C-HACK: A DATA SCIENCE TUTORIAL AND HACKATHON FOR UNDERGRADUATE STUDENTS IN CHEMICAL ENGINEERING

Evan A. Komp[1], Brenden Pelkie[1], Nida Janulaitis[1], Michael Abel[2], Ivan Castillo[2], Leo H. Chiang[2], You Peng[2], David C. Beck[1] and Stéphanie Valleau[1]
1. University of Washington • Seattle, WA 98195
2. The Dow Chemical Company • Lake Jackson, TX 77566

INTRODUCTION

Data science is becoming increasingly ubiquitous in nearly every industrial field[1,2] and has found traction in chemical engineering[3–5] (ChE), yet instruction is sluggish in primary education and only recently beginning to diffuse out of computer science coursework in higher education.[6] Current strategies include incorporating novel techniques into existing coursework and targeted approaches such as data science specializations, but the norm is not yet matching the need in industry.[7–9] Resources for students to seek out on their own, such as online coursework or articles, often describe methodologies in the contexts of traditional computer science and do not provide case studies relevant to chemical engineers. Data science is an umbrella of topics and tools that need to be discussed in the context of their application.[10] The disparity between the applications of data science to ChE topics and their joint instruction calls for more opportunities for students to learn and apply data science to ChE problems.

From a pedagogical standpoint, active learning has received a lot of recent discussion.[11–14] The stated primary objective of active learning is to engage the student in the learning process. This goal seems intuitive and straightforward, but it is difficult to prove with evaluated outcomes that active learning methods are always the optimal choice. Improvements in learning outcomes are associated with more subtle design choices, such as cooperative settings, instruction during problem solving, and centering learning around an inquiry or problem.[15] Subsets of active learning methodologies, such as Inquiry Based Learning (IBL)[16] and Problem Based Learning (PBL),[17] focus on providing students with an exploratory task through which they discover key concepts and effective strategies for themselves. The core necessities for these types of methods to be beneficial are to provide students with sufficient motivation and give them an opportunity to collaborate.[18] We have already seen project-based coursework being used to teach data science;[19] the challenge remains integrating it in the contexts already being explored by ChE students. One method that could be used is "coopetition,"[20–22] where students work in teams on a problem with access to learning resources.

While designing and incorporating active learning of data science into chemical engineering curricula would improve students' skillsets, reforming degree tracks requires a lot of planning and can be slow. Optional learning opportunities,[23,24] such as the herein presented chemical engineering hackathon (C-HACK), provide students with intermediate agency in their learning of this subject matter. C-HACK is an optional 2-week active learning event designed to teach students data science in the context of chemical engineering. The event was conducted in 2021 and 2022. The primary objectives of the event were:

- Objective 1: Provide an opportunity for undergraduate ChE students to gain data science skills, regardless of initial level of comfort with the material, specifically targeting groups typically under-supported in computing.
- Objective 2: Give students an opportunity to work in teams to solve a current industrial problem in a limited amount of time using the skills from Objective 1.

Objective 1 stemmed from two subgoals. First, there was no bias or discrimination in terms of registration and participation in the event such that those who were not yet comfortable with writing code and working with data, including individuals more likely to feel disconnected from STEM and computing,[25] including women, still signed up. The second subgoal was that all students were able to tackle the problems in Objective 2. To fulfill the first subgoal, no experience was required, and the material was prepared assuming

participants had no data science background. Contemporarily, participants needed only an internet connection. To address the second subgoal, the first week of the event consisted of a series of interactive tutorials ranging from the basics of Python™ coding to some machine learning topics that could be used for Objective 2. Python was chosen as the coding language to teach data science as it is extensively used for data science applications and has an active open-source community.[26,27]

Objective 2 aimed to provide students with an experience in team dynamics so that they could develop soft skills sought after by industry employers. To fulfill this goal, the second part of C-HACK consisted of a timed hackathon using real data provided by Dow, Inc. ("Dow"), where students had the opportunity to work with industry representatives. The hackathon portion of the event is a reward style coopetition where participants compete in teams to tackle a problem. To recognize student achievement, participants received official certificates for their work, and top teams won monetary prizes and awards.

An overview of the event, including the organizational period, the tutorials, and the hackathon, is shown in Figure 1. In the following section we describe the detailed format of the event and assess our objectives as measured through participant self-evaluation surveys.

EVENT DETAILS

The fully virtual event took place over the first two weeks of the students' winter quarter. An overview of the event is shown in Figure 1. In what follows we describe the organizational period of the event (Figure 1a), such as how committees were formed to create the content, how the schedule was created, and how the event was advertised and incentivized. We then discuss the tools used to facilitate learning and introduce data science. Next, we detail the tutorial content and format (Figure 1b), and finally the structure of the hackathon and the problems that were presented to the participants (Figure 1c).

