Insights from Teaching "Data Analytics for Chemical Engineers"

A.J. Medford

Assistant Professor

School of Chemical & Biomolecular Engineering

Georgia Institute of Technology

Acknowledgements

- GT School of Chemical & Biomolecular Engineering
  - Martha Grover
  - David Sholl
  - Fani Boukouvala
  - Carson Meredith
- GT Professional Education
  - Jennifer Wooley
  - Jacky Ko
  - Fatimah Wirth
  - Zaid Sewer
- Teaching assistants
  - Gabriel Gusmao
  - Sihoon Choi
  - Ben Comer
  - Ray Lei

Engineers will not be replaced by artificial intelligence, but engineers who know how to work with artificial intelligence will replace those who don’t.

- Someone Else

Talk Outline

- **Data Science for the Chemical Industry (DSCI) program overview**
  - **Insights and Survey Results**
- Data Analytics for Chemical Engineers course overview
  - Insights and Survey Results

Background: DSCI Certificate Program

- Online Graduate Certificate in ***"Data Science for the Chemical Industry" (DSCI)***
- Open for external participants and (under)graduate students
- First class: Fall 2019 (~50 students)
- External participants from 3M, Dow, Mosaic
- Amazing interactions between professionals, undergraduate and graduate students
- 2 core courses: designed by and for Chemical Engineers
  - *"**Data Analytics for Chemical Engineers**" – AJ Medford*
  - ***"Data-Driven Process Systems Engineering"** – Fani Boukouvala*
- 2 elective courses: PhD in Machine Learning (GT), MS in Analytics (GT), MS in Cybersecurity

https://chbe.gatech.edu/data-science-certificate

DSCI Program Format

12 credit hours;

4 courses

Online courses are offered in three terms:

Fall (Aug. – Dec.),
Spring (Jan. – May),
Summer (May – Aug.)

Courses are 16 weeks long in Fall & Spring, 11 weeks in Summer

100% online courses

Lessons can be viewed at any time during the week once released

Instructional team will have weekly live office hours to answer questions

Online proctoring system used for exams

Students may take off a semester when needed

- Flexible program ensures accessibility for industry professionals

DSCI Program Goals

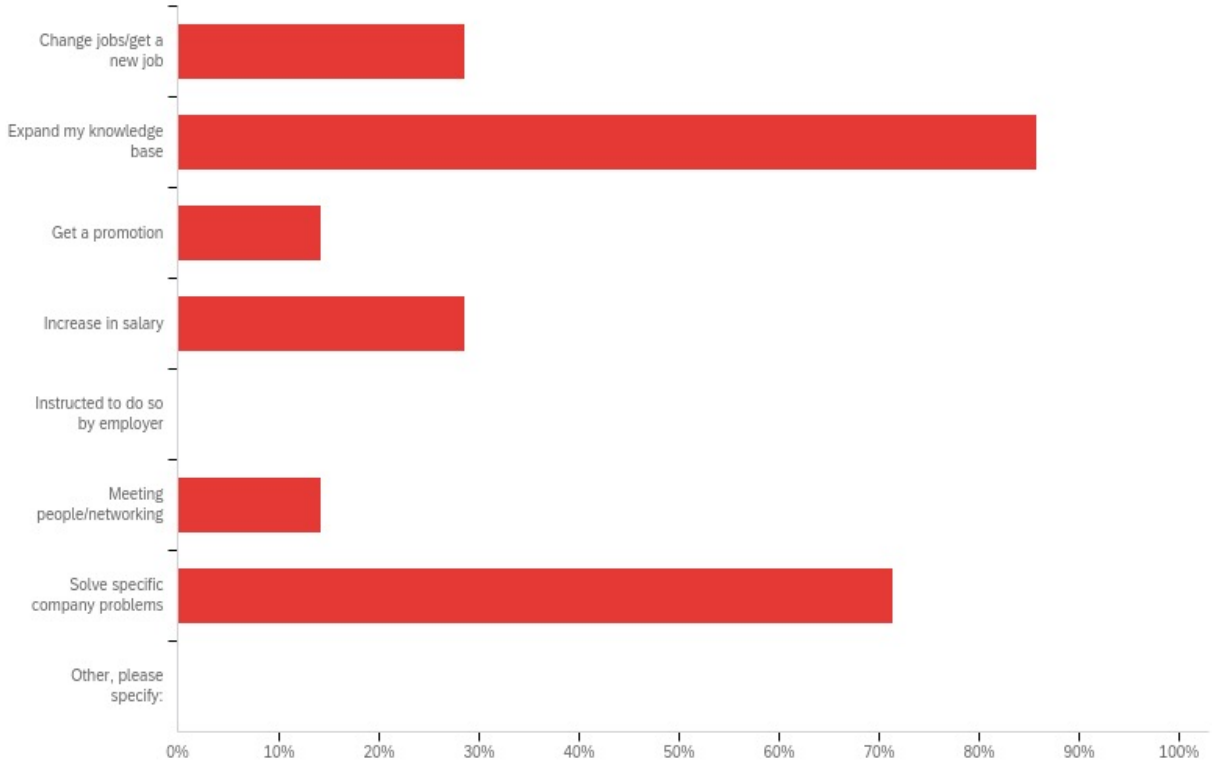
- Accessible to students with little or no background in programming, optimization, statistics
- Provide chemical engineers with “literacy” in data science, machine learning, and computational techniques
- Designed to be able to train new engineers or re-train current engineers
- Graduates will not be experts in these topics, but will be able to effectively formulate problems, apply basic techniques and communicate with experts

Technology and Infrastructure

- Python programming language
  - numpy, scikit-learn, matplotlib, etc.
- Jupyter Notebooks for lecture notes and assignments
  - Notes posted to Github
- Canvas + Vocareum
  - Tools for hosting assignments and auto-grading
- Peer grading
  - Homeworks use peer grading to reduce workload and help students see multiple perspectives
- Virtual proctoring
  - HonorLock virtual proctoring tool records screen/room (but websites can be white-listed)

