11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.20096

Teaching Health Data Science through Experiential Learning: The Datathon Approach

Oscar Perez-Concha, Mark Hanly
Centre for Big Data Research in Health, University of New South Wales, Sydney, NSW, Australia.

How to cite: Perez-Concha, O.; Hanly, M. (2025). Teaching Health Data Science through Experiential Learning: The Datathon Approach. In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. https://doi.org/10.4995/HEAd25.2025.20096

Abstract

This paper describes the design and implementation of an annual health data science datathon aimed at teaching practical skills in health analytics to undergraduate and postgraduate university students. Datathons provide an experiential learning environment where students work in teams to analyse real-world health data and propose solutions to current challenges in the field. The steps to design and execute the datathon are outlined, covering data, partnerships, task design, and rubric development. We discuss the educational benefits of formalising research questions, working in groups, programming, and presenting findings. Growing access to electronic health data combined with the availability of free and opensource analysis platforms and libraries, and cheap cloud compute power make datathons a universally accessible learning experience.

Keywords*: Health data science; datathon; experiential learning; student engagement.*

1. Introduction

Health data science is an emerging discipline sitting at the nexus of (i) medicine and health, (ii) computer science, and (iii) statistics and artificial intelligence (AI). This highly applied and inherently interdisciplinary subject attracts a broad range of students with diverse educational and career trajectories, ranging from working health professionals with vast domain knowledge but limited coding skills to undergraduate computer science students familiar with cutting-edge algorithms, but lacking contextual experience.

Experiential learning has been identified as an effective approach to engage diverse student cohorts (Amat et al., 2021; Rohani et al., 2024). Hackathons and datathons are examples of pedagogical approaches which offer such experiential learning (Silver et al., 2016). A hackathon is a social coding event where groups of individuals come together to work intensively (often competitively) on a practical problem over a short period of time. These events emerged from

the tech industry and have gained popularity as an educational tool internationally and across multiple disciplines (Garcia, 2023). Datathons follow a similar model to hackathons, but with a stronger focus on the analysis and interpretation of data. While academic interest in datathons has lagged that of hackathons, data-orientated events are catching up in popularity (see Figure 1), likely facilitated by the increasing availability of large open data repositories (Anslow et al., 2016).

The practical nature of these events forces students to engage with pragmatic challenges which are often de-emphasised in more traditional pedagogical approaches, for example navigating data documentation and handling missing or ineligible data. The team-based nature of these events also encourages students to build soft-skills including teamwork, communication, project management and community engagement (Anslow et al., 2016). A recent systematic review of 30 studies describing results from hackathons and datathons has confirmed their educative benefits, including enhanced learning and increased student engagement and motivation (Oyetade et al., 2024). Datathons can also be relatively low-cost events, making them widely accessible learning experiences (Kuter & Wedrychowicz, 2021). This accessibility is facilitated by the widespread availability of free, opensource analysis platforms and libraries, opensource data, including synthetic electronic health records (Kuo et al., 2024), and free cloud compute power for students.

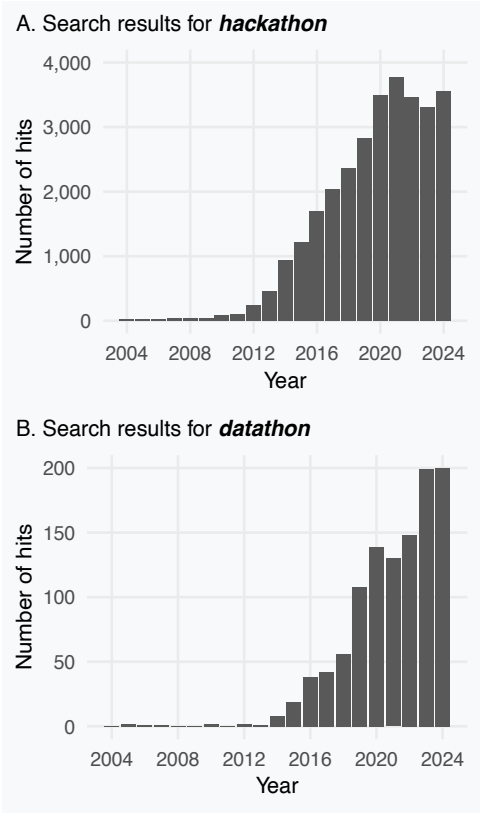

Figure 1. Number of Google Scholar search results for (A) Hackathon and (B) Datathon by year, 2004 - 2024.

