USING THE DATASQUAD MODEL TO JOIN PRACTICAL "PEOPLE SKILLS" WITH DATA SCIENCE EDUCATION

Deborah Wiltshire and Paula Lackie
GESIS-Leibniz Institute for the Social Sciences
Carleton College
deborah.wiltshire@gesis.org

With the extremely rapid developments in the field, teaching data science in a sustainable way is quite challenging. We know there are demands for skilled data scientists, but we also know there is a gap between what can be covered in the usual curriculum and the real world of data practitioners encounter in their post-college lives. The DataSquad model is empowering on several levels: for beginning students it provides a foothold to more robust data literacy through working with more advanced peers on documentation. For those with more technical experience, they can develop human skills through project management, peer mentoring, and problem solving. This paper outlines how the early success of the Carleton DataSquad can contribute to tackling the students' challenge of needing experience before they can get a job.

INTRODUCTION

Researchers and staff in data support services have long observed a gap between the skills taught in data-related academic curricula and the skills students require in the workforce. These observations are widely supported in academic literature that speaks of a global data skills shortage and consensus exists that standard data curricula often do not prepare students adequately for data-related employment (Miller, 2014; Stanton & Stanton, 2020; Thompson & Kellam, 2016). Academic data-related programs often focus on specific skills like programming, algorithms, and database design, but miss the more nuanced ways in which these tools are used in a real-life work context (Stanton & Stanton, 2020). This matters because the realities of research data preparation, management, and analysis are very different in practice than in theory. There simply is not enough time in a term to cover everything, so this gap between the curricula and experience persists.

As the research landscape becomes ever more complex in terms of data, methodologies, programming languages and software – as well as funder requirements and compliance – this shortfall in skills will become more problematic as we see a steady stream of graduates lacking the practical, 'real data' experience highly prized in the labor market and research teams throughout higher education. In their recent paper, Stanton and Stanton (2020) concluded that having a specific data-related degree or programming qualification is no longer sufficient for people entering into entry level data roles; the two main criteria employers seek when filling data-related roles are a wider range of skills and prior practical experience. This reveals that gap again: students may receive high quality data skills training in their academic programs, but they seldom get practical and diverse data-related work experience.

Miller (2014) argued that data-related curricula should include a wider array of important factors for *responsible* use, such as data governance, confidentiality, and ethics; these topics should be interwoven with traditional data science education (e.g. programming languages, algorithms, and analysis.) Stanton and Stanton (2020) added additional core human skills like problem-solving and abilities to work in a team cross-functionally and independently, while they also highlighted the rising importance of prior work experience. In their analysis of job advertisements, they found that nearly 80% of employers had specified the need for prior experience (Stanton & Stanton, 2020, p.148).

We know that practical data-experience is both difficult to attain and also essential for students trying to get a start in their careers. To further compound the challenges of gaining real data experience for minoritized students, authors like D'Ignazio and Klein (2020) and Young, et al. (2022) have raised the issue of lack of diversity in data-related work. For example, Young et al. (2020) found that between just 10-15% of data scientists and machine learning researchers were women and D'Ignazio and Klein found that 80% of Artificial Intelligence professors are male. Both studies pointed to the potential problems caused by such a lack of diversity – inherent biases that exist within society (such as sexism or racism) can be re-enacted and amplified within data collection, management and analysis processes, (a problem exaggerated greatly with the boom in A.I. and large language modeling). So they argue

In: Kaplan, J. & Luebke, K. (Ed.) (2024). *Connecting data and people for inclusive statistics and data science education. Proceedings of the Roundtable conference of the International Association for Statistics Education (IASE), July 2024, Auckland, New Zealand.* ©2025 ISI/IASE.

efforts are needed to ensure that marginalized groups are represented in data-related roles as well as in the data itself.

These problems are not new: the data skills gap and the lack of opportunity for students from minoritized groups have long been the subject of discussion among those providing data support services throughout the academic world. It is also recognised, however, that to tackle these solely with academic programs is not practical. Alternative or supplementary solutions are needed that can offer opportunities to a more diverse pool of students to learn and practice the hard and soft skills needed in their future employment. A tall order perhaps, but there are several approaches globally working toward solutions. We discuss two established approaches here - the capstone course and Work-Integrated Learning (WIL).