DSCI ChBE Courses are “vertically integrated”

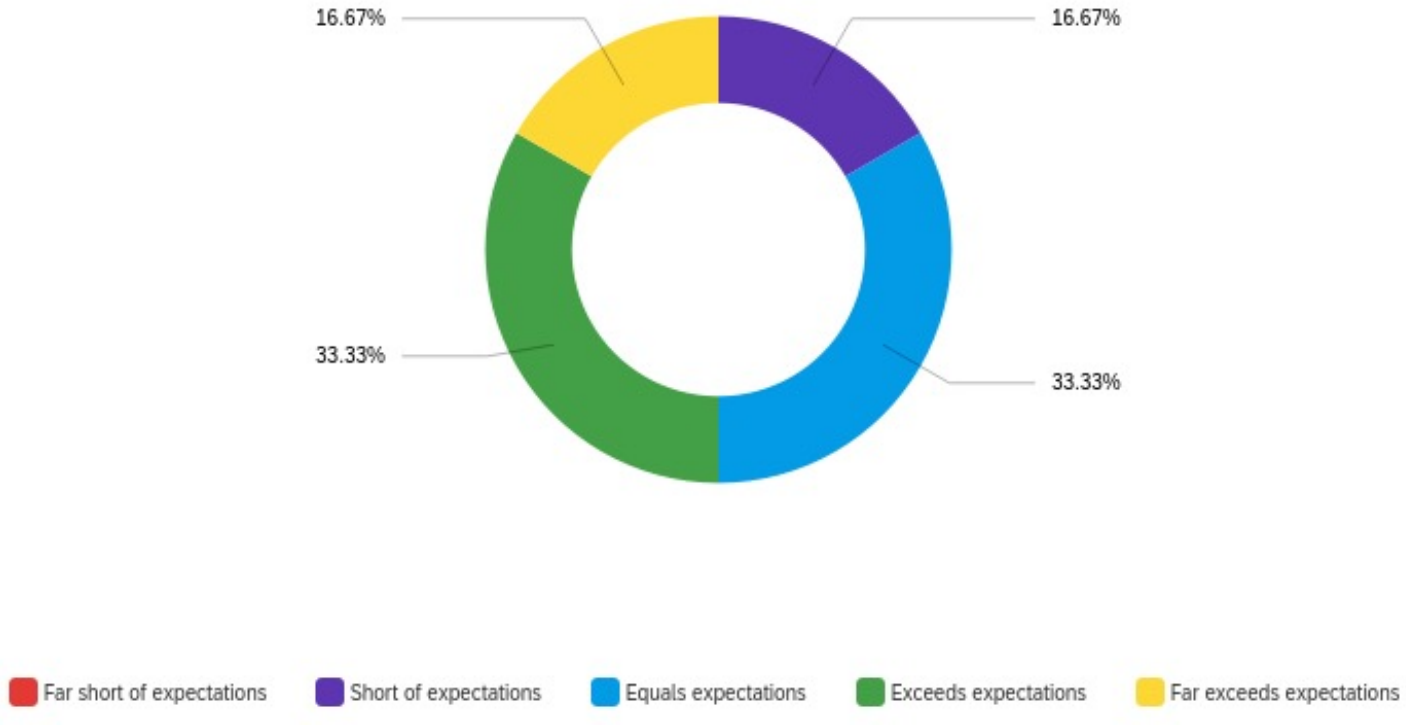
- Three distinct types of students
  - Undergraduate students (mostly junior/senior)
  - Graduate students
  - Industry professionals
- Various motives for taking the course
- Wide range of skill/experience levels



Both ChBE courses use a project-based structure

- Graduate/industry students define their own problem
  - Custom data from a real research project (data must be “open”)
  - Data from existing benchmarks, paper, or open repos (UCI, Kaggle, etc.)
- Problems must be “chemical engineering related” (broadly defined)
  - Process engineering datasets
  - Spectroscopic analysis
  - Prediction of materials properties
  - Pharmaceutical prediction
  - Water/air quality analysis
- Project groups are “vertically integrated”
  - ~1 industry, 2-3 grad, 2-3 undergrad
  - Mandatory weekly meetings and regular group evaluations
- 30% of final grade determined by project

Assessment from industry professionals (n=6)

