THE SYDNEY DATA STORIES: TOWARDS AN INCLUSIVE, INTERDISCIPLINARY APPROACH TO LARGE, DIVERSE, FIRST YEAR COHORTS

Di Warren, Samantha Clarke
School of Mathematics and Statistics, University of Sydney
Educational Innovation Team - DVC (Education) Portfolio, University of Sydney
diana.warren@sydney.edu.au

While first-year undergraduate cohorts in data science tend to be increasingly large and diverse, the approach to teaching can be somewhat siloed and not capitalize on the wealth of data stories across an institution. What could a more inclusive approach look like? How can we leverage the interdisciplinary nature of data science to design a curriculum which engages students from many different fields and backgrounds? Our study focuses on the Sydney Data Stories, which is a large, collaborative project across the University of Sydney. Colleagues across the university brought their stories into the lecture theatre, showcasing data and their insights from their field. We outline the storytelling shape of the curriculum, and then consider three case studies, with findings from student data.

INTRODUCTION

Over the past few decades, spearheaded by the influential Guidelines for Assessment and Instruction in Statistics Education reports (GAISE College Report ASA Revision Committee, 2007, 2016), the undergraduate statistics course has been transformed by focusing on critical thinking with real data. This, however, begs the question: which data? For statistics courses taught within a discipline (like Psychology or Engineering), the data tend to be field-specific. For larger "service" courses, the data may be chosen with a view to its affordances (Chick, 2007), or whatever the lecturer has available from their own research or from open-source data.

In this paper, we suggest another approach, exemplified in the Sydney Data Stories project. Leveraging the inherent interdisciplinary nature of data science, we invite colleagues across the university to bring their stories into the lecture theatre, thereby showcasing data and insights from their field. This brings immediacy and currency of data science into the classroom, leads to a current, diverse curriculum that engages students from many different fields and backgrounds, and forges a shared interest in and responsibility for data science education across the institution. Moreover, it models the relevance and transference of data science to all domains.

CONTEXT

Our study focuses on the flagship unit in data science at the University of Sydney - DATA1001 Foundations of Data Science (including the smaller Advanced version, DATA1901). The unit was launched in 2018 and has quickly grown from annual enrolments of 741 in 2018 to 2497 in 2024, as shown in Table 1.

Table 1. DATA1001 annual enrolments from 2018 to 2024

Year	2018	2019	2020	2021	2022	2023	2024
Annual Enrolments	741	1325	1609	1911	2096	2221	2497
% increase		78.8%	21.4%	18.8%	10%	6%	12.4%

The cohort is very diverse, with approximately 48% female and 52% male students, including 8.5% with disabilities and 8.4% first-in-family, commonly identified as "at risk" students across the institution (Advanced Analytics Planning and Enterprise Data, 2024). DATA1001 is a foundational first-year unit and is taken by students from many different majors and degree programs. For example, of the 49 majors in the Science degree, 34 strongly recommend that their students take DATA1001 as part of their 12cp of 'Science core', from Medical Science to High Performance in Sport.

Consequently, there is no mathematics prerequisite for DATA1001, and no prior statistical understanding is assumed. As a proxy for general academic ability for domestic students, the Australian Tertiary Admission Rank (ATAR) - which is a number between 0.00 and 99.95 that indicates a student's

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.*

position relative to all the students in their age group - is shown in Table 2 indicating the diversity of the cohort.

Table 2. DATA1001 domestic cohort by ATAR band in 2023 S2

ATAR Band	98-99.95	95-97.95	90-94.95	85-89.95	80-84.95	70-79.95	60-69.95	< 60
Size	108	144	232	188	116	79	12	24

CURRICULUM DESIGN

As the aim of DATA1001 is for students to become "statistical storytellers", the entire curriculum is designed explicitly around storytelling. This approach is enabled by the "Sydney Data Stories," which is a large, collaborative, campus-wide project led by the DATA1001 teaching team since 2018.

Learning through storytelling

Learning through storytelling is a well-established pedagogy, as it allows concretizing, assimilating, and structurising of cognitive thought (Evans & Evans, 1989). O'Donnell (2015) argued that all curriculum is storytelling: "Sequencing of selected learning experiences to aid consequential meaning making is also a fitting definition of curriculum" (p. 3).

