Statistics and Probability Letters 136 (2018) 42–45

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Statistical science in the world of big data☆

Nancy Reid

Department of Statistical Sciences, University of Toronto, 100 St. George St., Toronto M5S 3G3, Canada

ARTICLE INFO

Article history:
Available online 21 February 2018

MSC:
62-07
62-02
68-02

Keywords:
Data science
Inference
Machine learning

ABSTRACT

This essay considers the role of the statistical sciences in the world of big data, data science, machine learning, and artificial intelligence, with a decidedly Canadian slant.

1. Introduction

From January to June 2015, a series of twelve workshops took place across Canada, as part of a six-month thematic program at the Fields Institute for Research in the Mathematical Sciences, organized by the Canadian Statistical Sciences Institute. Most of the activity was in Toronto at the Fields Institute (Fields, 2015), which also provided the main funding. Allied workshops were held at the Pacific Institute for Mathematical Sciences in Vancouver, the Centre de Recherches Mathématiques in Montreal, and the Atlantic Association for Research in the Mathematical Sciences in Halifax. I chaired the program organizing committee and the international advisory committee, but the success of the effort is really due to the hard work and varied contributions of the committee members.

One positive outcome of this program was that I spent considerable time discussing with my colleagues, and thinking about, the place of statistical science in what was then the booming area of big data. This essay summarizes some of my thoughts on this, with the advantage of hindsight, and informed by the many changes that have taken place at remarkable speed since then.

Some highlights of the thematic program are described in Franke et al. (2016).

2. Big data

When we started planning the proposal submission for the thematic program in the summer of 2013, everyone was talking about big data, and calling it "Big Data". In fact we spent some time worrying that it might be risky to include this in the title of the thematic program, in case the phrase might be out of date before the program got started in 2015. As it turned out, it was not, although it had already decreased a few levels on the "Gartner hype cycle" (Gartner, 2014). The 2015 version of Gartner's cycle, published after our program ended, did not even include "Big Data"; a small footnote indicated that it had been removed as an emerging technology because it was now well established, as reported for example in Woodie (2015).

☆ Supported by the Natural Sciences and Engineering Research Council of Canada.
E-mail address: reid@utstat.utoronto.ca.
URL: http://www.utstat.toronto.edu/reid.

https://doi.org/10.1016/j.spl.2018.02.049
0167-7152/

However the phrase big data is still widely used, and continues to generate interest, as this special issue demonstrates. And as the program of workshops progressed, I became much less concerned about what we called it, as it was clear that the data are here to stay, and, for the most part, really are big. That is not necessarily the most interesting or most demanding aspect of the new world of data everywhere, but there is no doubt that in fields of study from anthropology to zoology there is more data being collected than at any time in the history of these fields. Of course big is in the eye of the beholder, or computer, and large amounts of data have been available in some settings for many decades: the *Liberty Digest* political poll of 1936, a spectacular failure, had a sample size of some 2.4 million (Harford, 2014); the first census of India was taken in 1872. Statisticians have long been accustomed to sampling as a means to gain insight from big data.

So while big data no longer qualifies, if it ever did, as an "emerging technology", it continues to get bigger, and to inform many changes in society, as noted by several papers in this volume. Ceri (2018) refers to the "Fourth paradigm" and Dryden and Hodge (2018) to a "new natural resource", for example.

It is, or was, popular to speak of the "four V"s; volume, velocity, variety, and veracity, and this can be a useful shorthand to highlight some of the challenges. From the point of view of statistical science, some of the more interesting problems arise in developing statistical methods for new types of data, such as images or networks; in finding ways to conduct valid inferences in very complex models, which is sometimes abbreviated as high-dimensional inference; and in developing and expanding our repertoire of methods for dealing with typically very small samples of very extreme observations. Each of these topics is included in this special issue: see for example Chung (2018) and Pluta et al. (2018) on images and networks, Giraldo et al. (2018) and Bivand and Krivoruchko (2018) and Fassò et al. (2018) on spatial data evolving over time. Bartolucci et al. (2018), Castruccio and Genton (2018) and Vieu (2018) all touch on various aspects of high-dimensional inference.

