Time for data science to professionalise

Data science must become a profession if it is to get to grips with the issues of data ethics, argues **Detlef Steuer**

Data scientists need an understanding of ethics. That much has become clear in recent years. Books and newspaper articles – even Netflix documentaries – tell of cases where data collection and data sharing practices may have undermined trust: earlier this year, for example, there were reports of an antivirus software company selling user data to advertisers (bit.ly/38HGhyh). There have also been stories of algorithms built with the noble intention of making decisions fairer and more efficient by removing human bias from the equation, only for the resulting decisions to end up biased by the data that the algorithms were trained on (reut.rs/37Czbud).

The importance of ethics for practitioners of data science was underlined by the publication of "A Guide for Ethical Data Science", jointly authored by the Royal Statistical Society (RSS) and the Institute and Faculty of Actuaries (IFoA). The guide states that "[data] science has the potential to be both beneficial and detrimental to individuals and/or the wider public", and that "[to] help minimise any adverse effects, members can seek to understand the potential impact of their work and consider any opportunities that may deliver benefits for the public" (bit.ly/3hEd8Yv). But such guides are only a first step to embedding ethics in data science. To fully achieve such an aim, the goal must be to professionalise data science.

I am convinced that the problems with applied data science exist because data science currently does not constitute a profession but is instead an occupation. In *Significance* in December 2018, Robert Langkjær-Bain wrote that "data are not just numbers without meaning or context",[1] and it is my contention that the "meaning and context" of data is not currently the subject of scientific investigation in the field of data science. This needs to change, and that is where the professionalisation of data science must lead us.

Data scientists with strong roots in statistics, or biostatistics, of course inherit the long tradition of ethical considerations in these fields. Nevertheless, statistics is only a subset of data science, as I will explain later.

Occupation versus profession

According to the philosopher Timo Airaksinen,[2] a profession differs from an occupation through several characteristics:

1. *Scientific training*, through which professionals know what is to be done by understanding the rational, epistemological foundations of professional action.
2. *Autonomy*, meaning that a profession can influence the social decisions that regulate its members' work and their related rights and obligations, separate from the requirements of any law or statute.
3. *Professional ethics*, meaning that professional advice cannot be unproblematically rejected, challenged, or ignored by the public when the public needs professional expertise. Therefore, existing professional ethics become a key issue when the public evaluates the potential influence of professional advice in relation to the quality of their life.

These characteristics are reflected in the code of conduct of the RSS (bit.ly/30QlvKJ), thereby defining statistics as a profession. For example, to quote from the code: "In general, the public has no ready means of judging the quality of professional service except from the reputation of the provider. Thus it is essential that the highest standards are maintained by all Fellows whenever they are acting professionally and whatever their level of qualification."

An occupation, distinct from a profession, is characterised mostly by work done according to given orders, and of being paid for the *result* of the work, not necessarily for the *skills and expertise* of the worker. Occupations do not necessarily have the independence of a profession, and may not require the broad and deep training that defines a profession.

Data science, in its current form, is not a profession. Writing in 2015, Michael Walker, president of the Data Science Association, stated that "the data science field is at the very beginning of becoming a profession" (bit.ly/2avfPax). It will have made further progress since those words were written. Still, I argue that not enough progress has been made to meet Airaksinen's definition of a "profession".

Part of the problem is the unresolved nature of the question, "What is data science?". Data science, as a concept, emerged because of a fundamental change in the availability of data. There was no grand vision for it. When it was expensive to gather and store data, it was necessary to consult with statisticians to carefully plan for the appropriate collection, management and analysis of data. But, nowadays, data is abundant. Each of us generates large volumes of information just by living our lives. This data is stored as a matter of course and can be accessed and analysed using freely available software tools. Not all of the analysts working with this data are formally trained statisticians. They are what we now call "data scientists", and they are doing "data science".

So, what is data science? There are competing ideas, but an oft-cited one was given by Donoho, who defined it as "the science of learning from data".[3] He went on to say that data science "studies the methods involved in the analysis and processing of data and proposes technology to improve methods in an evidence-based manner".

According to Donoho, what is called "greater data science" consists of six major elements of data analysis: data gathering, preparation, and exploration; data representation and transformation; computing with data; data modelling; data visualisation and presentation; and science about data science.

Donoho's final point – the "science about data science" – is essential to define the scientific training that should be required of a data scientist's work. However, the definition of "greater data science" fails to link data science with the two other characteristics of Airaksinen's "profession": autonomy and professional ethics.

Dr Detlef Steuer is a lecturer in data analysis and R at the Helmut-Schmidt-Universität, Universität der Bundeswehr, Hamburg.

Data science versus statistics

If statistics is a profession, why not data science? In answer to that question, it should be noted that data science does not directly equate to statistics. Statistical methods may form a part of data science, but data science also draws on parts of computer science and other disciplines.

