TEACHING PROCESS DATA ANALYTICS AND MACHINE LEARNING AT MIT

Moo Sun Hong, Weike Sun, Brian W. Anthony, and Richard D. Braatz
Massachusetts Institute of Technology • Cambridge, MA 02139

INTRODUCTION

Driven by increases in data and increased capabilities in data collection, storage, and reduced computational costs, companies use data to streamline operations, improve reliability, and optimizes processes.[1-4] These trends drive demand for scientists and engineers with expertise in *process data analytics*, that is, computer-aided tools for the analysis of process data.[1-4] Many companies have built the hardware and software infrastructure to bring the data into one database for easy access, with efforts in that direction at chemical engineering-related companies being well documented.[1-4] As an example of a specific industry that hires large numbers of chemical engineers, a recent edited collection describes process data analytics practices at pharmaceutical companies, namely, GlaxoSmithKline, Biogen, Bristol-Myers Squibb, Pfizer, Shire, Novartis Pharma AG, Merck & Co., and Eli Lilly and Company.[3] Two companies that design and build data-associated infrastructure for chemical engineering-related industries are OSIsoft and Aspen Technology.[5-6] The former is the developer and supporter of the PI System™ data management platform, which is a suite of software products for data collection, storing, searching, and visualizing data.[5] The latter has developed a suite of competing products while also offering software that would use that data as inputs, such as mechanistic modeling and supply chain management software.[6]

Plant data are being correlated to off-line product quality specifications, which in turn are being connected to the supply chain.[3] This process understanding has been used to (1) troubleshoot problems in plant operation causing off-spec product (e.g., raw materials and operator error); (2) optimize feed streams (e.g., selection of raw materials or mixtures from multiple suppliers); (3) propose process or control design changes to reduce operational problems, reduce waste materials and energy, and increase profitability; (4) design predictive maintenance schedules; and (5) facilitate continuous improvement practices.[1-4]

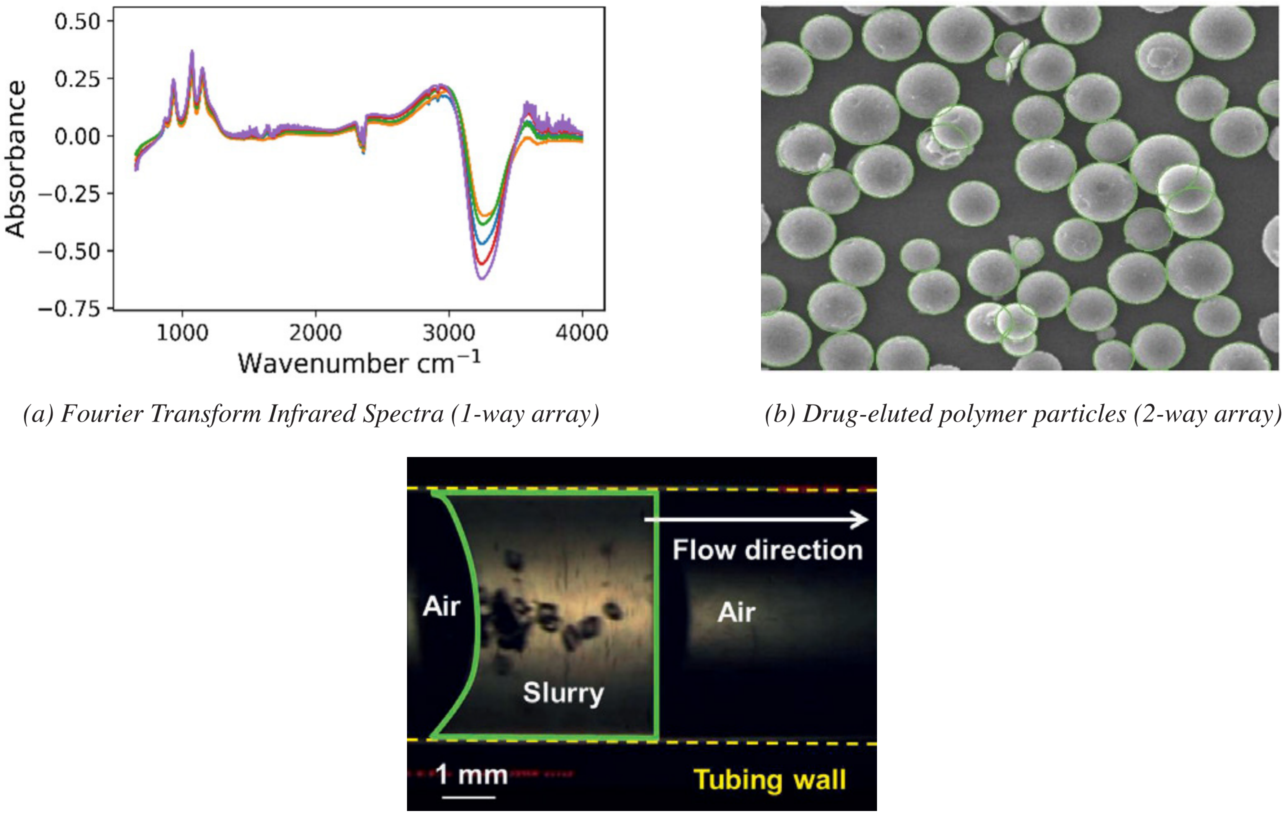
Data analytics tools such as control charts, principal component analysis (PCA), and partial least squares (PLS) have been important tools for manufacturing processes in industry for decades.[1-4] New data analytics and machine learning tools have become readily available that can generate improved results over the industry-standard techniques, and new datasets present different attributes and challenges that are not addressed using classical techniques (Figure 1).[1-4]