In statistics education, the power of storytelling has been widely recognised (Chick & Pierce, 2010), with Pfannkuch et al. (2010) positing that "Language and the telling of data stories have fundamental roles in advancing the GAISE agenda of shifting the emphasis in statistics education from the operation of sets of procedures towards conceptual understanding and communication" (p. 1). In a data science context, storytelling is not a framework imposed on the syllabus content; rather data science content has a natural storytelling shape, as shown in Figure 1. The data itself has an original story (its context), and then applying statistical thinking and computational skills to the data allows the emergence and creation of new sub-data stories (insights from the data). These sub-stories can be reframed as research questions, which provide a natural scaffold for learning statistical and computational literacy.

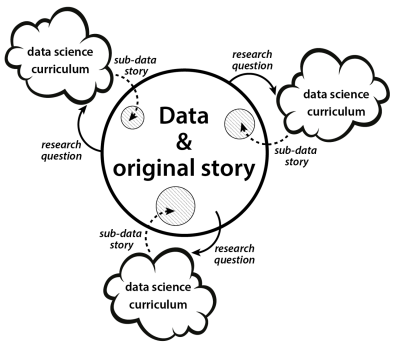

Figure 1: Storytelling Model in Data Science Education

An interdisciplinary approach to sourcing data stories

In the Sydney Data Stories, we invite colleagues across the university to bring their stories into the lecture theatre. See examples of recent data stories in the Appendix, from fields such as teenage psychiatry, road safety, and real estate economics.

The Sydney Data Stories capitalizes on the inherently interdisciplinary nature of data science. See Figure 2, which is an adaptation of Conway's seminal Data Science Venn Diagram (2010), with data science education positioned in the centre, and the domain knowledge sphere arising from colleagues. By interdisciplinarity, we mean the integration of knowledge from multiple domains. The discipline of data science requires knowledge of mathematics and statistics, computational skills, as well as the domain knowledge required to contextualise the data being studied.

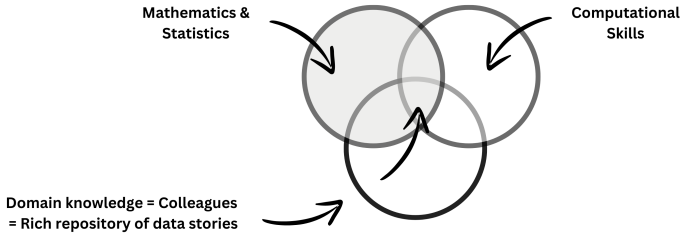

Figure 2. Interdisciplinarity in Data Science Education

As our colleagues are the discipline experts in their data (the ultimate "data dictionary"), the "Sydney Data Stories" approach enables students to understand the story behind the data, the nature of the variables, and how to find more understanding of it. Students may even re-meet the presenters as their lecturers in other units (e.g., Biology or Business), which fosters the integration and synthesis of their learning. Further, this collaboration promotes increased ownership of data science education across the institution.

Building curriculum around data stories

The Sydney Data Stories contributes an endless supply of interesting, current data stories to be embedded in the data science curriculum. Storytelling provides a cohesive, immersive way to carry the curriculum design, whereby students start in a story, stay in the story for theory, and then end with the story. Ideally, each data story is aligned to the learning outcomes, and introduces the week's material, providing the context and motivation for the learning activities, as shown in the Appendix. We built learning activities based on the researcher's specific story and data, including masterclasses, lab exercises, and data projects.

Each Topic (week) has five parts, as shown in Figure 3.

- Imagine (Data story) – Starting with the data story, allows it to be enabled (O'Donnell, 2015), as students connect with the domain first, and then watch a personal story by the researcher.
- Discover – Students learn the concepts, using the same domain as the data story.
- Challenge – Student do open-ended questions in groups based on the data story.
- Explore – Students start with guided exercises in RStudio, and then try open-ended exploration using real data related to the story.
- Evaluate – A revision quiz which has questions from each of the four learning activities, including the data story in Imagine.

