DATA SCIENCE FOR ALL: A STROLL IN THE FOOTHILLS

Jim Ridgway, Rosie Ridgway, and James Nicholson
School of Education, University of Durham, DH1 1TA, UK
Jim.ridgway@durham.ac.uk

Data science presents both opportunities and threats to conventional statistics courses. Opportunities include being at the bleeding edge of data analysis, and learning new ways to model phenomena; threats include the challenge of learning new skills and reviewing fundamental assumptions about explanation, prediction and modeling. Powerful data visualisations makes it easier to introduce students to fundamental statistical ideas associated with multivariate data. Data science provides methods to tackle problems that are intractable using analytic methods. Students need to learn how to model complex problems, and to understand the problematic nature of modeling – and they need to consider the practical and ethical implications of their (and others') work. Here, we offer a stroll into the foothills, along with aphorisms and heuristics for data analysts.

AN HISTORICAL INTRODUCTION

The history of statistics tells us important things about ways in which data science might evolve, which in turn helps us think about how best to help young people acquire skills relevant to their daily lives and to their future careers. Porter (2017) outlined the early history of the Statistical Society of London (the forerunner of the Statistical Society which became the Royal Statistical Society in 1887), which chose as its motto *aliis exterendum* – to be threshed out by others. He reported that initially, expressions of opinion were prohibited. One assumes that the initial focus was on the development of mathematical models, rather than direct application. The motto did not last long, and was dropped within a year. Pullinger (2013) offers a stark contrast in his brief history of the Royal Statistical Society. The founders had very diverse backgrounds; they included: a leading politician, an historian and political activist, and Charles Babbage – mathematician, astronomer, engineer, and inventor of the computer. The underpinning rationale was to study problems of real social relevance, create solutions, and to act politically to implement these solutions. So statistics was invented to address real problems, not simply to create tools that others can use to solve their problems. Early work in statistics led to the creation of a beautiful and powerful theory for modelling data, based around the Normal distribution. If data matched the strong assumptions made by the model, handling large data sets became mathematically tractable, and a raft of other benefits followed (e.g. being able to estimate population parameters to known degrees of accuracy, from small samples). Sadly, the model has now taken on a life of its own, and dominates the statistics curriculum; 'threshing' receives far too little attention.

CURRICULUM CRITIQUES AND OPPORTUNITIES

There have been a number of criticisms of current statistics curricula (e.g. Cobb, 2015; Ridgway 2015). Critics have argued that the current curriculum, too often:

- is stuck in the (early) 1900s
- over-values tractable mathematical models
- uses a simple-to-complex pedagogy
- focuses on generalizing from small samples to populations
- resists algorithmic thinking
- and fails to reflect the data revolution.

Current curricula often ignore data-heavy resources and current methods of analysis that affect the daily lives of almost everyone, such as those associated with social media and its many uses, and pattern recognition (and the use of automated decision making in many aspects of life).

There are opportunities for, and challenges to, invigorating teaching New methods of analysis are available; there are important contexts that relate to social upheaval which require urgent attention and action, such as global warming, migration, and poverty; there are new sources

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. iase-web.org

of all sorts of evidence such as open data (e.g. numbers in the European Union Open Data Portal, text in the New York Times archive, music in the Million Song Data set, images in Google's YouTube-8M); big data (including data from social media such as Topsy, and Likebutton); new players, notably data scientists, journalists who engage with data, fact checkers, and generators of fake news (e.g. websites that imitate reputable sources such as the Guardian newspaper), and new audiences, notably informed and uninformed citizens, policy makers and politicians.

EVIDENCE INFORMED DECISION MAKING

A number of authors (e.g. Wild and Pfannkuch, 1999) map some of the stages involved in evidence informed decision making. It is common to identify phases such as: problem exploration; data collection and management; analysis; decision making and action. So what should students learn? *Problem exploration* is perhaps the most important phase in problem solving. This can be facilitated by data visualisations where students can see salient features of the phenomena in question. *Data collection and management* now involves consideration of a much wider range of sources – e.g. from social media (including sounds and images) and from sensors (personal, internet of things, ecological etc.). Large-scale open data from official sources provide easy entry points to discussions on key statistical ideas such as: measurement issues (what is *poverty*?); concepts of central tendency (how is the very high 'average' household wealth in the USA compared to other countries compatible with relatively very low median household wealth?); use of indices (how could one possibly measure *sustainable development*?). It is possible to introduce multivariate thinking early in the curriculum, using authentic data. Ideas such as effect size, non-linearity and interaction can be grasped readily in the context of authentic data, such as educational attainment as a function of ethnicity, sex and poverty, or the incidence of sexually transmitted disease as function of age, sex, disease and time. Similarly, ideas of long-term change embedded within seasonal change can be grasped by appropriate dynamic visualisations. Data visualizations continue to be created, and there is an implicit assumption that naïve users can interpret novel visualisations with little effort. Sutherland and Ridgway (2017) argue against this, and assert that the ability to interpret and deconstruct novel visualizations is an essential component of statistical literacy, and should be taught. There is a good deal of guidance to hand (e.g. The Financial Times' Chart Doctor) and potentially engaging classroom activities can be developed – for instance by simulating *Tableau* competitions, where participants have to create an interactive display of a complex data set, under timed conditions, and are then offered a critique of their product. *Analysis* and *Decision making and action* are more problematic in the context of data science and require some reflections on the nature of prediction, explanation, and modeling.

