LOOKING BACK – LOOKING FORWARD; STATISTICS AND THE DATA SCIENCE TSUNAMI[1]

Jim Ridgway, James Nicholson, and Rosie Ridgway
School of Education, University of Durham, UK
Jim.ridgway@durham.ac.uk; j.r.nicholson@durham.ac.uk; r.a.ridgway@durham.ac.uk

The discipline of statistics arose from pressing needs to address a variety of social and scientific problems. The founders of the Royal Statistical Society in the UK, and the American Statistical Association were very diverse in their backgrounds and interests, but shared a common purpose – namely, to address difficult and interesting challenges. They also acted in similar ways, by working across disciplines, and inventing mathematics and models suited to new problems. Computer scientists have also addressed real-world problems, have pioneered interesting and exciting approaches to handling new sorts of data (e.g. from sensors and social media) and have developed new analytic tools (notably, tools based on machine learning); their work is having dramatic (and sometimes unexpected) impacts on society. Early encounters between statisticians and computer scientists often resembled 'turf wars' – with claims that statistics was fast becoming redundant, and that computer scientists' ignorance of core statistical concepts such as sample bias would prove fatal to their entire enterprise. The problems that beset the start of the twentieth century have not gone away; modern societies face a wide range of existential threats such as global warming and nuclear war. As before, collaboration across disciplines, and the creation of new modelling tools are needed to address these problems. Here we begin by drawing lessons from the development of computer science in its earliest days, focussing on Babbage's Analytical Engine. We then highlight key epistemological differences between traditional statistics and traditional computer science, such as the role of theory and the use of 'black-box' models. We argue the case for the development of the Epistemological Engine – a tool for analysing and improving the processes of knowledge creation and utilisation that will require the skills of both statisticians and data scientists. We conclude by identifying competences and dispositions relevant to students of statistics and data science, drawing on both contemporary developments and the earliest days of computing.

KEYWORDS: Modelling; Turf wars; Epistemology; Black-box; engineering

> Those who view mathematical science not merely as a vast body of abstract and immutable truths, whose intrinsic beauty, symmetry and logical completeness… entitle them to a prominent place in the interest of all profound and logical minds, but as possessing a yet deeper interest… this science constitutes the language alone through which we can adequately express the great facts of the natural world… will regard with especial interest all that can tend to facilitate the translation of its principles into explicit practical forms. Lovelace (1843, p2).

LESSONS FOR YOUNG MINDS FROM HISTORIES

Let us visit Victorian England for *some lessons for young minds* – these include lessons that go beyond sciences and technologies themselves. We derive our lessons from the lives of Charles Babbage (CB) and Ada Augusta King, Countess of Lovelace (AL) in the period 1833-1852. Some of the ideas here derive from Padua (2016); we are grateful to her for providing links to obscure but important references. CB is often credited as the designer of the first computer; AL as the first programmer. CB's Difference Engine was designed to create tables of numbers relevant to astronomy, navigation and mathematics, and was based on calculating successive terms in a given series using the method of differences which was built into a mechanical system of cogs and levers. His even more brilliant insight was the idea of a general purpose device with a store, a mill (CPU), a printer, and operation cards (for

[1] Citation: Ridgway, J., Nicholson, J., and Ridgway. R. (2019). Looking back, looking forward: statistics and the data science tsunami. (pages to come). Proceedings 62nd ISI World Statistics Congress, Kuala Lumpur, Malaysia. International Statistical Institute, The Hague, Netherlands. In Special Topic Session 515 Training Data Scientists at University for Rewarding Careers.

the program and numeric input) that would not need to be reset mechanically for each new table. Both the Difference Engine and the Analytical Engine were to be driven by steam. CB received huge amounts of state funding for the Difference Engine (more than the cost of building 2 battleships (a measure of research funding no longer used in the UK)). He realised that a successful Analytical Engine would do everything and more than the Difference Machine was capable of, and so devoted his energies to the Analytical Engine. A fully-functioning version of the Difference Engine was never built. The development of the Analytical Engine was never properly funded.

