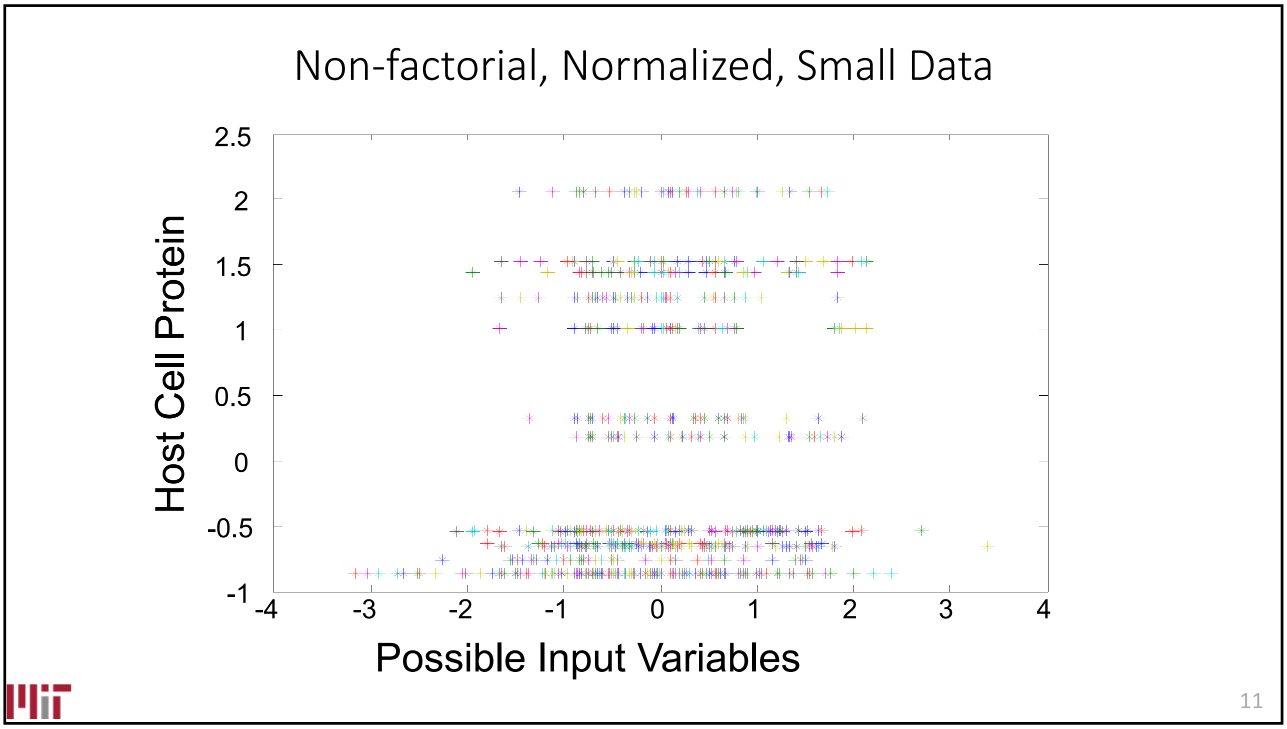
Example Application: mAb manufacturing modeling

- Application: model critical quality attributes in a monoclonal antibody manufacturing (mAb) process at Biogen
- Modeling goal: understand the parameters that affect production
- Elastic net outperformed the methods commonly applied in biopharma

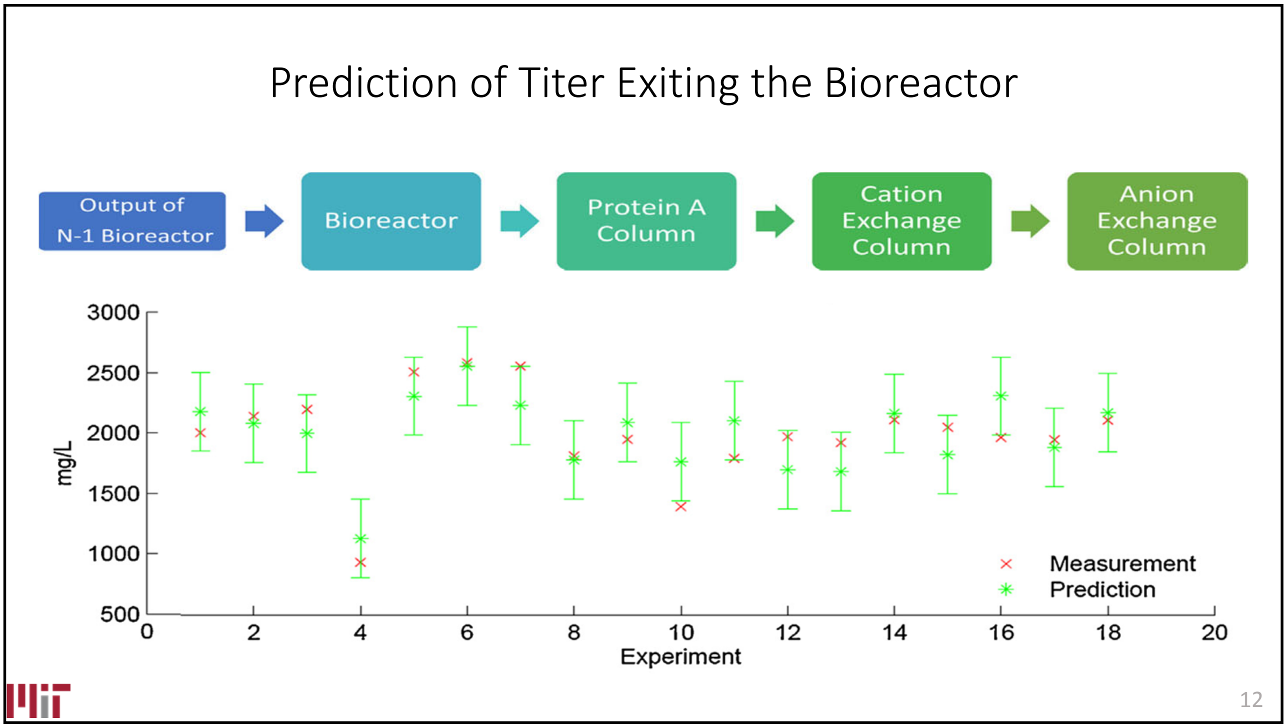
Production-Scale Data for a mAb



Shukla and Thommes, Recent advances in large-scale production of monoclonal antibodies and related proteins, *Trends in Biotechnology*, 2010.



