Mathematics, Statistics and Data Science

The process of extracting information from data has a long history (see, for example, [1]) stretching back over centuries. Because of the proliferation of data over the last few decades, and projections for its continued proliferation over coming decades, the term *Data Science* has emerged to describe the substantial current intellectual effort around research with the same overall goal, namely that of extracting information. The type of data currently available in all sorts of application domains is often massive in size, very heterogeneous and far from being collected under designed or controlled experimental conditions. Nonetheless, it contains information, often *substantial* information, and data science requires new interdisciplinary approaches to make maximal use of this information. Data alone is typically not that informative and (machine) learning from data needs conceptual frameworks. Mathematics and statistics are crucial for providing such conceptual frameworks. The frameworks enhance the understanding of fundamental phenomena, highlight limitations and provide a formalism for properly founded data analysis, information extraction and quantification of uncertainty, as well as for the analysis and development of algorithms that carry out these key tasks.. In this personal commentary on data science and its relations to mathematics and statistics, we highlight three important aspects of the emerging field: Models, High-Dimensionality and Heterogeneity, and then conclude with a brief discussion of where the field is now and implications for the mathematical sciences.

Models

Mathematical models provide a conceptual framework within which to interpret data. A well-established connection between models and data is provided by the statistical approach in which the primary task is the inductive inference from data to draw conclusions about unknown model parameters or structures. This is the process of blending models with data. The manner in which the model and the data are linked, and the relative belief in the accuracy of the model and the data, play important roles in this blending process. One class of examples are the complex models arising from Newton's laws, such as those governing the Earth's atmosphere for use in weather prediction. This field (at current levels of computer resolution) involves models with billions of state variables, which are confronted with datasets of millions of measurements at regular intervals several times each day – this is the process of data assimilation [2]. These problems, although increasingly data rich, are very model-driven, with a belief in Newton's laws providing a very strong constraint on the task of interpreting the data. At the other extreme are problems that are primarily data-driven and in which the model arises from the data rather than being a constraint – for example deep learning in image classification [3]. Matrix completion for the Netflix problem [4] is another example of a primarily data-driven application in which the model is not based on any fundamental modelling principles. Between these two extremes of data-driven and model-driven inference are numerous applications in biology and the social sciences in which cartoon models are used, such as the SIR models [5] describing the transmission of infectious diseases or continuum flow models for crowds [6]; in these disciplines, the models are significant constraints on the data but do not have the pivotal position that Newton's laws play in some application areas. In this discussion of model-driven versus data-driven procedures, and the spectrum in between, it is important to appreciate that whilst purely data-driven procedures might be appropriate for the task of forecasting or prediction, they do not provide the additional "mechanistic" or "causal" insights that arise when incorporating data into Newton's laws, for example. Indeed, in fields that are currently data-driven, it can be expected that mechanistic models will emerge as the data provides information about the fundamental mechanisms at play. In particular, this suggests that whilst the mathematical models of the last few centuries are "pencil and paper" models, the next few decades may open up new paradigms for mathematical modelling based around "machine-learnt" models that reside in computer memory and are organised around fundamental principles that emerge from the data.

High-Dimensionality

The topic of "high-dimensionality" arises in two important ways: through the size of the dataset and through the size of the model, as indicated in the examples described above. For high dimensional models, important questions relate to the ability of algorithms to scale to arbitrarily high, even infinite dimensional, formulations [7, 8] of the statistical inference problem. Another key development in high dimensional statistics is based on the concept of sparsity, which has proven to be remarkably successful in many applications, including the highly celebrated compressed sensing methodology [9, 10] and its noisy version, in which stochastic error terms are superimposed on the observed signal. The mathematical underpinning and understanding of high-dimensional statistical inference (see [11] and [12]) has evolved almost simultaneously with practical applications.