- *Technologies change – and contemporary technologies can present barriers to brilliant ideas*
- *Funding streams depend on delivering what you (more or less) promised*

If the state of technology was a limitation, what of the state of mathematics? Hot topics of the day included early explorations of non-Euclidean geometry (imagine that! – but why bother?), and imaginary numbers (again – surely impossible?). William Frend was a university mathematician and sometime tutor of AL who did not believe in negative numbers; he ridiculed the idea of zero. Along with some other mathematicians, he rejected the idea of using undefined symbols in algebra. Boolean algebra was first set out in Boole's (1854) *The Laws of Thought* – two years after the death of AL. By way of balance, CB held the most prestigious chair in mathematics in England; AL was a talented mathematician in her own right. These provide the conflicted context for AL's extraordinary insight that a machine could manipulate arbitrary symbols, and that these arbitrary symbols could refer to anything. "The distinctive characteristic of the Analytical Engine… the executive right-hand of abstract algebra... is in this that… [it]… weaves *algebraical patterns* like the Jacquard loom weaves flowers and leaves" (Lovelace, 1843, p3). "Supposing… the science of harmony and musical composition were susceptible of such expression and adaptations, the engine might compose elaborate and scientific pieces of music of any degree of complexity or extent." (Lovelace, 1843, p2).

- *Don't always believe what your tutors tell you cannot or should not be done*
- *Become an excellent mathematician*
- *Explore the art of the possible*

Women's scope for action in Victorian England was severely limited. Concerns were raised about the serious dangers to the mental health of both Mary Somerville (the mathematician after whom Oxford University's first college for women was named) – friend and correspondent on things mathematical with AL – and for AL herself, from studying mathematics to a high level. AL published just one paper. This paper began with a translation of Menabrea's (1842) paper on a talk by (his friend) Babbage in 1840 about the Analytical Engine. At the time, few women wrote original articles, but DID occasionally translate and summarise men's work. Babbage suggested she write notes on the article (and says she was responsible for "the algebraic working out of the different problems" (Babbage, 1864, para 136). Babbage himself published extensively on a wide variety of subjects. He discussed the Analytical Engine *ad nauseam* with friends and colleagues, but published little about it (despite huge volumes of notes and diaries).

- *Don't accept cultural norms that restrict your thoughts and actions*
- *Publish!*

What of the social dimension? CB (and therefore AL) was friendly with Michael Faraday, Charles Darwin, The Herschels, Mary Somerville, Augustus De Morgan, Florence Nightingale, Elizabeth Gaskell, Tennison, and the Duke of Wellington (and sundry other politicians).

- *Cultivate smart people from a wide range of disciplines*
- *Travel and learn*
- *Talk about your ideas*

So much for computer science. What of the history of statistics? The logo of the Royal Statistical Society (RSS) is a wheatsheaf, carried over from its predecessor the Statistical Society of London (SSL). The SSL originally adopted the motto *aliis exterendum* – to be threshed out by others. This conveys a clear world view from the world's first professional association of statisticians – that the primary function of statistics is to gather and organise resources that others will transform into something useful. Pullinger (2013) paints a very different picture. The motto was dropped after a year. He points to the diversity of the founders of the RSS (which included CB – mathematician, mechanical engineer, astronomer, and philosopher) and to their commitment to study practical problems and to find (and implement) solutions with direct social benefit – inventing new mathematics when needed. This

tension between gatherers and analysts, and between theoreticians and practitioners, articulated by Lovelace in the introductory paragraph, mirrored in both mathematics and statistics, is alive and well.

It is captured in some critiques of statistics curricula. Cobb (2015) and Ridgway (2015) argue that introductory courses over-value tractable statistical models, resist algorithmic thinking, and devote far too little time to realistic problems. This critique begs two questions: 'whose realistic problems?'; 'what models are missing'? In the early days of the RSS, the answer to the question about 'whose problems' might well have been 'everyone's' – illustrated via pioneering work in meteorology, health, genetics, agriculture and economics, and often associated with the invention of new mathematics. The extent to which this tradition of conducting pioneering work with practical applications, and inventing appropriate supporting mathematical structures, has continued can be judged by inspecting the list of past RSS presidents (see RSS, 2019).

