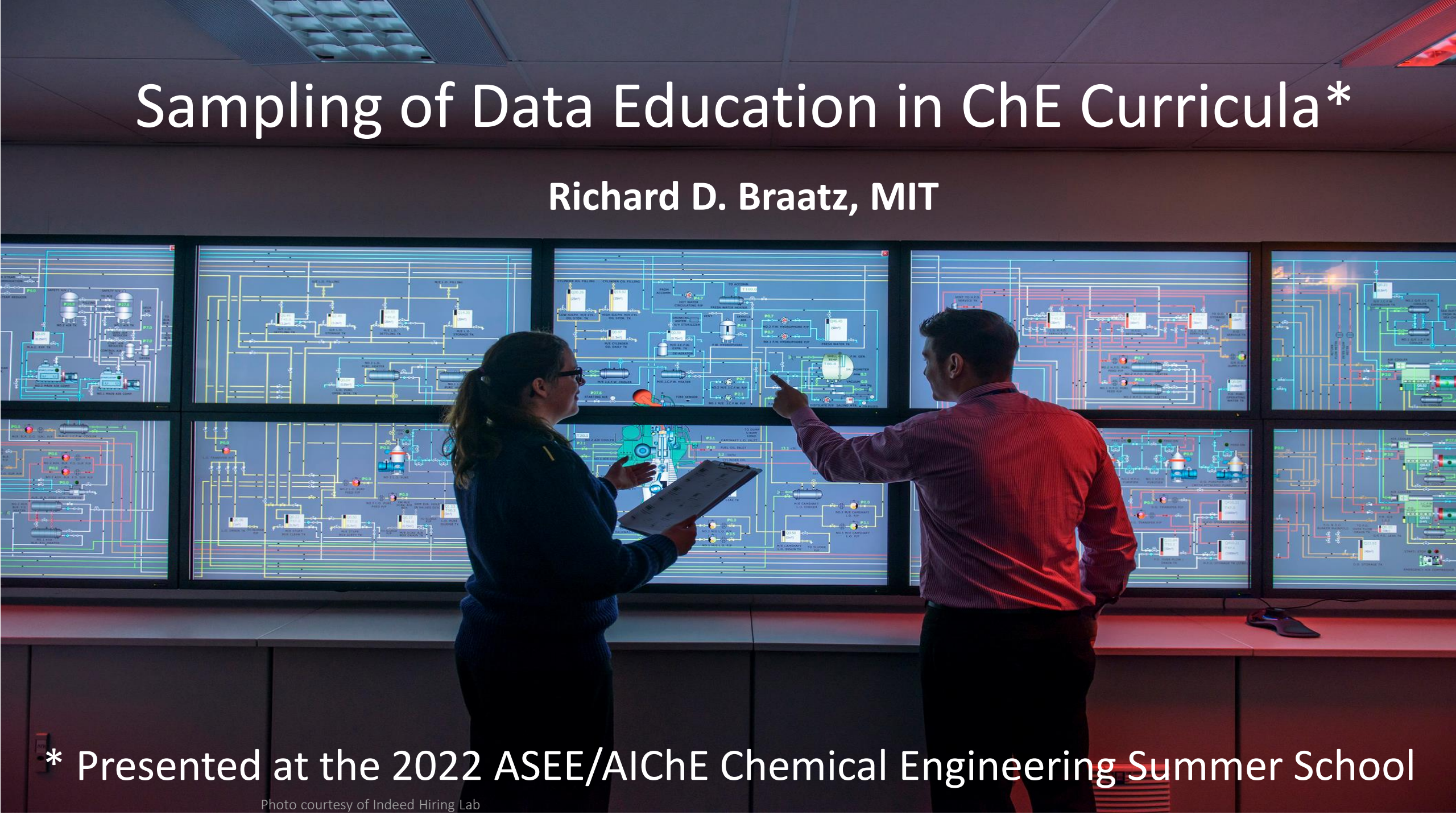
Sampling of Data Education in ChE Curricula*

Richard D. Braatz, MIT



* Presented at the 2022 ASEE/AIChE Chemical Engineering Summer School

Photo courtesy of Indeed Hiring Lab

Purpose and Goals of This Presentation

- The ***purpose*** is to provide materials and best practices to help new faculty in teaching applied statistics and data analytics

- The ***goals*** are to
  - learn about various subtopics within applied statistics and data analytics to assist in deciding what to include within the time allotted to this area within their curriculum
  - become familiar with available teaching materials and software resources
  - learn some best practices for imparting useful skills to undergraduate students
  - learn how the topics and skills connect with (and support) other topics in the chemical engineering curriculum

Why ChEs needs to learn process data analytics

- Companies are using data to streamline operations, improve reliability, optimize processes
- Driven by huge increases in data and reduced computer costs
- Driving demand for scientists and engineers with expertise in analytics

