DARE
ARC Training Centre in
Data Analytics for Resources & Environments

Australian Government
Australian Research Council

Data Science in University Teaching

This is the response from DARE to: https://amsi.org.au/amsi-ssa-data-science-review/

General response

To provide background and context to this response, DARE (Data Analytics in Resources and the Environment) is an ARC Industrial Transformation Training Centre. As part of the program, DARE has delivered three semesters of teaching at PhD level in Probabilistic Models for Complex Data (DATA6810) and Computational Inference for Machine Learning (DATA6811). These units were aimed at making sure all our HDR candidates started at the same level of knowledge in Data Science, regardless of their background. The units combined a grounding in Bayesian statistics with higher level computational skills.

As a multidisciplinary Centre, DARE brings together academics across Mathematics and Statistics, Computer Science and three applied domains, Minerals, Biodiversity and Water.

DARE agrees that currently the discipline of "Data Science" is ill-defined and can mean different things to different people. We also agree that in many cases "Data Science" is associated with computer science and data wrangling rather than with mathematics and statistics.

As an example, Data Science is often interpreted in the context of a company wanting to improve its revenue with recommender systems and persona segmentation for their marketing. In this case there are readily available standard machine learning tools that can be applied. As such a good knowledge of computer science will carry more weight, because the focus is to work efficiently within the company's pre-existing computing infrastructure and knowledge ecosystem.

Another example might be an organisation, which due to an increased ability to monitor systems and gather sensing data is talking about using Data Science to develop dashboards to quickly interpret the increasing volume of data. This is another typical interpretation of Data Science, but the approach involves data management and visualisation, not Data Science per se.

However, it needs to be recognised that some of the most exciting Data Science has been developed in tandem with applications in domain science. A clear example of this is the huge contributions from Astronomy to developments in Data Science.

Defining Data Science as a discipline is a pressing need in academia and industry, but can be difficult due to the vast diversity of perspectives.

Defining Data Science will assist practitioners and stakeholders to better understand the training, skills and developments in this important science area.

Data Science is clearly multi-disciplinary and should be defined in relation to its foundations in Computer Science, Mathematics and Statistics as well as the different domain applications.

DARE believes that the discipline of Data Science is multidisciplinary and should integrate foundations, algorithms and applications.

The foundations for Data Science consist of understanding of data management, modelling and statistical inference. These are generally provided by grounding in Mathematics and Statistics, but there is a clear intellectual overlap with Computer Science, particularly in machine learning, but also in statistics. More generally speaking, Computer Science skills are needed to be able to extend the algorithms to different applications.

However, it is the combination of skills in computer Science and Mathematics and Statistics that provides the strength of the Data Science discipline. A further strength is the close link of Data Science to applications in domain specific areas. Mathematicians and computer scientists need to work in collaboration with the experts to interpret the data and build better models and algorithms. This in turn leads to actual decision making, building on the tools and the theory developed in data science.

Without a solid understanding of the statistics, there is a risk of overreliance on "packaged solutions" in the application of Data Science (as the example above suggests). However, dealing with larger and larger datasets requires increased computing knowledge and scripting, simply to manage the data. In return, more and more complex problems require new statistical insights to discover causal relationships and deal with complex distributions. In some way, "clever mathematics" needs to collaborate with the "brute force computer science" to solve increasingly complex problems in the increasingly connected world.

One could argue that Computer Science can provide innovation in faster, better and scalable tools and algorithms for Data Science. Statistics then provides the insights, the understanding, and the innovation in the underlying theory for Data Science. However, as highlighted earlier, there is a huge intellectual overlap between the two sub-disciplines, and therefore there is also huge opportunity for close collaboration in appointments, teaching and research.

Furthermore, computer science includes the important sub-discipline of human computer interaction. This deals with the deeply human issues of making use of statistical and algorithmic tools to enable people to make decisions that are informed by data.

The strong integration of Data Science with domain applications also creates enormous opportunities for multidisciplinary collaborations, which will drive further innovations in all associated sub-disciplines.

Specific questions from feedback form

1. What is the role of your university department within the university's Data Science offering?

Reply:

As part of the program, DARE has delivered three semesters of teaching at PhD level in Probabilistic Models for Complex Data (DATA6810) and Computational Inference for Machine Learning (DATA6811). These units were aimed at ensuring our HDR candidates started at the same level of knowledge in Data Science, regardless of their background. The units combined a grounding in Bayesian statistics with higher

level computational skills. These units are currently transformed into micro-credentials for professionals and students.

2. Have there been recent changes to your department, staffing, degree or subject offerings, which have occurred as a result of Data Science initiatives?

Reply:

DARE is an ARC Industrial Transformation Training Centre focusing on Data Science and uncertainty quantification for applications in Minerals, Biodiversity and Water (https://darecentre.org.au). DARE incorporates multiple Universities and will run from 2020 – 2025, and therefore was an investment from industry, government and the Universities into Data Science. DARE has employed several new postdocs and research engineers in the area of Data Science and aims to promote Data Science to the community.

3. What are the most effective models for serving teaching in Data Science? Is there value in joint teaching by academic and industry statisticians?

Reply:

Please see our extended reply. As DARE sees Data Science as multidisciplinary incorporating Mathematics and Statistics, Computer Science and Domain applications, the teaching needs to reflect this. The best approach is to provide the teaching in context of case studies. This can involve a domain problem early on in the course, leading to understanding and training of the required Math & Stats knowledge and Computer Sci knowledge, leading at the end of the course to a cap-stone application of the knowledge to the domain problem. This scaffold can be applied to a whole degree, a single unit of study or to a short course.

Integration with industry is exciting for students and all stakeholders. This leads to new conversations and partnerships between academics and industry and provide direct links for students into industry. As academics, we can be too focused on our own sub-discipline, and interactions with industry can open up new avenues for research and teaching.

However, this is not without challenges. As with any collaboration, setting and defining clear expectations for the industry partner and the academics involved is a crucial step for a successful collaboration.

4. How can the mathematical sciences benefit from opportunities offered by the growth of Data Science in Australia? How can the role of mathematics and statistics be more centred in Data Science teaching?

Reply:

Please see our extended reply. As is highlighted Data Science should integrate foundations, algorithms and applications.

The foundations for Data Science consist of understanding of data management, modelling and statistical inference. These are generally provided by grounding in Mathematics and Statistics, but there is a clear intellectual overlap with Computer Science, particularly in machine learning, but also in statistics. More generally speaking, Computer Science skills are needed to be able to extend the algorithms to different applications.

However, it is the combination of skills in computer Science and Mathematics and Statistics that provides the strength of the Data Science discipline. A further strength is the close link of Data Science to applications in domain specific areas. Mathematicians and computer scientists need to work in collaboration with the experts to interpret the data and build better models and algorithms. This in turn leads to actual decision making, building on the tools and the theory developed in data science.

The discipline of Mathematics and Statistics is very much part of Data Science, but it needs to recognize its strong intellectual overlap and interactions with the other required skills to develop the overall Data Science discipline skill set.

Maths & Stats in public messaging could advertise this integration, multidisciplinarity and the foundation skills in Maths & Stats more to make a stronger case within the Data Science discipline. This will increase the ability to capitalize on the opportunities that currently exist.