Capstone courses are projects that form the culmination of educational programs and are mostly found in American higher education programs. These are similar in design to the dissertation or thesis that usually caps academic programs in other parts of the world, but the capstone course may be more focused on practical or industry-oriented projects than purely academic ones and they may include classroom time with the practical or applied elements (Howe, 2010). The final project submitted in a capstone course is often a report written in the professional writing format of the discipline or industry. Capstone courses, at least originally, were mostly delivered by individual academic disciplines and often by an individual staff member (Hauhart & Grahe, 2015).

Through these applied projects, perhaps in industry settings, students are exposed to opportunities to practice and develop the soft or human skills that will be required in the workplace. These might include work skills (e.g. problem solving, project design and management, budget management), interpersonal skills (e.g. asking for help, giving and receiving information and feedback, difficult conversations, dealing with criticism) and individual skills (e.g. understanding professional ethics and integrity, responding proactively to change, time and stress management; McCormack et al., 2011; Stwalley, 2024). The design and implementation of capstone courses may vary across disciplines and across time, but they share the common aim of helping students transition their classroom learnt knowledge into the workplace (Inamdar & Roldan, 2013).

Another key type of gap-spanning program is Work-Integrated Learning (WIL). Particularly popular in countries like the UK and Australia, WIL can be defined as, "Learning outcomes achieved through activities which are based in, or derive from, the context of work or the workplace" (Connor & MacFarlane, 2007, pg. 7). The philosophy is similar to the Internship experience but with some tweaks, micro-placements of just a few days, hackathons, or consulting to name a few (Kay, et al., 2019), but they share the same ethos of moving students from passive listening to active implementation (Jackson, 2015). A core component of WIL (and a difference from the usual internship process) is the intent to enable *all* students equal access to workplace learning (Moriña & Biagiotti, 2022). A recent scoping review of WIL by Lasrado, et al. (2023) identified the inclusion of a number of disadvantaged groups including those with disabilities, mental health challenges, caring responsibilities, international students, and those from lower socio-economic groups. For these groups, WIL can be a gateway to acquiring vital skills and experiences. For example, through WIL international students can experience the working culture in their host country (Vu, et al., 2022) and students with disabilities can be supported to push through the stigma they may face and to test out how they might cope with work-based challenges (Lawlis, et al., 2024).

In this paper we introduce a small grassroots project we believe could contribute to this sphere. The DataSquad emerged organically out of challenges in providing sufficient data support to researchers and administrative offices at Carleton College. Its informal and incremental origins means that it has not yet been formally studied or examined as a Capstone or WIL style program, but we argue it is flexible enough to offer something new to help address the gap between formal education and employment. This paper marks the starting point for such an examination, and aims to gain clarity on where the DataSquad could be situated in pedagogical research and to ascertain whether it holds potential for further expansion.

A CREATIVE SOLUTION TO AN INTERCONNECTED PROBLEM - THE DATASQUAD ORIGIN STORY

The origins of the DataSquad can be traced back to the early 1990's at Carleton College in the US. The undergraduate college of around 2000 students had long valued and provided true research experience. The drive to offer students practical data experience stemmed from a challenge faced by the college – an ever-increasing demand for expert data support that often outstripped the capacity of those tasked with providing it. This is a scenario played out across academia, with librarians and data support staff regularly encountering researchers and research teams needing support beyond simple data discovery. This presented staff at Carleton with a problem - they may have had the skills to help researchers with data management or wrangling, but they often lacked the time to offer comprehensive support across their campus. The early recognition of this problem came from the hiring of Paula Lackie as one of the first to specialize in academic technology as a career; she also brought data skills and sought to address the data support needs of the campus. As computer science programming and accessing data grew in popularity, students also found interest in digging into data challenges. Paula quickly realized she could harness the students' enthusiasm to help her provide data support services to the rest of the campus. This proved highly beneficial for all: the students increased their skills through their student employment with her, and she could attend to more challenging projects.