PREDICTION, EXPLANATION AND MODELING

It is possible to predict events accurately without understanding the underlying phenomena. For example, the phases of the moon are perfectly predictable, even if one believes in an earth-centric universe. It is also possible to have a complete explanation of a phenomenon, and yet be unable to predict events – simulations in chaos theory provides examples where very simple algorithms give rise to chaotic behavior.

Styles of Prediction

In the world of finance, one can identify two distinct approaches to predicting stock exchange movements, namely 'chartists' and 'modelers'. Chartists focus simply on applying mathematical techniques to data on past performance in order to predict future performance. Modelers set out to understand phenomena and to make predictions appropriately. For example, if one believes (on the basis of some evidence) that the developed world is moving towards an 'hourglass' distribution of wealth – a large group of poor people, a group of the super-rich, and a squeezed middle class - then one would invest in companies that tailor products to the emerging sectors – 'sub-prime' loan companies and 'dollar shops', and vendors of luxury goods, respectively, and would dis-invest in companies that serve middle-class clients.

Breiman (2001) argues that there are two cultures of data analysis. He asserts that the majority of statisticians use relatively simple models where outputs depend on a small collection of inputs that can be described in words; models have both predictive and explanatory properties. A

minority of statisticians adopt algorithmic modelling, where AI techniques such as neural nets and boosted decision trees are used to map inputs and outputs; here, there is little attempt to establish functional relationships between inputs and outputs. Natural language processing (e.g. Google translate) by such methods, after decades of failure using analytic methods, illustrates this approach. Anderson (2008) famously argued for the exclusive use of algorithmic modeling as follows: 'Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves'. Both styles of modeling have enjoyed successes; next, we address the problems that can arise with both classes of models, as a prelude to offering aphorisms and heuristics for modelling.

All models are wrong, but some are useful (Box & Draper, 1987, p424)

Challenges for analytic models

Every statistician knows that causality cannot be deduced from correlation. However, it is not difficult to find implied causality when statisticians use multiple regression and state 'variable *x* accounts for n% of the variance'. For example, in a very carefully written document from an authoritative source, we read: "In Belgium…students' socio-economic status explains… about 19% of the variation in student performance" (OECD (2016), p212). Socio-economic status does not explain anything – it is simply correlated with educational attainment – as other parts of the same document are at pains to explain. So the language used by statisticians when using simple analytic models can carry closet causality, when the model itself simply shows the predictability of one variable from others.

Predictions via simulation often allow users to explore 'what-if' scenarios by changing parameter values. Population projections that allow users to change assumptions about fertility rates, death rates, and migration, provide an example. These simulations can be useful pedagogical devices for introducing core concepts in modeling.

Models of systems are more complex, and can fail for a number of reasons. If one considers simple dynamic systems, models can fail because a key element has been omitted, or has been introduced after a model has been created – for example, the Brexit vote will disrupt the predictions of models of economic development in the UK and Europe; models of climate change that assumed all major industrialized countries would conform to the Paris agreement will need to be revised if the USA reneges on its promises.

Challenges for algorithmic models

Algorithmic models are particularly susceptible to the 'what you put in determines what you get out' problems to which every kind of model is vulnerable; recently examples include Google's automated image labelling systems classifying some African-Americans as gorillas, and Drive.ai's car navigation system not responding to the voice of Carol Reiley - its female designer. Using training data that includes family names, residence, and prior education is likely to lead to racist decisions, even if ethnicity is not included in the training data. Angwin *et al.* (2016) found that software used to assess the likely recidivism of criminals was twice as likely to identify black defendants (wrongly) as being a high risk. Predictive policing is similarly problematic. A useful thought experiment is to apply the methods (e.g. Risk Terrain Modelling) for exploring and perhaps remediating the causes of crime often associated with black communities, such as illegal use of drugs, to exploring and perhaps remediating the causes of crime associated with white communities such as fraud. It is likely that sites such as expensive tailors, top-end restaurants and champagne bars are frequented disproportionally by fraudsters, and should be visited frequently by the police, and perhaps be closed down, in order to reduce white-collar crime.