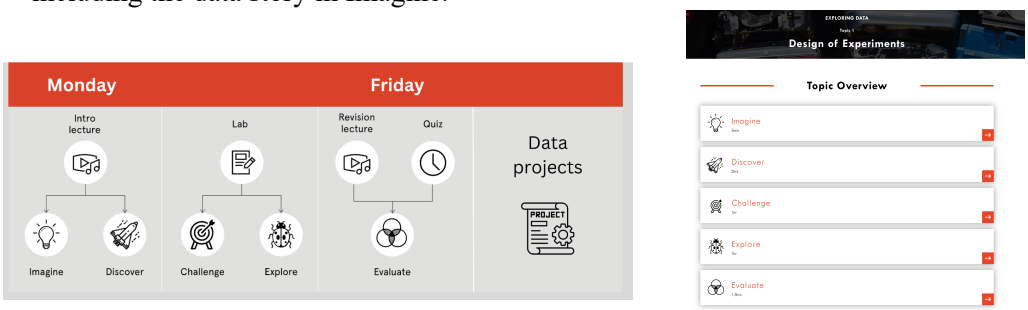

Figure 3. Weekly model for building learning activities in DATA1001 based on the Imagine data story, with exemplar for Topic 1

CASE STUDIES

Here we outline three examples of how we move from the data story to curriculum, with a particular focus on Study 1.

Study 1: Teenage mental health

The learning outcomes for Topic 1 (week 1) are LO1 (the importance of statistics) and LO2 (study design) – see the Appendix.

- Imagine (Data Story): A/Prof Elizabeth Scott is an expert in youth mood disorders, service developments for youth mental health, as well as sleep and circadian dysfunction. She introduced the complexity of data collection for studying teenage mental health, including ethics and randomised controlled trials. In addition, students were encouraged to consider how to manage their own mental health at university.
- Discover: The lecturer explained the importance of statistics and two types of study design (randomised controlled trials and observational studies) in the mental health area, including the relationship between an anti-acne drug (Roaccutane) and depression, as shown in Figure 4.
- Challenge: Students invented a medical trial, using a randomised controlled experiment, identifying all participants and identifying possible confounders and bias.
- Explore: An overview of RStudio without real data, to ensure low cognitive load, and students met their project group, with emphasis on healthy mental health practices.
- Evaluate: A revision quiz.

Figure 4. Discover lecture outline for Topic 1, including a data story on depression

Study 2: Indigenous marine biology

The learning outcomes for Topic 5 (week 5) are LO5 (linear regression) and Graduate Quality (Cultural Competence) – see the Appendix. The Imagine Data Story was presented by Dr. Mitch Gibbs, who is a Thunghutti man through kinship of the Dunghutti nation, studying the effect of climate change on oysters. Gibbs introduced the importance of linear modelling, as well as how cultural knowledge systems affected his research. The subsequent learning activities are based on Sydney air pollution data, as well as a Shiny App based on Gibbs' oyster data research (https://garthtarr.shinyapps.io/experiment/).

Study 3: A rare butterfly collection

Dr. Jude Philp is the Senior Curator of the Macleay Collections at the University of Sydney, which includes a rare butterfly collection. After a Masterclass with Philp, very messy butterfly data were given to the Advanced students (DATA1901) as a research project, which led to a Shiny App-based installation in the Chau Chak Museum on campus. See Warren & Clarke (2019) and Warren (2022) for discussions on the data projects. Statistical storytelling through data projects, not only helped students develop statistical literacy, but also develop transferable communication skills.

OUTCOMES FROM STUDENT DATA

Despite the diverse background of the DATA1001 cohort, with approximately 44% having basic or no mathematical background from high school, the design of DATA1001/1901 seems to be accessible and equitable, with a high proportion of students passing even with no mathematical background (over 70%). Moreover, for domestic students, the correlation between their marks in the Extension 1 (Advanced) or Mathematics (Basic) high school units and DATA1001 were only 0.16 and 0.042 respectively, suggesting a very low association.

Searching for the commentary on "data stories" in the Undergraduate Student Surveys (USS) qualitative feedback, and coding by common ideas, we found references to "data stories" were predominantly positive. Some of the following themes emerge.

Engaging

Basing the data science curriculum on data stories from the institution enabled the acquisition of statistical literacy to be accessible and compelling to students with a broad range of backgrounds. By statistical literacy, we mean the ability to problem-solve with data and appropriately communicate findings in context – that is, becoming a data storyteller.

When there is a story attached, and something to solve statistically with that data story, everything becomes more engaging. (Research survey student feedback, Dec 2018)
The data stories are so interesting! It made data science seem relevant to life outside of university and future careers. (USS Feedback, 2022 S1)
The data stories in each new lecture made the content (which could have been quite dry otherwise) much more interesting, relevant, and enjoyable to learn about. It felt more interactive and I found myself looking forward to the conclusion of each data story as I was curious about the outcome! (USS student feedback, 2022 S1)

Inclusive

While students may have initially preferred data stories from their own field, many appreciated a range of stories.