Moo Sun Hong *is a postdoctoral researcher in the group of Professor Richard D. Braatz at the Massachusetts Institute of Technology (MIT). He received an M.S. in Chemical Engineering Practice and Ph.D. from MIT. His research is in advanced biopharmaceutical manufacturing systems. Honors include the AIChE PD2M Award for Excellence in Integrated QbD Practice, the AIChE Separations Division Graduate Student Research Award, and the Jefferson W. Tester Award from the MIT Chemical Engineering Practice School. ORCID: 0000-0003-2274-5030*

Weike Sun *received a B.S. from Tsinghua University and a Ph.D. from MIT. She is an expert in process data analytics who is the primary author of multiple open-source software packages including ALVEN for algebraic learning via elastic net for static and dynamic model identification and SPA for smart process analytics, which automatically selects and applies process data analytics and machine learning methods based on domain knowledge, the specific data characteristics, and nested cross-validation procedures.*

Brian W. Anthony *is Director of the MIT Master of Engineering in Manufacturing Program and Co-Director of the Medical Electronic Device Realization Center. He received a B.S. from Carnegie Mellon and M.S. and Ph.D. from MIT. He has over 25 years of commercial, research, and teaching experience in product realization and advanced manufacturing. He has over 20 patents and received an Emmy from the Academy of Television Arts and Sciences for sports broadcast technical innovation. ORCID: 0000-0001-6346-5276*

Richard D. Braatz *is the Edwin R. Gilliland Professor at the Massachusetts Institute of Technology (MIT) where he does research in data analytics, design, and control of advanced manufacturing systems. He received an MS and PhD from the California Institute of Technology and was on the faculty at the University of Illinois at Urbana-Champaign and a Visiting Scholar at Harvard University before moving to MIT. He is a member of the National Academy of Engineering. ORCID: 0000-0003-4304-3484*



(a) Fourier Transform Infrared Spectra (1-way array)

(b) Drug-eluted polymer particles (2-way array)

(c) An in-line stereomicroscope image from the real-time video (3-way array) of a crystallization process.