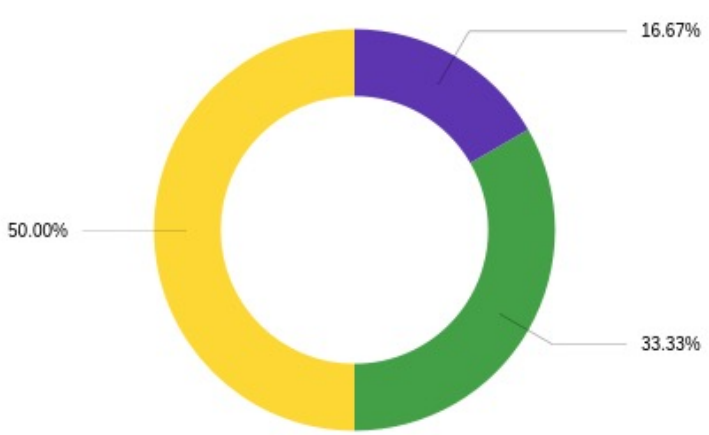


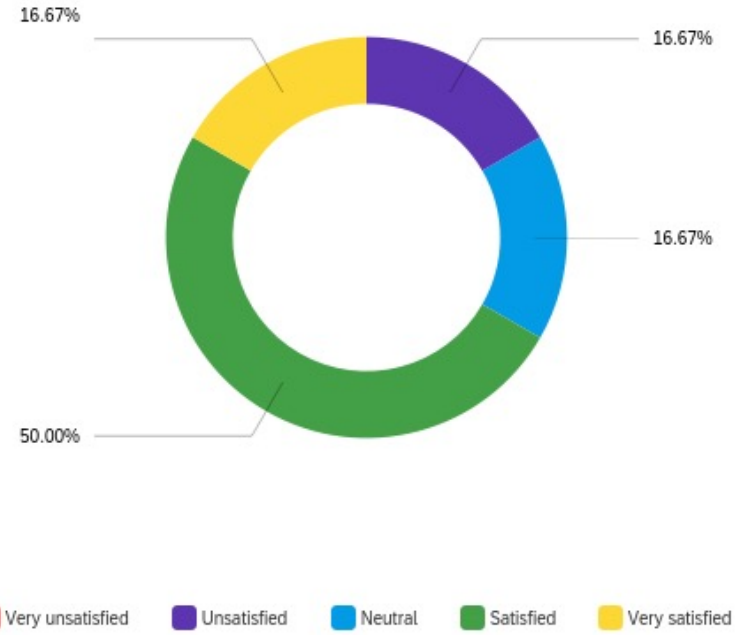
On a scale of 1 to 10, with 10 being the highest score, Rate Your Satisfaction with CHBE 6745 – Data Analytics for Chemical Engineers

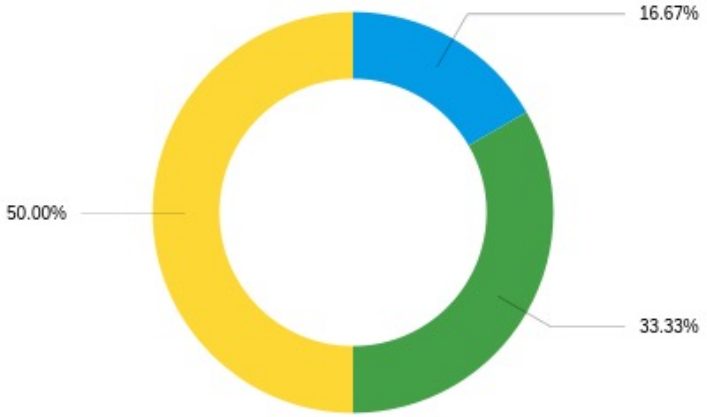
Rate your experience in the Online Graduate Certificate in Data Science for the Chemical Industry.

Online format is largely successful

Overall, Rate Your Satisfaction:







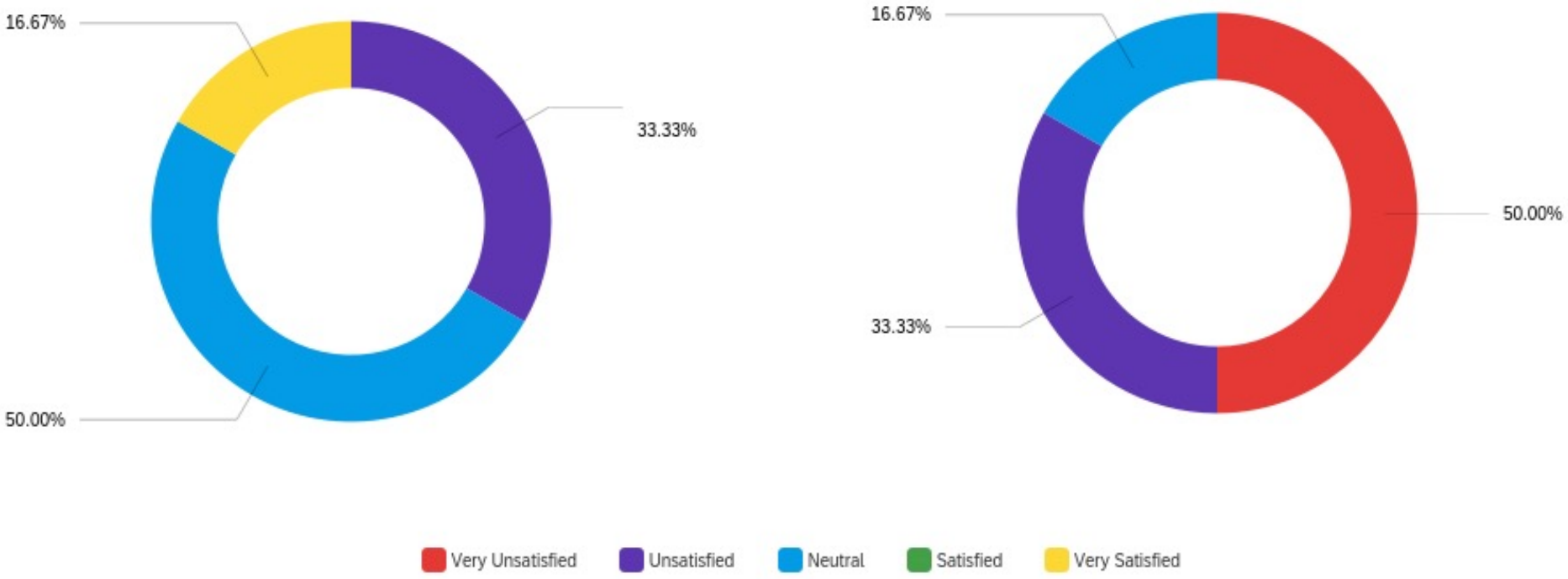
Online Lectures

Assessments
(quizzes, mid-term exams, projects, etc.,)

Number of Opportunities to Interact with the Course Instructor(s)

…but there is room for improvement

Overall, Rate Your Satisfaction:



Number of Opportunities to Interact with Fellow Students

Course Proctoring Tool (Honorlock)