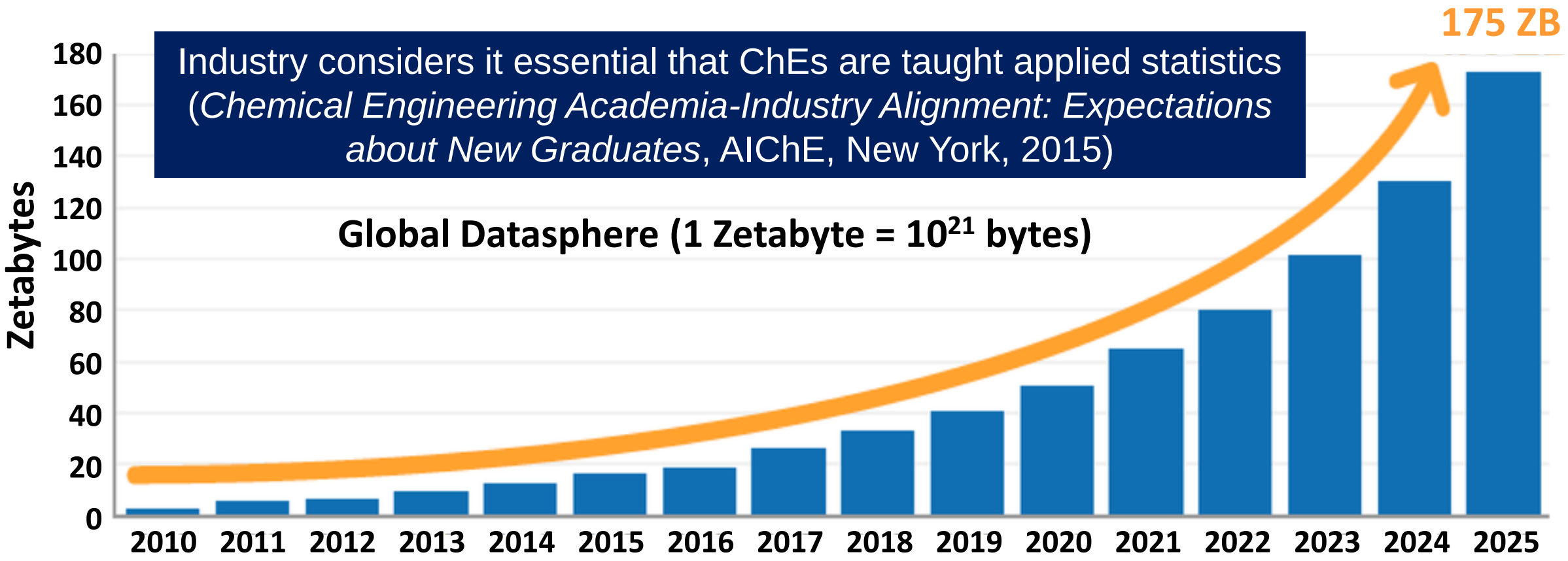
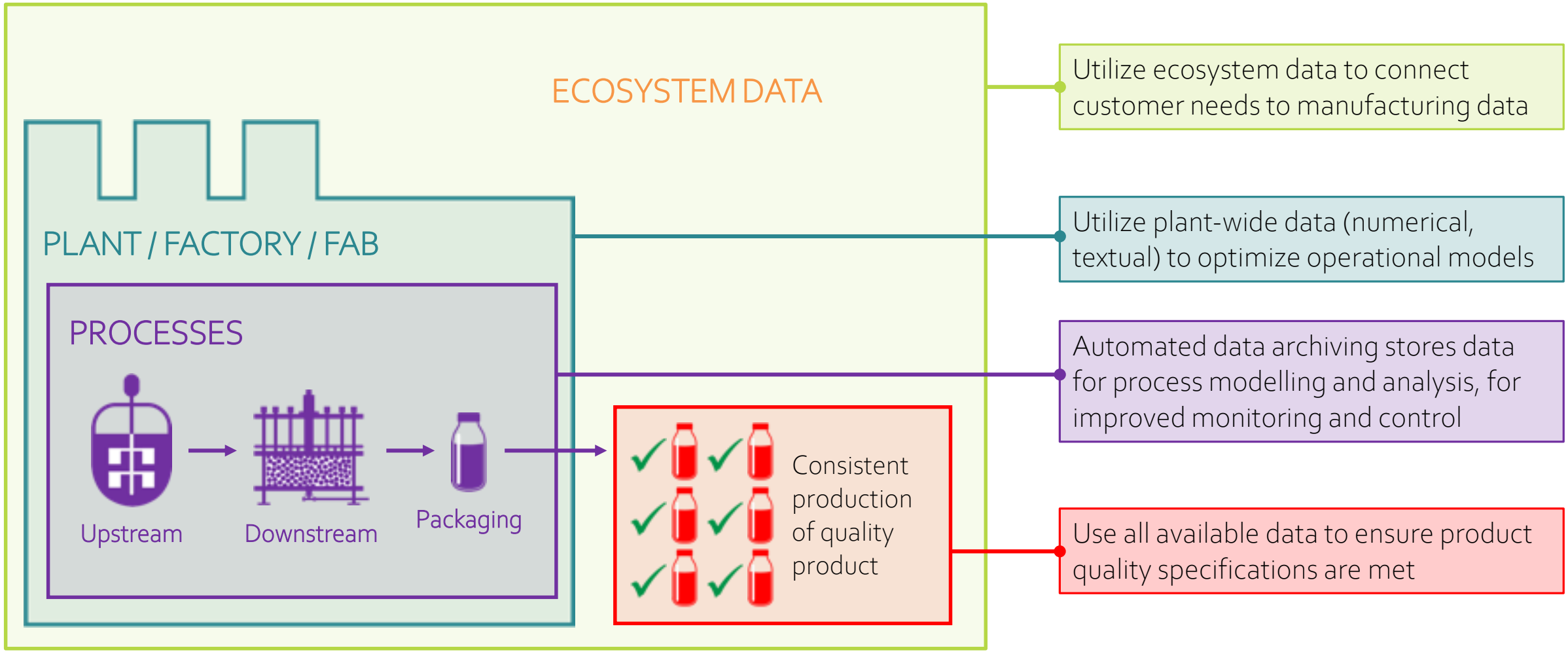


