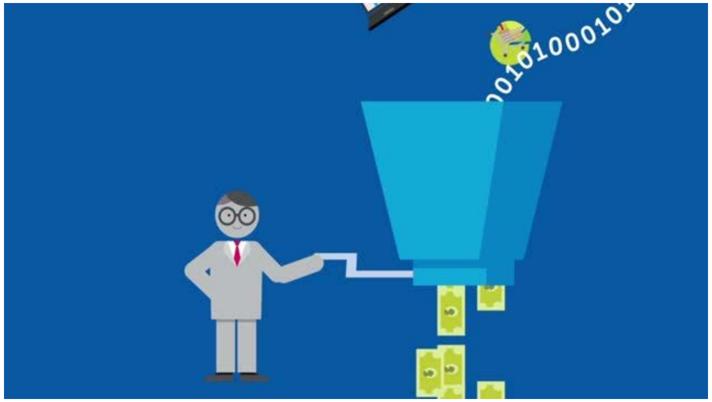
Data Science Education Using Real Data

Photo courtesy of Indeed Hiring Lab

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**

- **<u>All</u> graduates should learn data analytics**

Figures courtesy of Opera Solutions and Indeed Hiring Lab

The Importance of Authentic Data

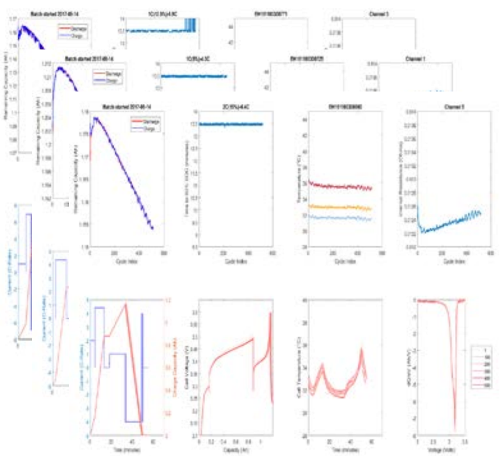
- **To produce graduates who excel at data analytics, activities must allow students to *practice***

- **Real data allows the experience to be *authentic*, so that students buy in and connect to the real world**

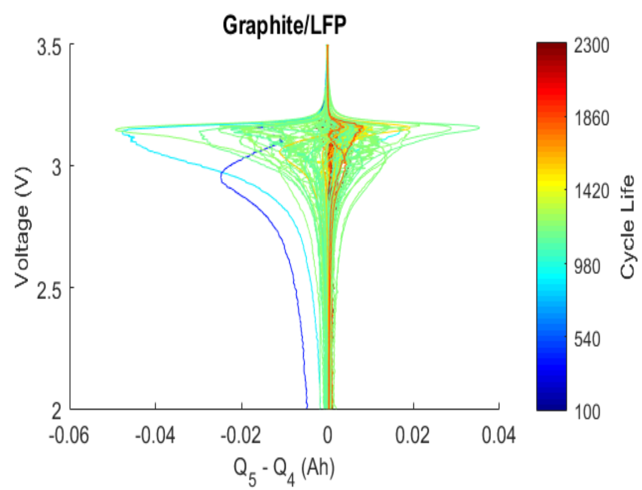
Authentic Dataset: Lithium-ion Batteries

- **Operational data for 50 commercial batteries from commercial cycler**
- **Can be used for modeling, prediction of battery cycle life**
- **https://github.com/petermattia/battery-parameter-spaces**

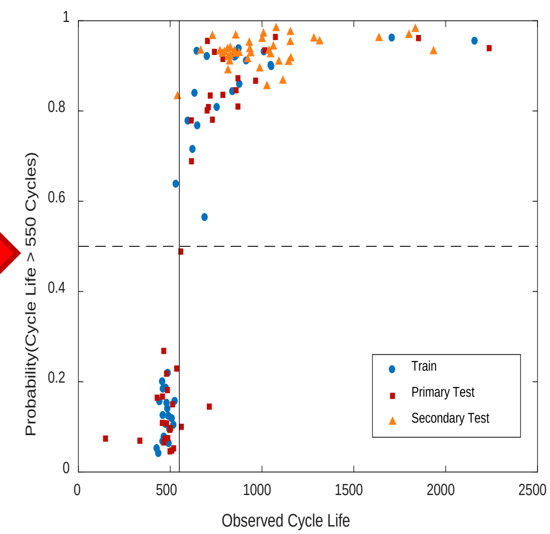
Experimental Cycling Data



Feature Engineering & Elastic Net



Classification Modeling



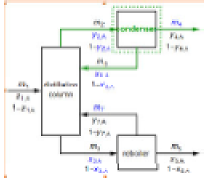
- **Will soon build a large public biomanufacturing dataset**