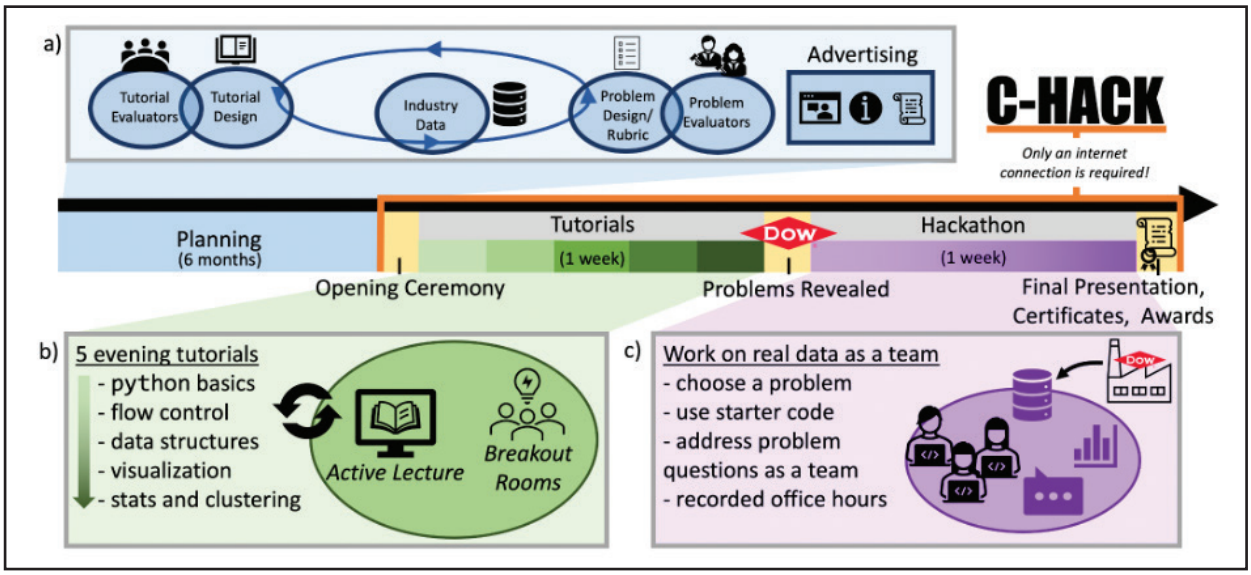

***Figure 1**. Overview of the two-week event. a) **Planning Period.** Committees were created to evaluate tutorial content, assist with tutorial delivery, and evaluate participants' work on their projects. The format and content of the event were designed to adhere to Objectives 1 and 2, including student incentives, software, scheduling, tutorial material, and hackathon problems based on industrial data. The event was advertised through fliers, virtual info sessions, and before core undergraduate class lectures. b) **Event Starts with Tutorials.** After an opening ceremony where the agenda, resources, and industry representatives were introduced, one week of evening interactive tutorials on Python and data science topics was conducted. c) **Hackathon for ChE Undergrads**. Two days after the last tutorial, representatives from Dow introduced their dataset, the story behind it, why it is important, and the coopetition problems to choose from. Participants had four days to use data science to work on the problems. Two office hour periods were open for students to ask organizers and industry representatives about the data and problems. The office hours were recorded and made available to all participants. After their work was turned in, participants gave 10-minute presentations. Rubric assessment by judges was conducted on the students' work and an award ceremony followed. The Dow Logo is a trademark of "Dow, Inc." ("Dow") or an affiliate of Dow.*

Event Planning

The event was organized over approximately six months to address the objectives. The organizational steps are shown in Figure 2. Industrial representatives were contacted to find a partner that could provide data, participate in the organization of the event, and interact with participants. Dow agreed to provide industrial data and organizational time. The team of organizers discussed an emphasis on team dynamic soft skills that industry looks for in potential hires, which led to Objective 2 and the hackathon structure. It was decided that students would work in teams to use data science to solve open-ended problems on the Dow data and present their work to judges (see section **Hackathon**). To provide motivation in the coopetition setting, 20 monetary and social incentives were included.[28] Participants received a free t-shirt and a mug, and those who turned in work were given an official certificate of completion. The certificate gave them the option to advertise their work on our website and mention their participation on their resume. Students could also receive monetary prizes and awards for exceptional work. The C-HACK event was conducted with these industry partners in 2021 and in 2022, with both years organized by the same methodology; however, the work presented in this manuscript details the specifics of the 2022 event.

We did not want the event to appeal only to those already familiar with data science in accordance with Objective 1, so the first part of the event was a series of tutorials to introduce and explore data science tools. To maximize participation, polls were given to undergraduate students to determine what logistic choices would be most appealing. These included determining what time during the academic quarter students would be most available, what times of day during which tutorials could be conducted given class schedules and student desires, rest periods between tutorials and the hackathon portion of the event, and how much time they would be willing to commit. Poll results contributed to the hackathon structure with one week of tutorials and one week of coopetition. We wanted these tutorials to be encouraged but not mandatory so that participants could join for any content they were unfamiliar with.

Volunteer committees were organized to help conduct the event (Figure 2b). To align with Objective 1, a committee of instructors was organized to produce and deliver the tutorial content, and a set of undergraduate tutorial evaluators volunteered to give feedback on quantity and clarity of tutorial content before it was delivered. A group of graduate students familiar with the tutorial content volunteered to assist with the tutorials. A committee of post-doc, faculty, and industry judges worked together to design and use a rubric to evaluate the final deliverables of the team projects. The number of organizers and approximate time commitments are shown in Table 1. Note that the number of people for some groups would scale with an event with more or fewer participants.