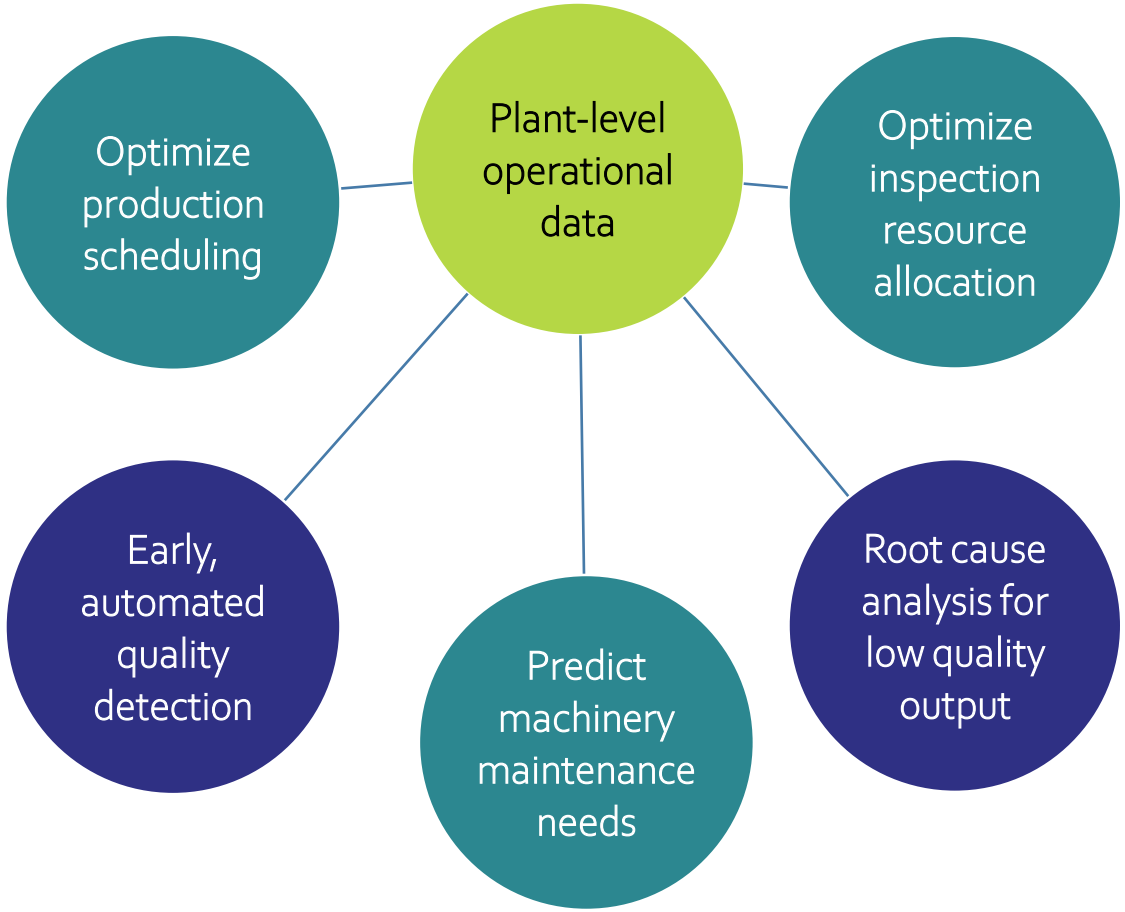
Figure adapted from Seagate

Data Analytics at All Levels in the Chemical Industry



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

Examples of Typical Data Analytics Applications



- 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 such as 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