There are important lessons here for science in general, as well as for data science and policy. According to Murgia (2017), more than 80% of the genome data that underpins genetic medicine comes from Caucasians (and a further 14% from Asians). It follows that analyses of gene-disease links that lead to medical interventions may be ineffective (and perhaps dangerous) for non-Caucasian groups. The same article asserts '650,000 African-Americans may have

undiagnosed Type 2 diabetes because of a genetic quirk that fools a commonly used diabetes blood test', and gives further example of misdiagnosis of heart disease in African-Americans, and of epilepsy in Asian countries.

Challenges to the enterprise of modeling

Systems themselves evolve over time. Ridgway (1998) describes 'macrosystemic change' and gives the example of the evolution of planet Earth from gas ball to liquid to solid surface to plant world to 'modern' biosphere. The system itself evolves through stages where new sorts of things become possible – for example, plants creating oxygen as a by-product of photosynthesis which makes it possible for animal life to evolve. Over short time periods, one might consider the likely macrosystemic change associated with: automating much of the work done by humans; automated decision making; climate change; mass migration; genetic modification against ageing; and the steady increase in countries with nuclear weapons. These will require major revisions to models of any aspect of human life.

A further challenge to modelling occurs when the phenomenon being modelled is unstable; there may be inherent instabilities where variables that are effective predictors of events cease to predict these events at a later date – as in the case of Google 'Flu (see Lazer et al, 2014). A complication is that models can become the targets of people who want to 'game' the system. For example, if it is demonstrated via sentiment analysis that certain words predict stock exchange movements, it would be unsurprising if traders created bots to flood (say) *Twitter* with these keywords before buying or selling stock in volume. This is analogous to Goodhart's law that the more a quantitative indicator is used for decision making, the more it will be subject to corruption, and the more it will cease to measure what it once did.

Mixing messy models and telling stories

Most big problems (global warming, human development, health and disease, social inequality) are multi-layered and multi-faceted. A variety of qualitatively different data is available. Decision makers have different world views and reward systems. Analyses based on different data sets, using different methods can each illuminate issues, but no one analysis is likely to produce results that, on their own, can guide policy. Consider the challenge of international migration. Governments create policy about migration (such as 'return to country of origin'), often making use of official statistics. We can ask what knowledge one needs to understand the phenomenon. There are some obvious questions about who migrates (including demographics, skill levels, vulnerability), migration routes (tracking movement through different countries), patterns of migration (temporary, circular, permanent), the reasons for the migration, the impacts on the source and destination countries (e.g. loss of talent, and match with labor market needs, respectively), and regions (integration; impact on local labor markets) and the impacts on the migrants themselves (such as the contractual obligations they face, discrimination). Current systems do not provide these data (EU, 2017), and some current data is flawed; for example, emigrants often do not de-register from their country of origin. The questions themselves are increasing hard to clarify – when is an 'economic migrant' different from a 'climate change' migrant? There are technical issues of getting evidence about hard-to-reach groups such as migrants living in camps.

So what sorts of data might be relevant? There is a need for data from a variety of sources, over extended periods of time. National Statistics Offices (NSO) have relevant data from surveys, but these data are not adequately harmonized across states, are often out of date, and fail to provide evidence on key issues. Administrative data could play an important role because people interact with health services, education etc. and create a longitudinal data trail. Even if ethical issues related to privacy can be overcome, problems with harmonization will remain. Big data offer the third major source of evidence. Phone and web traffic analyzed by place and language, social media including google searches can provide information on changing population demographics and important trends in health. None of these ideas is new; the notion of 'collaborative economies' or 'information ecosystems' where companies and NSO share data (or private companies make all their data available to NSO) are being discussed actively. NSO do make use of social media (e.g. in Finland, to explore the effects of a radical social policy; in the Netherlands, to get better data on the activities and locations of companies). The Turkish government has used Facebook to identify

'dissidents', and set up large numbers of fake accounts to change sentiment. The Governance Lab at New York University offers interesting examples of the use of multiple sources of data to support government policy making and implementation.

Hermeneutics refers to the process of interpreting text. Text very rarely has a single unambiguous meaning; text cannot be understood without reference to the viewpoint and motivation of its writer/s, its history and cultural context, the language in which it was written and any subsequent translation, and the cultural and interpretive intentions of the reader. One can make the same argument about data. To understand data, one needs to understand the viewpoints and motivations underlying its collection, its history and cultural context, the recording and recoding processes it has been subjected to, and ones' own cultural background and interpretive intentions. Analyzing data-informed stories from NSO, users of administrative data, and big data users can highlight the need (and foster) skills in deconstructing 'data' in the ways that students in the humanities and social science learn to deconstruct text. Inventing and using their own measures of some phenomenon is likely to foster the idea that citizens need to be involved in the choices of measures used to report the success of governments.