Talk Outline

- Data Science for the Chemical Industry (DSCI) program overview
  - Insights and Survey Results
- **Data Analytics for Chemical Engineers course overview**
  - **Insights and Survey Results**

A brief history of “Data Analytics for ChE”

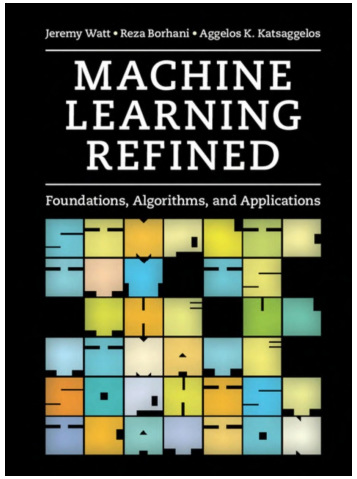
- S2018 – Advanced Data Analysis for Chemical Engineers
  - In person grad/undergrad ChBE elective
- F2018, 2019 – Data Analytics for Engineers
  - In person undergrad elective, co-taught with Eva Dyer from BME
  - Cross-disciplinary with all engineering schools
- S2020 – Data Analytics for Chemical Engineers
  - In person grad/undergrad ChBE elective (half ad-hoc online)
- F2020, F2021 - Data Analytics for Chemical Engineers
  - Online grad/undergrad/professional ed ChBE elective

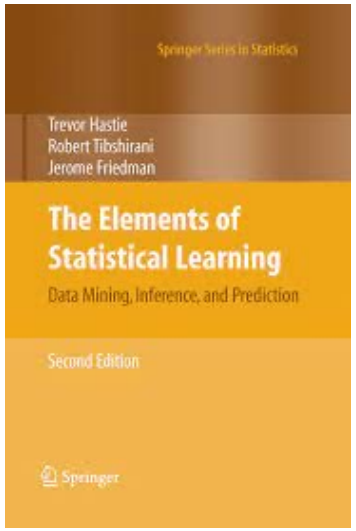
“Data Analytics for ChE” - Topics Covered

- Numerical methods (set the tone)
  - Python, Linear algebra, optimization (2 weeks)
- Regression
  - Linear regression, kernel regression, regularization, cross-validation (3 weeks)
- Classification
  - Perceptron/Logistic regression, SVMs, kNN, decision trees
- Data management (an “intermission” in the middle)
  - Data wrangling/cleaning, Pandas, APIs
- Exploratory data analysis
  - Dimensional reduction (PCA, manifold methods), clustering (k-means, mean-shift)
- Feature Engineering (Advanced/additional methods)
  - Supervised dimensional reduction, symbolic regression, feature selection, time series

“Data Analytics for ChE” – Texts and Resources

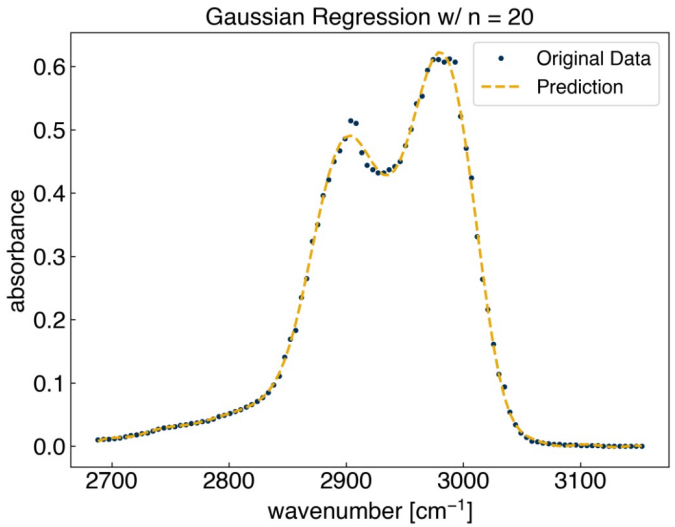
- Main Textbook:
  - “Machine Learning Refined” Watt, Borhani, & Katsaggelos (2016) (1st edition) Partially available: https://github.com/jermwatt/machine_learning_refined
- Supplementary Texts:
  - “Scientific Computing with Python” Johansson (2015) Available: https://github.com/jrjohansson/scientific-python-lectures/
  - “The Elements of Statistical Learning” Hastie, Tibshirani, Friedman (2008). Available: https://web.stanford.edu/~hastie/Papers/ESLII.pdf
  - “Numerical Linear Algebra” Trefethen & Bau (1997). Partially Available: http://people.maths.ox.ac.uk/~trefethen/text.html
- Github notes (amalgamation of blog posts)
  - Freely available: https://github.com/medford-group/data_analytics_ChE

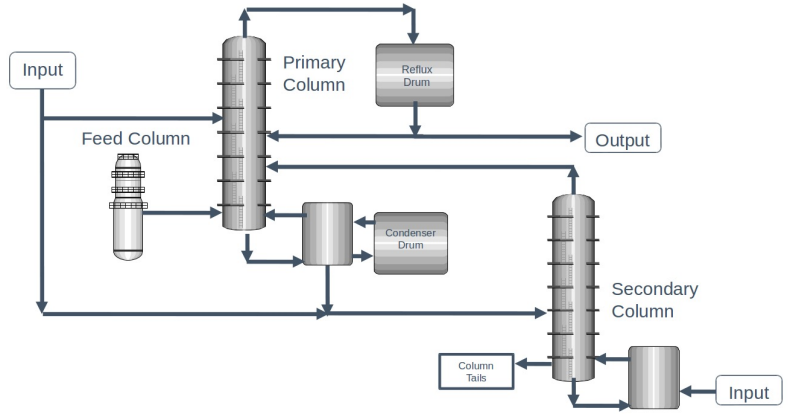




Why is it “for chemical engineers”?