Spectrum of Data Education in ChE Curricula

- a few lectures in a process design course
- a course in statistics taught from the mathematics department
- a course in applied statistics taught by a chemical engineering department
- a course in experimental design from an industrial engineering department
- an applied statistics four-week module taught as part of an applied mathematics course
- an introduction to applied statistics and data analytics four-week module taught as part of a process systems engineering course

Sample Topics in Statistics/Data Analytics Courses in ChE

- Experimental design: generate data so that the model will be good enough
- Linear/nonlinear regression: least-squares, response surface methodology
- Uncertainty quantification: estimation of confidence & prediction intervals
- Chemometrics: handling correlated data (e.g., PLS, PCA, FDA)
- Statistical process control: do data indicate that the process is under control? which variables are likely associated with the fault? how to classify new data based on historical data?
- Machine learning: sparse vs. dense models, construction of sparse models, lasso & elastic net, neural networks

Example: Data Education at the University of Buffalo

- 14 weeks to juniors
- Lecturer: David A. Kofke (ChE)
- William Navidi, *Statistics for Engineers & Scientists*
- Sampling and descriptive statistics, probability, error propagation, common distributions, confidence intervals, hypothesis testing, factorial experiments

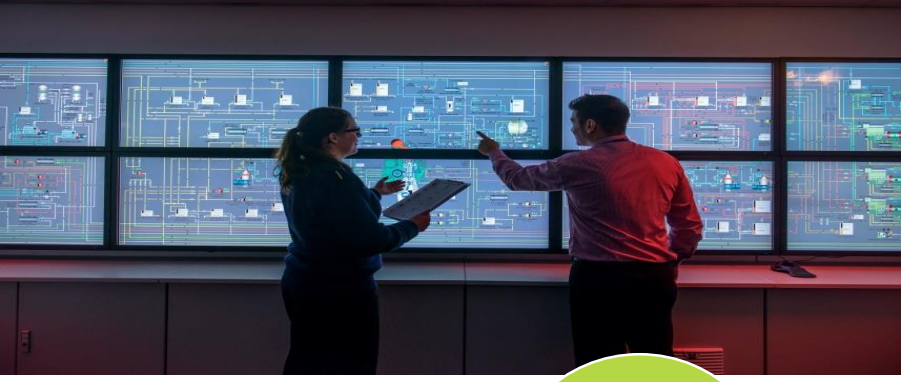
Example: Data Education at UT Austin

- 16 weeks to juniors
- Lecturer: Keith Friedman (ChE)
- R.A. Johnson, *Statistics & Probability for Engineers*
- Linear regression, JMP, simple distributions, confidence intervals, ANOVA, hypothesis testing, design of experiments, statistical process control

Example: Data Education at Brigham Young University

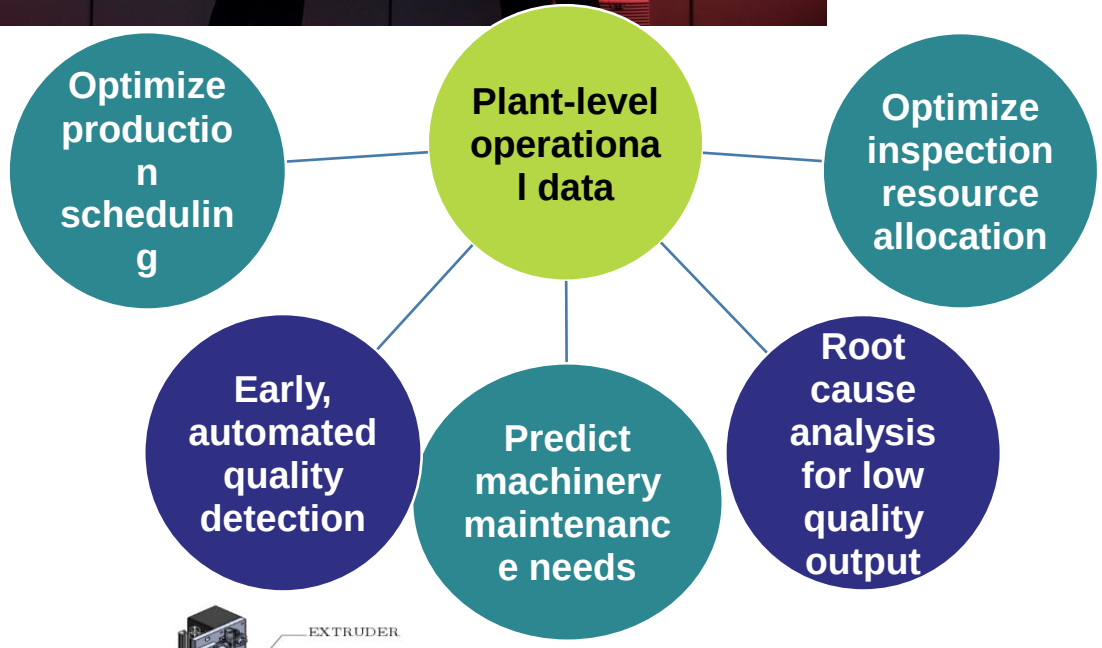
- 14 weeks to sophomores, 14 weeks elective for seniors & graduate students
- Lecturers: Matt Heiner (Stats), Larry Baxter (ChE), John Hedengren (ChE)
- William Navidi, *Statistics for Engineers & Scientists, Lecture Notes and Online for Seniors/Grad Students*
- The scientific method; probability, random variables, common discrete and continuous random variables, central limit theorem; confidence intervals and hypothesis testing; completely randomized experiments; factorial experiments