Related to the question of valid inference is the difficulty of correctly assessing the variability when the number of observations is very large; Cox (2015) shows how crucial the assumptions of independence are for the usual formulas, but assumptions of independence are especially unreliable in very large sets of data; a point noted also by Faraway and Augustin (2018) and Bühlmann and van de Geer (2018). Many of the classical statistical approaches do not scale well from a computational point of view, which suggests fruitful avenues for collaboration with computer scientists and engineers (Lau et al., 2018; Ceri, 2018). Strategies common to many analyses of big data include dimension reduction, assumptions of sparsity, regularization, and methods for the analysis of complex dependencies.

Something that emerged repeatedly during our series of workshops, in talks across many domains, was that statisticians are drawn to complex, typically high-dimensional, models, emphasizing inferential methods. While this is a difficult area, it is reasonably well-studied in the statistical literature, and a number of promising advances have been made. Computer scientists, on the other hand, are better equipped to deal with size, speed, and scalability. Data owners, particularly in business applications, are very cognizant of the computational issues, so that statistical ideas can get lost in the machine. Data science seems to be the field that brings these strands closer together.

3. Data science

By the time our program ended in mid-2015 data science was coming to replace big data as a short-hand for the world of lots of data. This has now become much more current, and in my view represents a multi-disciplinary field that includes aspects of applied mathematics, computer science, statistics, and subject-matter applications. Although it has been argued that statistical science is data science (Yu, 2014) and that departments of statistics should rename themselves (a few indeed have already done so), I disagree. Jenny Bryan summarized this succinctly in a talk in our program in which she pointed out that the statement "statistics is 'just' data science" is in direct contradiction to the statement "data wrangling is not statistics" (Bryan, 2015).

Data science programs at universities are being developed at a rapid pace, in direct response as far as I can tell to demands from government and industry, and by and large these involve a blend of courses from mathematics, computer science and statistics, with additional emphasis on workflow, reproducible research, communication, and visualization. The fact that so many groups have fairly quickly converged on fairly similar programs (see, e.g., DeVeaux et al. (2017) and James (2018)), suggests that there is broad consensus on how to train someone to seek employment as a data scientist. Implicit in many of these programs, but not usually explicit, is the notion that data science should be usefully applied to 'real' data in consultation with the scientists, social scientists, and humanists who are gathering this data. At the same time many disciplines are themselves training data scientists within their own fields, and this has moved much beyond traditional subfields of the physical and social sciences.

Data science as a research field seems much more difficult to pin down. As is familiar to applied statisticians of all stripes, it is not clear if data science is something one studies, or something one does. The National Science Foundation has made an explicit effort to define the research field through its TRIPODS program (National Science Foundation, 2016), and the Alan Turing Institute in the UK devoted its inaugural year to a very broad series of scoping workshops (Blake and Olhede, 2016) to explore the research domain that could be data science. These are the only broad national research initiatives focussed on research in data science that I am aware of, although there are a great many centres for data science and/or big data in universities and research institutes around the world.

For the moment it seems useful to consider the research field of data science as a blend of statistical modeling and inference, data management, computing at scale, optimization, communication and visualization. This field will not survive on these topics alone, though, because like statistics and applied mathematics, theory and methods need to develop in close

collaboration with application areas: Shi (2018), Meng (2018) and Smirnova et al. (2018) stress the importance of deep engagement in the subject matter. What is perhaps new is that these application areas range over a much broader set of scholarly disciplines than in the past. Also new is the vastly increased interest from industry and government in putting troves of data to use, even if the particular use may sometimes be rather vague. And a thoroughly welcome development, in my view, is the remarkable quality of visualizations informing modern journalism.

