Data science, statistical investigations, team sport and assessment

In recent years, the term data science has been used so much it’s a wonder it hasn’t made it onto any shortlists of Oxford Dictionaries’ Word of the Year. Universities and more recently, school authorities, have been anxiously scrambling to set up programs and courses in data science. Business forums have seen discussions on the difference between data science and data analytics, while the stampede to define data science has seen a word-smithing smorgasboard.

Although it is a truism that the jobs of the future may not be able to be imagined today, the current furore over data science is important for many reasons across education levels, across disciplines and across workplaces. It is important for statistics education and the lessons from statistics education are important for data science. It is also an opportunity for statistics and statistics education to grasp and hold.

Wearing a number of hats, both nationally and internationally, I have listened to much about data science over recent years. In October, wearing my ISI (International Statistical Institute) presidential hat, I attended the second United Nations World Data Forum (UNWDF) in Dubai. Because the UNWDF is particularly focussed on the UN SDG’s (Sustainable Development Goals), much is concerned with official statistics and data for development and citizen benefit, so that statistical literacy, sometimes called or confused with data literacy, also features. However, a session which included comments of particular relevance to education, both school and university, was “Data Scientists: What are they?” with speakers leading data science teams in large organisations across industry, business and government, including communications, securities, information technologies, and official statistics. The UNWDF sessions tend to be discussion panels, often conducted in interview format, so I cannot refer you to papers or reports, but if you have a spare 75 minutes, the recorded session is at https://undataforum.org/WorldDataForum/sessions/ta6-08-data-scientist-what-are-they/. Here I paraphrase a few comments highly relevant to education from my rapidly scribbled notes, with apologies to the speakers for extracting just a little from an extensive session:

• Data science is everywhere and not new
  o A label for work being done for years
  o What’s changed is recognition of what can be done and bringing this out of the back room
  o Need to know what can & can’t be done with data

• Data Science is not a person – it’s a team. Diversity essential
  o Need specialists: statisticians, data engineers/computer scientists, machine learning optimiser, subject matter person
  o Need: curiosity in problems; identifying and posing problems so they can be tackled; investigative and problem-solving; bring together data, data problem-solving, technical; challenge team
  o But all members need some statistical foundation

• Data science gives what, statistics gives why/understanding

There are many pointers in the above to be explored, including that statistics is fundamental to data science. The statistical community must not only emphasize this but also avoid any attempt at substitution of the term data science for statistics.
The emphasis that data science is a ‘team sport’ and that diversity of expertise is essential are echoed by employers who emphasize they do not want a production line of hybrid graduates, but want sufficient diversity of graduates with a combination of balance + specialisation to work in effective teams. Hence it is of critical importance that university data science programs include foundations in statistics and relevant information technology, but then allow, and encourage, significant diversity. Such ideas are not new, particularly in quantitative working environments, and particularly those involving statistics. It has long been known that combining statistics and/or mathematics with at least some foundation in another area increases career opportunities, and none more so than statistics with information technology (IT) – a combination which has long opened amazingly rich and diverse careers. However, the IT component needs to be an enabler to tackle anything in IT, and the statistical learning needs to reflect the practice of statistics.

A challenge at school – and indeed also at introductory tertiary – level is the question of how much foundation programming/coding is needed in a technologically and data rich world. This is reminiscent of the long-standing and still-ongoing question of how much mathematics is needed for statistics, and much can be learnt in data science from looking at successes and failures in statistical education, in which success tends to be associated with balance in the purposeful development of thinking and immediate and future learning.

Qualities such as curiosity in problems, problem-solving, communication and teamwork have often been termed generic skills. Generic skills are what you learn while you’re learning something else – they must be embedded in constructive, student-centred learning in contexts relevant to students. For example, teamwork needs to be learnt within contexts in which a team is needed. Investigative skills are learnt in many disciplines, but the combination of qualities and skills identified for the practice of data science are very much those of the statistical investigation process.

There has been long-time advocacy from statisticians and statistics educators on authentic learning of the full statistical investigation process, including: problem elicitation and preparation for tackling statistically; preparing data (including planning, collecting, sourcing, identifying, organizing, validating); exploration of models and assumptions as well as actual analysis; and presentation of findings. Such advocacy has also emphasized: data-driven concepts and statistical thinking; real contexts and complex ‘large’ data; technological and data systems know-how; and student ownership and constructivism.

Therefore, surely statistics learning is exactly what is needed in data science. However, despite some excellent work by many worldwide, there has not been sufficient penetration and implementation of such advocacy, and it is important for data science to reflect on why. There are too many reasons to mention here, but fear is a contributing factor in many. And a big one in which fear plays a major role is assessment: fear of open-endedness; workload fears; fear of creativity and diversity; fear that students “won’t do it right”; fear of uncertainty and variation. An ongoing challenge to leaders in statistical education is to provide exemplars and evidence to allay such fears. For example, even with very large cohorts, workload from inclusion of authentic group statistical investigations can be offset by changing exam and test styles to focus on knowledge and procedures, because the creativity, problem-solving, statistical thinking, statistical practice and communication are assessed within the
investigations. An interesting point here is that multiple-choice questions in introductory statistics tend to be highly curriculum-specific and dependent on local culture/conditions, whereas criteria and standards for investigations tend to be more universal, although they need exemplars. This curriculum-specificity of multiple-choice questions is why questions designed and tested for a specified curriculum may be rejected elsewhere as inappropriate or even misleading.

In both prize-winning papers of 2018 announced in this issue, there are open-endedness, potential for debate and different views in real and complex contexts, but which are still readily accessible and relevant to school and introductory tertiary students.

There are lessons for data science from the past three decades of effort in statistical education and there are many prospects for statistics in data science. Data science gives opportunities to renew push for authentic learning that reflects practice of ‘greater statistics’ and ‘greater data science’.

Helen MacGillivray