What is data science?

"Data science", can best be understood as a retronym from the job title "data scientist". Scientists at LinkedIn and Facebook came up with this title in 2008, to describe a new generation of highly productive data analysts in the tech industry (Patil 2011). The demand for data scientists grew quickly. The 2012 article named "Data Scientist" as "The Sexiest Job of the 21st Century" (Davenport and Patil 2012). According to this article: "... what data scientists do is make discoveries while swimming in data ... At ease in the digital realm, they are able to bring structure to large quantities of formless data and make analysis possible. ... Data scientists' most basic, universal skill is the ability to write code."

At first it was not clear which part of this description was distinctive of data scientists. Maybe the new thing was the data (big, complex, varied) or the analysis strategies (machine learning, artificial intelligence). A common response in universities was to design vocational masters programmes with training in big data and machine learning.

Other authors and institutions began to take a broader view; the changes in data and analysis were not fundamental to the nature of data science - what had changed was the *culture* of data analysis, a culture that treated programming as a fundamental tool in finding, cleaning, exploring and analyzing. In 2013, two large grant-giving bodies awarded $37.8 million to support "data science environments" at Berkeley, the University of Washington and New York University, with broad themes including "Tools and software", "Reproducibility and open science", and "Physical and intellectual space" (see About MSDE). In 2015, David Donoho, an eminent statistician and computational scientist, distinguished between "Lesser data science" – restricted to big data and machine learning – and "Greater data science" – a wider field of the science of learning from data (Donoho 2015). He argued, convincingly, that data scientists had rediscovered J.W. Tukey's "data analysis" (Tukey 1962) (see appendix "Data science and statistics"). He saw great risks in concentrating on Lesser Data Science: "Choosing in this way is likely to miss out on the really important intellectual event of the next fifty years".

As the interpretation of data science matured, so did its position in the curriculum. Berkeley found that many undergraduates from different fields were taking introductory classes in statistics and computer programming, and concluded that students were responding to a general need for training in coding for data analysis. In 2015, they pioneered an entry level undergraduate course "Foundations in Data Science". It is open to students in all schools, including the arts, and has no requirements for university courses in mathematics or programming (see data8 policies). It has become the fastest growing program in the University's history, with 2000 students projected to enroll for 2017-18 (Data Science Environments 2017). A 2018 report from the US National Academies of Sciences, Engineering and Medicine argues that there should be "development of a basic understanding of data science in all undergraduates" (National Academies of Sciences and

Medicine 2018).

The basic tools for data science training are high-level programming languages, especially `R` (a statistical programming language) or `Python` (a general purpose language widely used in science, data science, and teaching) (Hardin et al. 2015).

In retrospect, the outline of the movement that we now call data science has been apparent for a number of years. Many of us have seen exceptional graduate students or post-docs who have vastly expanded the range of data they can work on, by using programming. Our training is starting to reflect this experience. The first year of Midlands Integrative Biosciences Training Partnership (MIBTP) programme explicitly teaches programming, and statistics through programming in `R`; student self-reflection and the quality of its graduates suggests that this training has had a major impact in their ability to do research. There is `R` and `Python` training in the Bioinformatics MSc; a requirement for `R` in the "Network Geographies" third-year module; `Python` in the "Urban Analytics" module, and a compulsory programming module in the MSc in Brain Imaging and Cognitive Neuroscience.

References

Data Science Environments, The Moore-Sloan. 2017. "Creating Institutional Change in Data Science." http://data8.org/su18/policies.html.

Davenport, Thomas H, and DJ Patil. 2012. "Data Scientist: The Sexiest Job of the 21st Century." *Harvard Business Review* 90 (10):70–76.

Donoho, David. 2015. "50 Years of Data Science." In *Princeton NJ, Tukey Centennial Workshop*. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

Hardin, Johanna, Roger Hoerl, Nicholas J Horton, Deborah Nolan, Ben Baumer, Olaf Hall-Holt, Paul Murrell, et al. 2015. "Data Science in Statistics Curricula: Preparing Students to 'Think with Data'." *The American Statistician* 69 (4). Taylor & Francis:343–53.

National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25104.

Patil, DJ. 2011. *Building Data Science Teams*. O'Reilly Media, Inc.

Tukey, John W. 1962. "The Future of Data Analysis." *The Annals of Mathematical Statistics* 33 (1). JSTOR:1–67. http://projecteuclid.org/euclid.aoms/1177704711.