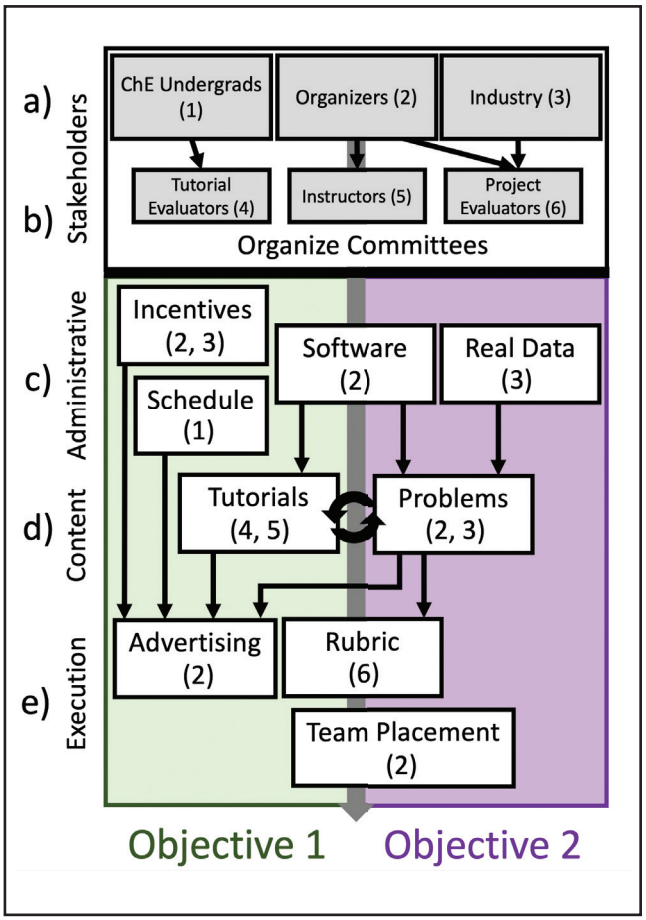

Figure 2*. Workflow of the planning of the event over about 6 months to address the objectives. a) Chemical engineering undergraduates, a group of organizers, and a group of industry representatives contributed to the event. b) Committees were created to help produce the event. A set of ChE undergraduate students volunteered to give feedback on tutorial material. Tutorial instructors, including a group of graduate students, volunteered to create and deliver to assist participants during the tutorials. Volunteer faculty, post-docs, and industry representatives evaluated hackathon problems. Box colors in a) and b) of the figure are a key for who contributed to the remaining organization. c) Administrative organization includes how students would be incentivized in a problem-based learning setting, what times worked best for students to maximize participation, what software would be easiest for participants to quickly be able to access and learn data science, and what industrial data could be used to create problems. d) The format and general content of tutorials was determined, problems were crafted based on the data and the material that could be covered, and tutorials were updated to best prepare students for the problems. e) The event was advertised to students, and final details such as a rubric to judge the final projects and how best to organize student teams were determined.*

TABLE 1
Organizers of the event and time commitments for event planning and execution.

Group	Number of Individuals	Approximate Time Commitment (After First Year)
Head Organizers	3	20 hr (15 hr)
Industry Representatives	3	6 hr
Tutorial Instructors (Faculty and Grads)	5	6 hr (3 hr)
Student Tutorial Evaluators (Undergrads)	10	45 min
Tutorial Helpers (Grads)	12	2 hr
Project Evaluators (Faculty, Post-docs, and Industry)	10	3 hr

The event was held entirely virtually, and the software tools used to conduct the event and facilitate learning were chosen to make it as easy as possible to participate (see section **Choice of Software**). Registration was free, easy, and took place through an online registration tool. Registration was open for any undergraduate student in the department of Chemical Engineering at the University of Washington as well as the Chemical, Biological, and Environmental Engineering department at Oregon State University. The event was advertised live and asynchronously on both campuses. This included fliers on campus, flash presentations before core classes, and open information sessions. We focused on motivating the students with what they would receive for participating, and the emphasis that no experience was required to align with Objective 1. A website was created as a hub for information on the event including the timeline, how to register, a code of conduct, and participation incentives. Those who registered had access to a group messaging platform to discuss the event and ask questions.

The basic content structure of the tutorials was determined to introduce data science tools on the 1-week timeline using the chosen software. To make tutorials as engaging as possible, we chose to have each be an interactive learning engagement followed by problems in small groups. See section **Tutorials** for details on the content and structure. We designed two problems for the hackathon portion of the event. One was more straightforward and the other was potentially more challenging to give teams of all levels opportunities to apply their knowledge by Objective 1. Both were open-ended questions about the data provided by Dow and aligned with their original industrial objective for collecting the data (see section **Open-Ended Problems**). A rubric was created to allow for uniform scoring of student's work and prizes to be awarded to top performing teams. Tutorials were updated in order to introduce concepts that might be helpful for working with these data and problems. See section **Connecting the Problems to the Tutorials** for details. We determined a strategy of team formation based on best practices (see section **Team Composition**).

We chose to conduct an opening ceremony before the tutorials where the timeline, format, and resources were discussed. During the ceremony, students could interact with the industry sponsors for the first time. The data and hackathon problems would be revealed two days after the tutorials ended, so that competition played no hindrance in learning and was only introduced once teams had been formed for students to work together within a coopetition setting. Documents provided to the participants, such as the agenda, instructions, and code of conduct, can be found in the event repository.[29]

Choice of Software

Jupyter Notebooks® [30] using Google Colab® were chosen to give participants access to data science tools through Python for both the tutorials and the hackathon. This means that students could run code without having to install any software; Colab requires only an internet connection and can even be used on a smartphone. Colab is also amenable to collaboration, allowing students to remotely work together in real time or asynchronously. This ease of use of this toolset makes it ideal for increasing participation among undersupported groups. Jupyter Notebooks are also a highly interactive format of coding, where the user can execute code snippets and view the results in real time.[31,32] This includes numerical results and visualizations that participants can use to guide their understanding and make meaningful conclusions. Google Drive® was chosen to share all documents and data as it is easily accessible with the participants' institutional emails and is connected to Colab. Zoom® was chosen to host all meetings as it was provided for free to students at both institutions and allowed for the recording of meetings for students to refer to later. A Slack® organization (group messaging service) was created for participants to communicate with each other and with organizers. With Slack they could discuss the format of the event or ask coding related questions. All of these tools can be accessed using just an internet connection.