The question of 'missing models' raises bigger issues. All models are simplifications of some reality, and the choice and applicability of any model depends on the phenomenon to be modelled, and the purpose to which the model will be put. "All models are wrong, but some are useful" (Box and Draper, 1987, p424). A problem with introductory statistics courses has been an over-emphasis on standard models (e.g. using the Normal distribution) developed to solve problems in a pre-computer age, and a focus on generalising from samples to populations. This is appropriate where data is expensive to collect, where small samples can represent populations (often the case in agriculture and medical trials - but not in situations where disaggregated data show different patterns), and where phenomena are stable over time (again, agriculture and some medical trials, but not social phenomena over time), and where there is little computational power. Even in favourable circumstances, models can be applied badly – see Ioannidis (2005) on *why most published research findings are false* and the Open Science Collaboration (2015) on failures to replicate 'well-known' results in psychology. These failures constitute a serious threat to the business of creating new and useful knowledge, and advancing progress in a number of academic disciplines. The failures themselves can be traced to poor practices of data collection, analysis and interpretation, which can be recognised, and remedied.

CONFLICTING EPISTEMOLOGIES

Breiman (2001) described two approaches to analysing data. He argued that most statisticians typically apply transparent models where a small collection of well-defined inputs are used to predict outputs - so models are used primarily to explain and also to predict (he argues that this leads to irrelevant theory and questionable conclusions). In contrast, a small proportion use algorithmic modelling; techniques such as neural nets and random forests are used to map inputs and outputs. The focus is primarily on prediction with little attempt to explain. This can be viewed as a 'data science' stance.

Ridgway *et al* (2018) map out some challenges for algorithmic models – notably that what you get out is determined by what you put in. So algorithmic models are strong on 'what is' but weak on 'what ought to be' and can have undesirable consequences when used for (for example) job selection or predictive policing. Perez (2019) provides further examples. These problems are exacerbated when the data set itself does not represent the population as a whole – for example drawing conclusions from (conventional) medical research that is based almost exclusively on Caucasians. This is a particularly problematic challenge for data science, where decisions about analysis are often based on pragmatism; a variety of models are applied to a data set, and the final choice of model is based on fit and the ability of the model to predict future events.

Statistics has been characterised by engagement with real-world problems; what of data science? Consider these examples of computer uses, software and devices:

- Google, Amazon, Facebook, Skype;
- recognition of individuals via face, fingerprint, voice, gait, patterns of key presses;
- tracking (via fitness trackers, credit card use, data from transport networks);
- speech recognition and language translation;
- medical diagnosis;
- detection of disease outbreaks via analysis of google search data;
- the Internet of Things – smart refrigerators, TVs, cars, and domestic robots;
- 'deep fake' videos;

- predicting crime and recommending custodial sentences;
- satnav; autonomous vehicles and weapons systems;
- mapping dwellings from aerial images, in remote settings;
- emotion detectors for classrooms and cars.

A striking feature of data science has been the variety of problems addressed, the kinds of data analysed and used, the range of novel models developed, and its direct effects (intended and unintended) on people's lives. Most of these developments can be described as 'engineering' – a useful product emerges from an analysis of an interesting challenge. The relationship between statistics and data science is analogous to the relationship between mathematics and engineering. Engineers don't do 'applied mathematics' they do 'engineering', and used mathematics where appropriate. Similarly, data scientists don't do 'applied statistics' they build things, and use statistics when they (think they) need to.

It is worth reflecting on the extent to which analytic models, *p*-values and effect sizes have contributed to the developments in computer science that have radically reshaped the modern world. For the practical examples listed above, the designers' ambitions are for 100% success, not for theoretical nicety, nor for performance that is 'significantly better than chance'.

DESIGNING THE EPISTEMOLOGICAL ENGINE

We are living in interesting times; new phenomena are emerging (associated with billions of people having internet access, much greater wealth and better health, worldwide). New sorts of data are available; there are new sorts of analytic tools; there are new creators of knowledge (notably technology companies) and new distributors, consumers and users of knowledge. The problems that beset the start of the twentieth century have not gone away; modern societies now also face existential threats such as global warming and nuclear war. There is a need for knowledge-generators to engage with problems that can be characterised as 'messy', 'complex', or 'wicked'. These problems are characterised as being ill-defined in terms of specifying relevant variables or measuring progress; they often involve interacting systems at different levels. For example, climate change is influenced by the actions of individuals (e.g. car choice and use), local structures (e.g. support for recycling), national structures (e.g. policies on house insulation and domestic solar power), and international initiatives (e.g. consensus on restricting carbon emissions). There is no 'right' level to work at; there are multiple ways to measure system states and the results of different initiatives.

