Rensselaer

Integrating Data Science Advances into Chemistry and Chemical Engineering Curriculums

Juergen Hahn

Department of Biomedical Engineering

Department of Chemical & Biological Engineering

Center for Biotechnology and Interdisciplinary Studies

Rensselaer Polytechnic Institute

Overview

- CACHE Perspective on Data Science
- Data Dexterity at Rensselaer
- Discussion

CACHE Perspective

- What is CACHE?
  - Computer Aids for Chemical Engineering (CACHE), is a 501(c)(3) not-for-profit organization
  - Purpose is to promote cooperation among universities, industry and government in the development and distribution of computer-related and/or technology-based educational aids for the chemical engineering profession

CACHE
Computer Aids for Chemical Engineering

CACHE Perspective

- What is CACHE?
  - Founded in 1969
  - Has 28 trustees: 7 from industry and 21 from academia
- CACHE's contributions
  - Conducts surveys on educational needs
  - Develops and disseminates class room material, tools, and case studies
  - Sponsors small, technically-focused conferences

CACHE Perspective

- CACHE convened a data analytics task force in 2018
- Task force came up with three tasks
  - Conduct survey of employers (are their needs being met) and faculty (where do/should students currently learn data analytics)
  - Hold a conference on Foundations of Process Analytics and Machine learning (FOPAM)
  - Identify opportunities for adding to CACHE's Teaching Resources web site

CACHE Survey

- CACHE survey on computing is conducted in two steps:
  - Survey of trustees (separate surveys for academia and industry)
  - Survey of faculty at dozens of chemical engineering departments and, separately their advisory council members
- Survey results for trustees are available; broader results will be disseminated in the fall
- While focus is on computing in general, some data science questions are included

CACHE Employer Survey Results

- Programming, while not directly related to data science, is a foundation for data science
- Question: Does your employer expect you to be competent in a programming language?
- Results evenly split between agree and disagree

CACHE Employer Survey Results

- Question: Do you write computer programs at work to solve engineering problems? If so what language do you use?
- Python received the most mentions followed by MATLAB, Visual Basic, and Fortran and some SQL & Pascal

CACHE Employer Survey Results

- The majority of respondents felt that new hires have sufficient training in computer applications and also in programming skills
- However, all the responses to the question "Your new chemical engineering hires have sufficient training in statistics and data analytics?" were negative

CACHE Employer Survey Results

- Question: Do your new chemical engineering hires have sufficient training in machine learning?
- There were no positive responses, but approximately half were unsure of where machine learning skills would be needed and the remainder felt that new hires did not possess the necessary skills

CACHE Employer Survey Results

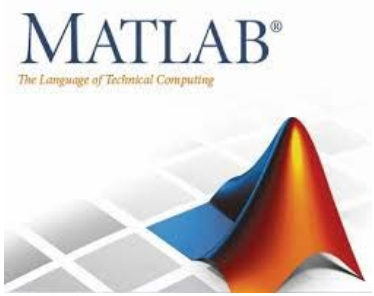
- Overall, respondents felt that new hires had the skills that they needed to be successful in their jobs, but responses for individual categories varied (see previous responses)

CACHE Academia Survey Results

- Question: Is computer programming taught to your chemical engineering students?
- Almost all responses were positive but the methods of delivery varies
  - Taught in computer science
  - Taught within the department
  - No separate course, but this is integrated into the curriculum

CACHE Academia Survey Results

- Question: If your students are taught a programming language, then please select the language that is currently being taught
- The vast majority of responses related to teaching MATLAB followed by Python.
- Only a few individual responses selected Visual Basic, Java, or C/C++



CACHE Academia Survey Results

- Question: What courses do your undergraduate students take that covers the topic of statistics and/or analytics?
- Approximately half of the responses pointed to the unit ops lab and the other half to modeling and statistics courses
  - Most of these courses were taught within the department, but approximately ¼ were taken in engineering
- Few departmental elective courses available

