INTERNATIONAL DATA SCIENCE IN SCHOOLS PROJECT

Neil Sheldon
Teaching Statistics Trust
mail@neilsheldon.net

This paper explains the background to the International Data Science in Schools Project, the development of the curriculum frameworks, the conceptual approach envisaged, the use to which the frameworks can be put, and the possible ways forward for the project. The paper also discusses, principally in a UK context: the challenges of implementation, appropriate methods of teaching, learning and assessment, resource implications, and the curricular relationship between data science and statistics. Finally, the paper takes a broader view, arguing for a 'data across the curriculum' approach, and distinguishing levels of competence by means of the terms data literacy, data skills and data science.

INTRODUCTION

The International Data Science in Schools Project (IDSSP) is the brainchild of Nick Fisher, a statistician from Australia. (Fisher worked for CSIRO for 30 years before founding his own company, ValueMetrics Australia, in 2001.) In January 2018 Fisher set up a curriculum team of statisticians and computer scientists, drawn from Australia, Canada, New Zealand, United States and United Kingdom. The present author joined the team from the UK, bringing experience in school teaching as well as an academic background in both statistics and computer science. The curriculum team is supported by a larger advisory group, drawn mainly from the same five English-speaking countries, but with members in the Netherlands and Germany too.

Working largely by email, but with occasional international meetings and regular consultations with the advisory group, the curriculum team drew up two frameworks, one for students and one for teachers. These frameworks present both a conceptual overview of an approach to teaching data science and a highly detailed description of possible content. The age range of students is envisaged as the last two years of secondary school, though a broader view would suggest that the materials are suitable for early tertiary education and for training in employment.

The frameworks document was published in September 2019. Publication of the frameworks marked the end of Phase 1 of the Project.

Phase 2 of the project is implementation. At the time of writing, implementation may yet take any of several different paths. A minimal way forward would be for educators to take the lead in progressing the frameworks. The frameworks are publicly available for curriculum developers to use and modify, wholly or in part as they wish, with no further input from the IDSSP team. The most ambitious programme would involve an IDSSP team developing extensive resources to support courses and qualifications in a variety of formats and suitable for different modes of delivery. This would include developing a MOOC for self-learning or a flipped classroom approach to teaching, and there would be an imperative to train teachers as well as teach students. Clearly this most ambitious version of Phase 2 would be very expensive. At the time of writing, possible funders are being approached, and there are good reasons to be optimistic.

BACKGROUND

In recent years, there has been an increasing recognition of the importance of data in commerce, across society as a whole and in education. In 2016, the Royal Statistical Society's Data manifesto put the importance of data like this:

> What steam was to the 19th century, and oil has been to the 20th, data is to the 21st. It's the driver of prosperity, the revolutionary resource that is transforming the nature of economic activity, the capability that differentiates successful from unsuccessful societies.

Some researchers (e.g. Donoho 2015) have traced the origins of data science back 50 years or more, but in the last 10 to 15 years its recognition as a discipline has grown. In schools, certainly in

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),
July 2020, Online. Voorborg, The Netherlands. ©2020 ISI/IASE iase-web.org/Conference_Proceedings.php

the UK, data science as such does not yet feature widely in the curriculum. However, there is increasing acceptance of the importance of using real data, real technology, real contexts and large data sets across the curriculum in the UK.

- Porkess, A world full of data (2013), identified data-rich statistical content across a range of subjects as diverse as biology, business studies and history.
- The Royal Statistical Society and ACME report, Embedding Statistics at A level (2015), covered similar ground, looking specifically at newly reformed qualifications.
- The A Level Content Advisory Board (2014) recommended to the UK government that statistics should, for the first time, become a compulsory part of A level mathematics, and said that 'The use of real large data sets should permeate the teaching, learning and assessment of statistics' with 'the use of technology in both teaching and assessment'. These recommendations were incorporated into the revised specifications for mathematics introduced for first teaching in 2017, though it is not yet clear that teaching and assessment have taken them fully to heart.
- Pittard, The integration of data science in the primary and secondary curriculum (2018), found a mixed picture in terms of how different subjects have incorporated data and technology.

OVERVIEW OF FRAMEWORKS

The IDSSP Curriculum Frameworks document (2019) identifies two imperatives arising from the growth of data in scale and complexity:

> … to ensure that there is an adequate supply of people entering the workforce who are equipped to handle the new challenges of learning from data.
> … an equally pressing need for people in our societies to be more capable of understanding, interpreting, critiquing and making decisions based on data as they cope with the vagaries of life.