Non-factorial, Normalized, Small Data



Prediction of Titer Exiting the Bioreactor

Output of N-1 Bioreactor → Bioreactor → Protein A Column → Cation Exchange Column → Anion Exchange Column

Prediction Error Using All Upstream Inputs

		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)**

Side Comment on Quality of Fit

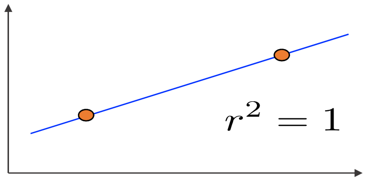
$$\mathrm{SSE} = \sum_i (y_i - \tilde{y}_i)^2$$

Use validation set

The coefficient of determination (R-squared)

$$\boxed{r^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}} \qquad \mathrm{SST} = \sum_i (y_i - \overline{y})^2 \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

An R-squared value of 1 is considered good, but can be a misleading indicator of the predictive value of the model



Nonlinear regression

- Refers to models that are nonlinear functions of the parameters
- Linear and nonlinear regression are often referred to as parameter estimation in chemical engineering and related branches of engineering
- For example, kinetic parameters are identified from experimental data
- The next slides discuss how to estimate parameters from dynamic data and to quantify the accuracy of the parameters

Parameter Estimation as an Optimization

# of measured variables

# of sampling instances

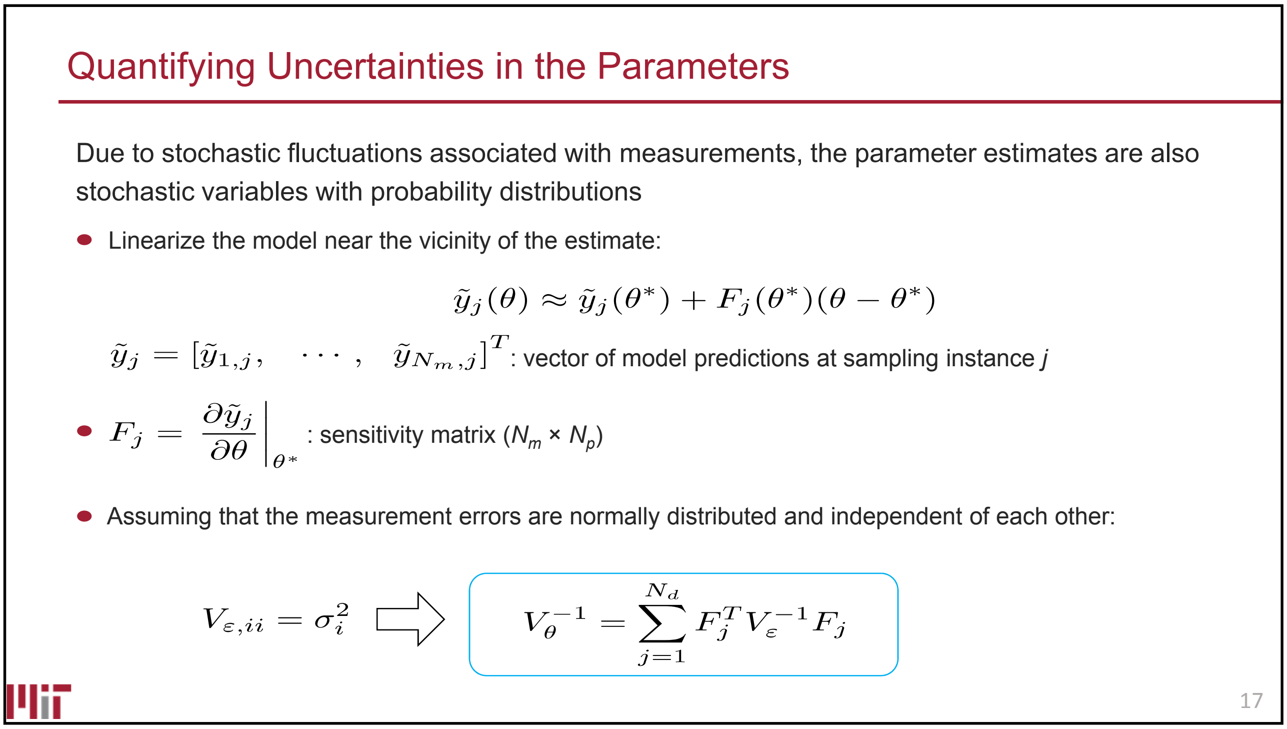
$$\min_{\theta} \sum_{i=1}^{N_m} \sum_{j=1}^{N_d} w_{ij} \left(y_{ij} - \tilde{y}_{ij}(\theta)\right)^2 \qquad \text{(C1)}$$

For the best estimates, w_{ij} should be set as $(\sigma_i^2)^{-1}$

The optimization (C1) is more general than used for sensor calibration, since (C1) can be applied to experimental data from dynamic processes and the model in (C1) can be nonlinear in parameters

$$Y = \begin{bmatrix} \vdots \\ y_{ij} \\ \vdots \end{bmatrix}, \quad \tilde{Y} = \begin{bmatrix} \vdots \\ \tilde{y}_{ij} \\ \vdots \end{bmatrix} = A\theta \Longrightarrow \boxed{\begin{aligned} &\min_{\theta} \left(Y - \tilde{Y}(\theta)\right)^T W \left(Y - \tilde{Y}(\theta)\right) \\ &= \min_{\theta} (Y - A\theta)^T W (Y - A\theta) \end{aligned}}$$

$$W(i,j) = w_{ij}$$



Quantifying Uncertainties in the Parameters

Due to stochastic fluctuations associated with measurements, the parameter estimates are also stochastic variables with probability distributions

- Linearize the model near the vicinity of the estimate:

$$\tilde{y}_j(\theta) \approx \tilde{y}_j(\theta^*) + F_j(\theta^*)(\theta - \theta^*)$$