Addressing 'wicked problems' is likely to involve working with multiple sources of messy data, and using a variety of analytic tools (see Ridgway *et al* 2018). Inter-disciplinary action is almost certain to be essential to success. However, scientists working with even relatively simple problems can make a mess of things. There are serious challenges to current methods of knowledge acquisition, illustrated by the very poor quality of much of the research funded at great expense in universities (see Ioannidis (2005); Open Science Collaboration (2015)). These cannot be explained away as the result of poor practice by a few individuals; they reflect systemic failure by some academic communities. There is an urgent need to analyse and improve the whole system associated with the creation and use of knowledge – in short, designing an Epistemological Engine (EE) has become a priority. The prime candidates for creating and building the EE are statisticians and data scientists.

Early encounters between statisticians and data scientists were often acrimonious; 'statistics' would be a casualty in 'the death of theory', and data scientists' ignorance of core statistical concepts such as sample bias and overfitting would prove fatal to their entire enterprise. The EE should be founded on techniques and skills used in both data science and statistics. Data scientists create open data repositories (e.g. https://registry.opendata.aws/), and have adopted a culture of sharing code – especially Workflows (e.g. https://github.com/) to facilitate a comparison of different analytical techniques and modelling assumptions. They use Common Task Frameworks wherein success is judged on terms of actual performance in analysis, not theoretical niceties. Statisticians bring sophistication about data acquisition (including synthesising and triangulating data sources), preparation, and exploration. They can contribute to analyses, data representation and communication, and can comment on issues such as the likely generalisability of findings. They bring considerable sophistication about modelling. Identifying the style of modelling being used by different researchers (explicitly or implicitly) should be automated in the EE.. Ridgway (1998) classifies styles of modelling, and describes

analytic models (such as those found in school physics), systems models (such as those found in school biology) and macrosystemic models – these are systems models where the system itself undergoes change. Macrosystemic models can be divided into two groups – models where the changes in the system are relatively predictable (e.g. ecological restoration; the life cycle of the butterfly) or unpredictable (Brexit; climate change and global political stability in the Trump era).

The EE should comprise a large tool collection. Sample tools include:

- Critical evaluation of specific studies, using criteria for evaluation such as those identified by Ioannidis (2005) and the Open Science Collaboration (2015), e.g. identifying weak effects using small samples, and testing multiple hypotheses until a 'significant' result is found;
- Identification of academic areas where there is insufficient sharing of data, code and workflows;
- Identification of academic areas that are paradigm-bound (i.e. characterised by analyses of rather few classes of data, and by the use of a small set of analytic tools);
- Tools for automated testing of code and workflows;
- Identification of results that are important for some theoretical claims, where the evidence base is weak (e.g. where there has been little replication across relevant populations);
- Automation of literature searches, and the conduct of meta-analyses;
- Creating semantic nets of academic papers in terms of both content and authorship in order to document the flow of discovery processes;
- Methodology classification systems, that support automated classification;
- Analogy generators, to suggest developments in fields other that the one in which a method or tool was developed;
- Methods for analysing large corpora of research in different fields to examine the epistemological assumptions made (including pragmatism).

Knowledge gaps

There are some glaring gaps in our knowledge that need to be remedied, we need: more formal theories of data analysis; more work on the cognitive psychology of data visualisation and interpretation; and more and better modelling of emotion, social behaviour, and cognition; better understanding of the processes of knowledge generation, distribution and use, and more tools for working with very large data sets.

COMPETENCES FOR STUDENTS OF DATA WRANGLING

So what do students need to know in order to work in this brave new world? Here, we offer some more lessons for young minds.

- *Be aware of the politics of technology: technologies are never neutral (e.g. cars cannot be driven by the very young or old, or the poor)*
- *Attend to unintended consequences (e.g. cyberbullying via social media) via 'what if' games*
- *Engage with moral issues (e.g. the dangers of the Panopticon)*
- *Be aware of epistemological issues: the nature of knowledge as conceived in different academic disciplines - how it is created, shared, learned, and used (and by whom, and for what purposes)*
- *Understand modelling and the limits of modelling, and the principles of model validation;*
- *Explore the reasons for the existence of data sets – adopt a hermeneutical approach*
- *Create a conceptual web to link between seemingly different methods*
- *Understand the principles underpinning different techniques (e.g. neural nets)*
- *Learn to represent the same problem in a variety of ways*
- *Become fluent in the use of major data repositories*
- *Share your code and workflows*
- *Invent and modify data visualisations (including dashboards)*

CONCLUDING REMARKS

The early history of data science holds important lessons; technology, mathematics and society are in a continuous state of rapid change. Students should be made aware that current knowledge will be superseded, and that there are social forces that can limit their creativity.