These imperatives have clear implications for school level education, and they lead to the project's two objectives:

> To ensure that school students acquire a sufficient understanding and appreciation of how data can be acquired and used to make decisions so that they can make informed judgments in their daily lives, as students and then as adults.
>
> To instill in more scientifically able school students sufficient interest and enthusiasm for Data Science that they will seek to pursue tertiary studies in Data Science with a view of making a career in the area.

The frameworks document recognizes that the wide variety of educational jurisdictions makes it impossible to devise a single course in data science that would satisfy all requirements. This is true even at the most basic level of how much time would be available within different curriculum structures for a course in data science. It is all the more evident when considerations such as different practices in delivery and assessment are taken into account. The frameworks are therefore intended to be used flexibly: a 'mix and match' approach to using the frameworks is encouraged.

Another aspect of the intended flexibility relates to the word 'Schools' in the project's title. The frameworks have clear potential uses in beginners' courses in tertiary education. In addition, they have already generated interest in the UK as possible benchmarks for the training of employees in graduate recruitment schemes at companies with a strong data focus.

Conceptual approach

At the heart of the project's interpretation of data science is the cycle of learning from data (fig. 1). The techniques and skills involved in the cycle are described as follows.

- *Problem elicitation and formulation*: questioning skills, listening skills, developing domain knowledge
- *Getting the data*: data harvesting, data wrangling, data management, experimental design, sampling
- *Exploring the data*: exploratory tools, data summaries, visualization, algorithms, coding
- *Analyzing the data*: developing and testing models, identifying aberrant phenomena, prediction, quantifying uncertainty, coding, algorithms, visualisation
- *Communicating the results*: presentation skills, plain language skills, presentation graphics, client focus

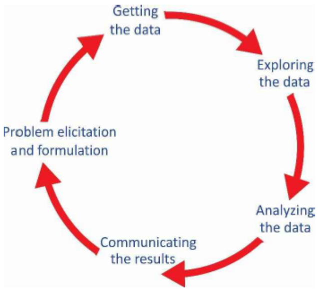
Getting the data
Exploring the data
Analyzing the data
Communicating the results
Problem elicitation and formulation

Figure 1 The basic cycle of learning from data

Content

The IDSSP frameworks are very comprehensive, largely because it is assumed that few people reading the document with a view to implementation will already have a good understanding of the scope and range of data science. The project's aim, therefore, is to provide a 'fairly detailed map of the Data Science landscape'. That detailed map is summarized below.

The frameworks are divided into two units. Unit 1, which could be envisaged as the introductory material suitable for a one year course, comprises seven Topic Areas, with aims as follows.

- *Data Science and Me*: to introduce students to the big ideas of data science, the importance of data in their lives, social and ethical issues, the data science learning cycle.
- *Basic techniques for exploration and analysis*. This heading covers three topic areas: tools for a single feature/variable, pairs of features/variables, and three or more features/variables. These topic areas aim to introduce students to appropriate tools, simple graphical displays and numerical summaries, so that they are able to find interesting stories and make discoveries in the data as quickly as possible. These topic areas are fundamental to many subsequent topics.
- *Graphical displays and tables*: to develop further students' understanding of appropriate choices for graphical display in learning from data and in presenting conclusions; to show how tables are used to convey exact values, or to present summaries of data that are too complex to be conveyed graphically.
- *The data-handling pipeline*: to give an introduction to the tools used to deal with data and to develop an understanding of data management issues.
- *Avoiding being misled by data*: to develop an understanding of how to critique data and data-based claims, bias, confounding and random error; to introduce good practice such as random sampling and randomization; to motivate the incorporation of uncertainty in estimation.

Unit 2 comprises 10 topic areas. It is envisaged that curriculum designers would select from these according to local needs in order to construct a second year of a data science course. The material in Unit 2 is rather more specialist than that in Unit 1. In particular it includes more formal approaches to statistical inference than are found in Unit 1.

The first four topic areas in Unit 2 are as follows.

- *Time series data*
- *Map data*
- *Text data*
- *Image data*

In each case, the aim is to develop basic understanding of their respective data types and skill in displaying, exploring, interpreting and presenting results.

The other six topic areas are as follows.