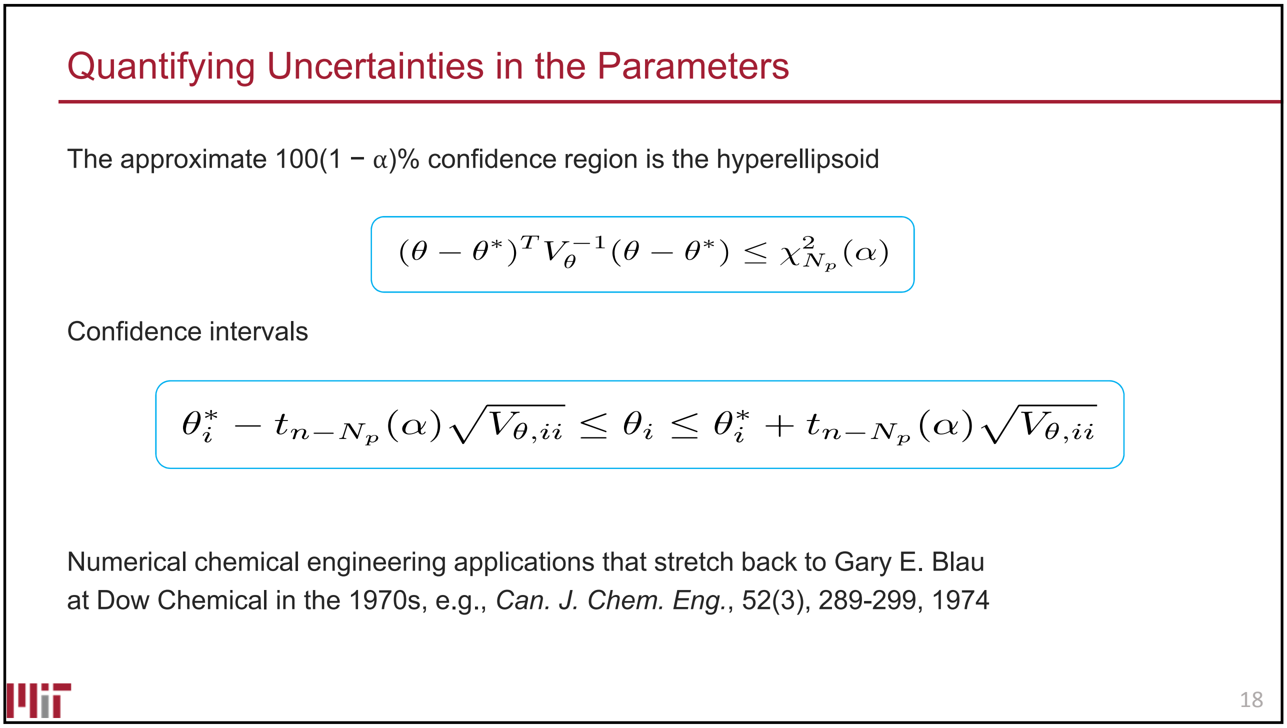
$\tilde{y}_j = [\tilde{y}_{1,j}, \quad \cdots, \quad \tilde{y}_{N_m,j}]^T$: vector of model predictions at sampling instance *j*

- $F_j = \left.\dfrac{\partial \tilde{y}_j}{\partial \theta}\right|_{\theta^*}$: sensitivity matrix (N_m × N_p)

- Assuming that the measurement errors are normally distributed and independent of each other:

$$V_{\varepsilon,ii} = \sigma_i^2 \Rightarrow \boxed{V_\theta^{-1} = \sum_{j=1}^{N_d} F_j^T V_\varepsilon^{-1} F_j}$$

17



Quantifying Uncertainties in the Parameters

The approximate 100(1 − α)% confidence region is the hyperellipsoid

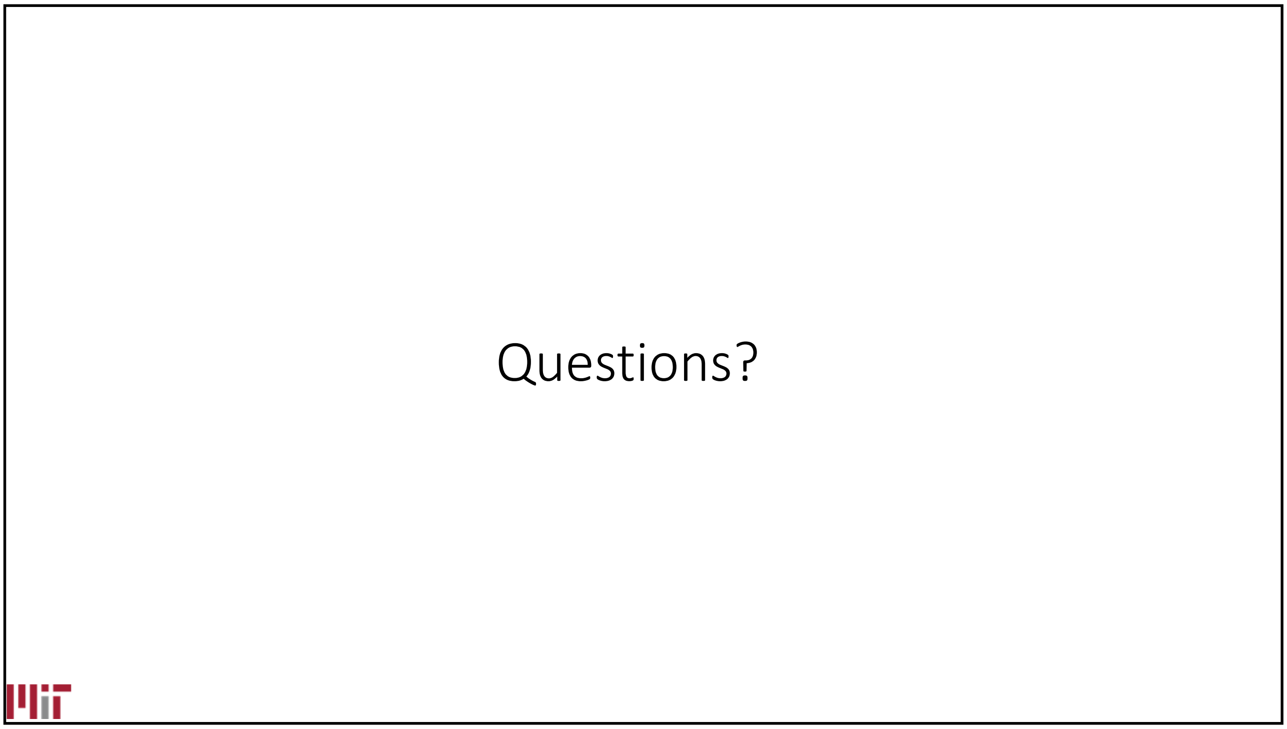
$$\boxed{(\theta - \theta^*)^T V_\theta^{-1} (\theta - \theta^*) \leq \chi^2_{N_p}(\alpha)}$$

Confidence intervals

$$\boxed{\theta_i^* - t_{n-N_p}(\alpha)\sqrt{V_{\theta,ii}} \leq \theta_i \leq \theta_i^* + t_{n-N_p}(\alpha)\sqrt{V_{\theta,ii}}}$$

Numerical chemical engineering applications that stretch back to Gary E. Blau at Dow Chemical in the 1970s, e.g., *Can. J. Chem. Eng.*, 52(3), 289-299, 1974

18



Questions?