Tutorials

Five two-hour tutorials were given over five weekdays in the evening on Zoom. Each consisted of 50 minutes of discussion-like active learning experiences, then a 20-minute break, followed by a set of problems that students could work on together in small groups. Attendance was not mandatory but was encouraged, so that individuals of all proficiency levels could develop their coding skills. The topics covered in each tutorial are outlined below:

- Introduction to Python – how to use Python in Jupyter Notebooks on Colab, Python syntax, variables, data types, and code commenting
- Flow control – conditional statements, loops, functions and basics of functional programming
- Data structures – organizing, storing, and working with data, NumPy [33] and Pandas [34] libraries
- Visualization – producing informative visual communications of data or results, Matplotlib [35] library
- Project dependent topic – statistics, regression, unsupervised clustering algorithms

Each tutorial was given in the form of a Colab notebook that was reviewed by the undergraduate tutorial evaluation committee. The specific content was chosen to give participants a pure introduction to coding and build up to using Python to conduct data science tasks, so that participants of any level could benefit. No 10-hour course can extensively explore data science using Python, so we focused on introducing specific tools and applications to enable students to work on the hackathon problems at a basic level and apply them elsewhere. Links to additional resources such as explanations and useful Python libraries were included to give participants more opportunity to explore. The tutorials were designed to be continuous such that each day built on previous days. To maximize continuity, we used the same dataset to explore different Python tools throughout the tutorial. The lecture notebooks contain pre-filled instructional content described by an instructor and active exercises the instructor completed with students. Questions, pauses, and tangents were highly encouraged to maximize inquiry and exploration by the students. The lectures were all recorded for participants who could not attend the tutorials. After each active learning portion and a break, students were split into small groups of 3-5 and given 50 minutes to solve coding problems related to the lecture. The small group format was chosen as it has been shown to provide the benefits of cooperation on learning.[36] Assistant instructors familiar with the material were available to give guidance and facilitate cooperation. Given that Zoom allows for screen sharing and the work is done entirely on a virtual notebook, students and instructors could display their work and progress. The tutorial notebooks are openly available in the event repository.[29]

Team Composition

We wanted the format of the teams for the hackathon to provide the best opportunity for participants to develop team dynamic soft skills. We chose to have a maximum team size of four participants. Participants were allowed to sign up as a team. In accordance with Objective 1, we also enabled individual sign-ups to encourage those without previous coding experience or interested friends to participate. We took special care to balance the solo participants who did not have teammates as follows. Each participant filled out a survey asking for their class level, gender identity, and comfortability with Python on a scale from one to five. An in-house Python tool was created to sort through participants and place them on open teams to minimize the standard deviation of all teams' "scores," which were computed as a weighted sum of participants' class level and Python skills. This is especially important because undergraduate students of any age were encouraged to participate, and we wanted to ensure that younger students had opportunities to learn from the soft skills of their older peers, and new coders from experienced coders. Finally, the tool enforces that no self-identified woman or non-binary individual was on a team alone, as these voices have shown to be overshadowed by men when singled out in STEM settings.[37] The in-house tool is rudimentarily similar to tools such as CATME, while also allowing pre-selected teams.[38] It may be worth incorporating these more sophisticated tools in the future. The Python tool is freely available in the event repository.[29] In summary, some participants signed up as a closed team, others as an open team of less than four, and others alone, where unassigned participants were added to open teams (or new teams) according to the Python tool.

Hackathon

Two days after the tutorials had concluded, the representatives from Dow announced a real dataset and two open-ended problems for teams to choose from. Documents describing the problems and the data were shared with the participants via Google Drive. A folder was shared uniquely with each team for them to work together in and turn in their work. Teams had four days to address the problem statements using Python and provide the work they completed in the form of a Colab notebook and a short one-page description of what they did. Finally, two days after the work was turned in, the students were asked to give a 10-minute presentation on their work to a collection of evaluators. The evaluators gave scores to teams based on their coding work and final presentation according to a rubric. Scores were normalized and averaged to award first through third place teams cash prizes for each of the two problems. Teams also received certificates for their project description and presentation. Detailed instructions and templates were provided

to the participants. These and the rubric can be seen in the event repository.[29]

Open-ended Problems

As opposed to highly structured problems, open-ended problems give students an opportunity to engage and develop their own problem solving skills. Problem solving is touted as the central theme of engineering, but it is unclear how to best design an open-ended problem to develop problem solving skills.[39] Not all engineering problems are the same, and so not all problems draw on the same set of proficiencies. Instead, they tend to be ill-structured, where not all elements of the problem are known with confidence, criteria for success may be variable or unknown, and do not possess a set solution path. Jonassen identifies problem types such as design, selection, and troubleshooting, among others.[40] To align with Objective 2, the hackathon problems should mimic the open-ended qualities listed above, with the only difference being that participants have access to professionals who have worked on the data previously.

The dataset given to the students is a set of measurements taken from 33 positions spatially distributed within a rectangular basin in Dow's East Waste Treatment Plant (EWTP). EWTP is a typical activated sludge aeration basin [41] used to treat organic pollutants before discharging effluent into receiving waters. Outfall at EWTP would intermittently experience elevated solids and total organic carbon, regulatory compliance parameters, which could result in production curtailment. The samples were collected via a drone, and measurements include temperature, suspended solids, metal concentrations, pH, among others, as well as bacterial abundance at taxonomic levels from kingdom down to species. Some classes of bacteria could not be identified so remained classified as "Unknown". The true identities of the bacteria and metals were replaced with encodings for proprietary reasons. Code and a starter notebook were provided to the participants to expedite data loading from file and aggregation to a desired taxonomic level.

The students were asked to use this information to explore relationships within the basin and comment on potential causes. In Problem 1 participants were tasked with producing a metric for basin performance and using this to visualize the operation of the basin over space. This provides a means for participants to make qualitative assessments based on their own metrics. In Problem 2 participants were asked to explore the distribution of bacterial abundance across the basin, identify any heterogeneity or clusters in the bacterial population, and identify any association between bacterial abundance and metal concentrations. Data science methods discussed in the tutorials could give insight on these tasks, but there was no "known" final answer that the students had to reach. Instead they had to make and justify their own conclusions. The detailed description of the basin, problem statements, and provided code are available in the event repository.[29]

These two problems fall in the category of troubleshooting; the waste basin is experiencing issues, and the participants are asked to investigate the cause. They are mostly unstructured in that there is no fixed solution path, and there is no strictly defined definition of success. Participants must explore the data and come up with their own strategy to understand the basin's function. We would like to note that the focus of this communication is not on the specific problem. The most important aspects of the problems are the use of industrial data to explore a real problem and the open-ended nature of the tasks. This gives teams the opportunity to use the tools available to them and their intuition to extract relevant information from the real world in IBL fashion, as they would if working on a team in industry. Readers looking to conduct a similar event should focus on finding an industry partner that can provide a dataset with these qualities, not the particular problem presented here; a willing industrial partner that does work related to the chemical engineering topics covered in the department will be able to help craft a good hackathon problem. The students had the chance to interact with the industry sponsors at the opening ceremony, and to chat about the specifics of the problems during Zoom office hours and project presentations.