Example: Data Education at MIT (Elective)



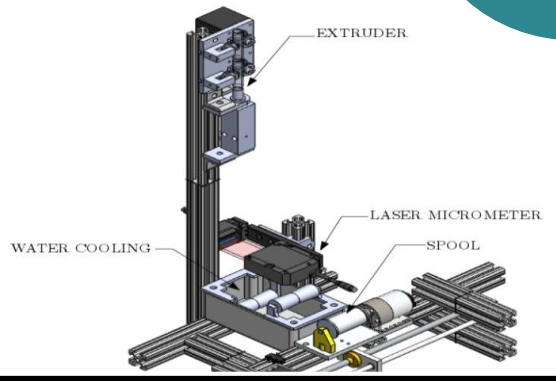
Course objectives:

- Learn data analytics methods & selection of methods for specific applications
- Gain experience in applying analytics methods to real, authentic datasets
- Provide enough understanding to be able to troubleshoot unexpected results

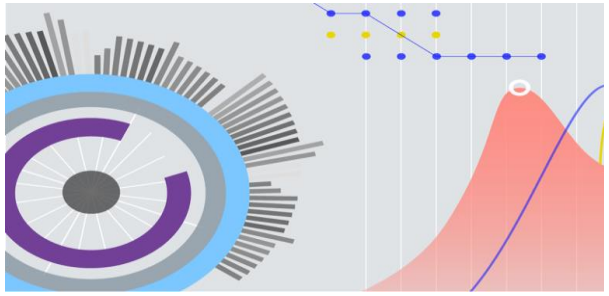


Topics include

- unsupervised, supervised, and semi-supervised learning
- correlation analysis, latent variable methods
- temporal data and time series analysis
- feature engineering, kernel methods
- neural networks
- ensemble learning, random forest
- real-time video, hyperspectral imaging



Acquisition, use, and storage of contextualized data



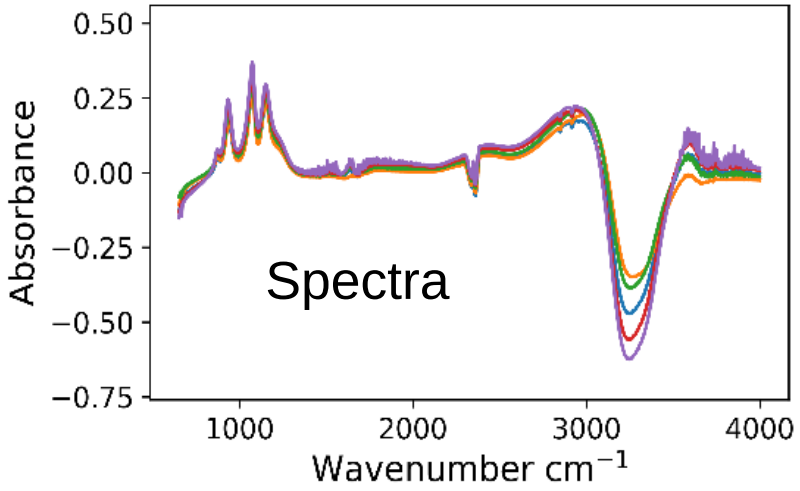
Analytics, modelling, simulation

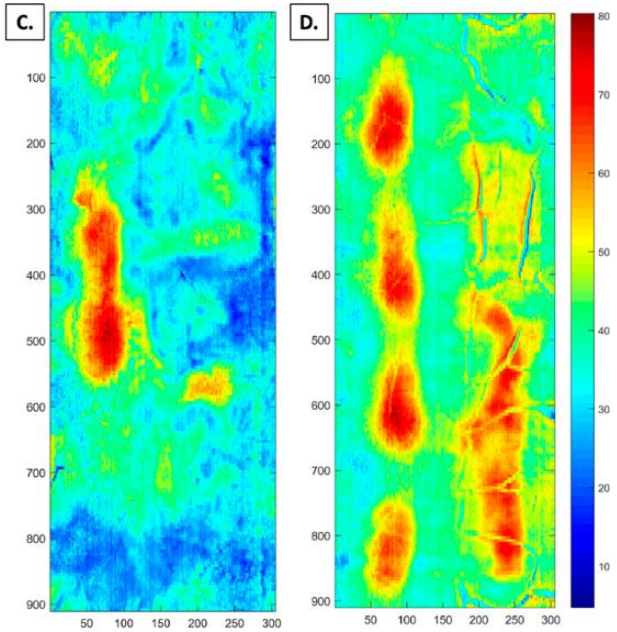
Learn from the Data – Information and Knowledge