CACHE Academia Survey Results

- Question: What courses do your undergraduate students take that covers the topic of statistics and/or analytics?
- Approximately half of the responses pointed to the unit ops lab and the other half to modeling and statistics courses
  - Most of these courses were taught within the department, but approximately ¼ were taken in engineering
- Few departmental elective courses available

CACHE FOPAM Conference

- Foundations Of Process Analytics and Machine learning is the new triennial series of CACHE-sponsored conferences
- Aim is to provide a forum for researchers from industry and academia to discuss the current status and future directions of data analytics and machine learning in the process industries.
- Held in Rayleigh, NC, August 6-9, 2019

CACHE FOPAM Conference

Home | Important Dates | Technical Program | Workshop | Registration | Lodging | Poster Sessions & Special Issue | Sponsors | Contact

FOPAM

Foundations of Process Analytics and Machine learning
The StateView Hotel—Marriott Autograph Collection in Raleigh, NC
August 6 - 9, 2019

FOPAM 2019

Conference Registration

Important FOPAM registration and general conference information is located in this document: **FAST FACTS - updated July 30** to include Lunches Wednesday - Friday

Agenda

The **Agenda** has been updated to include Lunch Wednesday through Friday

Conference Organization

Meeting Chairs

Richard D. Braatz
MIT

Thomas A. Badgwell
ExxonMobil

Phillip Westmoreland
North Carolina State University

CACHE FOPAM Conference

- FOPAM topics:
  - Data analytics and high-throughput data processing for process health monitoring, estimation, and diagnosis
  - Machine learning for process and control design
  - Computational chemistry/materials analytics (including machine learning for molecular design)
  - Reinforcement learning and deep reinforcement learning for process operations and control
  - Synergies between process, materials, business, and supply chain data analytics
  - Data analytics education for chemical engineers

CACHE FOPAM Conference

- Summary of integrating data analytics into the curriculum discussion at FOPAM
  - A multi-pronged approach is needed to train everyone in process data analytics
  - There was interest in replacing Laplace transforms in the control course with data analytics (1/3 dynamics, 1/3 time-domain PID and plant-wide control design, 1/3 data analytics) but no textbook is available
  - Different programs would have different strategies in the future (based upon brief survey)

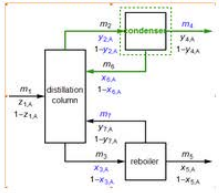
CACHE Teaching Resources

- CACHE is organizing teaching resources that can be found on the web.
- Resources include syllabi, schedules, computer-aided tools, interactive simulations, screencasts, concept questions, textbook information, useful links, and in some cases, complete course notes.
- Some resources are developed by CACHE, but the majority is contributed by the community (links to external resources are evaluated and updated 1-2 times per year)