As the service slowly evolved, by 2014 the DataSquad emerged as a way to give students practical data experience for wherever their future took them. The DataSquad sits within Academic Technology (a division of IT) and maintains its independence from specific academic disciplines. At its head is Paula Lackie, who oversees the recruitment and supervision of all DataSquad students. Paula takes responsibility for assessing project needs and matches these with student skill levels, then co-designs the projects with students. Throughout the life of these projects, Paula is on hand to guide the students where needed, but the data and documentation work remains in student hands.

A small sample of projects include:

- Converting assignments for a new Professor in Economics whose statistical language is Stata, but her students only had experience in R. A Squad member rewrote the assignments to be appropriate for R and provided well documented code and output, to match the assignment process. All to the specifications of the faculty member.
- Creating an interactive Sankey diagram for a Career Center website to match Alumni career fields with their majors when they graduated. Curious students can drill-into the visually displayed data until they find specific volunteer/alumni coaches for their post-college lives.
- Numerous data cleaning and reshaping projects for faculty research projects. Handing off the drudgery work of data preparation to eager DataSquad students freed faculty up to focus on the analysis stages of their work when they can shift from teaching to research.
- Constructing text transcripts of broadcast news from JSON data provided through an API for a Political Science research project to evaluate bias in news resources. The source data were in 5-minute slices of transcribed news across 11 stations. The resulting working data were more than 500,000 text files chronicling news coverage of events such as major elections, coverage of the Pandemic of 2020, and the murder of George Floyd and subsequent trial of Officer Chauvin.

The DataSquad model is flexible; at Carleton College it consists of several formal student roles, each providing their own learning path. In addition to a staff director, student roles include:

- **Project Management Intern:** co-hosted by the Career Center and Academic Technology. The Intern is responsible for helping to coordinate the other students in the DataSquad and acting as a liaison between them and the DataSquad lead. The Intern is essential to the full-time staff member's ability to maintain the Squad; they can manage some busywork (like access rights), but more importantly, they also can be a source of reliable information regarding the current student experience. They gain experience in leading a team, tracking employees, and other formal project management skills.
- **Technical Writer:** Technical writers are responsible for writing and/or managing the blog, website content, technical reports, or project reports. Their role is to translate sometimes complex technical work into plain language summaries for sharing the teams' work with non-technical audiences. This role also gives beginners a pathway to gaining programming-adjacent experience while also building on extremely valuable communication skills.
- **Lead Data Scientist** and **Assistant Data Scientist:** These roles involve applying their data science skills (e.g. data cleaning, re-shaping, processing, database design, algorithm

> development, etc.) to solve problems for 'clients' from the College or beyond. They all start at the Assistant level and as their technical skills develop *and* they exhibit attention to detail and an interest in developing managerial skills, they may progress to the Lead Data Scientist role. This simulates a career progression. The Leads get opportunities to develop leadership skills through learning to think and communicate beyond their own experience with challenging problems, parsing projects so they can be tackled in teams, and are accountable for project documentation. (Following FAIR data principles.)

Other parallel learning objectives for Squad members are the lifecycle of data and of projects; the importance of consistency, clarity, and thoroughness in documentation; and learning to judge their own capacities and focus. Having defined the roles (and reiterating their responsibilities) and providing guidance in ways that all members of a team can see gives everyone repeated opportunities to absorb these key employment skills, much of which boil down to how to communicate effectively and be a contributing team member. In this way it is appropriate to draw an analogy between running a DataSquad and coaching a team sport.

On top of practice in data wrangling, students are introduced to the areas of ethics and data security. All Squad members must sign a confidentiality agreement form. With any potentially sensitive project, students describe the data paths in their research processes as well as their own thoughts on potential security issues. Random audit checks are made by the Project Manager to assure they are actually following their own protocols. This focus on data security is a key component of the professional development aspect of the DataSquad and is a skill much in demand.