- *Machine learning, supervised*: to develop understanding of classification and prediction systems, and to learn how to apply basic tools in practical settings.
- *Machine learning unsupervised*: to develop understanding of contexts in which it is of interest to identify groups in data, and to learn to apply basic tools in practical settings

- *Recommender systems*: to learn about situations in which recommender systems are used, the sorts of data that are collected to develop these systems, and some methods for building such systems.
- *Interactive visualization*: to learn how interactive visualization can enhance steps in the learning from data cycle, particularly exploring data and communicating results.
- *Inference using bootstrapping*: to introduce confidence intervals in a random sampling context using simulation methods (bootstrap resampling).
- *Inference using randomization tests*: to introduce concepts of significance testing through randomized experiments and simulation methods.

CHALLENGES FOR DELIVERY

A substantial course in data science, of the type envisaged by the IDSSP frameworks, creates many big challenges for its delivery. The response to challenges of this magnitude can often be to say that delivery is impossible. A more positive approach is to take the challenges on and overcome them. This approach is essential if a data science course is to be successful at school level.

Resources: teacher capability

There is very little doubt that teachers able to deliver a data science course of the type envisaged are in short supply. Hence there is an IDSSP framework for teachers, designed to give a broader and deeper understanding than is appropriate for students. If Phase 2 of the project can raise sufficient funding, then the teaching materials produced will cater for the training and accreditation of the teacher workforce to empower them to deliver data science courses with confidence and enthusiasm.

Resources: technology

A data science course cannot be delivered in anything but an artificial form unless suitable technology – broadly speaking, a computer and appropriate software – are available to students throughout their learning. Certainly, in the UK, it is uncommon for students to work with computers as a tool always on the desk every lesson. A far more common model is for students to have an occasional lesson in a dedicated computer facility. But such an occasional model would simply not work in a data science course. There are very few topics which would *not* need continuous hands-on access to technology. Perhaps the introductory topic ‘Data Science and Me’ is the only candidate for technology-free teaching.

An important consideration in the design of the frameworks was the software to be used and the amount of coding required. As the framework document puts it, ‘… a computer with reasonable visualization capability … is as essential to learning how to work with data … as is access to modern laboratories for studying biology, chemistry or physics’. The frameworks assume no prior knowledge of any software packages, but it is expected that ‘students will acquire competence with at least one (freely available) software language or package (probably Python or R)’. In Unit 1, students would mostly use ‘point and click’ systems or modify code that has been provided. There might be more formal coding in Unit 2. It is worth remarking that there is intended to be minimal overlap between a data science course based on these frameworks and a computer science course of the type studied at school. Consequently, there would be no obstacle to students studying both.

Assessment

Continuous hands-on access to technology in the teaching and learning raises important questions about assessment. In the UK, most subjects, including mathematics and statistics, are assessed by paper-based examinations taken at the end of the course. This assessment model is doubly inappropriate for data science.

Firstly, it is hard to justify having a traditional paper-based assessment for a course which has involved learning with technology always to hand. Assessing *without* technology can only be artificial as it deliberately avoids engagement with what and how students have learned.

Secondly, it is also hard to justify testing skills acquired across a two-year course only at the end of that course. So consider, for example, a topic such as map data. Students will have spent perhaps two or three weeks learning to use specialist software to analyze and interpret map data, before then moving on to another topic. The appropriate assessment is surely to measure the students’ capabilities at the end of this two- or three-week period, not to wait until many months later and measure what they have remembered.

EXPRESSIONS OF INTEREST SINCE THE FRAMEWORKS WERE PUBLISHED

In the time since their publication, the IDSSP frameworks have attracted considerable attention and support – not always from predictable directions.

Endorsements

The frameworks have been endorsed by:

- Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers
- American Statistical Association
- BCS The Chartered Institute for IT
- International Statistical Institute
- New Zealand Statistical Association
- Royal Statistical Society
- Statistical Society of Australia
- Statistical Society of Canada
- Teaching Statistics Trust

Publications

The Teaching Statistics Trust publishes the journal *Teaching Statistics*, edited by Helen MacGillivray. At the end of 2020 the journal will produce a special e-book edition called 'Teaching Data Science and Statistics'. At the same time, the journal will change its title to *Teaching Statistics and Data Science*. This change of name is intended to mark the continuity between the two disciplines, but also to mark a change in emphasis towards data science.