The educational benefits of datathons are clear, but there are also many practical challenges to their successful delivery. While there are many open datasets available, an appropriate datathon dataset must balance sufficient complexity with being reasonably analysis-ready given the short timeframe (Mougan et al., 2024). Cultural diversity and varied skill levels among team members can be a challenge to effective teamwork (Oyetade et al., 2024). There are also the everyday logistical challenges of planning and promoting the event, securing an appropriate venue, providing IT support, engaging industry partners, and the financial cost of these activities (Kuter & Wedrychowicz, 2021).

This paper describes our experience organizing two annual datathons at the University of New South Wales, Sydney (UNSW) in 2023 and 2024. The datathons aimed to provide undergraduate and postgraduate students with an experiential learning activity where they could apply the methods they have been learning to real-world problems. We outline the design and execution process, highlighting the educational impact and lessons learned.

2. The Datathons

The learning objectives of our datathon events were to engage students in applied health data science through experiential learning; strengthen teamwork and communication by working in groups; build technical skills in data analysis and programming; foster critical thinking by tackling real-world data challenges; and provide exposure to the complexities of working with real-world health data. These objectives aimed to ensure that students not only developed technical proficiency but also gained essential soft skills and a deeper appreciation of health data science in practice.

To illustrate how these objectives were achieved, we describe the execution and outcomes of our datathons held in 2023 and 2024.

2.1. Year 1 (2023)

The first UNSW Health Data Science datathon took place in May 2023. This one-day event saw the participation of 22 Masters of Science in Health Data Science students working in six teams, five on campus and one hybrid online/on-campus team. The event focused on the Antiretroviral Therapy in HIV dataset, a realistic synthetic dataset comprising demographic details and longitudinal clinical data on viral loads, CD4 counts, and drug regimens for 8,916 patients with HIV (See Figure 2). This is a free and opensource dataset, generated using AI as part of the Health Gym project, which aims to provide clinically realistic datasets while avoiding the usual disclosure risks associated with health data (Kuo et al., 2022).

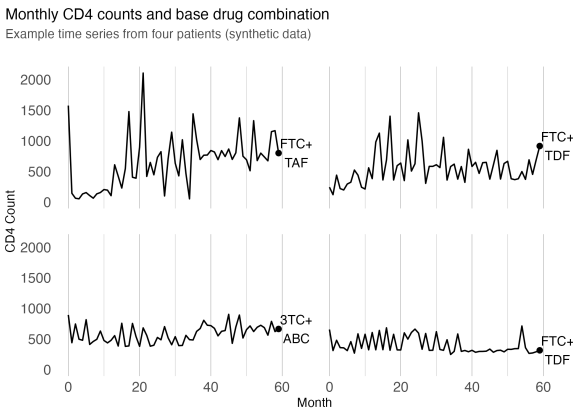

Figure 2. CD4 counts over time and base drug combination for four randomly selected patients from the Antiretroviral Therapy in HIV dataset.

Teams were challenged to pose a research question and develop a solution using their health context expertise and analytic skills. Five experts were on hand to guide the teams, aiding them

in crafting their research questions and executing their proposed solutions effectively. This included applied researchers with content expertise in HIV medications and machine learning, and Health Informaticians from New South Wales Health, Sydney Local Health District. At the end of the day, each team gave a short presentation, summarizing their research question, analytic approach and results. A team of independent judges assessed the teams based on a structured rubric (described below) and winning teams were awarded small cash prizes (max $150 AUD per student) in the form of gift vouchers.