The fact that the words data science have been used to describe a course, a program, a job, and a technology, as well as a new field of research, means there is a great deal of uncertainty about what determines the research field. Collaboration and a renewed focus on communication are possibly the most important hallmarks. Blei and Smyth (2017) emphasize this collaboration and its impact on science.

4. Machine learning, deep learning and artificial intelligence

Machine learning is a distinct sub-field of computer science with a clear research agenda and a relatively long history; many statisticians will have been introduced to topics in machine learning via Hastie et al. (2009) or Bishop (2006). Deep learning, as explained for example in talks by Brendan Frey and by Machine Learning Workshop (2015), refers to both computational and modeling aspects of neural networks with a very complex architecture. Deep learning seems to be the breakthrough that has led to rapid improvements in voice recognition, natural language processing, and image analysis, for example. Artificial intelligence is, in principle, a broader but vaguer concept that is usually used to describe computational methods that in principle draw inspiration from the study of human intelligence.

In Canada both government and industry are making quite large investments in research in artificial intelligence, particularly in Montreal, Toronto and Edmonton, and there is a great deal of interest in the potential of this research to bring innovation to many areas of society, including health care (see, e.g. Vector Institute (2017)), finance, energy, manufacturing, and so on.

The highly visible successes of deep learning in applications that are widely used in mobile devices have generated a great deal of interest, excitement, and funding for artificial intelligence, and it is easy to feel that statistical science is missing the wave. At least some of the excitement is somewhat mis-placed; as a relatively new technology, AI is perceived to be able to solve all data problems, but as consulting statisticians are well aware, clients often assume their data require a particular solution, when deeper probing can clarify that there are many other, possibly simpler, options available. Statistical methods will continue to be essential for progress in the world of big data, and classical statistical concepts are more relevant than ever. For an elegant application of classical work in experimental design to high-dimensional model selection, see Cox and Battey (2017).

5. Conclusion

We designed the thematic program as a blend of foundational themes and applications-oriented themes. In the former we chose to emphasize machine learning, high-dimensional inference, optimization and visualization; for the latter social policy, health policy, environmental science and networks. Of course as usual the distinctions between the applications and the foundations were blurry, and many other application areas were touched on in various presentations.

What emerged from our experience was another important set of cross-cutting themes that I would call essential components of data science, in addition to those mentioned above. One is a renewed emphasis on workflow, which is at least partly a reaction to the concern about reproducible research (Wilson et al., 2017). Another cross-cutting theme is communication, which includes visualization, but also the recognition that researchers have a responsibility to ensure that the insights obtained from big data are used to inform decision making and policy, especially in the public realm (Azzone, 2018; Sharples, 2018). Perhaps the most important cross-cutting theme related to big data is privacy, which covers all aspects of the data life-cycle, from collection, to re-use. One aspect of this is statistical disclosure limitation, which the national statistical agencies have studied for many decades; another is the notion of differential privacy, motivated by principles of cryptography and secure communication. More recently privacy concerns have expanded to include concerns of algorithmic fairness, particularly with deep learning algorithms, which can be quite opaque even to their developers (Shah, 2017; O'Neil, 2016).

Statistical science encourages us to be cautious, and statisticians are sometimes cautious to a fault, which has the potential to limit our effectiveness in multidisciplinary collaborations. However, some cautionary voices are no doubt needed, especially in view of the immense public interest in the promise of big data. There has in the past several years been a remarkable increase in the reporting of statistical ideas and data-driven discoveries in the popular press, and a parallel increase in public interest in statistical and data science. This is very promising for the development and growth of statistical science, and for the active engagement of statisticians with the new fields that emerge, whatever they are called.

Acknowledgments

I would like to thank the Fields Institute for Research in the Mathematical Sciences for support for the thematic program described here, and my colleagues on the organizing committee: Yoshua Bengio, Hugh Chipman, Sallie Keller, Lisa Lix, Richard Lockhart and Ruslan Salakhutdinov. Helpful conversations with Raymond Ng, Mary Thompson, Don Fraser, Sofia Olhede and Bin Yu have also framed my thinking on many of the issues around big data and data science.