Curriculum development bodies

MEI, a curriculum development body in the UK, has developed a pilot programme in data science, informed by the IDSSP. The materials have been released for schools to use during Covid-19 closure. Details are available here: https://mei.org.uk/data-science

Examinations bodies

At the time of writing, two international examinations bodies, motivated by the IDSSP work, are actively considering the possibility of having a qualification in data science. Again, details are not yet in the public domain, and these initiatives are commercially sensitive.

Parallel developments

The Royal Society published its report 'The dynamics of data science skills' in 2019. Following that there have been several UK-focused meetings on data science and its relationship with statistics.

Employers

Several large UK employers operating in data-rich environments, all with an international reach, have expressed interest in the frameworks as a basis for training their graduate recruits in data science. At least initially, the frameworks would be used as a guide to what trainees need and for benchmarking their progress. If the full Phase 2 of the IDSSP takes place, then these employers would be keen to use the resources developed.

BROADER IMPLICATIONS

It is clear that interest in data science is growing. The importance of the subject is increasingly being asserted, and there is a recognition that data education should be one of the foundations of learning alongside the "three Rs" – Reading, wRiting and aRithmetic. This is not just a matter for school students. In higher education, in employment, and most importantly in everyday life an understanding of data is essential in order to play a full part.

The IDSSP sets out a framework for a *specialist* course in data science, one that can only realistically ever be for a minority. The IDSSP vision may be necessary in order to produce the experts in data science that we need, but it is not sufficient. There is a need for a coherent programme to teach the non-experts how to understand and learn from data too. And this broader conception of data education would benefit from an agreed nomenclature for the different levels of expertise. One possible hierarchy is as follows.

- *Data literacy*: the fundamental ability to understand and interpret data that we would ideally want all members of society to have in order to play their full part as citizens.
- *Data skills*: the particular techniques of analysis, interpretation and communication that are required in data-rich subjects. The particular skills will vary from one subject to another, just as the types of data do. (Think, for example of map data in geography versus text data in literature.)
- *Data science*: the deep understanding of a broad range of data types and the appropriate techniques for problem formulation and solving, data acquisition and management, exploring and analyzing data, and communicating the results.

DATA SCIENCE AND STATISTICS

Broad acceptance of these frameworks for data science has the potential to feed back into the way in which statistics is taught. Specifically, a data science approach is conceptually richer than statistics courses typically are, it is more practically focused, and it harnesses technology in a way that school-level statistics rarely achieves.

One description of data science is statistics for the technological age. Arguably, statistics courses, certainly of the kind found in schools and colleges in the UK, would benefit enormously from a stronger focus on real data and real technology, with teaching, learning and assessment all being more practically based.

CONCLUSION

The IDSSP frameworks have their origins in the growing recognition of the importance of data science. The frameworks attempt to identify the core content in data science, as well as suggesting conceptual approaches to pedagogy. The challenge now is in taking this work further – either by educators and trainers across the world taking the IDSSP frameworks and using them as they see fit, or by the IDSSP team raising the funds necessary in order to develop a rich set of resources for free worldwide use.

More broadly, these developments in data science should feed back into more ambitious and effective approaches to data literacy, data skills, and statistical education generally.

REFERENCES

Donoho, D. (2015). 50 Years of Data Science. *https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf*

IDSSP. (2019). Curriculum Frameworks for Introductory Data Science. *http://idssp.org/files/IDSSP_Frameworks_1.0.pdf*

Pittard, V. (2018). The Integration of Data Science in the Primary and Secondary Curriculum. *https://royalsociety.org/~/media/policy/Publications/2018/18-07-18-The%20Integration%20of%20Data%20Science%20in%20the%20primary%20and%20secondary%20curriculum.pdf?la=en-GB*

Porkess, R. A world full of data. *https://www.rss.org.uk/Images/PDF/influencing-change/A-world-full-of-data.pdf*

Royal Society (2019). Dynamics of data science skills. https://royalsociety.org/topics-policy/projects/dynamics-of-data-science/

Royal Statistical Society (2019), Data manifesto. *https://www.rss.org.uk/Images/PDF/influencing-change/2016/RSS_Data%20Manifesto_2016_Online.pdf*

Royal Statistical Society and ACME (2015). Embedding Statistics at A level. *https://www.rss.org.uk/Images/PDF/publications/embedding-statistics-at-a-level.pdf*