2.2. Year 2 (2024)

Following the success of the 2023 datathon, the second UNSW Health Data Science Datathon expanded to include over 50 UNSW students from 12 different schools and research centers for a two-day event in December 2024. There were 13 teams in total, with one team again working as a hybrid online/on-campus team. Participants analyzed infectious disease data from EPIWATCH®, an AI-driven system that provides near real-time global monitoring of a range of infectious diseases and syndromes by scanning open-source online news sources. (MacIntyre et al., 2023). Each record in the dataset represents a news article mentioning an infectious disease or syndrome (See Figure 3). Minimum data includes the disease or syndrome(s) in question, the timing and location of the report, as well as the title and link to the original online article.

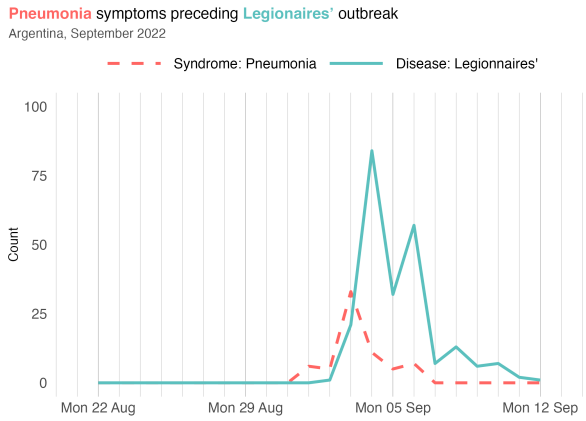

Figure 3. EPIWATCH® data showing counts of news articles mentioning pneumonia syndromes or Legionnaires' Disease in Argentina, August-September 2022. Note that publicly available online reports of unexplained pneumonia symptoms preceded the identification of Legionnaires' disease as the source of the outbreak.

On day 1 of the event, participants were challenged to use the EPIWATCH® data to design a data science pipeline (including data processing and analysis) with the aim of predicting emerging infectious disease outbreaks. On day 2 of the event, the EPIWATCH® data was incrementally updated with an additional day of data every 5 minutes over the course of 90 minutes. This exercise was designed to simulate the experience of real-time surveillance and allow students to test and demonstrate the effectiveness of the data cleaning and analysis pipelines they had developed on day 1. The larger number of teams in this running meant that it was not practical for every team to present their work. Instead, the judging panel made an initial selection of eight teams based on qualitative review of their progress and their performance in the simulation exercise. The shortlisted teams presented their

approaches formally and were reviewed based on a structured rubric. Again, winning teams were presented with small cash prizes in the form of gift vouchers.

3. Designing the datathon

3.1. Data and Partnerships

In our experience, acquiring relevant and high-quality datasets is the most practical starting point to designing a datathon event. The available data dictate the types of questions or challenges that can be posed, the access requirements and computational needs for participants, and the necessary domain knowledge input that will be needed. Suitable datasets for a datathon must meet several requirements to ensure the success of the event: 1) **Complexity**: The dataset should be challenging enough to push participants to think critically and apply advanced analytical methods; 2) **Big Data Characteristics**: The dataset should include a sufficient number of records or a volume of information that allows for the application of algorithms and enables participants to draw meaningful conclusions; 3) **Relevance and Motivation**: The dataset should be interesting and relevant to participants, aligning with current health challenges. This ensures that students remain motivated throughout the datathon and feel a sense of purpose by recognising the potential impact of their analyses.

Given the inherent sensitivity of patient data, finding suitably large and accessible datasets was a particular challenge. We overcame this by drawing on opensource, synthetic data in 2023, and EPIWATCH® surveillance data, comprising news article data rather than patient data, in 2024. For both editions, the data was released a week before the event, along with a data dictionary and a brief online session. This session provided an overview of the data and included a Q&A on the dataset and the datathon.