Since 2014, over 97 students have been student employees on the DataSquad at Carleton College, providing cutting-edge data support services across the institution (Lackie, 2019; 2021a; 2021b). These undergraduate students may have started in their first semester or in their final year; they have been 68% non-white and 31% non-male (including non-binary and trans). Sixty percent came from 20 different countries and many of the US-students were first-generation Americans and/or first-generation college students. (For comparison, in 2024 the college had 2007 enrolled students, of which 30% non-white, 50% non-male, and 11% had come from 59 different countries.) In recent years, the students have mostly self-identified as computer science majors, though many switch to other disciplines as they gain more experience. Psychology, NeuroScience, Statistics, and Mathematics have also been popular majors for DataSquad members. That is not the full story, however: the DataSquad also has a good representation from Studio Art, Music, Dance, History, Linguistics, Sociology, and Economics.

This model is empowering on several levels: students get that elusive experience of working with non-scaffolded, messy, real-world research projects. They also have repeated opportunities to experience and practice the soft skills employers value: learning to navigate working with others, peer-mentoring and useful collaboration, communicating within teams and with "clients," and managing projects especially with documentation for reproducibility! Indeed, DataSquad alumni have reported continuously it was their experience on the Squad their employers identified as the key to getting their post-college job, into graduate school, onto a research project of their choice, etc.

CAN THE DATASQUAD MODEL BE ADAPTED FOR OTHER DATA SERVICES?

Although formal research to assess the DataSquad model has not yet been conducted, at international meetings and conferences, presentations on the DataSquad model have attracted interest, both within and outside the US. In 2020, a small group of academics and data support specialists from the US, Canada, and from the UK Data Service formed a small working group to explore ways to adapt the DataSquad model to diverse academic environments. Since then, tentative plans to broaden the model have started with the launch of a DataSquad at UCLA Libraries, managed by Tim Dennis, Director of the Data Science Center (Dennis, 2021), and another within the Computational Research office of the University of Minnesota Libraries.

While still largely aspirational, collaboration between the DataSquads and their managers may take the form of a networked *DataSquad International*, (loosely following the pattern defined by carpentries.org). This kind of multi-level collaboration could facilitate greatly sharing of resources like job descriptions, student handbooks, web site design, and other overhead that can be taxing when starting up a new program. Additionally, cross-DataSquad collaboration could enhance opportunities for students and managers from all backgrounds. Different disciplines, support organizations, levels or styles of study, and experience. The point is this model empowers the students to see themselves as members of a community of data professionals and also support managers in the overhead of mentoring of a constant stream of novices in a constant stream of projects.

What are the barriers to adopting this model? A recent survey of around 100 academic and data institutions (mostly from North America and Europe) offering data support services found that although 70% of them employ students to facilitate the delivery of their services, a lack of staff time to train and supervise students coupled with the transient nature of student employees are the primary barriers to employing students in the technically-intense world of data support (Isuster & Rod, forthcoming). These obstacles are tightly interconnected, as one respondent stated, "*It's hard to imagine how [students] would be able to help…the learning curve is steep*" (Isuster & Rod, forthcoming).

The time required to supervise students working on data projects was also highlighted as being a significant burden on top of the frequently experienced combination of high workloads and low resources. The solution to address these obstacles is partially in recognizing that the "service" of the DataSquad may be as much about training the next generation of data professionals as it is about helping serve the explicit data needs of an institution. Indeed, as students are trained and then get snatched up by researchers and other offices, it helps to see this not as a loss, but a success. Those research teams will get the benefit of a more skilled student who will enlighten their peers and may even teach the lead researchers a bit about data lifecycle and FAIR in the process.

Additionally, DataSquad students add value to a professional staff. It is a common problem among data support professionals that it is difficult to keep up to date with the latest developments. Recruiting students fresh from courses that teach new methods and software is a cost effective and relatively efficient way of regularly cycling in up-to-date knowledge. We recognise that this requires a mind-set shift, (along with the idea that training the DataSquad students is part of the service profile), that would benefit from support at the top of the organizational structure.