Figure 1. *Example Data Sets. Image (a) is of spectra in which each spectrum is a vector of absorbance as a function of wavenumber. Image (b) is a grey-scale image of biodegradable polymer microspheres that contain a drug that releases over a prolonged period of time when injected into the body. Image (c) is a snapshot of a real-time grey-scale video of a slug of slurry between two air slugs moving through a tube, in which the particles in the slurry are organic crystals and the placement of the slurry slug between air slugs results in mixing of the slurry as it moves down the tube. Images (a) and (c) are adapted from References 7 and 8 respectively.*

Modern data include 1-way arrays (e.g., infrared spectra), 2-way arrays (e.g., two-dimensional particle size distributions), 3-way arrays (e.g., black-and-white video, hyperspectral images), and 4-way arrays (color videos).[9] While methods have long been available and applied to 1- and 2-way arrays, and suboptimal methods have been available for 3-way arrays, algorithms and software have become available for handling the higher order data structures while producing more accurate results.[9] With the price of a color charge-coupled device (CCD) camera at less than $100, imaging is more widely used in chemical and biological systems.[9] These images provide grey-scale video with dimensions x, y, and t (color adds an RGB* dimension).[9] Applications of real-time imaging include controlling the amount of seasoning with images of snack foods,[10] real-time process control of flexible systems,[11] and measuring the crystal size distribution.[12]

Teaching chemical engineers to be effective using data analytics is as important today as in any time in the history of the discipline, and advances in sensor technologies, machine learning, and associated software have enabled applications that were not formerly possible.[1-4] This article describes experiences with teaching process data analytics and machine learning over the last few years, including in (1) a new chemical engineering elective course in which the students come from chemical engineering, mechanical engineering, and engineering management, and (2) an undergraduate chemical engineering concentration in process data analytics with courses coming from a mixture of multiple disciplines including chemical engineering, computer science, economics, and management. Along the way, the article will describe the challenges in teaching data science to chemical engineers as well as scientists and engineers working in the chemical and biotechnology industries, and also offer strategies for overcoming those challenges.

* RGB = red-green-blue

COURSE DESIGN

In Fall 2019 we designed a Process Data Analytics course for students to (1) learn a wide variety of process data analytics methods and know how to evaluate the tradeoffs amongst methods for specific applications; (2) gain experience applying these methods to specific real applications at scale, including in the laboratory; and (3) understand how the data analytics methods work and why algorithms sometimes produce unexpected results. The course covers tools for solving problems based on data, ways to choose algorithms for a specific problem, numerous application examples, and tricks of the trade. Topics include:

- unsupervised, supervised, and partially supervised learning (12 hours)
- correlation analysis, latent variable methods (5 hours)
- temporal data and time series analysis (6 hours)
- feature engineering, kernel methods (3 hours)
- neural networks, deep learning (3 hours)
- ensemble learning, random forest (5 hours)
- big data, video analytics and hyperspectral imaging (3 hours)

The suggested reading materials include References 7–15 and are heavily weighted to references that describe data analytics applications to real process datasets.

A challenge in applying process data analytics and machine learning methods is that a substantial level of expertise is required to select the best data analytics tool for a particular application.[14] Tools come from various fields, including applied statistics, analytical chemistry, operations research, and computer science.[15] Current chemical engineering curricula do not prepare students with the background needed to understand all of the valuable tools in a deep theoretical level, and so practitioners tend to use only the methods that they are most familiar with, which can produce suboptimal results.[14] To train students to be the most effective in analyzing data, extensive information is provided on how to select the best methods for each application, in order to allow the user to focus on the goals of the analysis rather than learning all of the mathematics underlying the methods.[14]

The course grade is based on seven assignments: five problem sets (50%) and two projects (50%). In each assignment the students are tasked with solving an authentic data analytics problem in which all data are collected from real processes. The topics included in the seven assignments are the topics listed in the above bulleted list of topics. In each assignment the students apply the specific data analytics methods covered in the lectures shortly after the methods are covered in the lectures. Sample questions for an assignment include:

1. Your objective is to construct models for estimating paracetamol concentration from ATR-FTIR spectra and temperature collected in a batch purification process.
2. Visualize the data and justify whether linear models are appropriate for this dataset. Report on whether any data are likely outliers or biased.
3. Given the amount of data, is it recommended to use ordinary least-squares regression to construct a model that depends on all of the absorbances? Why or why not?
4. Apply ridge regression, lasso, and elastic net to the temperature and the absorbance at all frequencies.
5. Compare and discuss the results from the application of the multiple data analytics methods.

Students are encouraged to discuss the assignments with the professors, teaching assistants, and other students in the class. Students are allowed to even ask others to help debug software, as long as the software that is turned in is the student's own. The actual codes tend to be relatively short since the students are allowed to use any of the large number of software programs and packages that have become available, which has grown exponentially in recent years.

Each lecture describes process data analytics problems with real data, and explains and demonstrates how to solve those problems. Some theoretical background for each data analytics method is provided so that the students understand the underlying principles to help guide method selection and to assist in interpreting and debugging results, but the course is not a "theory course," that is, the course does not present any theorems or proofs or ask the student to do any theoretical derivations. Most of the time in the twenty lectures is spent on the solution of data analytics problems with the MATLAB® commands used to generate the results given. Each of those lectures describes methodologies and their application. Five lectures are entirely or primarily focused on specific applications, primarily in fiber extrusion, (bio) pharmaceutical manufacturing, and lithium-ion batteries. Most students use MATLAB in their problem sets and projects, since most students have experience with MATLAB from past courses, but the students are allowed to use alternatives such as Python® and R®. All course materials are electronic. Students are expected to read the course materials before class and to read the materials again before doing homework. When the number of students in the course is sufficiently high in a particular offering, a senior graduate student is assigned to the course as a teaching assistant. At a minimum, the course is assigned a graduate student to grade assignments.

The data analytics and machine learning methods used in the course are available in base MATLAB or its Statistics and Machine Learning and System Identification toolboxes. Compared to Python and R, MATLAB is easier to learn, has much shorter computational times, has more options for analysis of time series data, and has the widest variety of engineering applications for applying the data-driven models, such as in feedback control design. Python and R have a wider suite of advanced machine learning algorithms available and are open source so that graduates can use the software in industry without needing licenses.

As mentioned above, this course only uses authentic datasets. For example, one of the datasets provided in the course is high-throughput battery cycling data for classification of battery lifetime (Figure 2).[13] In this data analytics problem, the lifetime of rechargeable batteries can be predicted from data collected during the first 110 cycles, and the batteries can be classified into long and short lifetime based on data collected during only the first 5 cycles, before capacity degradation has occurred. The students are given Reference 13, which contains the application of two machine learning methods, namely, the elastic net for the prediction problem and a logistic map for the classification method. The students are tasked with applying multiple machine learning methods to these data analytics problems, with the objective of obtaining improved performance over the published results. This posing of the problem in terms of a challenge is motivating for the students. For Project 1, the students apply data analytics and machine learning methods covered in the first half of the class and, for students who choose to use this same dataset for Project 2, the students apply methods covered in last half of the class.