- Combine toy datasets with chemical-engineering relevant datasets for examples and demonstrations
- Chemical engineering datasets used:
  - IR spectroscopy data (regression, 1D for easy visualization)
  - Dow chemical process data (regression/time series, high-dimensional data) [Thanks to Leo Chiang and Ivan Castillo!]
  - Perovskite prediction data (classification, high-dimensional data, categorical variables) [Bartel et al., 10.1126/sciadv.aav0693]
- Show examples of industry-relevant data (Dow), data typical of experimental measurements (spectra), and data from computational predictions (perovskites)
- Allow students to bring their own ChE datasets in for projects

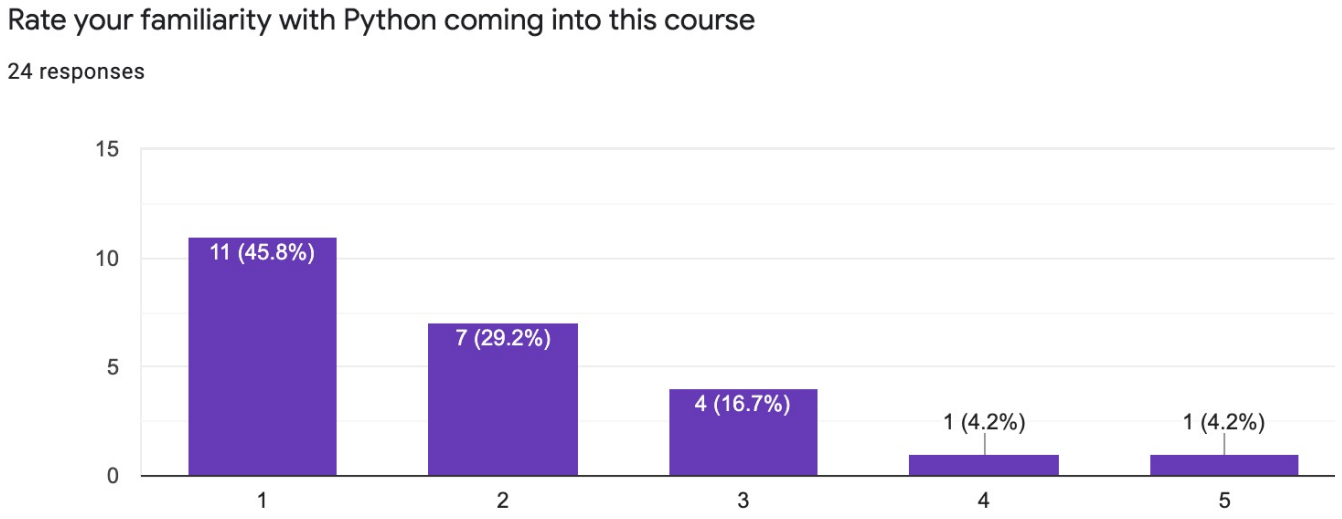


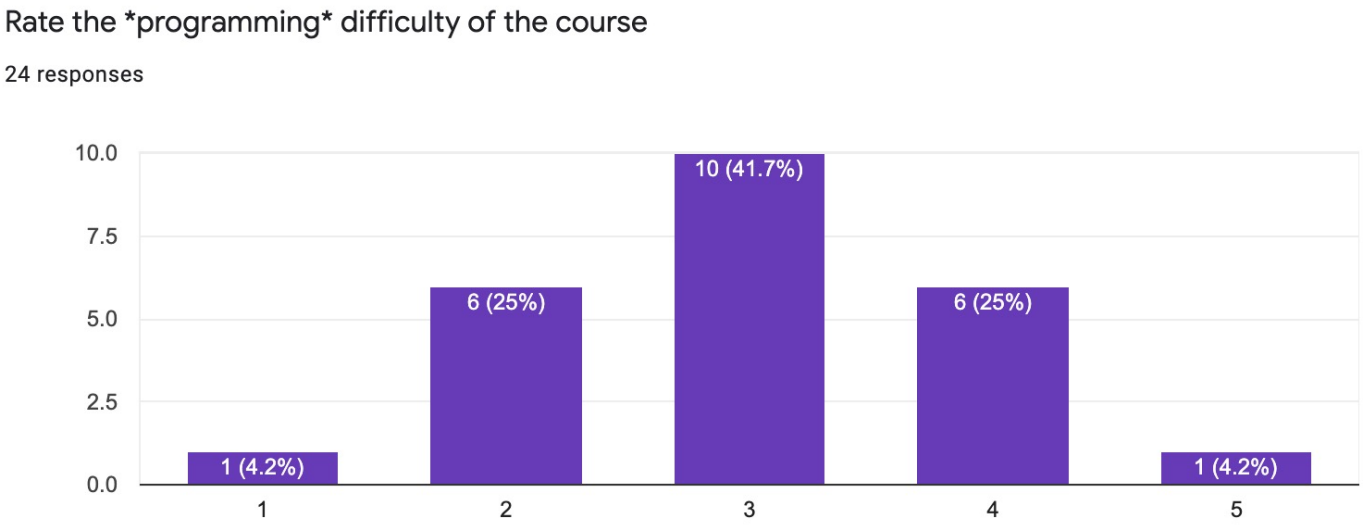


Extreme diversity in prior experience

- Prior experience in Python is not correlated with total experience!
  - Undergrads who have only had 1 programming course
  - Undergrads double-majoring or minoring in computer science
  - Grad students in experimental groups who haven't coded in years
  - Grad students in computational groups who regularly write Python
  - Industry professionals who haven't coded since they learned FORTRAN
  - Industry professionals who actively work in data science roles
- Strategies:
  - Keep the course inclusive. It should be accessible (if challenging) for students with no prior Python experience.
  - Encourage students to work together. Try to nucleate groups with at least 1-2 students with a lot of prior experience.
  - "Inverted point weighting" on exercises. Give lots of points for very easy problems, and few points for very difficult problems.
  - Use projects to keep strongest students engaged and cultivate collaboration between students of varying skill levels.

