Winter 2021

# The Data Science Corps Wrangle-Analyze- Visualize Program: Building Data Acumen for Undergraduate Students

Nicholas J. Horton
*Amherst College*

Benjamin Baumer
*Smith College*, bbaumer@smith.edu

Andrew Zieffler
*University of Minnesota*

Valerie Barr
*Mount Holyoke College*

**Harvard Data Science Review • Issue 3.1, Winter 2021**

# The Data Science Corps Wrangle-Analyze-Visualize Program: Building Data Acumen for Undergraduate Students

**Nicholas J. Horton[1], Benjamin S. Baumer[2], Andrew Zieffler[3], Valerie Barr[4]**

**[1]Amherst College, [2]Smith College, [3]University of Minnesota, [4]Mount Holyoke College**

**Published on:** Feb 25, 2021

## Introduction

We congratulate Kolaczyk, Wright, and Yajima on their innovative statistics practicum that places "practice" at the center of data science education (Kolaczyk et al., 2021, this issue). Their year-long practicum course focuses on the data science life cycle with engagement with external partners and university consulting projects. We agree that training postgraduates in practice needs to be foregrounded in the curriculum in order for students to develop necessary depth in data science practice.

## What Are the Essential Skills?

In their provocative article in *The American Statistician*, Nolan and Temple Lang (2010, p.97) challenged the statistics profession by asking whether their graduates had the skills that the Boston University practicum fosters. They asked several questions, two of which are particularly relevant to this discussion:

- Do we provide students with the essential skills to engage in statistical problem solving and keep abreast of new technologies as they evolve?
- Overall, are we doing a good job preparing students who are ready to engage in and succeed at statistical inquiry?

What are those *essential skills*? The National Academies of Science, Engineering, and Medicine's (NASEM) *Data Science for Undergraduates* consensus report (NASEM, 2018) outlined key components of data acumen:

1. mathematical foundations
2. computational foundations
3. statistical foundations
4. data management and curation
5. data description and visualization
6. data modeling and assessment
7. workflow and reproducibility
8. communication and teamwork
9. ethics

It is reassuring to see how these components are woven into the major topics in Figure 2 of Kolaczyk et al. (2021).

Ensuring that data science practitioners develop at least a basic level of mastery of each of these is critical for their ability to extract meaning from data. Developing practicum sequences may allow many institutions to be able to answer Nolan and Temple Lang's questions in the affirmative. We hope that many do proceed to replicate such models in their graduate curricula.

## Data Science Education Isn't Just for Graduate Students Anymore

We strongly believe, however, that such real-world training and experiences cannot be reserved just for graduate students. The same primary argument made by Kolaczyk et al. (2021)—that graduate students need a richer understanding of the interplay of theory and practice than we have historically offered—also applies to undergraduate students. Not including these experiences at the undergraduate level is a missed opportunity given the prevalence of computation and data practices required in the contemporary workforce. This is especially true given computation's role as a gateway to STEM fields (Holdren & Lander, 2012) and to more lucrative employment opportunities, especially for women and other underrepresented groups (Melguizo & Wolniak, 2012; U.S. Department of Commerce, 2017).

Data science has seen dramatic growth and increased interest, as highlighted in the NASEM (2018) report. Many institutions now offer an introductory data science course (e.g., Adhikari & DeNero, 2015; Baumer, 2015; Çetinkaya-Rundel & Ellison, 2021; Donoghue et al., 2021; Hardin et al., 2015; McNamara et al., 2017). Such courses provide an overview of key skills, including data wrangling, data visualization, and workflow and reproducibility, including the use of version control tools (Beckman et al., 2021). Other courses (e.g., the University of California, Berkeley, Data 100 course) pick up where the introductory courses leave off.

While many of these courses and programs teach students relevant data science skills, we can expect coursework to develop students' data acumen only so far. It is unclear whether coursework alone is enough to provide students with the experiences with data and computing they need to be successful in tomorrow's workplace. To that end, we describe aspects of a National Science Foundation (NSF)-funded project that aims to provide these types of experiences to undergraduate students.

## Data Science Corps: Wrangle-Analyze-Visualize (DSC-WAV)

DSC-WAV is an NSF-funded data science workforce development project that provides undergraduate students (primarily sophomores and juniors) with opportunities for unstructured, authentic data science experiences (https://dsc-wav.github.io/www). Projects are solicited from local governmental and not-for-profit organizations that give students repeated practice solving real-world data science problems that matter.