Making the course content so diverse means that a wide variety of students can engage with different aspects of the course content, whilst demonstrating the interdisciplinary nature of data science. (Research survey student feedback, Dec 2018)
Data stories are so engaging and broad. [It's] very interesting to see data science applied in all different aspects of research and life. (USS Feedback, 2022 S1)
The variation between weeks was captivating. Looking at data stories across different fields. (USS Feedback, 2022 S1)

Wholistic

As data stories allowed engagement with broader issues, such as values and ethics, they fostered the development of non-discipline-specific skills. For example, one student wrote: *"By learning about statistic[al] application[s] on social problems, like the impacts of cigarettes are having on health…we are able to find interesting trends [that] can then be used to find solutions to or reduce problems that people are being affected by." (Research survey student feedback, Dec 2018).*

Overall, in the USS Feedback, the data storytelling pedagogy seems to be viewed positively, as expressed in these exemplars:
The use of real-life data, examples and stories to help illustrate points clearly." (USS Feedback, 2023 S1)
Data stories were useful in contextualising the content we were learning and showing how it can be put into practice. (USS Feedback, 2023 S1)

One student, however, expressed a preference for a more classic theory-driven approach: *A lot of time is spent fleshing out data stories ... it would be more effective if the most important theoretical principles were laid out first before diving into examples*. (USS Feedback, 2023 S2)

CHALLENGES AND STRATEGIES

Building capacity in the teaching team

Gould (2010) helpfully identified some of the challenges in implementing a GAISE-style data-centric approach to the statistics curriculum, including the need to upskill instructors to teach technology, as discussed in Pfannkuch et al. (2010) and more recently by da Ponte and Noll (2018).

In response, we have developed a mentoring team-teaching ecosystem at scale, which leverages the increasing cohort sizes in fast-growing data science units. It involved recruiting top-performing

third-year data science students ("demonstrators") to assist lead tutors in running lab classes. This has many advantages, including the professional development of tutors and demonstrators, and fostering a culture of continual improvement in learning activities and assessments, as outlined in Warren et. al. (in press).

Privacy of data

In a related sub-unit (OLET1632 Shark Bites and other Data Stories), a researcher reported on the Australian Shark Incident Database. Some more sensitive variables in the database are not available for public release, however, given the potential for unhelpful reporting in the public media. As a result, we asked the students to sign a non-disclosure agreement before being given the data, and the data were only available during class time. This allowed the students an authentic experience in data ethics.

Managing cognitive overload

A data-centric approach required management of cognitive overload and differentiation in the class, as there are many different areas of knowledge and competence required in data science – from digital literacy, teaching students to ask research questions, and unlearning common mistakes with interpretation of data. Underlying all these challenges is the student's ability to self-learn (tell their own story with data), as much data science is learnt by engaging with the community of "global storytellers" (e.g., www.kaggle.com, stackoverflow.com).

One strategy we deployed is scaffolding the learning of coding through explicit "Coding Milestones". This prevented instructors from accidentally using more advanced coding in their demonstrations, and allowed students to identify coding threshold concepts, as well as celebrate their development in coding skills.

CONCLUSION

Our project suggests implications for interdisciplinary curriculum design, as well as changing our relationships with researchers from other disciplines - leveraging their data story for motivating diverse cohorts. Our project aligns with the Roundtable's focus on inclusive approaches to teaching and learning that (1) support and motivate students with diverse needs, (2) acknowledge and validate indigenous and cultural knowledges related to data, and (3) take an interdisciplinary perspective. It is a holistic, engaging, and inclusive approach to data science education.

ACKNOWLEDGEMENTS

We would like to warmly acknowledge the reviewers as well as the presenters of the Sydney Data Stories. This research is based on an Ethics project at the University of Sydney: 2017/995

REFERENCES

Advanced Analytics Planning and Enterprise Data. (2024). DATA1001 student data [Unpublished raw data]. University of Sydney.