Connecting the Problems to the Tutorials

To provide the participants with knowledge of potentially useful tools related to these problems, the tutorial content was created with the problems in mind. Specifically, Tutorial 4 gave particular emphasis to visualizing data in space and in creating three dimensional plots. Tutorial 5 covered some basic statistical topics and discussed unsupervised machine learning, including basic clustering algorithms. Lastly, the dataset used throughout the tutorials was a record of storm events in the United States from the National Oceanic and Atmospheric Administration (NOAA) archive,[42] which was formatted in latitude, longitude, and space similar to how the problem dataset was formatted in depth, width, and length of the basin. These choices ensured that the participants were introduced to tools that might be helpful when addressing the problems.

RESULTS AND DISCUSSION

In total, 107 students from both campuses (22 OSU, 85 UW) participated at some level, and 67 in 19 teams completed their team-based activities in the coopetition. Before and after the event, students were asked a series of eight questions via a survey to understand how well we accomplished the event goals, listed in Table 2. A PDF of the survey can be found in the event repository.[29] Qualitative questions (all

but the first question) were given on a scale from 1 to 5. For the first question ("Gender: How do you identify?") students were given options of "Female", "Male", "Transgender", "Gender variant/non-conforming", "Prefer not to disclose," and "Other." These options were not mutually exclusive. We refer to those that identified as male as men and those that identified as female as women. Abbreviations for the survey questions are listed in column two and used to refer to the questions hereafter. The objective that the questions aimed to measure is listed in column three.

A total of 79 participants responded to the survey before the event; 30 identified as men, 47 identified as women, and two identified as "Other" or "Prefered not to disclose." Given these responses, we cannot meaningfully comment on the effect of the event on nonbinary individuals, but if we compare the ratio of women to men to the overall population in ChE at the University of Washington (46%), we can see that feedback on the event was preferentially provided by women. It is unclear if the distribution of gender identities for those that filled out the surveys is representative of the 107 total participants.

For the first survey, the responses to the questions by women and men is shown in Figure 3. There was room for improvement by both men and women for all topics; however the students were least comfortable with the three python related topics (Data Types, Flow Control, and Python Visualization).

A total of 39 participants responded to the survey after the event concluded. The responses to the questions among the overall population before and after the event are shown in Figure 4. We see that while Objective 2 responses did not substantially change, responses to Objective 1 questions related to Python showed significant improvement due to the event. The magnitude of these improvements in terms of effect size is shown in Table 3. The p-value

TABLE 2
Questions asked to participants and the goal they measure.

Question Asked	Abbreviation	Primary Objective
"Gender: How do you identify?"		Objective 1
"I understand what variable data types are."	"Data Types"	Objective 1
"I understand the concepts of flow control in Python."	"Flow Control"	Objective 1
"I am comfortable with using Python to visualize and analyze scientific data for problem solving."	"Python Visualization"	Objective 1
"I am comfortable working in teams to solve problems by dividing the work in subtasks and communicating my progress."	"Team Projects"	Objective 2
"I am comfortable working on open-ended problems (where there is no single correct answer) to find a solution by brainstorming with my team."	"Open Ended"	Objective 2
"I am comfortable writing and orally presenting work I have done on open-ended problems."	"Communication"	Objective 2
"I feel I have opportunities to apply what I am learning in my degree on problems similar to what I will encounter in the future."	"Degree Applicability"	Objective 2

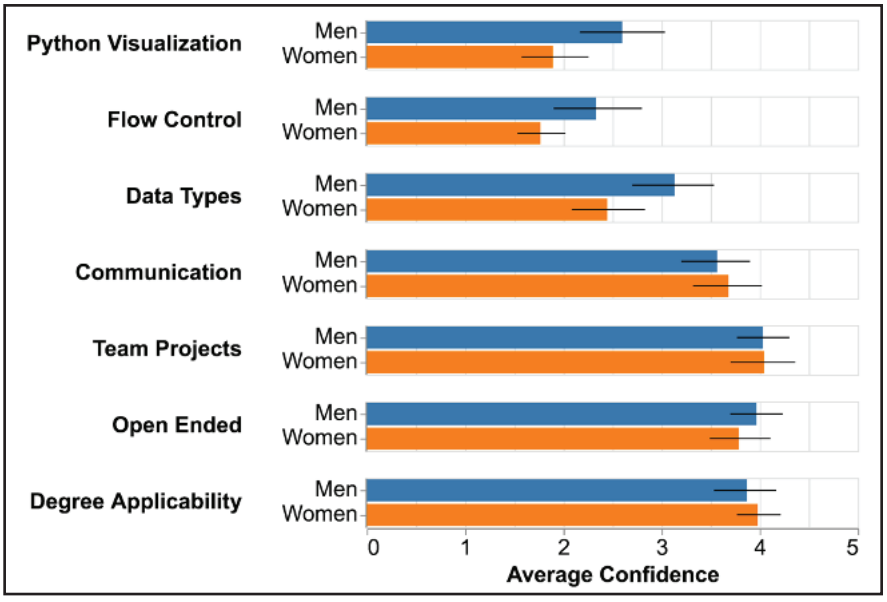

Figure 3. *Average response to the seven qualitative questions asked of participants before the event, by self-identified man or woman. The black lines are one standard deviation. The responses for the three Python related questions show the most room for improvement.*

indicated by the two-sided unequal variance t-test is shown in the third column, where statistically significant changes in response are emphasized. In order to remove the possibility of survey bias, we also computed the significance for the subset of participants that responded to both surveys in the fourth column. While the average change in response was positive for each question, only Objective 1 questions have improvements greater than 1.0 effect sizes.

