BIG DATA & SOCIETY

Commentary

Big Data & Society
January–June 2017: 1–6
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053951717703997
journals.sagepub.com/home/bds
SAGE

Big Data is not only about data: The two cultures of modelling

Giuseppe Alessandro Veltri

Abstract

The contribution of Big Data to social science is not limited to data availability but includes the introduction of analytical approaches that have been developed in computer science, and in particular in machine learning. This brings about a new 'culture' of statistical modelling that bears considerable potential for the social scientist. This argument is illustrated with a brief discussion of model-based recursive partitioning which can bridge the theory and data-driven approach. Such a method is an example of how this new approach can help revise models that work for the full dataset: it can be used for evaluating different models, a traditional weakness of the 'traditional' statistical approach used in social science.

Keywords

Big Data, statistics, recursive partitioning, modelling culture, social sciences

Introduction

Much has been discussed about the potential of Big Data in shaping social scientific research, well summarised in this journal by Kitchin (2014) and, at the same time, there has been a wide discussion about the dangers of Big Data; in particular, the potential simplifications of human agency and context of data production (Boyd and Crawford, 2012; Schroeder, 2014; Tinati et al., 2014). While the discourse of risks has been fairly articulated, discussion concerning Big Data's innovations and contributions has been less detailed and has often remained limited to the level of generic promises of boosting the descriptive and predictive power of social sciences. In this commentary, I would like to focus on how the availability of data unlocks a different approach to analyse social science datasets, with a brief discussion of model-based recursive partitioning which can bridge the theory and data-driven approach. It is a similar case to the 'renaissance' of social network theory and analysis: thanks to the large increase in availability of relational datasets about human interactions and relationships, network scientists have been able to develop a new wave of theoretical concepts and techniques (Barabasi and Posfai, 2016). Overall, the discussion about Big Data should be considered in the wider context of two more significant changes of the past two decades: 1) the crisis of measurements in social sciences; 2) the rise of competing paradigms between traditional statistical methods and algorithmic and machine learning approaches. Both points will be explored in this commentary in order to argue that the contribution of Big Data and associated methodologies should be discussed in the context of existing methodological conundrums. At the same time, Big Data should be considered in terms of innovation and hard fought progress in analytical methods, born in a data-rich environment that makes data actionable, is essential.

Taking the two points above into consideration will position the debate about the usefulness and validity of Big Data research in the context of 'computational social science': the interdisciplinary investigation of the social universe on many scales, ranging from individual actors to the largest groupings, through the medium of computational techniques in contrast to traditional statistical methods (Lazer et al., 2009). This

Department of Sociology and Social Research, University of Trento, Italy

Corresponding author:
Giuseppe Alessandro Veltri, Department of Sociology and Social Research, University of Trento, Via G. Verdi 26, 38122 Trento, Italy.
Email: giuseppe.veltri@unitn.it

umbrella definition is somewhat long and will be refined later as we examine many topics involved in the practice of CSS and the variety of computational approaches that are necessary for understanding social complexity. For example, the 'many scales' of social groupings involve a great variety of organisational, temporal and spatial dimensions, sometimes simultaneously. In addition, computation or computational approaches refer to numerous computer-based instruments, as well as substantive concepts and theories, ranging from information extraction algorithms to computer simulation models (Alvarez, 2016; Cioffi-Revilla, 2013).

The methods crisis in social sciences

The 'thirst' for methodological innovation in social sciences that might have led to the current hype around the 'data revolution' could be due to the enduring crisis that has characterised most of the widely-used existing techniques. Surveys are exemplary in this case; a pillar methodology across so many different disciplines that is suffering a long lasting crisis due to the increased difficulty in response rates, sampling frames and limited capacity in capturing variables that are more and more important proxy in sociological information; for example, precise geographical location (Burrows and Savage, 2014; Savage and Burrows, 2007).

Similar considerations concern the in-depth interview; another important instrument of data collection in social science. One criticism concerns the translation of a technique developed before the advent of digital media and the question related to the implications of interviews carried via computer-mediated communication. While much literature focused on the interviewer's bias, there is not much acknowledgment of the increasing scientific literature of human bias in memory recollection, sensitivity to contextual elements in activating heuristics and the process of rationalisation (Kahneman and Tversky, 2000; Podsakoff et al., 2003) that affects interviewees. Increasingly, self-reported surveys and interviews measuring human motivations and behaviour are under scrutiny and compared to more 'organic' sources of data (Curti et al., 2015). This is not to say that Big Data does not raise a substantial amount of concern regarding the tendency to consider these as 'organic': the current debate is the kind of critical social science that should accompany all methods used by social scientists. Perhaps for historical reasons, the artificial nature of traditional methods has been long forgotten until recently, when their capacity of generating quality data has become increasingly more problematic.

Such limitations are even more clear if we consider two further aspects: first, the large majority of social science data from surveys and interviews are cross-sectional without a longitudinal temporal dimension (Abbott, 2001); second, most social science datasets are coarse aggregations of variables because of the limitations in what can be asked from self-reported instruments. Big Data are forcing innovation on both accounts, moving from static snapshots to dynamic unfoldings and from coarse aggregations to high resolutions of data. The interesting byproduct of these innovations is the possibility of an increased focus in the social sciences on processes rather than structures. For the first time, we can obtain longitudinal baseline norms, variance, patterns, and cyclical behaviour. This requires thinking beyond the simple causality of the traditional scientific method into extended systemic models of correlation, association, and episode triggering. Once more, network analysis is a good example here: the availability of longitudinal relational data sparked the recent methodological and theoretical innovations about networks' dynamics (Barabasi and Posfai, 2016).