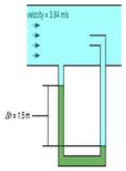
CACHE Teaching Resources

Intro to Chemical Engineering

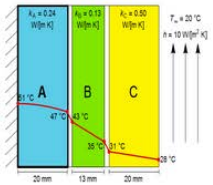
Material/Energy Balances



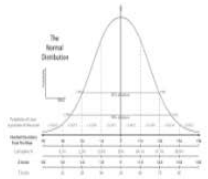
Fluid Mechanics



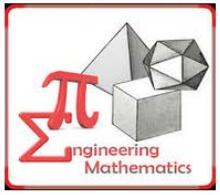
Heat Transfer



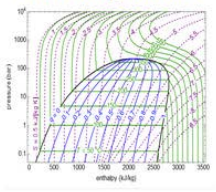
Statistics



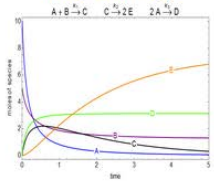
Engineering Mathematics



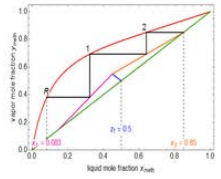
Thermodynamics



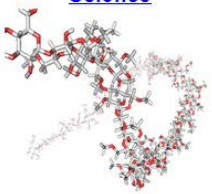
Kinetics/Reaction Engineering



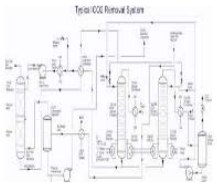
Separations/Mass Transfer



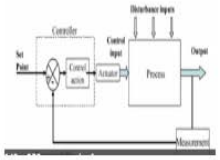
Material Science/Polymer Science



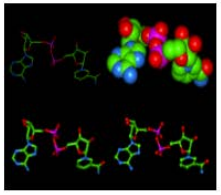
Process/Product Design



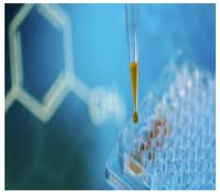
Process Control



Molecular Modeling



Bioengineering



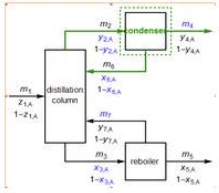
Safety

Conventional/Renewable Energy

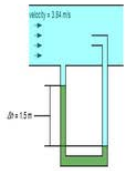
CACHE Teaching Resources

Intro to Chemical Engineering

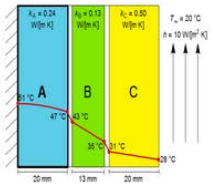
Material/Energy Balances



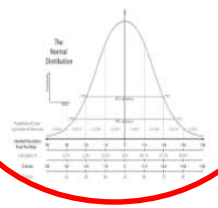
Fluid Mechanics



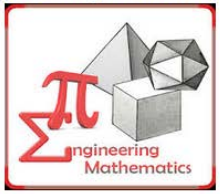
Heat Transfer



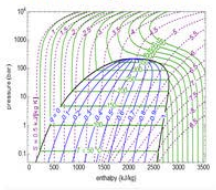
Statistics



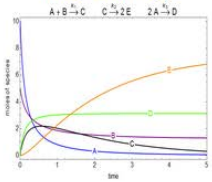
Engineering Mathematics



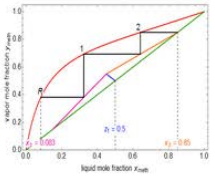
Thermodynamics



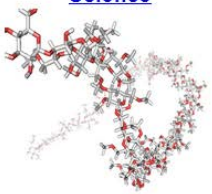
Kinetics/Reaction Engineering



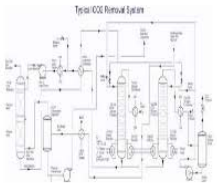
Separations/Mass Transfer



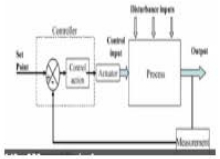
Material Science/Polymer Science



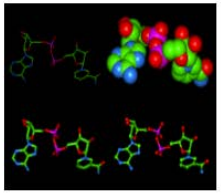
Process/Product Design



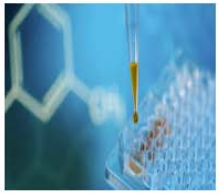
Process Control



Molecular Modeling



Bioengineering



Safety

Conventional/Renewable Energy

This is where data science resources will be found

CACHE Teaching Resources

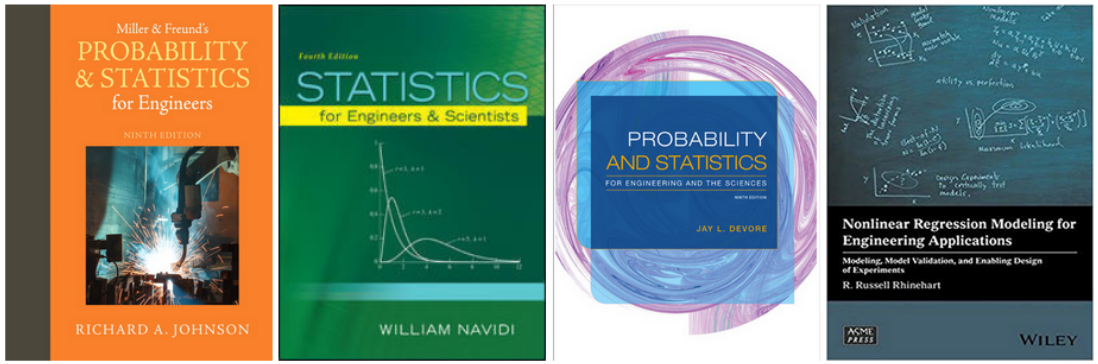
Statistics

This page is managed by **Dr. Richard D. Braatz** (**Massachusetts Institute of Technology**).