Finally, to assess the change in comfortability of individuals, especially among women who are traditionally underrepresented in computer science and STEM,[37,43] we consider the participants (n=34) who responded to both surveys. Of these, 19 self-identified as women and 15 self-identified as men. Figure 5 shows the distribution of the difference in students' responses between the two surveys. The distribution of improvement, represented by positive change, is skewed most positive for Objective 1 questions. On average, the improvement for women is greater than the improvement for men, indicated by vertical lines plotted over the histogram. For Objective 2 questions, we note little to no improvement or even a reported decrease in comfortability by men for some questions. While they are not statistically significant, it is an interesting observation. This may be due to some teams having to rearrange or drop out due to participants not having time to contribute. We do note that teams that completed the projects scored well on their rubric-evaluated oral presentations, at 70.0% on average. The subsections of this score (oral presentation, teamwork, and project results) directly correspond to three Objective 2 questions. See the rubric in the event repository for details.[29]

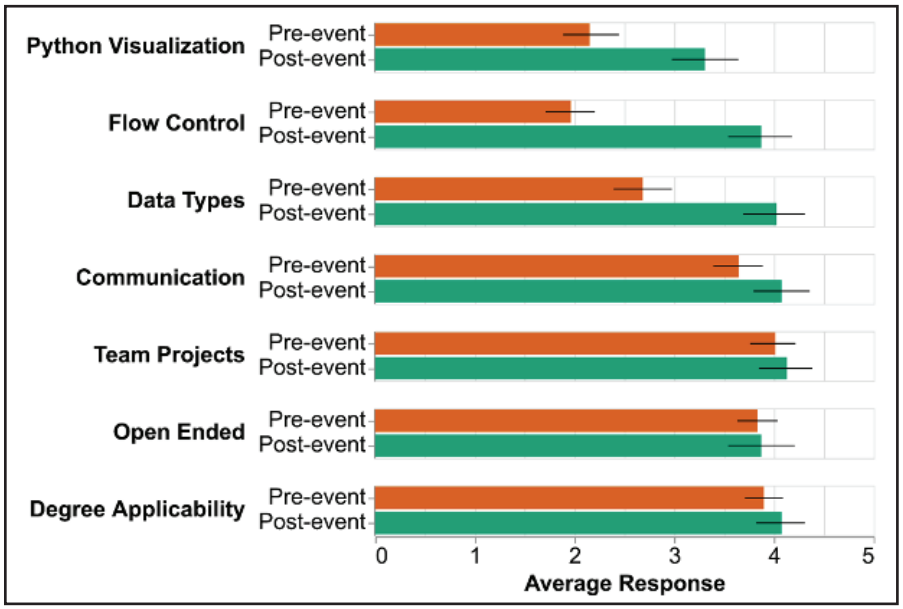

Figure 4. *Average response of all participants to the qualitative questions before (79 response) and after (39 responses) the event. Black lines are one standard deviation. Among the Objective 2 questions, we notice no improvement within confidence. For Objective 1 questions, we see significant improvement before and after the event.*

TABLE 3
Improvement in confidence for six qualitative questions after the event. Statistically significant changes in response are emphasized in bold text.

Question	Improvement[a]	p-value all response	p-value for students who responded to both
"Data Types"	1.34	**8.52 E-9**	**9.64 E-4**
"Flow Control"	1.91	**1.20 E-14**	**5.34 E-9**
"Python Visualization"	1.16	**1.02 E-6**	**1.31 E-3**
"Team Projects"	0.11	5.29 E-1	4.86 E-1
"Open Ended"	0.04	8.57 E-1	5.98 E-1
"Communication"	0.43	**3.08 E-2**	2.61 E-1
"Degree Applicability"	0.21	2.81 E-1	1.0

[a]Number of effect sizes improved, where effect size is computed as the average of standard deviations of the pre and post populations.

CONCLUDING REMARKS

We have shown that a learning-centered coopetition is an effective method for introducing Python and data science in the context of chemical engineering problems. Participants of the event reported significantly increased confidence with Python-related topics after the event, with a more pronounced reported effect for those that identify as women. This is encouraging, as women have traditionally had less support in data science related topics. Interactive tutorials on coding basics using relatable datasets help prepare students to work in teams to explore real chemical engineering problems with data science. While we did not measure substantial improvements in self-evaluated team soft skills, a coopetition provides students with an opportunity to work and communicate problems in a fashion similar

to how they would in industry. Additionally, students performed well on their rubric evaluated final presentations, which corresponded to this objective. The presence of industry representatives and monetary rewards gives students additional incentive to work through the problem. Optional events such as these allow students to learn (and prove that they have learned) data science in the context of real chemical engineering problems. C-HACK hence helps to fill the increasing demand for students with both data science and chemical engineering skills in industry.

ACKNOWLEDGMENTS

The authors would like to acknowledge funding by the Micron Foundation that made the event and this research possible. We also acknowledge Dow for providing real data to be used in the coopetition. The event would not have been possible without a number of organizers and volunteers from the University of Washington and Oregon State University and UW's eScience Institute, who facilitated learning and scored projects.