It is true that working with students in this way is not without challenge and here the initial lessons learned at Carleton are invaluable. For example, it is essential to be clear when communicating and insist on honesty in a way that acknowledges the benefits of learning from mistakes. Anecdotally, this early post-covid era has new challenges; students are reticent to admit when they do not understand while also exhibiting a deep fear of failure. Though always a good idea, now it is even more important to model how to deal with failure as integral to progress and reiterate what has been learned in the process. Relatedly, it is very important to hire very carefully for the intern role as this person will likely be an excellent sounding board for facilitating cross-generational communication and help interpret current student expectations.

DISCUSSION

One of the disadvantages we see with the capstone course model, is that it is discipline specific, and being closely tied to the taught program or degree, there seems little scope to open up to a broader range of students. In contrast, DataSquad is discipline-agnostic, bringing in students from across the university. This enables the DataSquad to provide at least some basic data skills for students who would normally have limited or no exposure to data related training. This matters because in today's data-driven world, all students could benefit from at least a basic exposure to data-related work. The WIL movement seems to offer a more inclusive approach, and DataSquad fits naturally within this ethos since it offers data-related work experience to all students who apply themselves, even Dance majors.

With an emphasis on helping students get the skills and resilience they need, we can also focus on supporting those students who are less likely to get hired in technical positions elsewhere. The diversity seen in the DataSquad cohorts so far is encouraging. If we can ensure that DataSquad cohorts continue to represent the diversity in our populations, this approach could support the diversity and equality we want to attain in data professions and analytical outcomes. Additionally, these students will gain the necessary and highly transferable and skills that do not nicely fit into a syllabus. The combination of all this experience, plus their motivation to solve actual problems from their campus, leads to at least a few more minoritized students gaining a higher foothold on their futures.

A key limitation of the DataSquad model is that its organic emergence was born from having data projects and eager student labor, but no additional budget. To survive, the model has continued to evolve to address changing conditions, but whilst anecdotally we feel that this model could be successfully expanded to other institutions, the drawback of limited resources and funding means that it remains largely unexamined. We argue that the model has merit within the WIL sphere and deserves further scrutiny, but greater resources are required to help drive the development of a more robust blueprint for other institutions interested in implementing a similar program. Next steps should also include a formal study of the impact, benefits and challenges of the DataSquad model.

Early indicators from a research study (Isuster & Rod, Forthcoming) have suggested an implicit indication that employing students is viewed primarily as a means of meeting the demand for data-related services and not as a service to the students themselves. Whilst this is not unreasonable, the DataSquad model pushes for a reframing of the use of students in providing data support. In viewing student labour simply as a burden with no benefit to the mission of the support unit, we miss the full range of opportunity the DataSquad model can offer. The vision of the DataSquad is to inspire and enable students from all backgrounds to access the opportunities offered by data-related careers, and to lead inclusion and growth of the data professions. Our focus is on driving the movement from consuming student labor, to reframing it as a bidirectionally beneficial relationship. We hope that in-depth research into the DataSquad model can provide support for this movement.

REFERENCES

Connor, H., & MacFarlane, K. (2007). Work Related Learning (WRL) in HE – a scoping study. *Centre for Research in Lifelong Learning, Glasgow, Caledonian University*. https://s3.eu-west-2.amazonaws.com/assets.creode.advancehe-document-manager/documents/hea/private/wrlreport_january2007_1568036946.pdf.

Dennis, T. (2021, May 20). *Using Student Workers in the Delivery of Data Service* [Video]. YouTube. https://www.youtube.com/watch?v=wQi2sFVKWrI.

D'Ignazio, C., & Klein, L.F. (2020) *Data feminism.* MIT Press.

Hauhart, R.C., & Grahe, J.E. (2015) *Designing and teaching undergraduate capstone courses (Jossey-Bass higher and adult education series)*. John Wiley & Sons.

Howe, S. (2010). Where are we now? Statistics on capstone courses nationwide. *Advances in Engineering Education, 2*. 1-27.