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

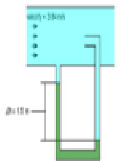
Teaching Resources in Statistics

Intro to Chemical Engineering

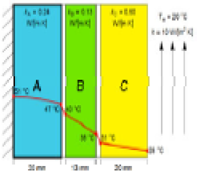
Material/Energy Balances



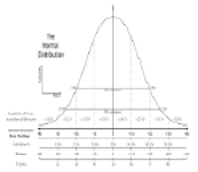
Fluid Mechanics



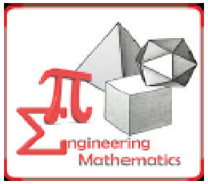
Heat Transfer



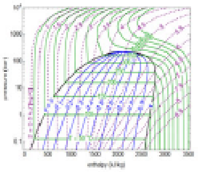
Statistics



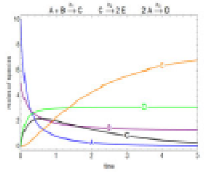
Engineering Mathematics



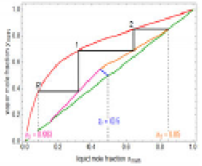
Thermodynamics



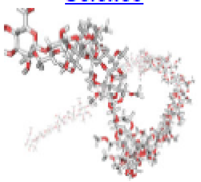
Kinetics/Reaction Engineering



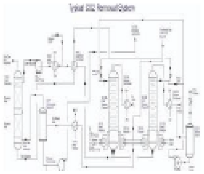
Separations/Mass Transfer



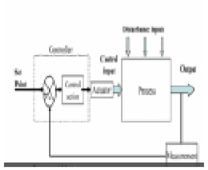
Material Science/Polymer Science



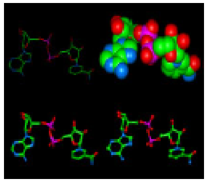
Process/Product Design



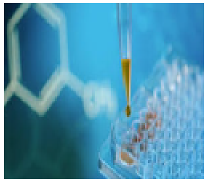
Process Control



Molecular Modeling



Bioengineering



Safety

Conventional/Renewable Energy

Teaching Topics

cache.org

syllabi, schedules, computer-aided tools, interactive simulations, screencasts, concept questions, textbook information, useful links, course notes

Sampling of Data Education in ChE Curricula

- **Undergrad data education ranges from**
  - **A few lectures in some chemical engineering course(s)**
  - **3.5 weeks in a chemical engineering course**
  - **Statistics and probability course taught by statistics/math faculty**
  - **Engineering statistics taught by a non-ChE engineer**
  - **Engineering statistics course taught by ChE faculty**
- **Graduate data education ranges from**
  - **None**
  - **Elective courses**
  - **Part of required courses**

Data Education in Undergraduate ChE Curricula

- **Universities sampled**
  - **University at Buffalo**
  - **University of Texas Austin**
  - **University of Massachusetts, Amherst**
  - **Massachusetts Institute of Technology**
- **A good coverage of different amounts and approaches used in ChE curricula**

Data Education in Undergraduate ChE Curricula

- **University at 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**

Data Education in Undergraduate ChE Curricula

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

Data Education in Undergraduate ChE Curricula

- **University of Massachusetts, 3.5 weeks to juniors**
- **Lecturer: Michael A. Henson**
- **Erwin Kreyszig, *Advanced Engineering Mathematics***
- **Probability distributions, confidence intervals, hypothesis testing, regression and correlation, factorial and fractional factorial experimental design, Matlab statistics**

Data Education in Undergraduate ChE Curricula

- Massachusetts Inst. Tech., small number of lectures to seniors in design and project courses
- Lecturers: numerous
- No textbooks
- Laboratory kinetic data and curve fitting

Data Education in a Graduate ChE Curriculum

- **MIT, 3 weeks (9 hours) to all graduate students**
- **Lecturers: Richard D. Braatz and James W. Swan**
- **Electronic lecture notes**
- **Probability theory, stochastic differential equations, parameter estimation, Monte Carlo methods, stochastic chemical kinetics**
- **Clear that most entering students do not have a basic understanding of probability and statistics**

Data Education in a Graduate ChE Curriculum

- **MIT, 3.5 weeks (10 hours) to most graduate students**
- **Lecturer: Richard D. Braatz**
- **Electronic lecture notes**
- **Statistical and model-based iterative experimental design, linear and nonlinear regression (parameter estimation), uncertainty quantification, control charts, chemometrics for sensor calibration and process monitoring, machine learning for construction of sparse models**

Data Education in a Graduate ChE Curriculum

- **Main goal: train students to be effective in translating data into making good decisions**
  - **Experimental design ⇨ generate data so that the model will be good enough**
  - **Linear/nonlinear regression ⇨ models for design & control**
  - **Uncertainty quantification ⇨ is the model good enough?**
  - **Chemometrics ⇨ handling correlated data**

Data Education in a Graduate ChE Curriculum

- **Main goal: train students to be effective in translating data into making good decisions**
  - Statistical process control ⇨ does data indicate that the process is under control?