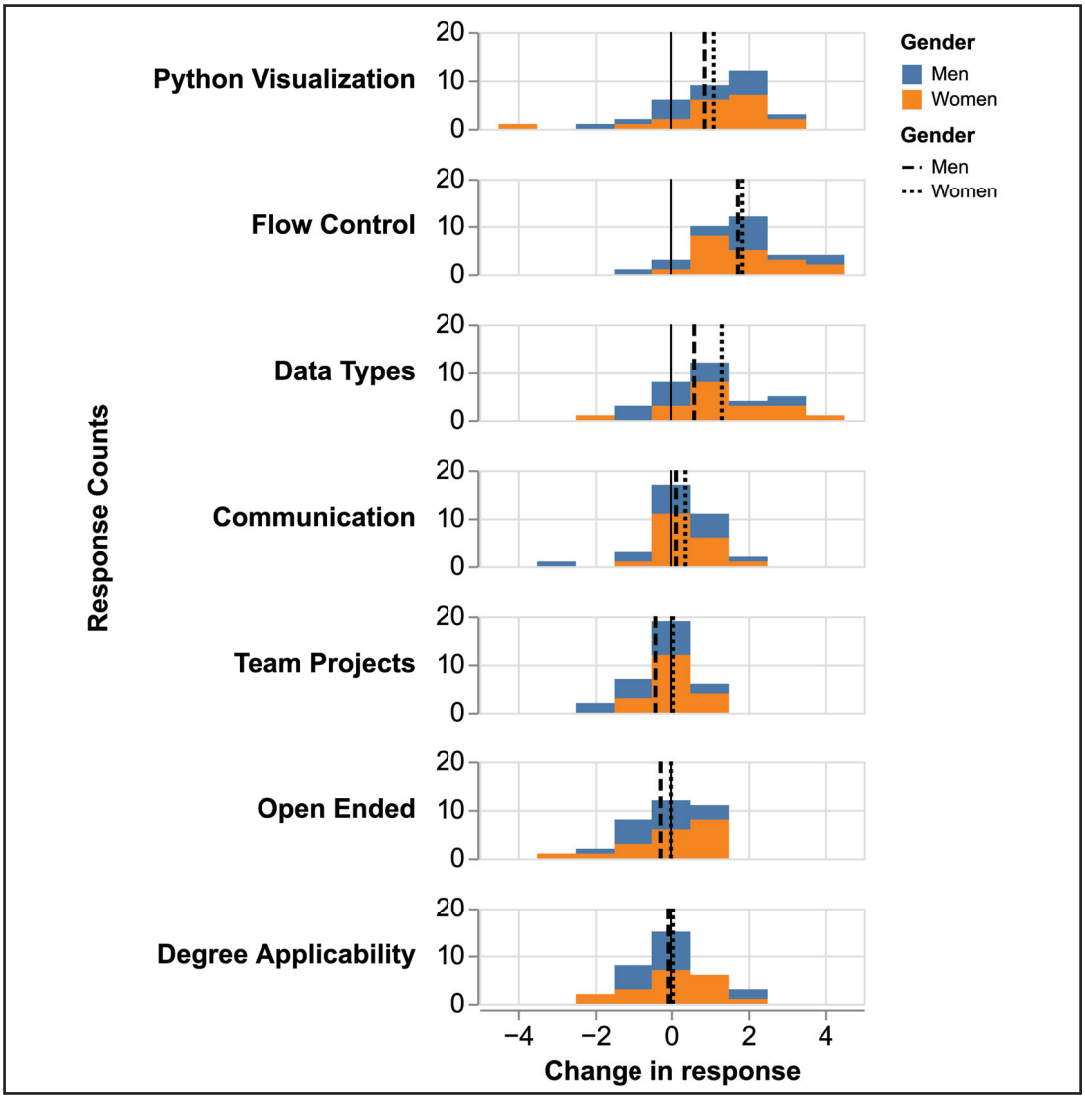

Figure 5. *Histogram of change in survey response, for each of the 7 qualitative questions by 34 participants that responded to both surveys. Histograms are stacked such that responses by women (orange) and men (blue) together compose the overall distribution. A positive change indicates that the participant reported improved confidence to the question after the event compared to before, while negative indicates that confidence was decreased by the event. A black vertical line is shown at 0, representing no improvement. The average change for men and women is shown as vertical lines. We see that for Objective 1 questions, where the average improvement is significantly more than 0, women reported a larger improvement than men in all cases. This is most significant for the data type question.*