Course data

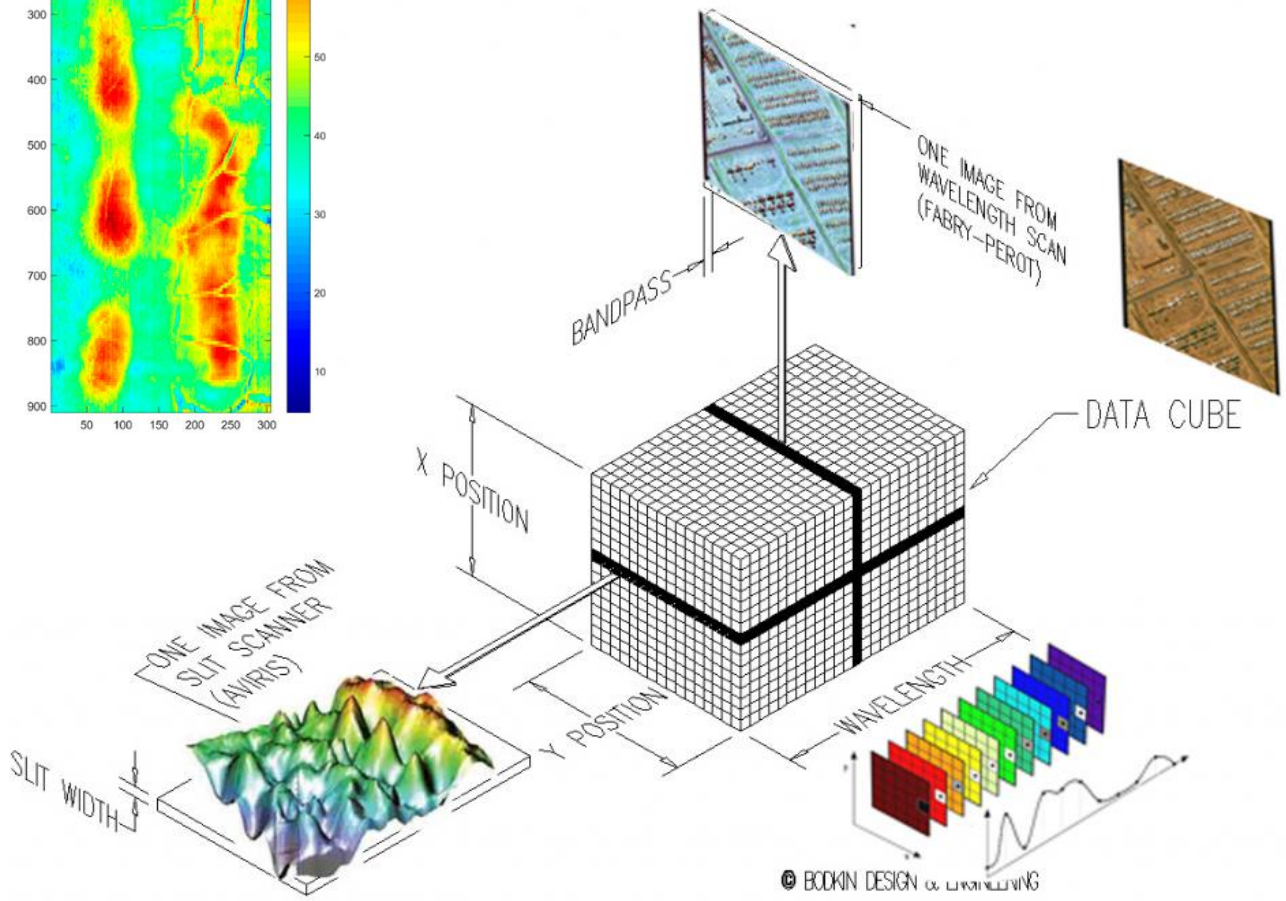
- Matlab-based problem sets and projects
- All electronic course materials