Technologies have created existential threats to humanity such as global warming and nuclear war. There is a pressing urgency to address such problems. Both statistics and data science have their roots in solving challenging problems, but have traditionally adopted somewhat different approaches. Statistics is characterised by sophisticated modelling using a small set of well-defined variables; data science is often a-theoretical. Data science adopts practices that should be applied across a wide range of disciplines, such as sharing data, code and workflows. Statistics is strong on discovery methods.

There is an urgent need to create an Epistemological Engine – a set of semi-automated tools to understand and support effective science. Statisticians and data scientists are the people best placed to create and maintain this Engine. We offer some ideas on the tool set that will comprise the EE, and some suggestions about the competences needed by future data wranglers.

And a final piece of advice for young minds: *make a wall poster of these words from Ada Augusta King, Countess of Lovelace…*

> "A new, a vast, and a powerful language is developed for the future of analysis… the theoretical and the practical in the mathematical world, are brought into more intimate and effective connexion with each other." (Lovelace, 1843, p3)

REFERENCES

Babbage (1864). *Passages from the Life of a Philosopher.* (https://en.wikisource.org/wiki/Passages_from_the_Life_of_a_Philosopher/Chapter_VIII

Boole, G. (1854). *The Laws of Thought. An Investigation of The Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*, Originally published by Macmillan, London. Reprint by Dover, 1958. Cited at https://plato.stanford.edu/entries/boole/#LawsThou1854

Box, G., and Draper, N. (1987). *Empirical Model-building and Response Surfaces.* New York: Wiley.

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.*, **16**(3), 199–231.

Cobb, G. W. (2015). Mere renovation is too little too late: we need to rethink our undergraduate curriculum from the ground up. *The American Statistician, 69*(4), 266–282.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine* **2**(8): e124. doi:10.1371/journal.pmed.0020124

Lovelace, A. (1843). Notes on a translation of Sketch of the Analytical Engine invented by Charles Babbage by L.F. Menabrea (1842).
https://en.wikisource.org/wiki/Scientific_Memoirs/3/Sketch_of_the_Analytical_Engine_invented_by_Charles_Babbage,_Esq./Notes_by_the_Translator. Downloaded 5 April 2019.

Menabrea, L. (1842). On the mathematical principles of the Analytical Engine. *Bibliothèque Universelle de Génève*, No. 82. October 1842. Translated by A. Lovelace. https://en.wikisource.org/wiki/Scientific_Memoirs/3/Sketch_of_the_Analytical_Engine_invented_by_Charles_Babbage,_Esq. Downloaded 5 April 2019.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science 348(6251)*. DOI: 10.1126/science.aac4716.

Padua, S. (2016). *The Thrilling Adventures of Lovelace and Babbage*. Padstow: Penguin

Perez, C. (2019). *Invisible Women: exposing data bias in a world designed for men.* London:Penguin.

Pullinger, J. (2013). Statistics making an impact. *J. R. Statistic. Soc. A, 176*(4), 819 – 836.

Ridgway, J. (1998). *The Modelling of Systems and Macro-Systemic Change - Lessons for Evaluation from Epidemiology and Ecology.* National Institute for Science Education Monograph 8. University of Wisconsin-Madison.
Retrieved from http://archive.wceruw.org/nise/Publications/Research_Monographs/Vol8.pdf.

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3). Retrieved from onlinelibrary.wiley.com/doi/10.1111/insr.12110/full.

Ridgway, J., Ridgway, R. and Nicholson, J. (2018) Data science for all: A stroll in the foothills. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward*. Proceedings of the Tenth

International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan. Voorburg, The Netherlands: International Statistical Institute.
http://icots.info/10/proceedings/pdfs/ICOTS10_3A1.pdf?1531364253
RSS (2019). Past Presidents
https://www.rss.org.uk/RSS/About/About_the_RSS/President_and_Vice_Presidents/Past_presidents/RSS/About_the_RSS/About_sub/President_and_vice_presidents/Past_presidents.aspx?hkey=eec573c6-b10e-4027-8846-9802a07a1d28 Downloaded 8 April 2019.