Part 1 Outline

1 - Introduction

1.1 Examples of typical data analytics applications

1.2 Unsupervised, supervised, and partially supervised learning

1.3 Least squares including sparse methods

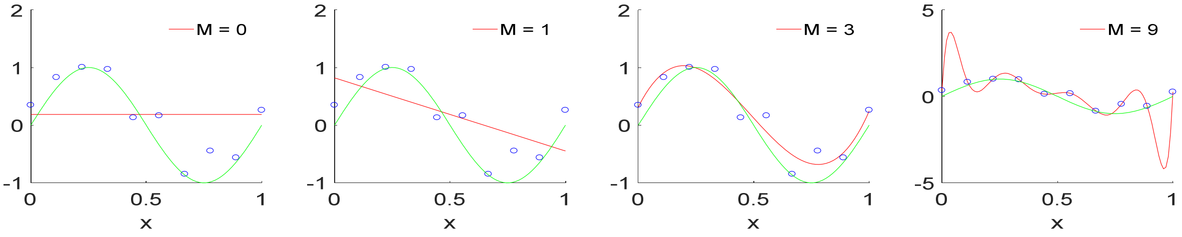
<u>1.4 Feature engineering</u>

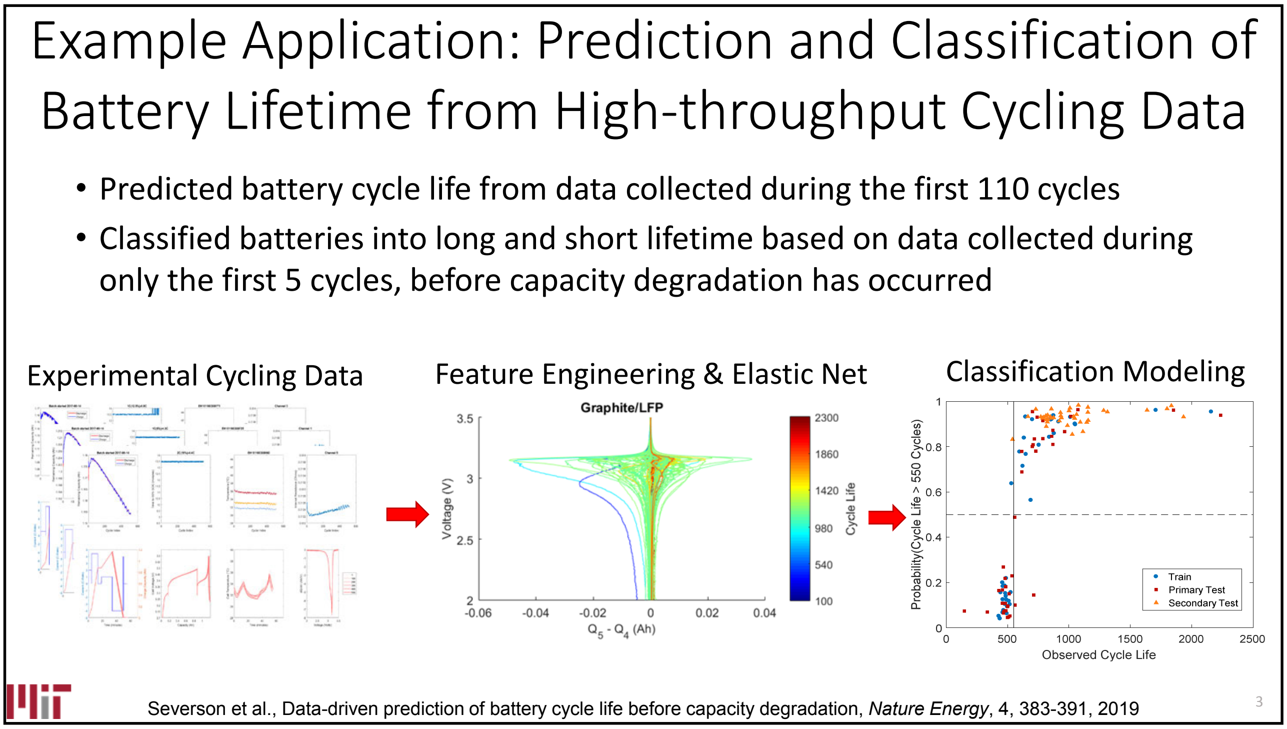
1.5 Kernel methods for nonlinear analytics

1.6 Neural networks and deep learning

Feature Engineering

- Transforming the raw data to concentration information content
- Some examples are using the coefficients of linear, quadratic, or cubic fits
- Apply sparse methods to remove features of low value in modeling





Example Application: Prediction and Classification of Battery Lifetime from High-throughput Cycling Data

- Predicted battery cycle life from data collected during the first 110 cycles
- Classified batteries into long and short lifetime based on data collected during only the first 5 cycles, before capacity degradation has occurred

