ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

On the role of statistics in the era of big data: A call for a debate

Piercesare Secchi

MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Paradigm shift in statistics
Big data analytics
High dimensional and complex data
Two cultures
Teaching statistics
Distributed inference
Data integration

ABSTRACT

While discussing the plenary talk of Dunson (2016) at the 48th Scientific Meeting of the Italian Statistical Society, I formulated a few general questions on the role of statistics in the era of big data which stimulated an interesting debate. They are reported here with the aim of engaging a larger audience on an issue which promises to change radically our discipline and, more generally, science as we know it. But is it so?

1. A sort of introduction

"Big data", is this the latest buzzword on the market or a credible indication of a paradigm shift which is changing science, not to mention statistics and the way we collect and analyze data? Without even a small answer to this big question, I was called to discuss the plenary talk *Probabilistic inference for big & complex data* delivered by David Dunson at the 48th Scientific Meeting of the Italian Statistical Society, held at the University of Salerno in June 2016. In the lack of profound insights, I thought I could at least formulate a few questions that would help me and my peers to size up some facets of the problem. These questions are here proposed again as an expedient to stimulate a discussion on the role of statistics in the era of big data, if such an era exists and will not soon disappear as a feeble rhetorical invention. First, however, a disclaimer is in order: by no means my questions cover the entirety of the debate on big data, which is by now already very rich even on the pages of statistical journals. Indeed I hope that other and more knowledgeable discussants will fill the gap and touch upon aspects of the involved relationship between big data analytics and statistics which I left, consciously or not, underground.

One thing however seems indisputable, the big data trade is generating a humongous array of ad hoc analytics with velocity, volume and variety competing with those that are said to capture the essence of the new data ecosystem. Be that as it may, unified theoretical frameworks for the statistical analysis – and inference – of big data are still missing, generating in the lay statistician the unspoken impression that we are dealing with alchemy, while chemistry is yet to come.

In *Pilgrim at Tinker Creek*, Dillard (1974) wrote:

> If we are blinded by the darkness, we are also blinded by the light. When too much light falls on everything, a special terror results.

Too many unfiltered new ideas hamper innovation. This is the starting point of the latest book by Verganti (2017), *Overcrowded: Designing Meaningful Products in a World Awash With Ideas.* In the dark we can light a candle, but what shall we do when we are blinded by too many lights? Verganti's dictum is to pursue innovation driven by meaning, placing the human

E-mail address: piercesare.secchi@gmail.com.

https://doi.org/10.1016/j.spl.2018.02.041
0167-7152/

Please cite this article in press as: Secchi P., On the role of statistics in the era of big data: A call for a debate. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.041.

back at the center. This could as well be a precious indication for the statistician lost in the forest of "Big Data Analytics" under the spell of automatic science.

In the next Section I will recall the questions which ignited the debate after Dunson's plenary talk at the 48th Scientific Meeting of SIS. Their original aim was to expose a few critical and general issues raised by the big data approach to science, not just statistics. Indeed, although not all the questions are immediately related to the challenges posed to statistics by the current demand of analytics for huge datasets, these issues necessarily engage statistics as the practice dedicated, by tradition, to the collection and analysis of data. In the concluding section, I will however touch upon a few theoretical challenges which I believe have the potential to become important in a future dominated by the analysis of big data.

2. The seven questions at the SIS meeting

2.1. Is there a role for statistics in the big data era?

I would describe statistics as the science of variability, meaning that the main goal of statistics is to develop paradigms, methods and algorithms for the mathematical exploration, elicitation and control of variability, and the uncertainty it generates. Inference and uncertainty quantification are at the core of statistics and they have generated correlated siblings like prediction, testing, controlling for dependence, confounding, randomization. Yet these fundamental ideas of statistics are not often considered the primeval sources of the big data enlightenment. Is this the beginning of the end for statistics as we know it?

2.2. After the big data deluge, where do we stand in the debate about the two cultures in statistical modeling?

In 2001 Leo Breiman wrote on *Statistical Science*:

> There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

After the big data deluge, where do we stand in the debate about the two cultures in statistical modeling?

2.3. Shall we abandon parametric statistics for good?

Big data are often characterized by huge sample sizes and a huge number of variables presenting very complex form of dependence. This is at the antipodes of the "received version" of statistics where datasets have small or moderate sample size and a reduced number of variables. In many big data applications, no parametric model is likely to capture the entire relevant variability occurring in the sample. Despite this, parametric models are still predominant both in the theory and in the practice of statistics. Is there a risk that in the big data era, parametric statistics will dangerously move from fitting models to data to fitting data to models? Could a truly non-parametric approach represent a point of contact between the two cultures evoked by Breiman (2001)? Indeed in non-parametric statistics, on the one hand, one assumes that the data are generated by a stochastic model, and on the other one, the stochastic model is treated as (almost) completely unknown.

2.4. Not always massive, but often complex…

The big in big data is frequently a shorthand for massive in volume, velocity and variety. But data could also be big in complexity. More and more often, our statistical analyses involve data objects that are not easily reduced to a Euclidean vector representation – the battlefield of classical Multivariate Data Analysis – without losing the information content they support. I am thinking about applications of statistics in modern science where the atoms of the analysis are functions and surfaces, positive definite matrices or tensors, manifold data, trees, networks, texts. Object Oriented Data Analysis (Wang and Marron, 2007) provides a framework and a useful mode of discussion, for approaching complex data challenges. In the big data era what will be the data objects of the statistical analysis?

2.5. Teaching the next generation of statisticians

Good big data analyses require a deep understanding of the "physical" phenomenon generating the data. Whether we are Bayesians or not, this should be reflected in the prior knowledge which dictates the statistical paradigm and model, the

Please cite this article in press as: Secchi P., On the role of statistics in the era of big data: A call for a debate. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.041.