The genesis of the project was our observation, consistent with reflections noted by Nolan and Temple Lang (2010), that much of our students' course-based problem-solving experience focused on clean problems with simple data sets. This often leaves them relatively unprepared for the reality of the data science applications they will face in professional settings. At the same time, many small organizations lack staff with adequate technical training. Much of the data science work that these organizations need does not require the kind of sophisticated machine learning or at-scale computing taught in graduate-level programs. Rather, the bulk of the work is composed of relatively mundane wrangling and visualization tasks that are approachable for our undergraduate students.

The DSC-WAV project, now in its second year, is working with its third cohort of students. In each cohort, each team of between two and five students works with a community organization on a data-focused problem. A faculty member, who lined up the project, also acts as supervisor for the team. To facilitate a team-based approach to data science, all project members participate in a structured set of onboarding activities (structured variously as a day-long onboarding workshop, two half-day workshops, or a series of online activities). The onboarding activities provide a brief introduction to the data-focused problem that the team will be working on, as well as one to two short tutorials related to tools/processes for the team-based work (e.g., GitHub). To help keep the project on track, one student on each team is selected for their prior experience and interest in community engagement, and serves as the primary student liaison with the organization sponsoring the project.

To help structure the work of the team, we adopted an agile development framework (specifically scrum, https://www.scrumguides.org/scrum-guide.html) with the goal of fostering an iterative approach to engage students with dynamic work over the course of a semester. The scrum framework utilizes a number of approaches that are relevant for data science practice:

- The work is organized into a series of short **sprints** to break up large tasks.
- Subtasks are organized into a **backlog** to identify priorities for that stage of the analysis.
- The team and stakeholders (faculty and community organization liaison) meet regularly to share results and make adjustments in advance of the next sprint.
- **Kanban project boards**, implemented using Trello or GitHub Projects, are used to review the backlog and team progress.
- **Code review**, implemented using GitHub pull requests, is included as a regular part of the process.
- **Sprint demos** are places where current results are presented and discussed in the context of the broader goals of the project.
- **Sprint retrospectives** are used to identify issues with the process and ways that the team might improve their work.

Based on our preliminary evaluation of the project and reviews of students' work processes and products, we have identified many benefits from the program. We have seen improvement in students'

oral and written communication, both among team members and in interactions with the client organization. Student participants also demonstrated growth in the working practices and tools emphasized by the project (e.g., using the agile process, working with GitHub).

We have also seen some of the same issues and challenges raised by Kolaczyk et al. (2021):

1. Despite the flood of data available, it is challenging to identify appropriate projects with community organizations (often due to the availability of staff at those organizations).
2. It is challenging for students to pose and answer questions at the desired level. The more practice they can get, the better.
3. Fidelity to the scrum structures is difficult to maintain during a semester, given the many competing responsibilities that students are juggling, and the reality that this project is not a full-time job.
4. Faculty support of the projects requires a major time investment.
5. There is a lot to cover in the onboarding meetings and supplemental training.
6. As alluded to in discussion of assessment in Kolaczyk et al. (2021), students struggle with the difference in nature between the collaborative project activities and a typical graded course based on individual effort. In our experience, this is particularly evident in use of sprint retrospectives and code review, where we find that undergraduates are hesitant to critique each other's work.

One big question for the DSC-WAV project is sustainability after the grant funding period. The grant has funded both faculty time and student participants (approximately 8–10 hours per week). To foster sustainability beyond the duration of the grant, we are considering a variety of models, including dedicated courses that mirror the structure but allow the program to be brought into the curriculum. One approach might be to replicate the DSC-WAV experience as a mid-level statistical consulting course, ideally integrated with the community engagement center that some baccalaureate institutions have.

It is also worth noting that the COVID-19 pandemic complicated our efforts. When institutions began remote learning in March 2020, the first cohort was nearing the end of their second sprint. Zoom was already part of the planning for the organization of the project, but it became the primary way that teams interacted. Although most of the teams transitioned well to remote work, it did not go perfectly, nor did it work for all team members. By the time Cohort 2 started (Fall 2020), which was fully remote, we were better able to anticipate problems and workarounds in the remote workflow.

## Closing Notes

We hope that the approach of Kolaczyk et al. (2021) is adopted widely at graduate programs in data science and related fields. Statistics graduate students need the experiences and training that they describe.