Chick, H. L. (2007). Teaching and learning by example. *Mathematics: Essential research, essential practice, 1,* 3-21. https://files.eric.ed.gov/fulltext/ED503746.pdf

Chick, H., & Pierce, R. (2010). Helping teachers to make effective use of real-world examples in statistics In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8)*, International Statistical Institute. https://iase-web.org/documents/papers/icots8/ICOTS8_2F2_CHICK.pdf?1402524969

Conway, D. (2010, September 30) *The data science Venn diagram.* Drewconway.com. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

da Ponte, J. P., & Noll, J. (2018). Building capacity in statistics teacher education. In D. Ben-Zvi, K. Makar, & J. Garfield (Eds.), *International handbook of research in statistics education.* Springer. https://doi.org/10.1007/978-3-319-66195-7_14

Evans, R. D., & Evans, G. E. (1989). Cognitive mechanisms in learning from metaphors. *The Journal of Experimental Education, 58*(1), 5-19. https://doi.org/10.1080/00220973.1989.10806518

GAISE Report ASA Revision Committee. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K–12 curriculum framework.* American Statistical Association. https://www.amstat.org/asa/files/pdfs/gaise/gaiseprek-12_full.pdf

GAISE College Report ASA Revision Committee. (2016). *Guidelines for assessment and instruction in statistics education College Report 2016*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

Gould, R. (2010). Statistics and the modern student. *International Statistical Review, 78*(2), 297-315. https://doi.org/10.1111/j.1751-5823.2010.00117.x

O'Donnell, M. (2015). Curriculum narratives: Learning as transition, transition as learning. In *STARS: Students Transitions Achievement Retention & Success 2015 proceedings*. https://www.unistars.org/papers/STARS2015/09D.pdf

Pfannkuch, M., Regan, M., Wild, C., & Horton, N. J. (2010). Telling data stories: Essential dialogues for comparative reasoning. *Journal of Statistics Education, 18*(1).

Warren, D. (2022). Mobilising the student's voice in Data Science education: the Great Barrier Reef data project. In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the gap: Empowering & educating today's learners in statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11)*, International Association for Statistical Education.

Warren, D., & Clarke, S. (2019). Choose your own adventure: Experiencing research through first-year group projects in data science. *ACSME, Sydney: The University of Sydney.* https://openjournals.library.sydney.edu.au/IISME/article/view/13477

Warren, D., Tarr, G., & Patrick, E. (in press). Promoting excellence and growth in data science education: Developing a mentoring ecosystem. In *5th International Conference for Mathematics Education proceedings.*

APPENDIX
Examples of "Sydney Data Stories" aligned with topics and learning outcomes in DATA1001/1901

Researcher	**Sydney Data Stories**	**Topic**	**Learning Outcomes / Graduate Quality**
A/Prof Elizabeth Scott Teenage Psychiatry	Imagine 1	1. Design of Experiments	LO1. Articulate the importance of statistics in a data-rich world, including current challenges such as ethics, privacy and big data LO2. Identify the study design behind a dataset and how the study design affects context specific outcomes
Prof Mike Bambach Road Safety	Imagine 2	2. Data & Graphical Summaries	LO3. Produce, interpret and compare graphical and numerical summaries, using base R and ggplot
A/Prof Danika Wright Real Estate Economics	Imagine 3	3. Numerical Summaries	
Dr Kylie Moulds Sports Science	Imagine 4	4. Normal Model	LO4. Apply the normal approximation to data, with consideration of measurement error
Dr Mitch Gibbs Marine Biology	Imagine 5	5. Linear Model	LO5. Model and explain the relationship between 2 variables using linear regression Graduate Quality (Cultural Competence). Actively, ethically, respectfully, and successfully engage across and between cultures. In the Australian context, this includes and celebrates Aboriginal and Torres Strait Islander cultures, knowledge systems, and a mature understanding of contemporary issues.
Dr Jason Chin Evidence Law	Imagine 6	6. Understanding Chance 1	LO6. Use the box model to describe chance and chance variability, including sample surveys and the central limit theorem
Andy Tran (Bioformatics)	Imagine 7	7. Understanding Chance 2	
A/Prof Mel Keep Instagram	Imagine 8	8. Sample Surveys	
Dr Lara Ford Immunology	Imagine 9	9. Hypothesis Testing 1	LO7. Given real multivariate data and a problem, formulate an appropriate hypothesis and perform a range of hypothesis tests LO8. Interpret the p-value, conscious of the various pitfalls associated with testing
Dr Llew Mills Addiction Medicine	Imagine 10	10. Hypothesis Testing 2	
Andy Tran (Bioformatics)	Imagine 11	11. Hypothesis Testing 3	