Inamdar, S. N., & Roldan, M. (2013). The MBA capstone course: Building theoretical, practical, applied, and reflective skills. *Journal of Management Education*, *37*(6), 747-770. https://doi.org/10.1177/1052562912474895.

Isuster, M., & Rod, A.B. (Eds.) (forthcoming), *Data culture in academic libraries: A practical guide to building communities, partnerships, and collaborations*. Association of College and Research Libraries.

Jackson, D. (2015) Employability skill development in work-integrated learning: Barriers and best practice, *Studies in Higher Education*, *40*(2), 350-367, https://doi.org/10.1080/03075079.2013.842221

Kay, J., Ferns, S., Russell, L., Smith, J., & Winchester-Seeto, T. (2019). The emerging future: Innovative models of work-integrated learning. *International Journal of Work-Integrated Learning, Special Issue*, *20*(4), 401-413. https://www.ijwil.org/files/IJWIL_20_4_401_413.pdf.

Lackie, P. (2021a, May 20). *Professional Development and the delivery of Data Support Services with Student labor* [Video]. YouTube. https://www.youtube.com/watch?v=qCyihGcnbKk.

Lackie, P. (2021b, May 20). *Managing Students in the delivery of Data Support Services* [Video]. YouTube. https://www.youtube.com/watch?v=xrq_yWMOQhY.

Lackie, P (2019, May 31). *Responding to Data Support Challenges with a Student DataSquad* [Video]. IASSIST. *https://doi.org/10.5281/zenodo.3620820*.

Lasrado, F., Dean, B. A., & Eady, M. J. (2023). Inclusive work-integrated learning in higher education: a scoping review. *Studies in Higher Education*, *49*(9), 1588–1609. https://doi.org/10.1080/03075079.2023.2271048.

Lawlis, T., Mawer, T., Andrew, L., & Bevitt, T. (2024). Challenges to delivering university health-based work-integrated learning to students with a disability: a scoping review, *Higher Education Research & Development*, *43*(1), 149-165, https://doi.org/10.1080/07294360.2023.2228209.

McCormack, J., Beyerlein, S., Bracklin, P., Davis, D., Trevisan, M., Davis, H., Lebeau, S., Gerlick, R., Thompson, P., Khan, M.J., Leiffer, P. & Howe, S. (2011) Assessing Professional Skill

Development in Capstone Design Courses. *International Journal of Engineering Education, 27*(6) 1308–1323.

Miller, S. (2014). Collaborative Approaches Needed to Close the Big Data Skills Gap. *Journal of Organization Design*, *3*(1), 26–30. https://doi.org/10.7146/jod.9823.

Moriña, A., & Biagiotti, G. (2022). Inclusion at University, Transition to Employment and Employability of Graduates with Disabilities: A Systematic Review. *International Journal of Educational Development*, *93*. https://doi.org/10.1016/j.ijedudev.2022.102647.

Stanton, W., & Stanton, A. (2020). Helping Business Students Acquire the Skills Needed for a Career in Analytics: A Comprehensive Industry Assessment of Entry-Level Requirements. Decision Sciences *Journal of Innovative Education*, *18*. https://doi.org/10.1111/dsji.12199.

Stwalley, R. (2024) Assessing Improvement and Professional Career Skills in Senior Capstone Design through Course Data. *International Journal of Engineering Pedagogy*, *7*(3), 130-146. https://www.learntechlib.org/p/207438/.

Thompson, K., & Kellam, L. (2016). *Databrarianship: The academic data librarian in theory and practice*. Association of College and Research Libraries. https://openlibrary.org/books/OL27450574M/Databrarianship.

Vu, T., Ferns, S. & Ananthram, S. (2022) Challenges to international students in work-integrated learning: a scoping review, *Higher Education Research & Development*, *41*(7), 2473-2489, https://doi.org/10.1080/07294360.2021.1996339.

Young, E., Wajcman, J. & Sprejer, L. (2023) Mind the gender gap: inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment, 38*, 391–414. https://doi.org/10.1111/ntwe.12278.