Syllabi, schedules

- **Sampling of Data Education in ChE Curricula**: An overview of strategies taken by various chemical engineering departments to educate students in applied statistics.
- **University of Buffalo**: Detailed syllabus for a one-semester course that covers theory and applications of probability and statistics appropriate for design of experiments and analysis of data.
- **University of Texas at Austin**: Detailed syllabus for a required one-semester course on applied statistics.
- **University of Massachusetts Amherst**: Syllabus for three to four weeks of applied statistics education as part of a course on mathematical modeling.
- **MIT**: Syllabus for several weeks of process data analytics education as part of a first-year graduate course on process systems engineering.

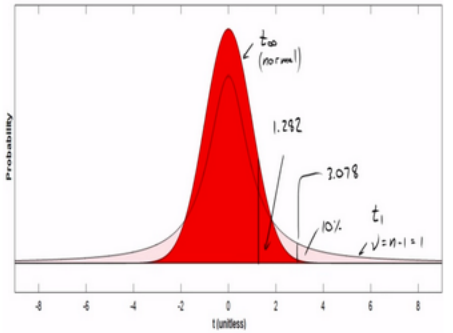
Textbooks

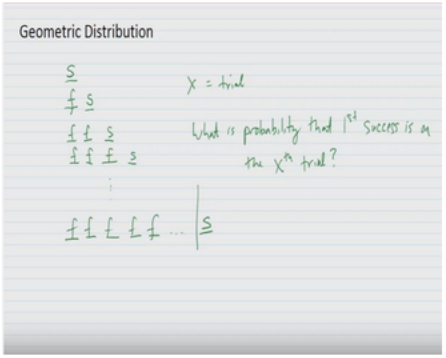


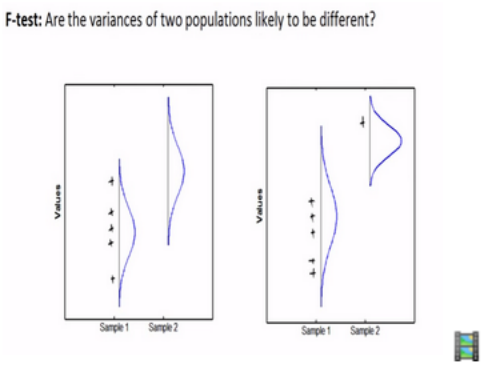
CACHE Teaching Resources

Screencasts

- **University of Colorado Boulder**: More than 20 short screencast videos on statistics, including examples, introduction to topics, software tutorials, and exam review problems. Here are three examples:







Software

- **Polymath (CACHE)**: Solves ODEs, linear equations, and nonlinear equations. **Screencasts** demonstrate Polymath use.

Useful statistics links

- **Wikipedia**
- **Virtual Laboratories in Probability and Statistics**: Interactive web-based resources for students and teachers.
- **Online Resource Materials on Multivariable Data Analysis**: Statistical methods, case studies, guidebooks, webinars.
- **Online Statistics and Probability Course**: Video lectures, slides, homework, and practice final
- **Online Graduate Applied Statistics Course**: Video lectures, slides, example problems

CACHE Teaching Resources

- Information on statistics and data analytics is not as abundant as teaching resources for other topics
- However, this will need to change in the future

CACHE and Data Science Summary

- CACHE is in the early stages of evaluating how to best support efforts in data science within the chemical engineering curriculum
- Ongoing efforts include
  - Survey of needs and offerings
  - Organizing a conference on machine learning
  - Updating Teaching Resources web site