Course covers a wide variety of process datasets

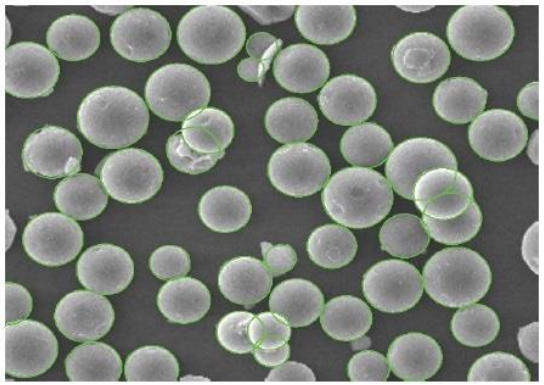


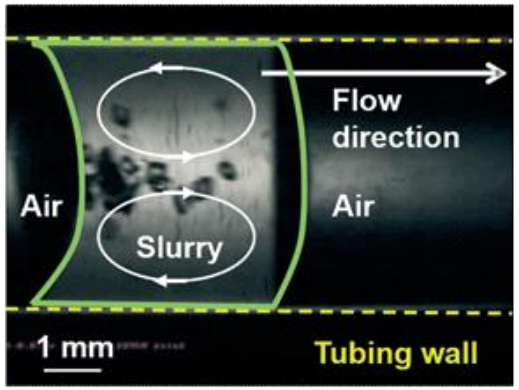


Hyperspectral Imaging



Optical Imaging and Video

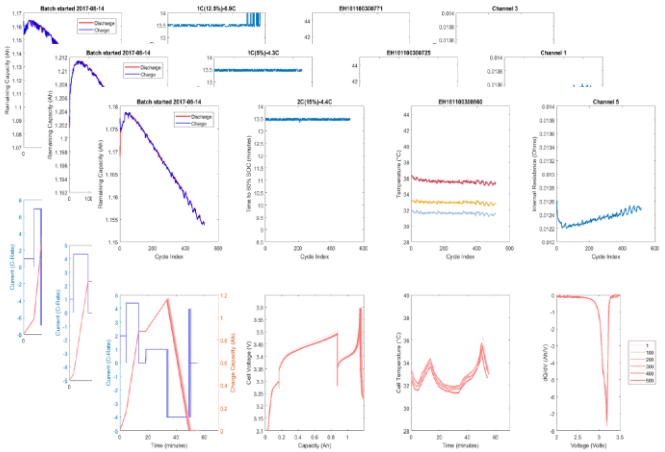




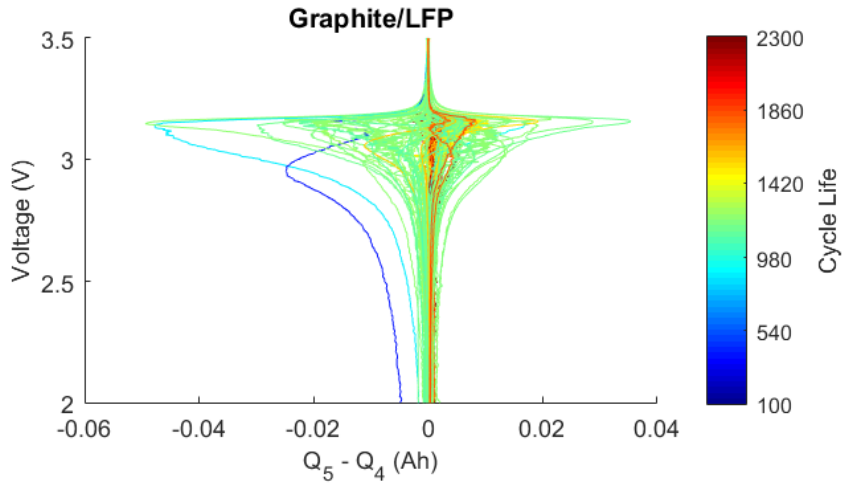
Course only uses authentic data

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

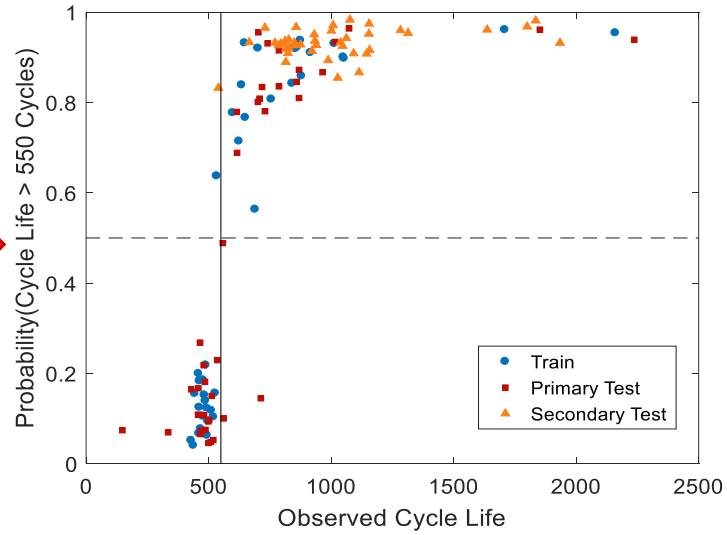
Experimental Cycling Data



Feature Engineering & Elastic Net

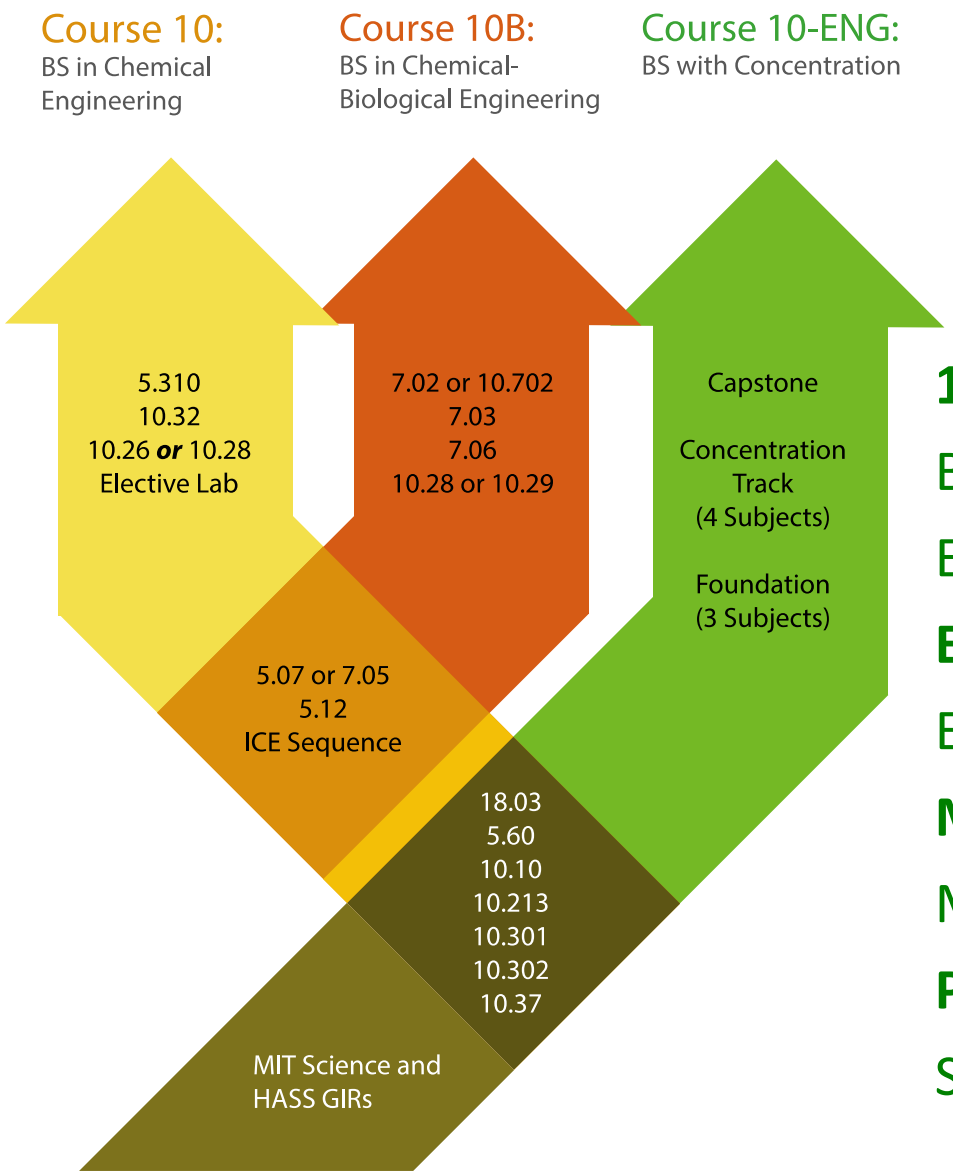


Classification Modeling



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

MIT Curriculum: A ChE Specialization in Process Data Analytics



10 ENG options

Biomedical

Energy

Engineering Computation

Environmental

Manufacturing Design

Materials Processing

Process Data Analytics

Society, Engineering, & Ethics

Course Decision to Make: Which Software to Use?

- Matlab: many ChE departments teach Matlab to all of their students (saves time); many online tutorials are available, http://www.mathworks.com/academia/student_center/tutorials/launchpad.html
- Python: open-source and listed in many ChE job opening descriptions
- Commercial statistics software, e.g., Minitab: allows graduates to be immediately productive in industry

- The best choice for your course depends on the content of previous courses in the ChE curriculum and the expertise of the teaching assistants

Challenges and Synergies

- Some ChE curricula do not include linear algebra → forces the course to include linear algebra or shy away from some methods, e.g., PCA/PLS
- Some ChE curricula do not include optimization → relatively easy to include basic concepts the course
- Some ChE curricula do not include computer programming → can manage by focusing on problem-solving with software, e.g., SAS JMP, minitab
- Data education integrates well with any earlier courses in linear algebra, optimization, and computer programming

Hands-on exploration of web-based resources

- **CACHE Teaching Resource on Data, Statistics, and Analytics**
  Syllabi, textbooks, screencasts, links to other resources
  https://cache.org/teaching-resources-center/statistics
- **LearnChemE Statistics Screencasts**
  >60 short screencast videos on statistics, including examples, introduction to topics, software tutorials, and exam review problems
  https://learncheme.com/screencasts/statistics/
- **Virtual Laboratories in Probability and Statistics**
  Interactive web-based resources for students and teachers
  http://www.math.uah.edu/stat
- **Online Statistics and Probability Course**
  Video lectures, slides, homework, and practice final
  http://www.lithoguru.com/scientist/statistics/review.html