COURSE ASSESSMENT AND TEACHING EXPERIENCE

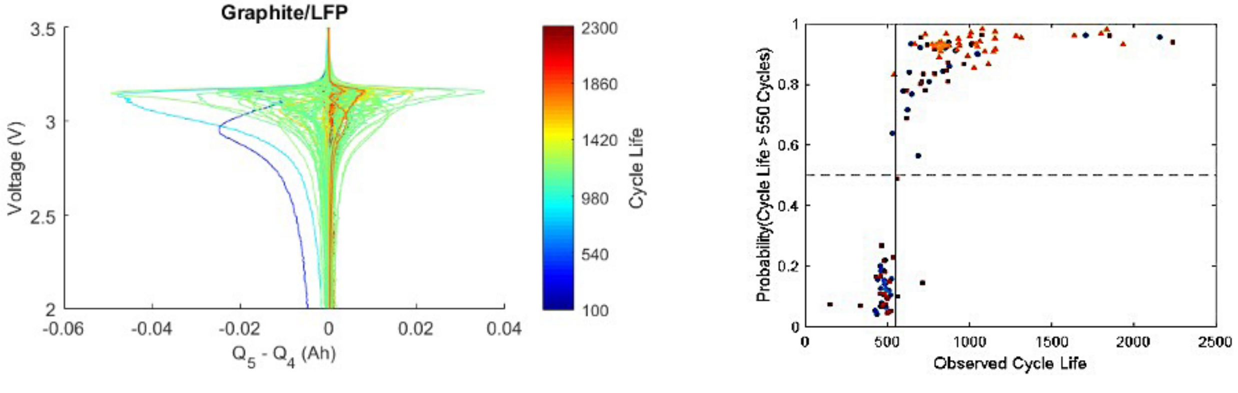
Student evaluations were collected each time that the course has been taught. All scores were out of a possible seven points. The average student self-assessment scores for Fall 2020 were 5.5, 5.6, and 6.3 for the course's three primary learning objectives respectively:

- "I understand a wide variety of process data analytics methods."
- "I know how to evaluate the tradeoffs amongst process data analytics methods for specific applications."
- "I have experience applying the process data analytics methods to specific applications."

The corresponding average scores were 6.7, 6.3, and 6.3 for Fall 2021 respectively. The average scores for the course material ranged from 5.9 to 6.7 respectively for

- "The expectations for the subject were clearly defined."
- "The learning objectives for the subject were met."
- "The subject's assignments contributed to my learning."
- "The subject's grading thus far has been fair."
- "I was satisfied with my overall learning experience in this subject."

The corresponding average scores ranged from 6.0 to 7.0 for Fall 2021.



(a)

(b)

Figure 2*. Machine learning-based prediction and classification of lithium-ion batteries into long and short lifetime. (a) Current and voltage data are transformed into voltage vs. change in charge between cycles 4 and 5, which are then converted into features, which are transformations of the data for subsequent analysis. (b) Classification results obtained by application of a logistic map to the features. Most of the data used to train the logistic map and test its performance are in the upper right quadrant, where the model correctly predicts that the battery has high cycle life (i.e., higher than 550 cycles), and the lower left quadrant, where the model correctly predicts that the battery has low cycle life. Plots are adapted from Reference 13.*

The average score for the instructors in Fall 2020 was 6.0 for "stimulated my interest in the subject" and 6.7 for "displayed thorough knowledge of subject material." For Fall 2021 the associated scores for these categories were 6.85 and 7.0 respectively.

Teaching the course to students of different technical expertise can be challenging. The mechanical engineering and management students who have taken the course rarely have expertise or interest in chemistry or spectroscopy, which we have addressed by providing more examples of the use of spectroscopy in all branches of science and engineering, including mechanical engineering. Another challenge is that some students who have signed up for the course lack basic knowledge of statistics or computer programming. In that situation we encourage the students to take basic foundational courses at MIT such as applied statistics.

A CHEMICAL ENGINEERING SPECIALIZATION IN PROCESS DATA ANALYTICS

The Massachusetts Institute of Technology offers both a B.S. in Chemical Engineering and a B.S. in Engineering (emphasis in chemical engineering), accredited by the Engineering Accreditation Commission of ABET.[16] The students in both programs are required to take basic courses in mathematics and the sciences and thermodynamics, Introduction to Chemical Engineering, Chemical and Biological Engineering Thermodynamics, Fluid Mechanics, Transport Processes, Chemical Kinetics and Reactor Design, two laboratory courses, and one capstone course.[16] The students are also required to complete four courses in a concentrated area of study, such as Energy, Environmental Studies, and Biomedical Engineering.[16]