However, due to the close relationship between data science and statistics, it is reasonable to compare the two.

Where statistics differs from data science is that ethical considerations are well known to statisticians. Statisticians have long-established standards for medical and pharmaceutical research, for example, which reach far beyond pure number-crunching. Ethical committees are an integral part of the process of data collection, and international and national consortia provide guidelines on how to ensure the well-being of patients that contribute their data.

Outside of this highly regulated environment, in the commercial world where data science most commonly operates, companies that collect and analyse data may feel (wrongly) that they have no real need to include any ethical considerations in their operations. But that stance has been weakened somewhat in the aftermath of the Cambridge Analytica scandal (bit.ly/2XxJvyJ) and other similar incidents. Ethical guides, like the RSS/IFoA's, are being drawn up. There are data science codes of ethics pushing for development (bit.ly/336kR9m), ethical frameworks adopted by governments (bit.ly/337a83J), and there is even a "Data Science Oath", modelled on the Hippocratic Oath (bit.ly/2Xy6WrD).

However, patchy acceptance and understanding of ethics is not the only reason why data science does not yet meet the definition of a profession. There is a second severe issue to consider: among those doing data science, there are many who do not know "what is to be done by understanding the rational, epistemological foundations of professional action".[2] In practice, one sees data scientists with all sorts of basic education, including some trained in three-month courses without any "science about data science" at all (to return to Donoho's definition)! For academically trained data scientists, outside of the few newly-built data science study programs, the scientific embedding of data science depends on the main subject of study, and there are at least two very distinct approaches that are taught:

1. *The inferential framework in statistics*, where data is seen as the result of some *data-generating process* in the world, and the goal is to make reliable statements about that world.
2. *The computational learning theory in machine learning*, where data is seen as examples, and the goal is to learn a general concept from these examples that is optimal in a certain sense, for example when applied with an algorithm to new examples.

So, there is no common, rational, epistemological foundation for science about data science in which *all* data scientists are trained.

Additionally – and linked to our earlier problem – there has been a lack of focus on ethics in data science education. The situation is improving. New study programs are starting to provide courses on ethics. But while this is certainly an important step towards a cultural change that integrates the social impact of a profession as part of the science (and training) of that profession, for data science there is still a long way to go.

Becoming a profession

Data science is in a focal position in our data-driven society. Algorithms determine many aspects of our lives and society as a whole. In some respects, data science is very close to being the engineering of the twenty-first century.

In a 2019 blog post for *Scientific American*, the political activist Ralph Nader wrote that engineers are "often the first to notice waste, fraud and safety issues" (bit.ly/2b8WSSK). The same should also hold true for data scientists, especially those with training in inferential statistics. Data scientists should be the first to notice shortcomings in the data to be analysed, to understand how inferential statistics should be performed, and how to properly interpret statistical tests and results. From this knowledge follows responsibility, and thus data scientists should be in a position to speak out if they see abuses of data and misinterpretations of results.

But too often they are not, and indeed some may lack a clear understanding of the responsibilities of data science to wider society. According to a December 2018 working paper from the RSS Data Science Section, data scientists "need to ask the right questions at all levels, discuss bias, and understand who might be harmed or disadvantaged".

Once data science becomes a profession, every individual data scientist could be supported by education and written guidelines on everyday ethical decision-making. Written rules of conduct for data science would help to establish a relationship of trust between data scientists, their clients, their employers, and society.

Individual voices trying to speak out would gain much more attention if representing a profession with autonomy and professional ethics. The status, reputation and power of any data scientist, and of data science as a whole, would be increased.

In cases of conflict of interest, an ethical guideline under the auspices of some professional society could offer an arbitration process. And last but certainly not least: if the expectations and responsibilities of data science would be clearly defined and formulated, it would be easier to fight back against improperly performed data science.

This is why it is imperative that data science becomes a profession. ■

Acknowledgement

I want to thank my colleague and friend Ursula Garczarek for thoughtful discussions on the topic. This article is based on our published paper, "Approaching Ethical Guidelines for Data Scientists".[4]

References

1. Langkjær-Bain, R. (2018) Data rights and wrongs. *Significance*, **15**(6), 12–15.

2. Airaksinen, T. (2009) The philosophy of professional ethics. In R. C. Elliot (ed.), *Institutional Issues Involving Ethics and Justice* (Vol. 1, pp. 201–215). Paris: Eolss Publishers.

3. Donoho, D. (2017) 50 years of data science. *Journal of Computational and Graphical Statistics*, **26**(4), 745–766.

4. Garczarek, U. and Steuer, D. (2019) Approaching ethical guidelines for data scientists. In N. Bauer, K. Ickstadt, K. Lübke, G. Szepannek, H. Trautmann and M. Vichi (eds.), *Applications in Statistical Computing: From Music Data Analysis to Industrial Quality Improvement* (pp. 151–169). Cham: Springer.