REFERENCES

1. National Academies of Sciences, Engineering, and Medicine (2018) Data Science for Undergraduates: Opportunities and Options. Data Sci. Undergrad. doi:10.17226/25104.
2. Diez-Olivan A, Del Ser J, Galar D, and Sierra B (2019) Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. Inf. Fusion 50:92–111.
3. Beck DAC, Carothers JM, Subramanian VR and Pfaendtner J (2016) Data Science: Accelerating Innovation and Discovery in Chemical Engineering. Wiley Online Library. doi:10.1002/aic.15192.
4. Venkatasubramanian V (2009) DROWNING IN DATA: Informatics and modeling challenges in a data-rich networked world. AIChE J. 55(1):2–8 (2009).
5. Chiang LH, Braun B, Wang Z, and Castillo I (2022) Towards artificial intelligence at scale in the chemical industry. AIChE J. 68(6). e17644.
6. Finzer W, (2013) The Data Science Education Dilemma. Technol. Innov. Stat. Educ. 7(2).
7. Kross S, Peng RD, Caffo BS, Gooding I, and Leek JT (2019) The Democratization of Data Science Education. 74:1–7. doi: 10.1080/00031305.2019.1668849.
8. Witt P, Hickman D, and Herron J (2021) Educational intensification: A partnership between industry and academia. Chem. Eng. Educ. 55(4):211–217.
9. Duever TA (2019) Data science in the chemical engineering curriculum. Processes 7(11):830.
10. Irizarry RA (2020) The role of academia in data science education. Harv. Data Sci. Rev. 2020–2020. doi:10.1162/99608F92.DD363929.
11. Robertson L (2018) Toward an epistemology of active learning in higher education and its promise. Misseyanni A, Lytras MD, Papadopoulou P, and Marouli C (Ed.) Active Learning Strategies in Higher Education. Emerald Publishing Limited, Bingley. 17–44. doi:10.1108/978-1-78714-487-320181002.
12. Hartikainen S, Rintala H, Pylväs L, and Nokelainen P (2019) The concept of active learning and the measurement of learning outcomes: A review of research in engineering higher education. Educ. Sci. 2019. 9: 276.
13. Johnson RT and Johnson DW (2008) Active learning: Cooperation in the classroom. Annu. Rep. Educ. Psychol. Jpn. 47:29–30.
14. Hernández-de-Menéndez M, Vallejo Guevara A, Tudón Martínez JC, Hernández Alcántara D and Morales-Menendez R (2019) Active learning in engineering education. A review of fundamentals, best practices and experiences. Int. J. Interact. Des. Manuf. IJIDeM 2019. 13:909–922.
15. Prince M (2004) Does active learning work? A review of the research. J. Eng. Educ. 93(3):223–231.
16. Aditomo A, Goodyear P, Bliuc AM and Ellis RA Inquiry-based learning in higher education: Principal forms, educational objectives, and disciplinary variations. Stud. High. Educ. 38(9):1239–1258 (2013).
17. Barell J (2007) Problem-Based Learning: An Inquiry Approach. Corwin Process. Thousand Oaks, CA. 179.
18. Ernst DC, Hodge A, and Yoshinobu S (2017) What is inquiry-based learning? Not. Am. Math. Soc. 64(6):570–574.
19. Saltz J, Saltz J, and Heckman R (2016) Big data science education: A case study of a project-focused introductory course. Themes Sci. Technol. Educ. 8(2):85–94.
20. Muijs D and Rumyantseva N (2013) Coopetition in education: Collaborating in a competitive environment. J. Educ. Change 2013. 15(1):1–18.
21. Tokunaga S, Martínez M, and Crusat X (2019) Coopetition: Industrial interplay to foster innovative entrepreneurship in energy engineering education. IEEE Glob. Eng. Educ. Conf. EDUCON April-2019.1063–1068.
22. Zhong B and Xia L (2022) Effects of new coopetition designs on learning performance in robotics education. J. Comput. Assist. Learn. 38(3):223–236.
23. Szeto A, Haines J, and Buchholz AC (2015) Impact of an optional experiential learning opportunity on student engagement and performance in undergraduate nutrition courses. Can. J. Diet. Pract. Res. 77(2), 84–88. https://doi.org/10.3148/cjdpr-2015-038
24. Seifried E, Eckert C, and Spinath B (2018) Optional learning opportunities: Who seizes them and what are the Learning Outcomes? Teaching of Psychology. 45(3):246–250. doi: 10.1177/0098628318779266.
25. Thébaud S and Charles M (2018) Segregation, stereotypes, and STEM. Soc. Sci. 7(7):111.
26. Kramer J and Srinath KR (2017) Python - The fastest growing programming language. Int. Res. J. Eng. Technol. 4(12): 354-357.
27. Lasser J, Manik D, Silbersdorff A, Säfken B, and Kneib T (2021) Introductory data science across disciplines, using Python, case studies, and industry consulting projects. Teach. Stat. 43: S190–S200.
28. Wang X, Wallace MP, and Wang Q (2017) Rewarded and unrewarded competition in a CSCL environment: A coopetition design with a social cognitive perspective using PLS-SEM analyses. Comput. Hum. Behav. 72:140–151.
29. Komp E. et al. (2022) CHACK_documents. https://zenodo.org/record/6512221#.Y27UEezMJ8Y
30. Kluyver T. et al. (2016) Jupyter notebooks – A publishing format for reproducible computational workflows. Position. Power Acad. Publ. Play. Agents Agendas. 87–90.doi:10.3233/978-1-61499-649-1-87.
31. Davies A, Hooley F, Causey-Freeman P, Eleftheriou I, and Moulton G (2020) Using interactive digital notebooks for bioscience and informatics education. PLOS Comput. Biol. 16: e1008326–e1008326.
32. Verrett J, Boukouvala F, Dowling A, Ulissi Z, and Zavala V (2020) Computational notebooks in chemical engineering curricula. Chem. Eng. Educ. 54(3):143–150.
33. Harris CR et al. Array programming with NumPy. Nature. 585:357–362.
34. Reback J. et al. pandas-dev/pandas: Pandas 1.0.3. (Zenodo, 2020). doi:10.5281/zenodo.3715232.
35. Hunter JD (2007) Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9(3):90–95.
36. Fittipaldi D (2020) Managing the dynamics of group projects in higher education: Best practices suggested by empirical research. Univers. J. Educ. Res. 8(5):1778–1796.
37. Baird CL (2018) Male-dominated stem disciplines: How do we make them more attractive to women? IEEE Instrum. Meas. Mag. 21(3):4–14.
38. Layton R, Loughry M, Ohland M, and Ricco D (2010) Design and validation of a web-based system for assigning members to teams using instructor-specified criteria. Adv. Eng. Educ. 2:1–28.
39. Olewnik A, Yerrick R, Simmons A, Lee Y, and Stuhlmiller B (2020) Defining open-ended problem solving through problem typology framework. Int. J. Eng. Pedagogy IJEP. 10(1): 7-30.
40. Jonassen DH (2000) Toward a design theory of problem solving. Educ. Technol. Res. Dev. 48: 63–85.
41. Tchobanoglous G, Stensel D, and Burton F (2002) Wastewater Engineering: Treatment and Reuse. 4th Edition Direct Textbook. McGraw Hill. New York, NY.
42. Storm Events Database. National Centers for Environmental Information. https://www.ncdc.noaa.gov/stormevents/ accessed on November 11, 2022.
43. Niler AA, Asencio R, and DeChurch LA (2020) Solidarity in STEM: How gender composition affects women's experience in work teams. Sex Roles. 82:142–154.❒