Data Dexterity at Rensselaer

- Rensselaer has a data dexterity requirements where students of all majors need to take at least two data intensive courses
- Requirement is modeled after communication intensive requirement, but now the focus is on data
- Schools and departments decide on how to structure their curriculum to meet data dexterity requirement

Data Dexterity at SoE at Rensselaer

- School of Engineering requires that all students taken one engineering-wide data intensive course (modeling and analysis of uncertainty) and one department-specific course
- Department specific-courses are either modeling or lab courses
- Task force on Data Dexterity was convened to look at core engineering course

Data Dexterity at SoE at Rensselaer

- Charge was to identify key (data/analytics) fundamentals, methods, techniques, and algorithms that should be taught to the School of Engineering undergraduate students, focusing primarily on the first data-intensive course that will touch all engineering students
- While there was a recent related report by the National Academies, this report was mainly framed from a Computer Science perspective

Data Dexterity at SoE at Rensselaer

- Analytic foundations refers to the fundamental science that enables data to be described theoretically and analyzed empirically:
  - (a) Mathematical foundations (e.g., set theory, probability, optimization)
  - (b) Computational foundations (e.g., algorithms, data structures, simulation)
  - (c) Statistical foundations (e.g., uncertainty, error, modeling, experiments)

Data Dexterity at SoE at Rensselaer

- Data representation and communication refers to the way data is managed, modeled and integrated into workflow
  - (a) Data management and curation (e.g., data preparation, management, privacy, cleaning, database design)
  - (b) Data description and visualization (e.g., consistency, exploratory data analysis, visualization, dashboards)

Data Dexterity at SoE at Rensselaer

- Data representation and communication refers to the way data is managed, modeled and integrated into workflow
  - (c) Data modeling and assessment (e.g., machine learning, sensitivity analysis, interpretation)
  - (d) Workflow and reproducibility (e.g., provenance, documentation, version control, collaboration)
  - (e) Communication and teamwork (e.g., needs analysis, reporting, presentation)

Data Dexterity at SoE at Rensselaer

- Ethical use refers to the use of data in one or more target domains as undertaken within some ethical framework
  - (a) Domain-specific considerations (hooks to domain, "connector" courses)
  - (b) Ethical problem solving (e.g., misuse, just use, privacy)

Data Dexterity at SoE at Rensselaer

- Implementation involves modifying existing Modeling and Analysis of Uncertainty Course to
  - Include a large data example which will be revisited throughout the course and ultimately results in a project
  - Include aspects that are currently not taught (e.g., visualization, intro to ethics)

Data Dexterity in Chemical Engineering

- Approach for data-intensive course(s) at the department is to integrate some elements of data handling into multiple courses
  - Intro to Computational Chemical Engineering
  - Chemical Process Dynamics & Control
  - Chemical Process Design

Data Dexterity in Chemistry

- Given that the chemistry department is in School of Science, their requirements are different than for engineers
  - General Chemistry II to include new data exercises and expand data treatment
  - Add one credit to existing Instrumental Analysis course to cover statistical data treatment and data sets

Data Science Minor at Rensselaer

- Rensselaer has recently introduced a minor in data science and engineering

RPI News

July 31, 2019

New Minor in Data Science and Engineering Offers Students Key Skills

All Rensselaer undergraduates will be able to pursue new minor beginning fall 2019

By Torie Wells

TROY, N.Y. — How can communities better mitigate the risks of natural disasters? How can health care be delivered effectively and efficiently? How can small businesses work with other entities to save money and time, in the delivery of services? The answers to such questions, which are sure to be faced by future leaders, will be found more easily through the use of data and mathematical models.

Beginning in the fall semester, all undergraduates at Rensselaer Polytechnic Institute will have the opportunity to pursue a multidisciplinary minor in Data Science and Engineering. The course of study will

Data Science Minor at Rensselaer

- The following courses need to be taken
  - Information Systems
  - Statistical Analysis
  - Design of Experiments
- Complemented by two elective courses in the areas of
  - Data Quality and Architectures
  - Inference and Learning

to be chosen from a list of 13 courses

Discussion

Thank you for your attention