Python is a surmountable challenge



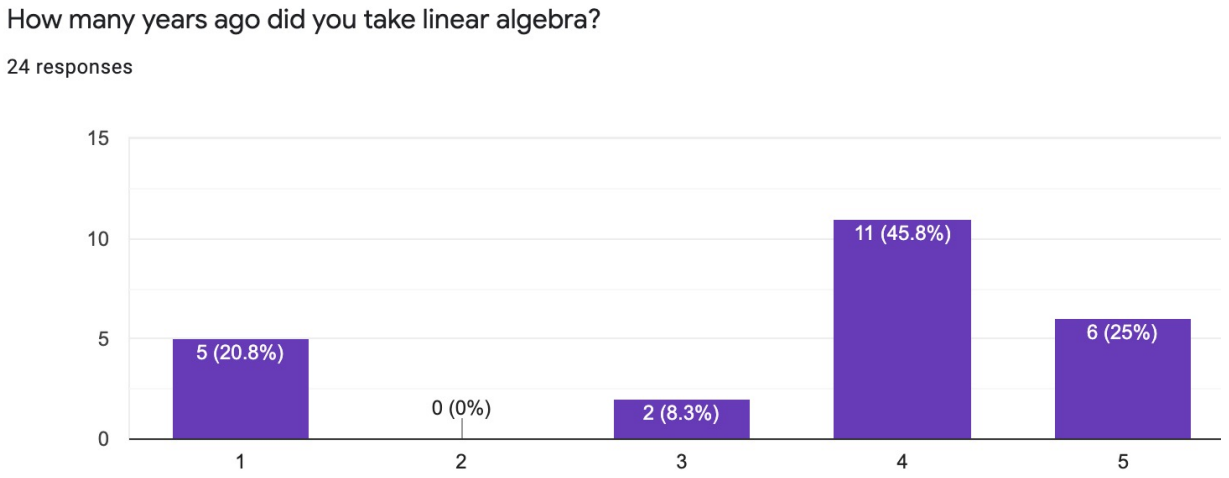


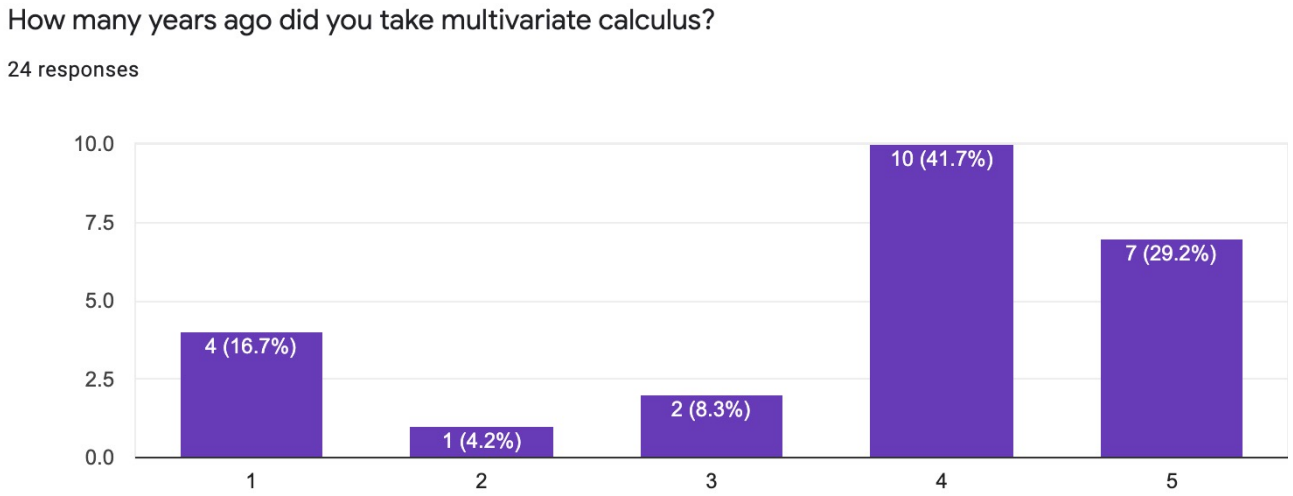
… but a Python primer can help a lot.

We now recommend that students not familiar with Python take a "Intro to Computing in Python" primer course.

"The class struck a good balance in which students that were weaker in Python/linear algebra were still able to keep up and learn a lot of new/useful skills while also challenging students that were stronger in Python/linear algebra. I really appreciated that, even though the course material moved fairly quickly and involved a lot of math that I haven't seen in years."

Students forget linear algebra & calculus





Impractical to delve deep into the math of ML models.

Pick a few mathematical points and explore them relatively thoroughly:

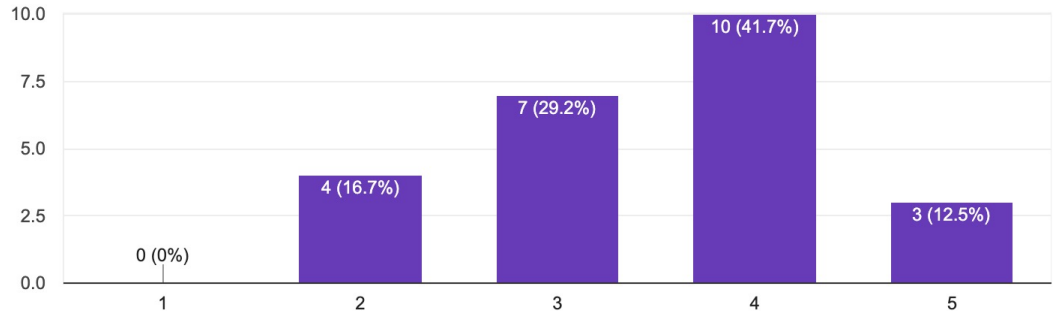
- Derivation of multi-dimensional ordinary least squares by minimizing sum of squared errors
- Creating a loss function that results in classification (softmax/logistic regression)
- Connection between eigenvectors and principal components

Primarily focus on the practical aspects of using implemented methods (e.g. scikit-learn) to solve problems.