In both cases, building partnerships with data custodians and domain experts was essential. In 2023, we partnered with the creators of the synthetic dataset, an HIV medication expert and data specialists from a local Sydney hospital to provide participants with insights into HIV treatment and the complexities of health data. In the 2024 datathon, we partnered directly with the EPIWATCH® team, who facilitated data access, shared their expertise, and mentored students in their analytic approaches during the datathon. These insights helped participants gain a deeper understanding of how real-world infectious disease surveillance systems operate and how health-driven analytics are used to monitor and respond to global health threats.

3.2. Research Question Design

Crafting research questions that align with both educational and practical objectives is essential. Examples include tasks involving disease prediction and anomaly detection, which foster critical thinking and technical skills. We adopted contrasting approaches in the 2023 and 2024

datathons. In 2023, formulating a research question was built in as part of the datathon challenge. This enabled participants to explore a variety of questions and analytical approaches, such as predictive modeling (e.g., estimating the probability of a patient having a low CD4 count given a specific drug combination) and causal inference (e.g., identifying factors influencing time-to-event outcomes using Cox regression) (Kuo et al., 2024). By leaving the research question open, participants were encouraged to think critically about the data's structure, its potential for answering different types of health-related questions, and how to design a question that can be translated into an operationalizable analysis.

In contrast, for the 2024 edition, the research question was less open ended, and students were required to focus specifically on predicting future outbreaks. While this provided a clear analytical direction, it also introduced an additional layer of complexity—participants first had to define what constitutes an outbreak. This was a non-trivial task, as outbreak definitions can vary depending on the disease. For instance, an influenza outbreak might be defined by a rapid increase in cases over a short period, whereas an Ebola outbreak could be characterized by even a small number of cases due to its severity. This ambiguity required participants to engage deeply with epidemiological concepts before proceeding with modelling, reflecting the real-world challenges of outbreak detection and biosurveillance.

3.3. Software and Infrastructure

In both iterations of the datathon, participants were provided with links to downloadable comma-separated values (CSV) text files containing the challenge datasets. These datasets had the dual advantages of being (i) reasonably small (42.6MB in 2023 and 15.3MB in 2024) and (ii) non-sensitive, meaning that neither additional computational resources nor secure analytic environments were required. Students worked on their own laptops and were allowed to use any analysis software they preferred. Popular statistical packages included Python and R.

3.4. Rubric Development

The open-ended nature of the challenge questions precluded the use of standard methods for quantifying predictive accuracy. Instead, we designed a structured rubric that offered a comprehensive evaluation of team performance. For example, for the 2024 event, assessment was divided into three equally weighted categories: analytic approach, surveillance quality, and communication of results. The analytic approach criterion assessed the robustness of participants' data handling and methodological choices, ensuring that data preparation and modeling techniques were both appropriate and reproducible. Surveillance quality focused on the effectiveness of outbreak detection, considering the timeliness and accuracy of predictions while minimizing false alarms. Finally, communication of results evaluated how well teams articulated their methods and findings through visualizations, presentations, and the overall clarity of their insights.

4. Outcomes and Lessons Learned

Participant feedback was overwhelming positive with 100% of post-event survey respondents agreeing with the statement *Based on your experience at the datathon, would you recommend it to other students?* Qualitative feedback highlighted the value of teamwork. For example, participants in 2004 wrote: 1) *"I very much enjoyed working with my team members to brainstorm how to convert complex and messy datasets into well-organised formats for data visualisation."*; 2) *"I particularly enjoyed working with my team and hearing others' approaches."* Respondents also positively emphasized the hands-on nature of the events: *"I loved getting to work with really complicated data based on real situations."*

A. Analysing the data.

B. Brainstorming with content experts.

Figure 4. Students participating in the UNSW Health Data Science Datathon, 2024.

Following the 2023 datathon, several teams co-authored a peer-reviewed academic paper with the Antiretroviral Therapy in HIV dataset developers and the datathon organizers, providing the students with a tangible output from the event (Kuo et al., 2024).