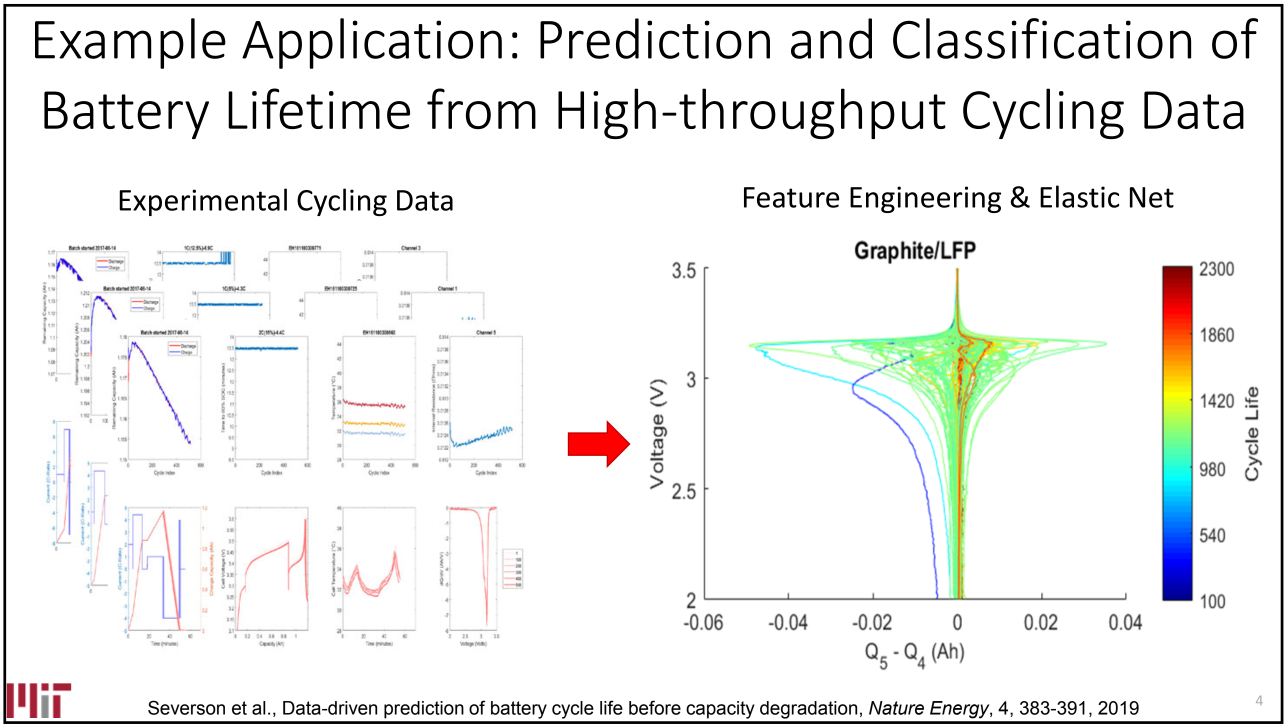
Experimental Cycling Data

Feature Engineering & Elastic Net

Classification Modeling

Severson et al., Data-driven prediction of battery cycle life before capacity degradation, *Nature Energy*, 4, 383-391, 2019

3



Example Application: Prediction and Classification of Battery Lifetime from High-throughput Cycling Data

Experimental Cycling Data

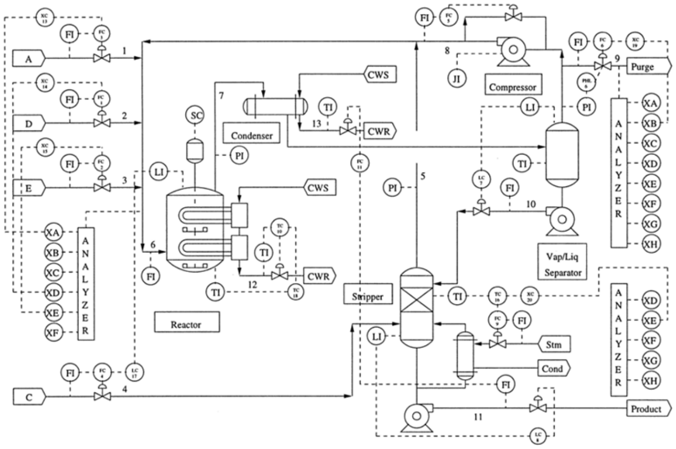
Feature Engineering & Elastic Net

Severson et al., Data-driven prediction of battery cycle life before capacity degradation, *Nature Energy*, 4, 383-391, 2019

4

Example Application: Process structural information

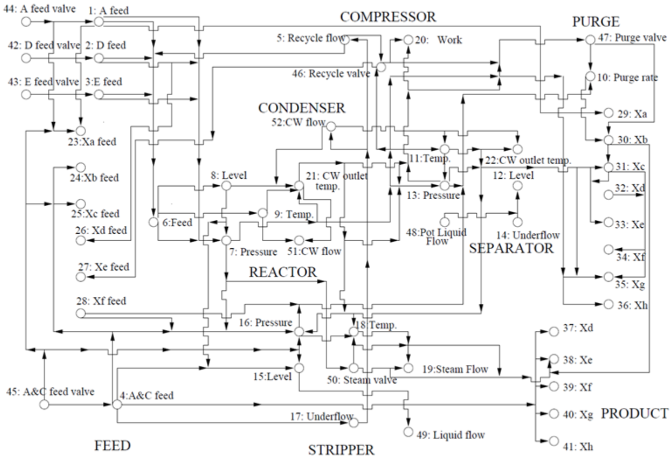
- Process modelers have prior information that can improve performance
- Causal features can be calculated using the process flow diagram
- Let’s demonstrate on the Tennessee Eastman plant with 21 sets of anomalous conditions



L.H. Chiang and R.D. Braatz. Process monitoring using causal map and multivariate statistics: Fault detection and identification. *Chemometrics & Intelligent Laboratory Systems*, 65:159-178, 2003

Example Application: Data → Causal Features

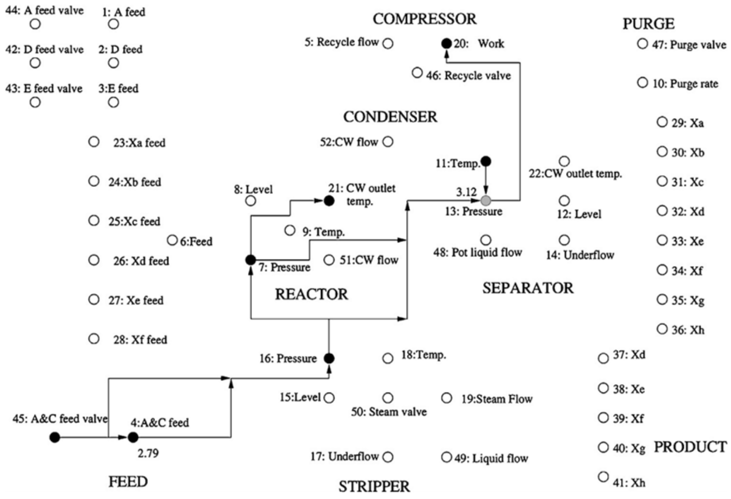
- Causal map can be constructed automatically from process flow diagram or P&ID
- Two distances: modified distance index and causal dependency
- Uses graphs & information theory



L.H. Chiang and R.D. Braatz. Process monitoring using causal map and multivariate statistics: Fault detection and identification. Chemometrics & Intelligent Laboratory Systems, 65:159-178, 2003

Example Application: Data → Causal Features

- Causal features characterize changes in single variables & variance relationships



L.H. Chiang and R.D. Braatz. Process monitoring using causal map and multivariate statistics: Fault detection and identification. Chemometrics & Intelligent Laboratory Systems, 65:159-178, 2003