However, we cannot afford for students to complete four years of an undergraduate degree and then have to complete a year-long master's program before they can do 'useful things' with data. While master's students will obviously have more depth and sophisticated abilities commensurate with their additional training, bachelor's graduates are increasingly being hired as analysts and data scientists, and need to have similar skills in order to be effective.

We also see the important role that two-year colleges can play in undergraduate data science education. There is a widespread development of new certificates and associate's degree programs at those institutions across the country. Our DSC-WAV project includes efforts that facilitate articulation between degree transfer programs as well as associate's-to-workforce programs.

We believe that programs such as the DSC-WAV are an important part of what is needed to ensure that students develop a deep and rich fluency with data. The activities that the students participate in reinforce almost all of the key components of data acumen outlined in NASEM (2018). While they do not allow us to confidently answer the questions posed by Nolan and Temple Lang (2010), we believe that the structured exposure to real-world problems is a useful step forward.

Other models and approaches complement and extend our approach. Reinhart and Genovese (2021) describe an innovative course that is a cross between software engineering, statistics, and data science that helps to develop complementary skills that many data scientists (at all levels) need for success. Lazar et al. (2011) detail a year-long undergraduate capstone in statistics with similar goals. Other innovative and sustainable models are needed.

We look forward to further discussions and sharing more details and insights from the DSC-WAV project in the future.

## Disclosure Statement

We acknowledge the support of NSF grants HDR DSC-1923388, HDR DSC-1923700, HDR DSC-1923934, and HDR DSC-1924017.

## References

Adhikari, A., & DeNero, J. (2015). *Data 8: The foundations of data science*. Data 8. https://data8.org

Baumer, B. (2015). A data science course for undergraduates: Thinking with data. *The American Statistician, 69*(4), 334–342. https://doi.org/10.1080/00031305.2015.1081105

Beckman, M. D., Çetinkaya-Rundel, M., Horton, N. J., Rundel, C. W., Sullivan, A. J., & Tackett, M. (2021). Implementing version control with Git and GitHub as a learning objective in statistics and data

science courses. *Journal of Statistics and Data Science Education, 29*(1), 1–35. https://doi.org/10.1080/10691898.2020.1848485

Çetinkaya-Rundel, M., & Ellison, V. (2021). A fresh look at introductory data science. *Journal of Statistics and Data Science Education, 29*, 1–11. https://doi.org/10.1080/10691898.2020.1804497

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching creative and practical data science at scale. *Journal of Statistics and Data Science Education, 29*(1), 1–22. https://doi.org/10.1080/10691898.2020.1860725

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Lang, D. T., & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "think with data." *The American Statistician, 69*(4), 343–353. https://doi.org/10.1080/00031305.2015.1077729

Holdren, J. P., & Lander, E. S. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. PCAST Report to the President. Executive Office of the President: President's Council of Advisors on Science and Technology, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

Kolaczyk, E., Wright, K., & Yajima, M. (2021). Statistics practicum: Placing 'practice' at the center of data science education. *Harvard Data Science Review, 3*(1). https://doi.org/10.1162/99608f92.2d65fc70

Lazar, N. A., Reeves, J., & Franklin, C. (2011). A capstone course for undergraduate statistics majors. *The American Statistician, 65*(3), 183–189. https://doi.org/10.1198/tast.2011.10240

McNamara, A., Horton, N. J., & Baumer, B. S. (2017). Greater data science at baccalaureate institutions. *Journal of Computational and Graphical Statistics, 26*(4), 781–783. https://doi.org/10.1080/10618600.2017.1386568

Melguizo, T., & Wolniak, G. C. (2012). The earnings benefits of majoring in STEM fields among high achieving minority students. *Research in Higher Education, 53*(4), 383–405.

National Academies of Science, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. https://nas.edu/envisioningds

Nolan, D. A., & Temple Lang, D. (2010). Computing in the statistics curriculum. *The American Statistician, 64*(2), 97–107. https://doi.org/10.1198/tast.2010.09132

Reinhart, A., & Genovese, C. R. (2021). Expanding the scope of statistical computing: Training statisticians to be software engineers. *Journal of Statistics and Data Science Education, 29*(1), 1–23.

https://doi.org/10.1080/10691898.2020.1845109

U.S. Department of Commerce. (2017). *Women in STEM: 2017 update* (Issue Brief No. 06-17). U.S. Department of Commerce, Office of the Chief Economist. https://www.commerce.gov/news/fact-sheets/2017/11/women-stem-2017-update

---