References

Azzone, G., 2018. Big data and public policies: opportunities and challenges. Statist. Probab. Lett. 136, 116–120. Special Issue on "The role of Statistics in the era of Big Data".

Bartolucci, F., Bacci, S., Mira, A., 2018. On the role of latent variable models in the era of big data. Statist. Probab. Lett. 136, 165–169. Special Issue on "The role of Statistics in the era of Big Data".

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer.

Bivand, R., Krivoruchko, K., 2018. Big data sampling and spatial analysis: "which of the two ladles, of fig-wood or gold, is appropriate to the soup and the pot?" Statist. Probab. Lett. 136, 87–91. Special Issue on "The role of Statistics in the era of Big Data".

Blake, A., Olhede, S., 2016. Report on the 2015/6 scoping programme. Alan Turing Institute UK. URL https://aticdn.s3-eu-west-1.amazonaws.com/2015/06/The-Alan-Turing-Institute-Scoping-Programme.pdf (Accessed: 01.10.17).

Blei, D.M., Smyth, P., 2017. Science and data science. Proc. Natl. Acad. Sci. 114, 8689–8692. http://dx.doi.org/10.1073/pnas.1702076114.

Bryan, J., 2015. New tools and workflows for data analysis. https://www.fields.utoronto.ca/video-archive/event/1277 (Accessed: 01.10.17).

Bühlmann, P., van de Geer, S., 2018. Statistics for big data: a perspective. Statist. Probab. Lett. 136, 37–41. Special Issue on "The role of Statistics in the era of Big Data".

Castruccio, S., Genton, M.G., 2018. Principles for statistical inference on big spatio-temporal data from climate models. Statist. Probab. Lett. 136, 92–96. Special Issue on "The role of Statistics in the era of Big Data".

Ceri, S., 2018. On the role of statistics in the era of big data: a computer science perspective. Statist. Probab. Lett. 136, 68–72. Special Issue on "The role of Statistics in the era of Big Data".

Chung, M.K., 2018. Statistical challenges of big brain network data. Statist. Probab. Lett. 136, 78–82. Special Issue on "The role of Statistics in the era of Big Data".

Cox, D.R., 2015. Big data and precision. Biometrika 102, 712–716. http://dx.doi.org/10.1093/biomet/asv033.

Cox, D.R., Battey, H.S., 2017. Large numbers of explanatory variables, a semi-descriptive analysis. Proc. Natl. Acad. Sci. 114, 8592–8595. http://dx.doi.org/10.1073/pnas.1703764114.

DeVeaux, R.D., Agarwal, M., Averett, M., Baumer, B.S., Bray, A., Bressoud, T.C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R.J., Sondjaja, M., Tiruviluamala, N., Uhlig, P.X., Washington, T.M., Wesley, C.L., White, D., Ye, P., 2017. Curriculum guidelines for undergraduate programs in data science. Ann. Rev. Stat. Appl. 4 (1), 15–30. http://dx.doi.org/10.1146/annurev-statistics-060116-053930.

Dryden, I.L., Hodge, D.J., 2018. Journeys in Big Data statistics. Statist. Probab. Lett. 136, 121–125. Special Issue on "The role of Statistics in the era of Big Data".

Faraway, J.J., Augustin, N., 2018. When small data beats big data. Statist. Probab. Lett. 136, 142–145. Special Issue on "The role of Statistics in the era of Big Data".

Fassò, A., Finazzi, F., Madonna, F., 2018. Statistical issues in radiosonde observation of atmospheric temperature and humidity profiles. Statist. Probab. Lett. 136, 97–100. Special Issue on "The role of Statistics in the era of Big Data".

Fields, 2015. Thematic program on statistical inference, learning and models in big data. http://www.fields.utoronto.ca/activities/thematic/thematic-program-statistical-inference-learning-and-models-big-data (Accessed: 27.09.17).