AN HISTORICAL CONCLUSION

The early history of statistics showed a tension between developing models, and engaging with practical problems. Statistics education can be criticized because it teaches about models *per se*, not problem solving in contexts where data is relevant; data science can be criticized for brash over-application of 'black-box' methods to complex contexts. Both criticisms need extensive qualification, but have some heuristic use. Data science offers students the excitement of the early days of statistics, and the opportunity to contribute new knowledge using novel data sets and novel methods, complemented by curated data sets and well-developed analytic techniques.

Data is being used increasingly by private organizations and by government to inform decisions. The workings of models themselves might be opaque (e.g. neural networks), but so too are the uses to which they are put. Students need to be aware of the potential benefits and dangers to individual and societies of data science, and in particular, the problems associated with automating decision making (see O'Neil, 2016). Human systems are complex, and involve stakeholders who are likely to have a variety of motivations and reward systems. An important analysis tool is the pre-mortem or 'system-gaming walkthrough' – if new performance measures, or new indicators of success are introduced, how might the new systems of rewards be 'gamed' by different agents in the system? The issue is then the extent to which this is undesirable, or perhaps fatal to the proposed change.

We conclude with some aphorisms and heuristics. Just as statisticians have aphorisms (e.g. 'don't confuse correlation and causality') and heuristics for data analysis (e.g. 'disaggregate your data, and think about Simpson's paradox), it will be useful to add some aphorisms and heuristics directly related to data science. Here are some starting points:

Aphorisms:

- You can predict things without being able to explain them
- You can explain things without being able to predict them
- Models can be unstable for a variety of reasons
- Systems can be unstable for a variety of reasons
- You need to be careful when you play with sharp tools
- Only cowards *hide behind* algorithms.

Heuristics for analysis:

- Always start off by looking at 'what is' – in a hermeneutical sort of way
- Always model a situation using at least 2 different representations
- Always model a situation using multiple sources of information
- Explain your methods and assumptions
- Calibrate on data sets that are as different as possible
- Review, remodel, recalibrate; go back 3 steps
- Before you advocate a change, do the 'system-gaming walkthrough'.

CONCLUSION

Within an overarching context of modelling, students should experience, use, and critique a wide variety of *data visualisations*, notably when setting out to understand phenomena; they should be exposed to major *data sources* and methods to access these sources, and to new *data analysis techniques*, applied to a wide variety of type of data. The aim is to provoke both an hermeneutic approach to data, and a relativist world view – data can be interpreted in many ways, but some are better than others. It is important to consider the consequences of decisions – technologies (and models) are never neutral – and to align these to values coherent with the ethos of universities in liberal democracies.

REFERENCES

Anderson, C. (2008). *The end of theory: the data deluge makes the scientific method obsolete*. Retrieved from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Angwin, J., Larson, J., Mattu, S., & Kirshner, L. (2016). *Machine Bias.* Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Box, G., & 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). *Statistical Science*, *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.

EU (2017). *Power from Statistics: data, information and knowledge*. Outlook Report: Brussels.

Goodhart, C.A.E. (1975). "Monetary relationships: a view from Threadneedle Street". *Papers in Monetary Economics* (Reserve Bank of Australia) I. cited in *http://en.wikipedia.org/wiki/Goodhart%27s_law*.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*, 1203-1205.

Murgia. M. (2017). *The data flow that can be deadly*. Financial Times, 1 Oct, p10.

Norvig. P. (2012). Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning. *Significance, August,* 30-33.

OECD (2016). *PISA 2015 Results (Volume 1): Excellence and Equity in PISA.* OECD Publishing, Paris. Retrieved from http://dx.doi.org/10.1787/9789264266490-en

O'Neil, C. (2016). *Weapons of Math Destruction*. London: Penguin.

Porter, T. (2017). *The Pursuit of Objectivity in Science and Public Life*. Paper presented at the Eurostat *Power from Statistics: data, information and knowledge* conference, Brussels.

Pullinger, J. (2013). Statistics making an impact. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 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.

Sutherland, S., & Ridgway, J. (2017). Interactive Visualisations and Statistical Literacy. *Statistical Education Research Journal 16*(1), 26-30. Retrieved from https://iase-web.org/documents/SERJ/SERJ16(1)_Sutherland.pdf.

Wild, C.J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248.