Given the need for chemical engineering graduates with computational skills, in 2018 we defined three computation-related concentrations: Engineering Computation, Manufacturing Design, and Process Data Analytics. The courses in the Process Data Analytics concentration are:

1. Engineering Computation and Data Science, which is an introduction to the formulation and solution of engineering problems using computation. The course emphasizes data analytics with topics including exploratory data analysis and visualization, filtering, regression, and building basic machine learning models (classifiers, decision trees, clustering).
2. At least one course from a list of statistics courses. The content of these courses is primarily in statistics, including Introduction to Probability and Statistics in Engineering, Statistical Thinking and Data Analytics, Statistics and Probability, and Fundamentals of Statistics.
3. At least one course from a list of data analytics courses. The content of these courses is primarily in the application of data analytics to specific classes/types of problems, including biomedical, signal processing, management, and the aforementioned course in process data analytics.

In addition, students can define their own concentrations, and several have defined those concentrations to include data analytics courses.

REFERENCES

1. Qin SJ and Chiang LH (2019) Advances and opportunities in machine learning for process data analytics. *Comput. Chem. Eng.* 126:465–473.
2. Qin SJ (2014) Process data analytics in the era of big data. *AIChE J.* 60:3092-3100.
3. Ferreira AP, Menezes JC, and Tobyn M, editors (2018) *Multivariate Analysis in the Pharmaceutical Industry*. Academic Press. London, UK.
4. Ündey C, Ertunç S, Mistretta T, and Looze B (2010) Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control. *J. of Process Control*, 20:1009-1018
5. OSIsoft. http://www.osisoft.com. Accessed on August 15, 2022.
6. Aspen Technology. http://www.aspentech.com. Accessed on August 15, 2022.
7. Togkalidou T, Fujiwara M, Patel S, and Braatz RD (2001) Solute concentration prediction using chemometrics and ATR-FTIR spectroscopy. *J. of Crystal Growth*. 231:534–543.
8. Jiang M and Braatz RD (2018) Low-cost noninvasive real-time imaging for tubular continuous-flow crystallization. *Chem. Eng. Tech.* 41:143–148.
9. Sun W and Braatz RD (2020). Opportunities in tensorial data analytics for chemical and biological manufacturing processes. *Comput. Chem. Eng.* 143:107099.
10. Bourg Jr WM, Bresnahan SA, Haarsma GJ, MacGregor JF, Martin PA, and Yu H (2006) Method for on-line machine vision measurement, monitoring and control of product features during on-line manufacturing processes. U.S. Patent 7,068,817.
11. Anthony BW and Chua F (2017) Computationally efficient optimal video comparison for machine monitoring and process control. *J. Manuf. Sci. Eng.* 139(10):101007.
12. Hong MS, Lu AE, Bae J, Lee JM, and Braatz RD (2021) Droplet-based evaporative system for the estimation of protein crystallization kinetics. *Cryst. Growth Des.* 21(11):6064–6075.
13. Severson KA, Attia PM, Jin N, Perkins N, Jiang B, Yang Z, Chen MH, Aykol M, Herring PK, Fraggedakis D, Bazant MZ, Harris SJ, Chueh WC, and Braatz RD (2019) Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*. 4:383–391.
14. Sun W, and Braatz RD (2021) Smart process analytics for predictive modeling. *Comput. Chem. Eng.* 144:107134.
15. Chiang LH, Russell EL, and Braatz RD (2001). *Fault Detection and Diagnosis in Industrial Systems*. Springer-Verlag. London.
16. Engineering (Course 10-ENG) (2022) Massachusetts Institute of Technology, Cambridge, Massachusetts. http://catalog.mit.edu/degree-charts/engineering-chemical-engineering-course-10-eng/. Accessed on August 15, 2022.❐