Franke, B., Plante, J.-F., Roscher, R., Lee, E.-S.A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M.M., Grosse, R., Hendricks, D., Reid, N., 2016. Statistical inference, learning and models in big data. Internat. Statist. Rev. 84, 371–389. http://dx.doi.org/10.1111/insr.12176.

Gartner, 2014. Gartner's 2014 hype cycle for emerging technologies maps the journey to digital business. http://www.gartner.com/newsroom/id/2819918 (Accessed: 27.09.17).

Giraldo, R., Dabo-Niang, S., Martínez, S., 2018. Statistical modeling of spatial big data: an apporach from a functional analysis perspective. Statist. Probab. Lett. 136, 126–129. Special Issue on "The role of Statistics in the era of Big Data".

Harford, T., 2014. Big data: are we making a big mistake? Financial Times, March 28. https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0 (Accessed: 28.09.17).

Hastie, T., Tibshirani, R., Friedman, J., 2009. Elements of Statistical Learning, second ed.. Springer.

James, G., 2018. Statistics within business in the era of big data. Statist. Probab. Lett. 136, 155–159. Special Issue on "The role of Statistics in the era of Big Data".

Lau, F. D.-H., Adams, N.M., Girolami, M.A., Butler, L.J., Elshafie, M.Z.E.B., 2018. The role of statistics in data-centric engineering. Statist. Probab. Lett. 136, 58–62. Special Issue on "The role of Statistics in the era of Big Data".

Machine Learning Workshop, 2015. Workshop schedule. http://www.fields.utoronto.ca/activities/workshops/workshop-big-data-and-statistical-machine-learning (Accessed: 27.09.17).

Meng, X.-L., 2018. Conducting highly principled data science: A statistician's job and joy. Statist. Probab. Lett. 136, 51–57. Special Issue on "The role of Statistics in the era of Big Data".

National Science Foundation, 2016. Transdisciplinary research in the foundations of data science. https://nsf.gov/funding/pgm_summ.jsp?pims_id=505347. Call for Proposals. (Accessed: 27.09.17).

O'Neil, C., 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Random House.

Pluta, D., Yu, Z., Shen, T., Chen, C., Xue, G., Ombao, H., 2018. Statistical methods and challenges in connectome genetics. Statist. Probab. Lett. 136, 83–86. Special Issue on "The role of Statistics in the era of Big Data".

Shah, H., 2017. The DeepMind debacle demands dialogue on data. Nature 547, 259. http://dx.doi.org/10.1038/547259a.

Sharples, L.D., 2018. The role of statistics in the era of big data: electronic health records for healthcare research. Statist. Probab. Lett. 136, 105–110. Special Issue on "The role of Statistics in the era of Big Data".

Shi, J.Q., 2018. How do statisticians analyse big data –our story. Statist. Probab. Lett. 136, 130–133. Special Issue on "The role of Statistics in the era of Big Data".

Smirnova, E., Ivanescu, A., Bai, J., Crainiceanu, C.M., 2018. A practical guide to big data. Statist. Probab. Lett. 136, 25–29. Special Issue on "The role of Statistics in the era of Big Data".

Vector Institute, 2017. Vector Institute to Collaborate with Peter Munk Cardiac Centre and University Health Network. https://s3.ca-central-1.amazonaws.com/vectorinstitute.ai/resources/Vector_Statement_PMCC-UHN_20170919.pdf (Accessed: 01.10.17).

Vieu, P., 2018. On dimension reduction models for functional data. Statist. Probab. Lett. 136, 134–138. Special Issue on "The role of Statistics in the era of Big Data".

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., Teal, T.K., 2017. Good enough practices in scientific computing. PLoS Comput. Biol. 13 (6), 1–20. http://dx.doi.org/10.1371/journal.pcbi.1005510.

Woodie, A., 2015. Why Gartner dropped big data off the hype curve. https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/ (Accessed: 27.09.17).

Yu, B., 2014. IMS presidential address: Let us own data science. http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/ (Accessed: 01.10.17).