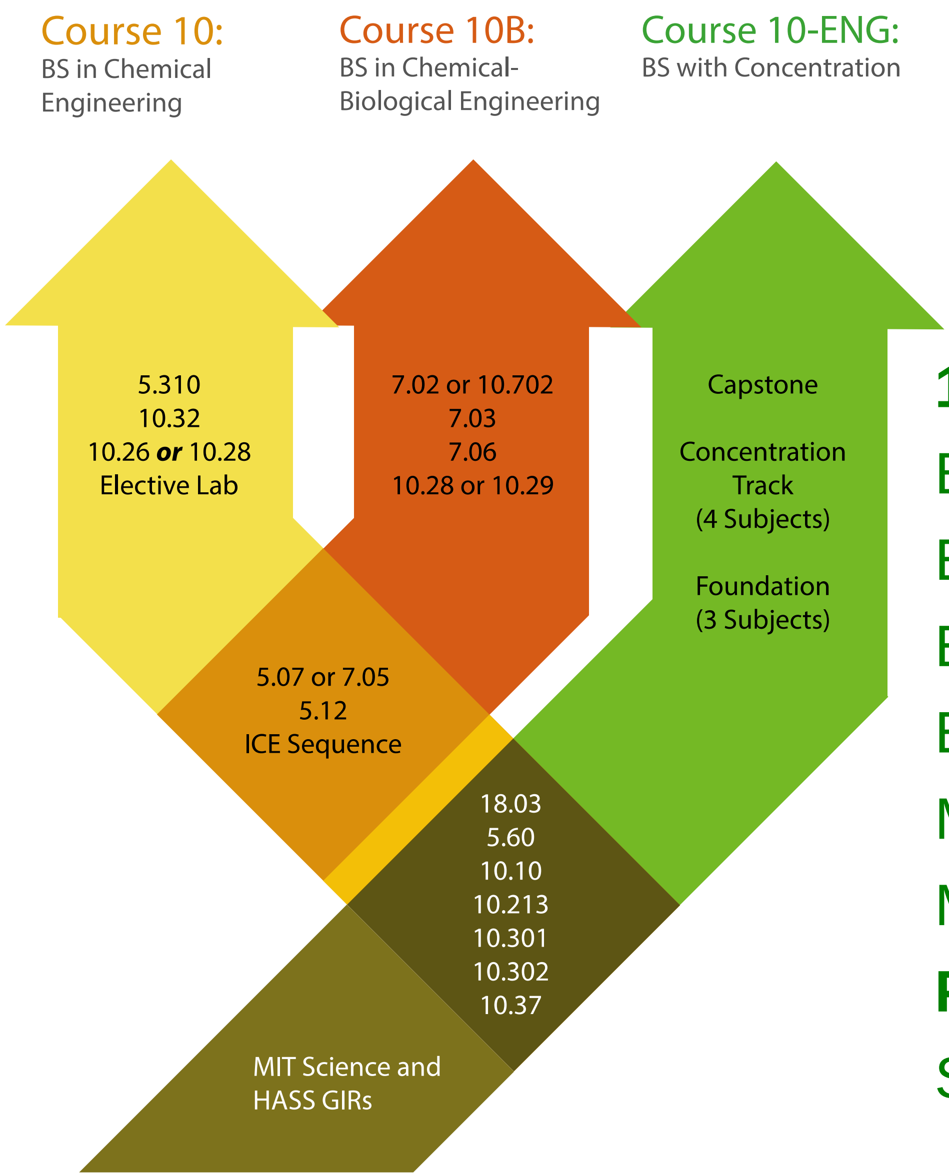
    ⇨ which variables are likely associated with the fault?

    ⇨ how do classify new data based on historical data
  - Chemometrics (i.e., principal component analysis, partial least squares) and Fisher discriminant analysis
  - Machine learning for construction of sparse models, e.g., sparse vs. dense models, lasso & elastic net methods

Data Education in a Graduate ChE Curriculum: Sensor Calibration, Regression, Uncertainty Quantification

- **Start with relating spectra to concentration**
- **Do linear and nonlinear least squares for constructing algebraic sensor calibration curves, using summation notation and matrix algebra**
- **Statistical process control: Shewart, CUSUM, EWMA, PCA-based T^2, 1D/2D contribution plots**
- **Do chemometrics for handling correlated data**
- **Do parameter estimation for nonlinear dynamic models, quantify uncertainties in parameters**

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