Big questions, informative data, excellent science

Adrian W. Bowman

School of Mathematics and Statistics

The University of Glasgow, U.K.

adrian.bowman@glasgow.ac.uk

February 15, 2018

Abstract

The expression *big data* is often used in a manner which implies that immediate insight is readily available. Unfortunately, this raises unrealistic expectations. A model which encapsulates the powerful concepts of statistical thinking remains an invaluable component of good analysis.

Keywords: big data, statistical models.

1 Introduction

In the public discussion which takes place in any sector – moral, political, social or scientific – themes encapsulated by keywords often arise. Those keywords may not always be

particularly well defined, but they become part of the common discourse. In the scientific arena, *big data* is a good example. As an expression of the size and complexity of the datasets which now arise routinely, and the challenging issues of analysis and interpretation which have to be confronted, the term is quite a useful one. Unfortunately, the expression *big data* is often used in a manner which implies that solutions to these issues are readily available. This raises unrealistic expectations.

A variety of other terms have become associated with *big data*, including *machine learning*, *artificial intelligence*, *data analytics* and *data science*. Criticisms are sometimes levelled that statistical thinking has not adapted to the new challenges and that these new modes of thinking are required to make progress. While it has to be recognised that much more powerful and sophisticated computational methods are now required, closer inspection reveals that statistical methods remain as a fundamental component of all these areas.

As ever, data are a means to an end. The perspective of this short paper is that good analysis focuses on *big questions* whose solutions are important, that the most important aspect of the data is not its size but the extent to which it is *informative*, and that a critical component of *excellent science* is the ability to weigh evidence in an appropriate manner. Statistical thinking lies at the heart of all of this.

2 Big questions

It is clearly a gross over-simplification, but possibly a useful one, to view the questions which arise in science and technology as falling into two broad categories.

1. What happens next?

2. What is going on?

The first of these might be broadened to include *What is that?*. The aim is principally one of prediction or classification. There are many situations where this is a crucially important question to answer well.

The second question asks about the process which underlies what we see in the data. Many scientific questions are essentially about the relationships between different variables. Where possible, identifying cause and effect would be an ideal outcome but sometimes we have to settle for associations. These relationships may be hidden inside very noisy data and they may not always contribute much to accurate predictions, but they tell us something about the processes at work.

Many of the techniques which fall under the banner of machine learning and artificial intelligence are focussed on classification and prediction. Image processing is a particular case where methods such as *deep learning* are proving very successful. Although the concept of neural networks is an old one, the ability to deploy structures with rich layers and train these on huge databases in a computationally efficient manner has proved very effective. Schmidhuber (2015) provides a recent review. There are two aspects of this which are worth noting. One is that the procedures are clearly dependent on the nature of the databases providing the training data. Classifying an image whose characteristics are very different from those in the database may have unanticipated consequences. That requires us to think hard about the data we use. The other issue is that the complexity of the trained neural network is often sufficiently great that it is very difficult to understand how the prediction or classification is being performed. Where the main criterion is accuracy of performance that may not be crucially important. However, it does underline the distinction between the two questions highlighted above.

In contrast, the concept of a *model* plays a central role in statistics. Excellent introductions to the thinking involved are given by Davison (2008), Freedman (2009) and many other authors. Dunson (2018) provides a more extensive discussion and comparison of the

worlds of machine learning and statistics. Further comments on the perspective induced by a model, and its particular role in answering the *What is going on?* question, will be developed in Section 4.

3 Informative data

The most important thing about data is not its size but the extent to which it is informative. There are cases where we have measurements from the whole population of interest – a population census comes close – but these are relatively rare. Some companies will have complete lists of customer transactions (telephone calls, bank withdrawals, etc.) but, even then, data from particular times do not exhibit all the types of behaviour which may occur in different economic and social contexts. In the census context, there remain important issues about missing data or underrepresented groups.

The issue of sampling, and what the data actually represent, remains crucially important. To take one example, housing patterns in the rental sector are very important indicators of societal issues and trends and transaction records held by particular rental agencies is an obvious source of data. However, the arrival of other agencies on the market, or campaigns by others to increase market share, can easily distort the picture reflected in a single agency. While relationships between some variables which characterise rental properties may be less affected, the volume of transactions clearly has to be treated with caution.

