Academic IT support for Data Science

Simon Price

Advanced Computing Research Centre, IT Services, University of Bristol, Beacon House, Bristol BS8 1QU, UK, simon.price@bristol.ac.uk

Keywords
Data Science, Big Data, Research Data Management, Research Software Engineering

1. Summary

Globally, over 500 universities now offer data science courses at undergraduate or postgraduate level and, in research-intensive universities, these courses are typically underpinned by academic research in statistics, machine learning and computer science departments and, increasingly, in multidisciplinary data science institutes. Much has been written about the academic challenges of data science from the perspective of its core academic disciplines and from its application domains, ranging from sciences and engineering through to arts and humanities. However, relatively little has been written about the institutional information technology (IT) support challenges entailed by this rapid growth in data science. This paper sets out some of these IT challenges and examines competing support strategies, service design and financial models through the lens of academic IT support services.

2. THE DATA SCIENCE LANDSCAPE

There is currently no universally agreed definition of "data science" and, much like the related term "big data", the term "data science" has been adopted by different communities to refer to substantially different concepts. In industry, data science is frequently associated with data analytics, data engineering and big data technologies such like Hadoop or NoSQL. These associations are also found in the university sector, where data science courses have emerged from an eclectic range of academic departments (Swanstrom 2016). However, research-intensive universities tend to emphasise the established academic disciplines of applied statistics, machine learning and computer science over specific technologies. Despite 52% of data scientists on LinkedIn having only earned the title in the past four years (RJMetrics 2015), a well founded academic claim has been made that data science as a practical applied discipline is at least 50 years old (Donoho 2015).

Setting aside industry's and academia's differing views on the core competences of a data scientist, few would argue against the multidisciplinary nature of data science nor that its impressive range of applications is entirely dependent on IT. Given the major investment in data science research and education by governments and industry, right across the spectrum of academic disciplines, this might seem like good news for financially hard-pressed academic IT support services.

The Alan Turing Institute, the UK's national centre for Data Science, was launched in January 2015 with an initial £52m (€67m) investment from government and industry (ATI 2015). Headquartered at The British Library and further funded through the UK's Engineering and Physical Sciences Research Council (EPSRC), the five founding partner universities of Cambridge, Edinburgh, Oxford, UCL and Warwick have each committed to major recurrent investment in Data Science within their institutions. In the United States, universities such as UC Berkeley, NYU, MIT and Stanford have major Data Science programs and the universities of Rochester and Michigan have both individually made investments on a similar scale to the UK's initial ATI investment.

Why then, with all this inward investment, is this not necessarily good news for the traditional university IT Services division?

3. IT CHALLENGES OF DATA SCIENCE

One challenge lies in the fact that effective research IT support, and data science research in particular, demands more than just the standardised, centralised, corporate-style IT support that universities have increasingly moved their IT divisions towards, albeit with worthy intentions of reducing operating expenditure by exploiting economies of scale. That 21% of the data science courses listed in (Swanstrom 2016) are delivered online suggests that this corporate approach to education IT might be working as well for data science as for other disciplines. However, research IT support for the application of data science across multiple disciplines is far more challenging.

Research data comes in a broad range of data types and formats from heterogeneous sources, including laboratory equipment, scanners, medical devices, sensors networks, digital collections or the web; it is processed through computational pipelines and scientific workflows; it is analysed and visualised using data science methods that may also require considerable time investment to gain an understanding of the data and domain together. The specialist IT knowledge required to support such involved and domain-specific work is more likely to originate from staff (or students) embedded in the research group itself rather than from generalist IT support staff working centrally.

In other words, the traditional academic IT Services division is highly unlikely to have enough staff with either a background in each specific research domain nor with up-to-date experience working as an applied data scientist. It is therefore not surprising that the money flowing into universities to develop data science research is being channeled into the operational expenditure of the research groups themselves rather than into IT Services. However, the irony is that the research groups do not have, and will never be able to afford, enough people with this rare combination of skills either.

4. CO-DESIGNING IT SUPPORT FOR DATA SCIENCE

In 2015, the University of Bristol started designing a data science institute, following its established model for multidisciplinary university-wide institutes. While its existing institutes address specific grand challenges, such as climate or health, the new one concerns a set of widely applicable methods, tools. A process of co-design between academics and professional services is exploring a broader model, the Data Institute, builds on Bristol's existing strengths in research data storage and management, advanced computing and research software engineering (Gardiner et al. 2012).

The main challenge Bristol or other universities in providing world-class support for data science research is to find economically sustainable ways of developing and bringing together teams of people, for the right period of time, who collectively have the right mix of skills to support academic IT aspects of data science embedded within research groups. From the IT side, a survey of all 190+ IT Services staff skills provided an up-to-date profile of relevant skills to inform training and future recruitment needs. On the academic side, a variety of staffing, advocacy and costing models are being considered to bridge the data science skills gap that neither academic IT nor the Data Institute can achieve alone. Options currently being explored include shared posts, facility-based versus fractional posts as a means of cost recovery from grants - to fund people and equipment. Also, potential synergies with corporate data storage, management and analytics are being examined: can we apply our data science expertise from research to bring wider benefits to the organisation.

5. REFERENCES

ATI website (2015). *The Alan Turing Institute - the UK's national institute for data science*. Retrieved February 12, 2016, from: https://turing.ac.uk/.

RJMetrics website (2015). *The State of Data Science*. Benchmark Report Series. Retrieved February 12, 2016, from: https://rjmetrics.com/resources/reports/the-state-of-data-science/.

Donoho, D. (2015). *50 years of Data Science*. Tukey Centennial Workshop, Princeton, NJ, September 18, 2015. Pre-print, Version 1.00, September, 18, 2015, Stanford University.

Gardiner, C., Gray, S., Price, S., Boyd, D., Knight, V., Steer, D. & Cregan, B. (2012). *Towards sustainable, enterprise-scale Research Data Storage and Management*. Digital Research 2012, University of Oxford.

Swanstrom, R. (2016) DataScience.Community website. *College & University Data Science Degrees*. Retrieved February 12, 2016, from: http://datascience.community/colleges.

6. AUTHOR BIOGRAPHY

Dr Simon Price is the Academic Research IT Manager for the University of Bristol and a Visiting Fellow in Computer Science, with research interests in Data Science, e-Science, e-Research and Digital Humanities. He leads the Research IT group in the Advanced Computing Research Centre, a joint venture between IT Services and the Department of Computer Science. Over the last two decades years he has led a wide range of academic IT projects and services at the University of Bristol, including strategic institutional projects like the data.bris Research Data Service, IT Services R&D, the Institute for Learning and Research Technology (ILRT), the Learning Technology Support Service, as well as the Bristol Online Surveys (BOS) national service as well as numerous academic software development projects. https://uk.linkedin.com/in/snprice