Essential to the success of both datathon iterations was the active participation of content experts from partner organizations, who provided students with valuable insights into data structures and their use in real-world settings. Another important contributor to the event's success was the early release of the dataset and the information session held one week prior, where domain experts were available to answer questions. This gave students time to explore the data in advance and seek clarification as needed. We found it practical to start with a relatively small one-day event with students from our own faculty before scaling up to a two-day event with students from across the university. Open-ended questions offered students an engaging challenge but also required the development of a structured rubric. In contrast, focusing on a more narrowly defined prediction or classification task could allow for automatic assessment using standard metrics, which may be preferable for larger events.

A major challenge in both iterations was securing a suitable dataset. In both 2023 and 2024, we were able to draw on data resources developed at UNSW which helped to facilitate access. To continue to provide suitable datasets for an annual event will require expanding beyond the

university to publicly or privately held data assets. A second major challenge was securing financial support, which in our case was provided through the UNSW Faculty of Medicine and Health.

5. Conclusion

The datathon model provides a relatively low cost and accessible model for teaching data science through experiential learning. Future iterations will aim to incorporate more diverse datasets and explore opportunities to collaborate with public and private data holders.

References

Amat, M., Duralde, E. R., Lam, B. D., Lipcsey, M., Persaud, B. K., & Celi, L. A. (2021). Hacking the hackathon: insights from hosting a novel trainee-oriented multidisciplinary event. *BMJ Innovations*, 7(3). https://doi.org/10.1136/bmjinnov-2020-000583

Anslow, C., Brosz, J., Maurer, F., & Boyes, M. (2016). Datathons: An experience report of data hackathons for data science education. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*. https://doi.org/10.1145/2839509.2844568

Garcia, M. B. (2023). Fostering an innovation culture in the education sector: A scoping review and bibliometric analysis of hackathon research. *Innovative Higher Education*, 48(4). https://doi.org/10.1007/s10755-023-09651-y

Kuo, N. I.-H., Perez-Concha, O., Hanly, M., Mnatzaganian, E., Hao, B., Di Sipio, M., Yu, G., Vanjara, J., & Valerie, I. C. (2024). Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project. *JMIR Medical Education*, 10(1). https://doi.org/10.2196/51388

Kuo, N. I.-H., Polizzotto, M. N., Finfer, S., Garcia, F., Sönnerborg, A., Zazzi, M., Böhm, M., Kaiser, R., Jorm, L., & Barbieri, S. (2022). The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Scientific data,* 9(1). https://doi.org/10.1038/s41597-022-01784-7

Kuter, K., & Wedrychowicz, C. (2021). Hosting a data science hackathon with limited resources. *Stat*, 10(1). https://doi.org/10.1002/sta4.338

MacIntyre, C. R., Chen, X., Kunasekaran, M., Quigley, A., Lim, S., Stone, H., Paik, H.-y., Yao, L., Heslop, D., & Wei, W. (2023). Artificial intelligence in public health: the potential of epidemic early warning systems. *Journal of International Medical Research*, 51(3). https://doi.org/10.1177/03000605231159335.

Mougan, C., Plant, R., Teng, C., Bazzi, M., Cabrejas Egea, A., Chan, R., Salvador Jasin, D., Stoffel, M., Whitaker, K., & Manser, J. (2024). How to data in datathons. *Advances in Neural Information Processing Systems*, 36.

Oyetade, K., Zuva, T., & Harmse, A. (2024). Evaluation of the impact of hackathons in education. *Cogent Education*, 11(1). https://doi.org/10.1080/2331186X.2024.2392420

Rohani, N., Gal, K., Gallagher, M., & Manataki, A. J. B. M. E. (2024). Providing insights into health data science education through artificial intelligence. *BMC Med Educ* 24(1). https://doi.org/10.1186/s12909-024-05555-3

Silver, J. K., Binder, D. S., Zubcevik, N., & Zafonte, R. D. (2016). Healthcare hackathons provide educational and innovation opportunities: a case study and best practice recommendations. *Journal of medical systems*, 40. https://doi.org/ 10.1007/s10916-016-0532-3