The statistical theory of sampling provides an invaluable framework within which these issues can be discussed, and sometimes addressed. Cochran (1977) provides a classic description while the RSS Presidential address of Smith (1993) raises more modern issues. Where the provenance of *big data* is unclear, the concepts of sampling provides a language which allows us to engage with, and consider the consequences of, the extent to which

data represent populations.

Then there is the situation where we have essentially all the information on variables, but for only a small number of cases. Genetics is a striking example of this; Laird and Lange (2011) provide a modern introduction. Pluta *et al.* (2018) provide a discussion of big data issues in this area. Brain imaging is another context where massive amounts of information are available, sometimes on small numers of patients or participants. Chung (2018) discusses the resulting issues in this context. In both of these application areas, the number of experimental units involved places an essential limit on the information which is present on variability at this level - a limit which no amount of clever analysis can overcome. Methods of estimation and inference which build in well justified assumptions about the context and characteristics of the data can be helpful here. Again, this is likely to be more feasible where there is the structure of a model within which suitable assumptions can be expressed. Quarteroni (2018) and Lau *et al.* (2018) discuss other settings where understanding of physical models can be very informative on the kind of modelling structure which it may be appropriate to adopt.

Augustin and Faraway (2018) and Dunson (2018) discuss further issues and examples of sampling issues in the context of big, and small, data.

4 Excellent science: the role of a model

It might reasonably be argued that the fundamental contribution of statistics is the concept of a *model* which is capable of capturing all the different aspects of scientific studies. This includes issues of sampling, a description of the relationship between variables, principles for the estimation of quantities (which may be more complex than simple parameters) and, crucially, principles of inference which allow us to reach justifiable conclusions. The language of statistical models, with a rich structure for the effects of variables and po-

tential interactions and the formulation of model comparisons, is an extremely powerful one.

While most attention is understandably focused on the structural component of a model, statistical models importantly allow rich expressions of the structure of the variability present in the data. Analyses which ignore the random effects present in repeated measurements or other grouping structures are at serious risk of reaching inappropriate conclusions. This powerful mechanism, described for example by Pinheiro and Bates (2009), is a further major contribution of statistics.

If the aim is excellent science, in the sense of understanding how processes operate, then statistical models provide enormously powerful tools for any situation where measurements exhibit variability. They enable us to draw clear conclusions and give clear expressions of uncertainty. While other forms of computational expertise are also required, the central place of statistical thinking remains unchallenged.

Acknowledgement

This work was supported in part by the Economic and Social Research Council through a grant (reference ES/L011921/1) to the Urban Big Data Centre based at the University of Glasgow.

References

Augustin, N. H. and Faraway, J. J. (2018). When small data beats big data. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear*.

Chung, M. K. (2018). Statistical challenges of big brain network data. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear.*

Cochran, W. G. (1977). *Sampling Techniques.* New York: Wiley.

Davison, A. C. (2008). *Statistical models.* Cambridge University Press.

Dunson, D. (2018). Statistics in the big data era: Failures of the machine. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear.*

Freedman, D. A. (2009). *Statistical models: theory and practice.* cambridge university press.

Laird, N. M. and Lange, C. (2011). *The Fundamentals of Modern Statistical Genetics.* New York: Springer.

Lau, F. D.-H., Adams, N. M., Girolami, M. A., Butler, L. J., and Elshafie, M. Z. E. B. (2018). The role of statistics in data-centric engineering. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear.*

Pinheiro, J. C. and Bates, D. M. (2009). *Mixed-Effects Models in S and S-PLUS.* New York: Springer.

Pluta, D., Yu, Z., Shen, T., Chen, C., Xue, G., and Ombao, H. (2018). Statistical methods and challenges in connectome genetics. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear.*

Quarteroni, A. (2018). The role of statistics in the era of big data: a computational scientist' perspective. *Statistics and Probability Letters, Special Issue on The role of Statistics in the era of big data, to appear.*

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks 61*, 85–117.

Smith, T. (1993). Populations and selection: limitations of statistics. *Journal of the Royal Statistical Society. Series A. Statistics in society 156*(2), 145–166.