Concepts are harder than math/programming

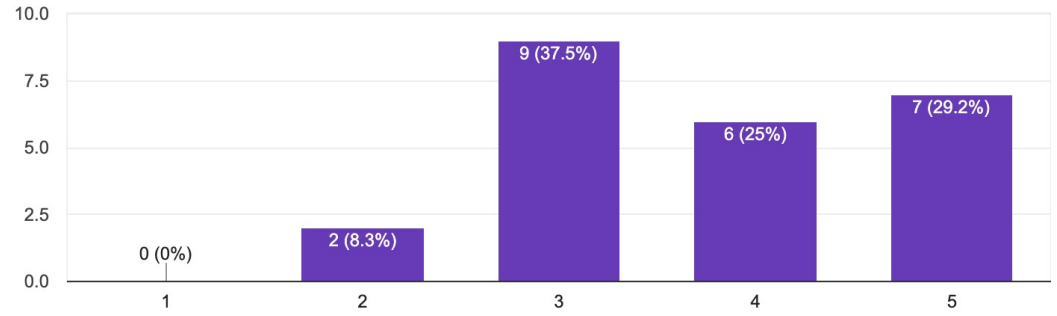
Rate the overall difficulty of the course

24 responses



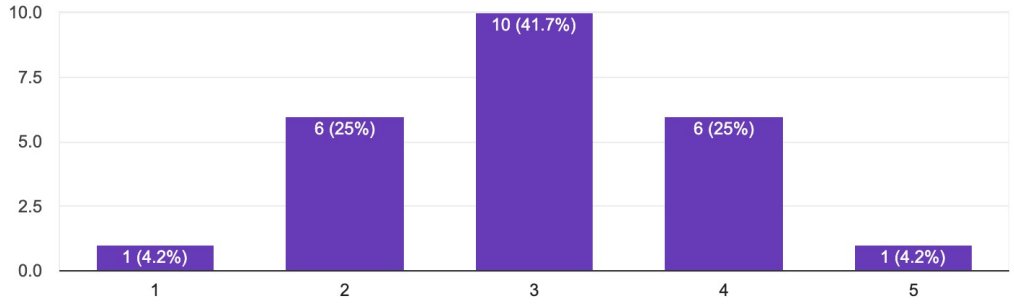
Rate the *conceptual* difficulty of the course

24 responses



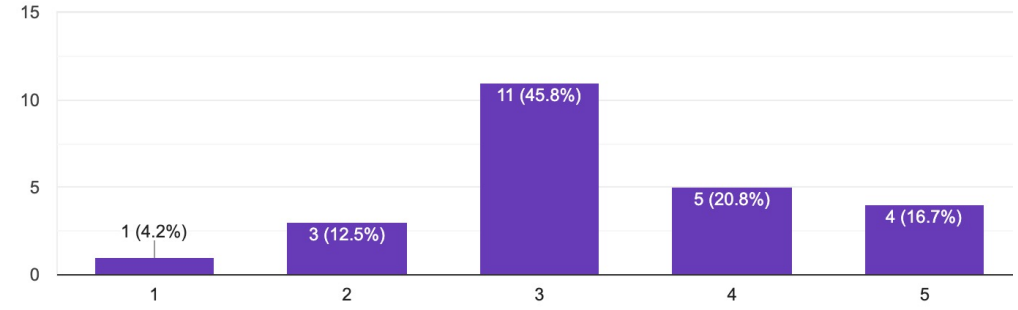
Rate the *programming* difficulty of the course