Example Application: Data → Causal Features

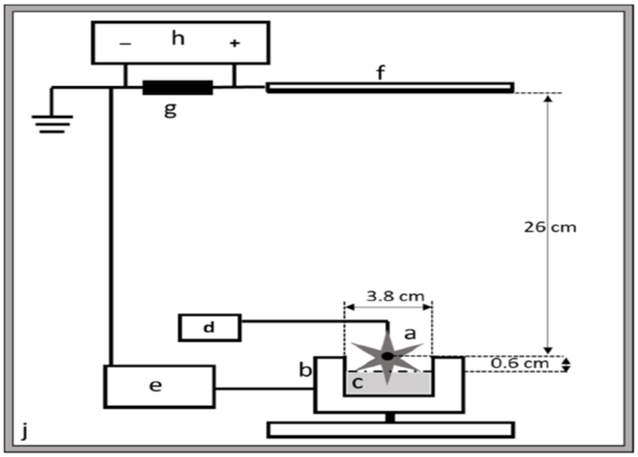
- Simple statistics applied to causal features defined by the process flow diagram had lower misclassification rates than more powerful methods applied to the raw data

	Fault				
Method	4	5	6	11	21
FDA1	0.196	0.034	1	0.331	0.815
FDA2	0.196	**0.020**	1	0.339	0.815
FDA3	0.176	**0.020**	0.941	0.290	0.980
PCA/FDA	0.176	0.024	0.993	0.316	0.830
FDA/PCA	0.208	0.034	1	0.335	0.726
DI/CD	**0.084**	0.055	**0.036**	**0.036**	**0.056**

L.H. Chiang and R.D. Braatz. Process monitoring using causal map and multivariate statistics: Fault detection and identification. Chemometrics & Intelligent Laboratory Systems, 65:159-178, 2003

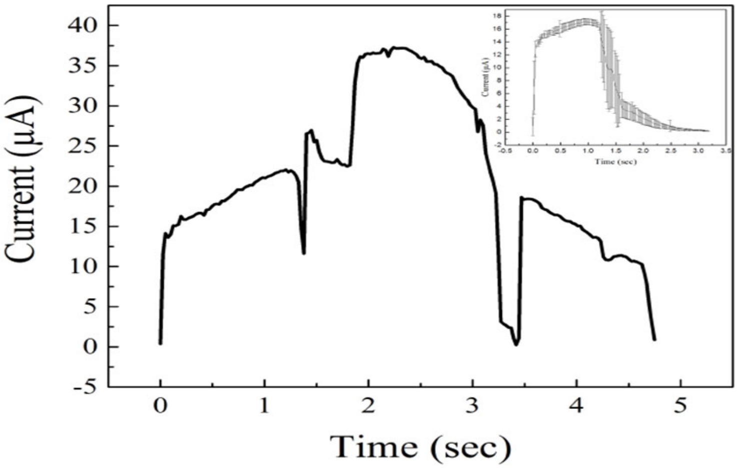
Example Application: Nanofiber Manufacturing via Free Surface Electrospinning

- Application: nanofiber production via free surface electrospinning
- Goal: define features that will help better predict quality and understand the parameters that affect production



I. Bhattacharayy, M.C. Molaro, R.D. Braatz, G.C. Rutledge, Free surface electrospinning of aqueous polymer solution from a wire electrode, Chemical Engineering Journal, 289:203-211, 2016

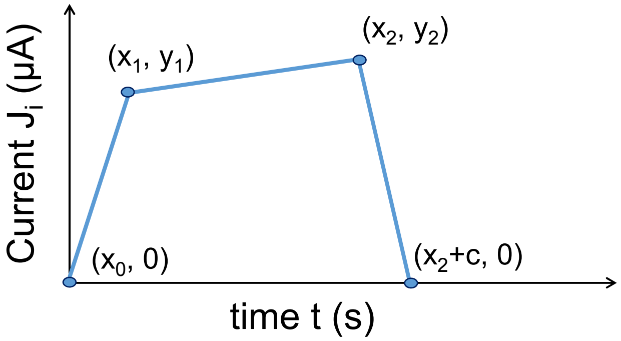
Example Application: Nanofiber Manufacturing via Free Surface Electrospinning



I. Bhattacharayy, M.C. Molaro, R.D. Braatz, G.C. Rutledge, Free surface electrospinning of aqueous polymer solution from a wire electrode, Chemical Engineering Journal, 289:203-211, 2016

Example Application: Nanofiber Manufacturing via Free Surface Electrospinning

- Features: number of jets and positions of transitions in each jet



I. Bhattacharayy, M.C. Molaro, R.D. Braatz, G.C. Rutledge, Free surface electrospinning of aqueous polymer solution from a wire electrode, Chemical Engineering Journal, 289:203-211, 2016

Example Application: Nanofiber Manufacturing via Free Surface Electrospinning

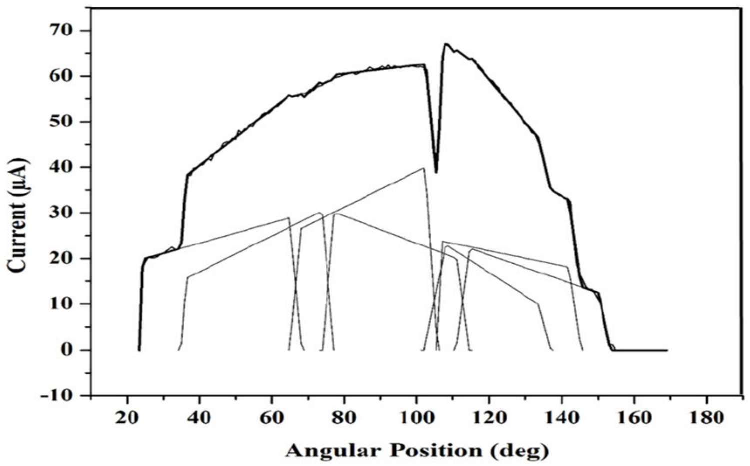
$$\min_{\theta} \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(Y(t) - J_{total}(t|\theta)\right)^2 - \log P(\theta)$$

$$\text{subject to } f_i(\theta) \le b_i \ \forall i, \qquad J_{total} = \sum_{i=1}^{n_{jets}} J_i(t)$$

$$J_i(t) = \begin{cases} 0, & t \le x_0 \\ \dfrac{y_1}{x_1 - x_0}(t - x_0), & x_0 \le t \le x_1 \\ \dfrac{y_2 - y_1}{x_2 - x_1}(t - x_1) + y_1, & x_1 \le t \le x_2 \\ \dfrac{-y_2}{c}(t - x_2) + y_2, & x_2 \le t \end{cases}$$

I. Bhattacharayy, M.C. Molaro, R.D. Braatz, G.C. Rutledge, Free surface electrospinning of aqueous polymer solution from a wire electrode, Chemical Engineering Journal, 289:203-211, 2016

Example Application: Nanofiber Manufacturing via Free Surface Electrospinning



→ can use features to predict product quality and improve understanding (see paper for details)

I. Bhattacharayy, M.C. Molaro, R.D. Braatz, G.C. Rutledge, Free surface electrospinning of aqueous polymer solution from a wire electrode, Chemical Engineering Journal, 289:203-211, 2016

Questions?