Early examples of high-dimensional problems arose in genomics in the late 1990s when relating disease status to the genetic profile of a person [13]. When measuring many biomarkers (expressions of many genes in the genome), for example around ten thousand in the early times of such applications, a model would typically involve (at least) one unknown parameter for every measured variable, encoding an unknown effect of the biomarker to the disease status. And thus, there is immediately an inference problem with ten thousand unknown parameters to be estimated from about one hundred people participating in a well-designed study. Successful models, classification and even causal inference techniques have been built in statistics and bioinformatics at the interface between molecular and computational biology. Nowadays, genetic profiles are measured with millions of biomarkers and, instead of a

well-defined single study, there are huge health databases containing information from very many people. New problems that arise include privacy issues and heterogeneity; the latter is discussed in the following paragraph.

Heterogeneity

With growing data volume, one might reasonably expect an increased sample size. But these large datasets that are now routinely available are usually not collected from well-designed experiments and they are often rather heterogeneous. They might exhibit unwanted time trends, variation or sub-population structures or arise from different sources like satellites, aircraft and weather balloons for weather prediction. When partitioning the massive data into fairly homogeneous groups (which, without further information, is a difficult task in statistical mixture or change point modelling), this often leads to severe high-dimensional problems in which the sample size within a homogeneous group is rather small in comparison to the dimensionality of the unknown model parameters. New avenues of investigation are required to tackle fundamental problems relating to heterogeneity in large-scale data. This is an area where new input is needed from mathematics and statistics because naive design and use of standard algorithms does not lead to accurate information extraction.

Where Are We Now?

Data science is clearly emerging as an identifiable research area of enormous importance, dealing with large-scale data problems in many application areas such as biology and medicine (epidemics, genetics and genomics, neuroscience), engineering (imaging, signal processing), geophysical sciences (climate, weather, the Earth's subsurface and numerous energy applications), the social sciences (economics, ranking and voting, crowd sourcing) and commerce (Amazon, Google, Netflix), to name just a few. Whether data science will become a distinct academic discipline in the way that computer science did in the 1950s remains to be seen. But, clearly, the subjects of mathematics and statistics have very close relations to data science, whatever form it takes. Statistics has a longstanding tradition and an established framework for quantifying uncertainties and this in turn helps to address substantial problems of replicability in data-driven science. Mathematics has a unique position to contribute to the foundations in information and data science. Whilst avoiding claims that mathematics and/or statistics should "own large parts" of data science, it is clear that we should embrace others who participate in the endeavour and the intellectual challenges that stem from doing so. Data science is certainly stimulating new and exciting mathematics and statistics. As a consequence, education in the mathematical sciences should incorporate more in-depth training in computing, mathematical modelling and statistical thinking, in the context of data-rich applications and theoretical paradigms.

References

[1] S.M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press/Harvard University, 1986.

[2] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 2015.

[3] J. Schmidhuber. Deep learning in neural networks: an overview. *Neural Networks,* 61: 85–117, 2015.

[4] E.J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9: 717–722, 2009.

[5] W.O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A*, 115: 700–721, 1927.

[6] R.L. Hughes. A continuum theory for the flow of pedestrians. *Transportation Research Part B*, 36: 507–535, 2002.

[7] M. Dashti and A. M. Stuart. The Bayesian Approach to Inverse Problems. Springer Handbook on Uncertainty Quantification, to appear. arxiv:1302.6989.

[8] S.L. Cotter, G.O. Roberts, A.M. Stuart and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28: 424–446, 2013.

[9] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52: 1289–1306, 2006.

[10] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52: 5406–5425, 2006.

[11] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer Verlag, Series in Statistics. 2011.

[12] T. Hastie, R. Tibshirani and M. J. Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman and Hall/CRC Press, Series in Statistics and Applied Probability, 2015.

[13] T.R. Golub et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286: 531–537, 1999.

Peter Bühlmann
(ETH Zürich, Switzerland) and
Andrew M. Stuart
(Warwick University, Coventry, UK)

A.M. Stuart is grateful to DARPA, EPSRC, ERC and ONR for financial support that led to some of the research underpinning this article.