24 responses

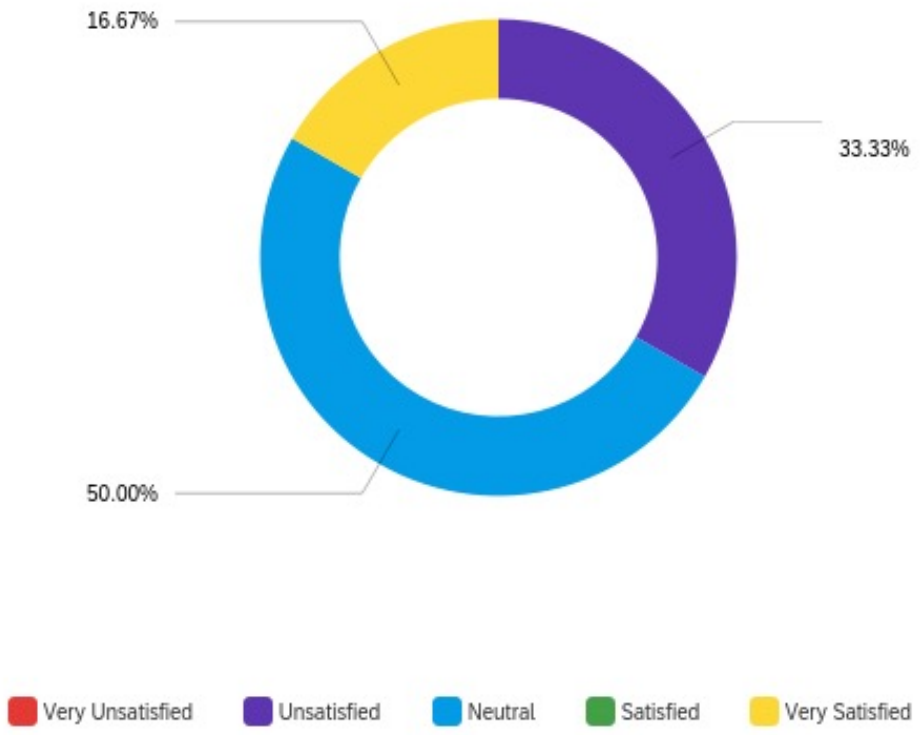


Rate the *mathematical* difficulty of the course

24 responses



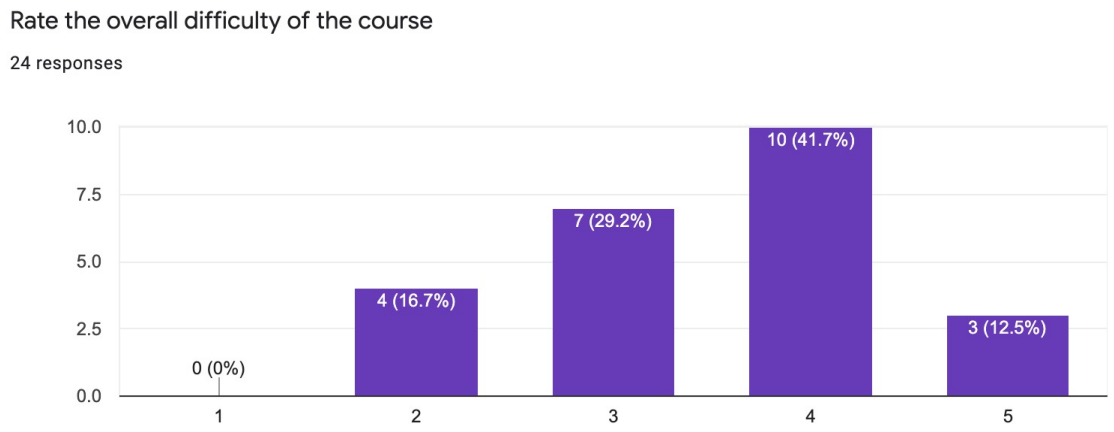
It is hard to get students to interact

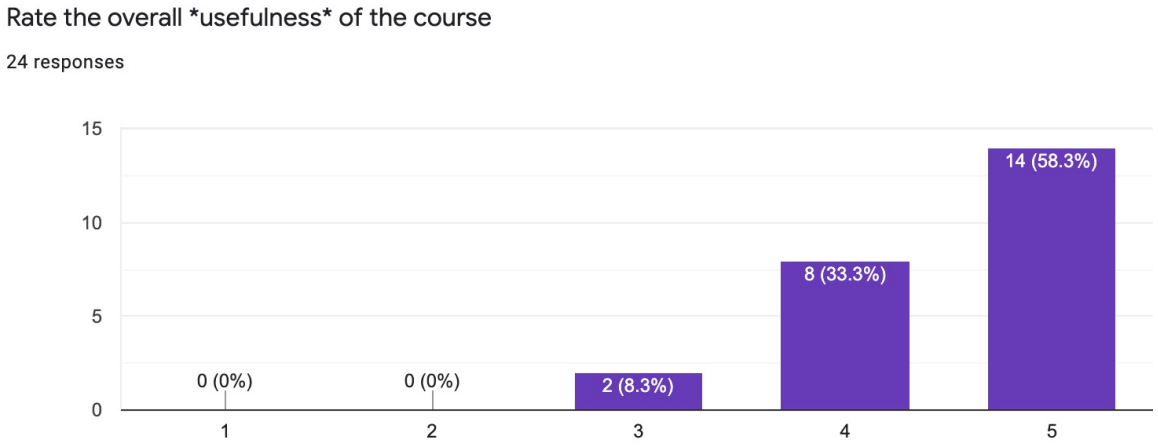


Number of Opportunities to Interact with Fellow Students

- Project and “support” groups
  - Assign groups in week 2, require weekly meetings and bi-weekly group evaluations
  - Groups are encouraged to work together more, and help each other on homeworks as needed
  - Some groups work out well, but many seem to interact as little as possible
- This was better, but still not great, even when the course was in person

The course is difficult but useful and rewarding





- Overwhelmingly positive feedback:
  - “So far, this has been one of my favorite courses I’ve taken at Georgia Tech.”
  - “I really like being able to see the real world impact of the different methods we're learning.”
  - “The end-of-term project is also a great way to put these concepts to use!”
  - “Best class I have ever taken.”
  - “The class struck a good balance in which students that were weaker in Python/linear algebra were still able to keep up and learn a lot of new/useful skills while also challenging students that were stronger in Python/linear algebra”

Good ChBE datasets are hard to come by

- Need datasets that:
  - Are curated and organized
  - Use standard and open data formats
  - Are open access
  - Ideally can be accessed via a single API
  - Ideally have some existing tutorial/background materials
- Existing repos/sources:
  - UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php)
  - Kaggle (https://www.kaggle.com/datasets)
  - NIST databases (e.g. NIST Webbook IR spectra)
  - OSI “case studies” (https://explore.osisoft.com/fandb_nurture/better-beer)
  - Literature papers
- Possible opportunity for ChBE community
  - https://cache.org/teaching-resources-center

Teaching online is bittersweet

- Advantages:
  - More accessible for students, **enables industry participation**
  - More flexible schedule for instructors
  - Allows students to work at their own pace (good for diverse skill levels)
- Disadvantages:
  - Lack of face-to-face interactions
  - Technology challenges with proctoring exams
  - Massive up-front cost to pre-record all lectures
  - **Difficult to incrementally modify content**

Engineers will not be replaced by artificial intelligence, but engineers who know how to work with artificial intelligence will replace those who don’t.

- Someone Else