Part 1 Outline

1 - Introduction

1.1 Examples of typical data analytics applications

1.2 Unsupervised, supervised, and partially supervised learning

1.3 Least squares including sparse methods

1.4 Feature engineering

1.5 Kernel methods for nonlinear analytics

1.6 Neural networks and deep learning

The Kernel Trick

- Feature spaces can be computationally expensive to calculate, however, many algorithms only depend on the inner product of the feature vectors
- The ‘kernel trick’ replaces the inner product with a call to the kernel function
- This changes the computational cost from O(d^3) to O(n^3) where d is the dimension of the feature space and n is the number of training points

Examples of Kernels

- A kernel matrix must be positive semidefinite
- Examples
  - Gaussian (radial basis function) kernel:

$$\kappa(\mathbf{x},\mathbf{z}) = \exp(-\beta\|\mathbf{x}-\mathbf{z}\|^2)$$

  - Polynomial kernel:

$$\kappa(\mathbf{x},\mathbf{z}) = (\mathbf{x}\cdot\mathbf{z}+c)^p$$

Kernel Methods in More Detail

- Nonlinear data is more likely to show linear pattern when mapped into a higher dimensional space (Cover's theorem), but high-dimensional mapping increases computational time
- The kernel trick gets around this while still having the benefit of high dimension

$$\boxed{k(\boldsymbol{x},\boldsymbol{y}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle}$$

- This idea applies to any method that can be expressed solely in terms of dot products

Examples of kernels:

Polynomial

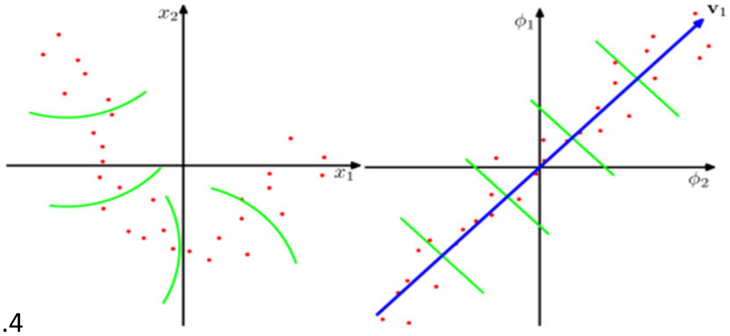
$$k(\boldsymbol{x},\boldsymbol{y}) = (\langle \boldsymbol{x},\boldsymbol{y}\rangle + c)^d$$

Hyperbolic tangent

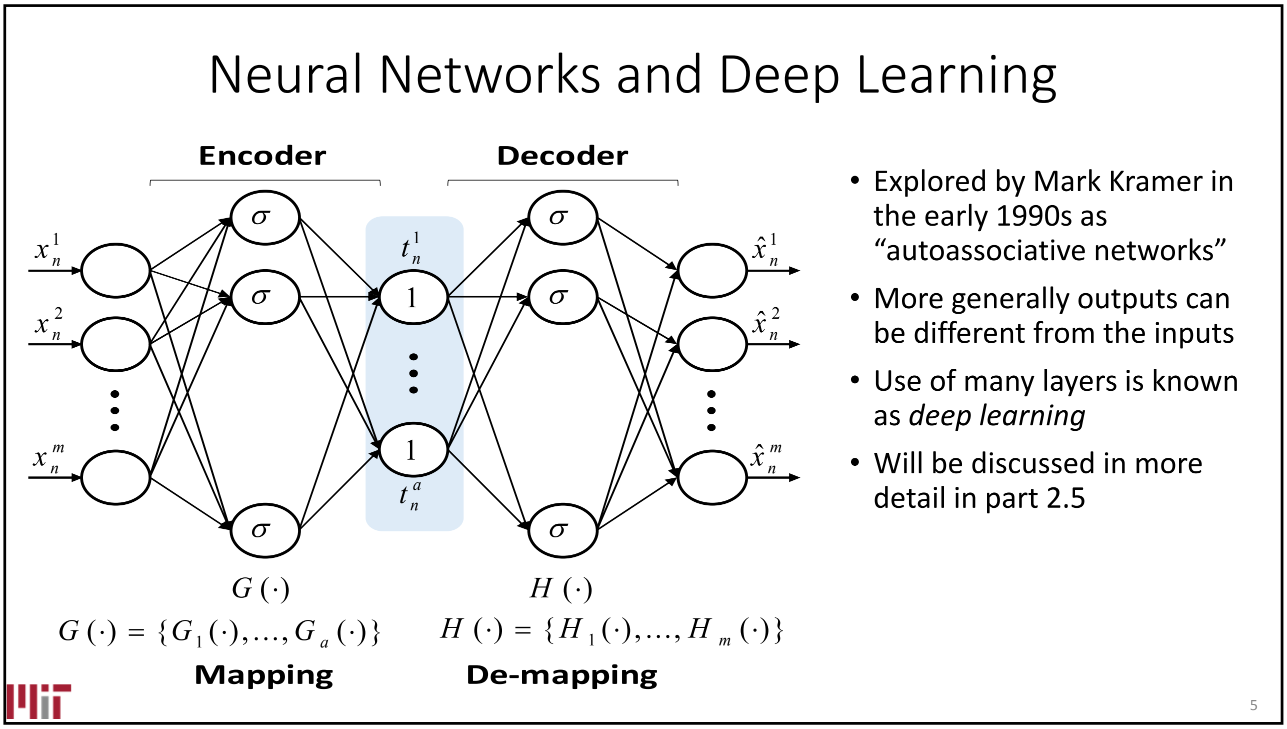
$$k(\boldsymbol{x},\boldsymbol{y}) = \tanh(\beta_0 \langle \boldsymbol{x},\boldsymbol{y}\rangle + \beta_1)$$

Gaussian

$$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2c}\right)$$



- Will be discussed in more detail in part 2.4

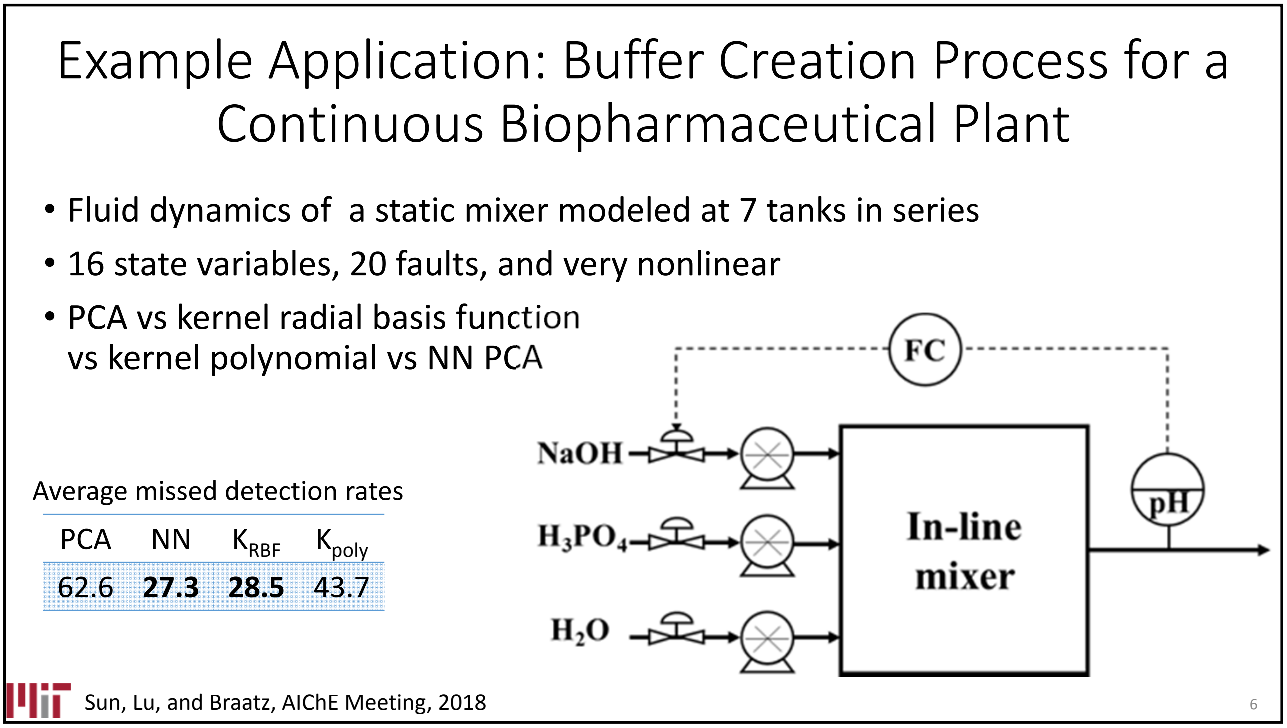


Neural Networks and Deep Learning

$G(\cdot) = \{G_1(\cdot), \ldots, G_a(\cdot)\}$ $H(\cdot) = \{H_1(\cdot), \ldots, H_m(\cdot)\}$

Mapping **De-mapping**

- Explored by Mark Kramer in the early 1990s as "autoassociative networks"
- More generally outputs can be different from the inputs
- Use of many layers is known as *deep learning*
- Will be discussed in more detail in part 2.5



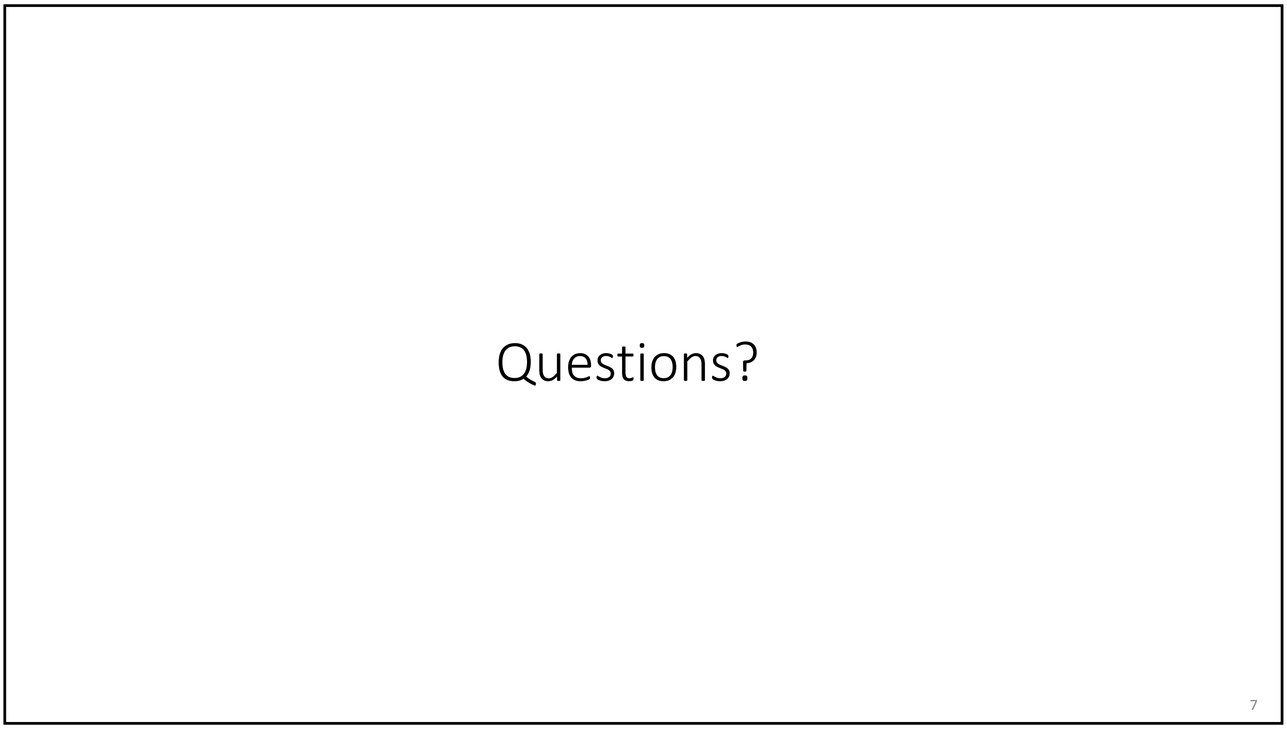
Example Application: Buffer Creation Process for a Continuous Biopharmaceutical Plant

- Fluid dynamics of a static mixer modeled at 7 tanks in series
- 16 state variables, 20 faults, and very nonlinear
- PCA vs kernel radial basis function vs kernel polynomial vs NN PCA

Average missed detection rates

PCA	NN	K_{RBF}	K_{poly}
62.6	**27.3**	**28.5**	43.7

Sun, Lu, and Braatz, AIChE Meeting, 2018



Questions?