However, any discussion about the limitations of a method is incomplete if not accompanied by the analytical and epistemological framework that it entails. A discussion about the epistemological status of Big Data is beyond the scope of this commentary. The focus, instead, will be on the analytical framework: a topic of discussion that has not received the same level of attention. It is a relevant discussion because it can simultaneously highlight the differences to past approaches and suggest the potential innovative nature of Big Data, its use in social science in terms of analytical approaches and not only in terms of access to the availability of data.

Two cultures of modelling

The role of Big Data and its impact on social science research needs to be addressed in the context of the 'computational and algorithmic turn' that is increasingly affecting social science research methods. In order to fully appreciate such a turn, we can contrast the difference between the 'two cultures of modelling' (Breiman, 2001; Gentle et al., 2012). The first is the 'data modelling' culture in which the analysis starts by assuming a stochastic data model for the inside of the black box of Figure 1(a) and therefore resulting in Figure 1(b). The 'algorithmic modelling' considers the inside of the box as complex and unknown. Such an approach is to find an algorithm that operates on x to predict the responses y (see Figure 1(c)). Borrowing from Breiman (2001), the data modelling approach is about evaluating the values of parameters from the data and after that the model is used for either information or prediction (Figure 1(b)). In the algorithmic modelling approach, there is a shift from data models to the properties of algorithms.

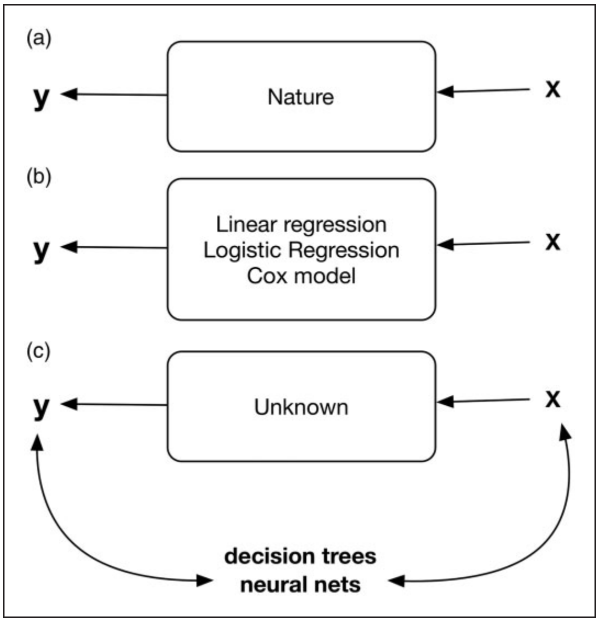

Figure 1. The two cultures of statistical modelling. a) the starting point, b) the data modelling approach, c) the algorithmic modelling approach.
Source: Breiman (2001).

While there is fascinating debate among statisticians and 'machine learning' researchers on the pro and cons of each approach, it is evident that the vast majority of quantitative analytical methods used in social science belong to the 'data modelling' culture. Among others, there are two large shortcomings to this 'traditional' approach as the complexity of datasets increases. First, assumptions about data are often violated: there is a multiplicity of models that are still a good fit for the dataset, even though they might express different relationships between variables. Second, we apply models that are 'global,' meaning that they are cast on the entire dataset without considering the possibility of 'local' models – different versions of the starting model that work better for a given segment of our dataset (that often means a segment of participants with given characteristics). Therefore, in both cases, often the risk is that we are producing conclusions about the one model selected and not the 'real one', which, in turn, can lead to irrelevant theory and questionable statistical conclusions. However, learning from the other modelling 'culture' can help us with this problem.

Bridging theory-based and data-driven approaches: Model-based recursive partitioning and Ockham's razor

One of the best examples of how social science can benefit from the analytical approaches developed in the context of Big Data is represented by the development of model-based recursive partitioning. This approach is an improvement on the use of classification and regression trees, the latter also being a method coming from the 'algorithmic culture' of modelling that has useful applications in social science but that is essentially data-driven (Berk, 2006; Hand and Vinciotti, 2003).

To summarise, classification and regression trees are based on a purely data-driven paradigm. Without making use of a predefined statistical model, such algorithmic methods search recursively for groups of observations with similar values to the response variable by building a tree structure. They are very useful in the context of data exploration and express their best utility in the context of highly complex and large datasets. Nevertheless, such techniques make no use of theory in describing a model of how data was generated and they are purely descriptive, although far better than the 'traditional' descriptive statistics used in social sciences when facing large datasets.

Model-based recursive partitioning (Zeileis et al., 2008) represents a synthesis of a theory-based approach and a data-driven set of constraints to the theory validation and further development. In extreme synthesis, this approach works through the following steps. First, a parametric model is defined to express a theory-driven set of hypotheses (e.g., a linear regression). Second, this model is evaluated to the model-based recursive partitioning algorithm that checks whether other important covariates have been omitted that would alter the parameters of the initial model. The same tree-structure of a regression, or classification tree, is produced. This time, rather than splitting for different patterns of the response variable, the model-based recursive partitioning finds different patterns of associations between the response variable and other covariates that have been pre-specified in the parametric model. In other words, it creates different versions of the model in terms of betas estimation, depending on different important values of covariates. (For the technical aspects of how this is achieved, see Zeileis and Hornik, 2007.) In other words, the presence of splits indicates that parameters of the theory-driven initial definition are unstable and that the data is too heterogeneous to be thus explained. The model does not describe the entire dataset.

For example, in Figure 2, using data from a current study, we have modelled (in a simplified fashion using a linear regression model) the relationship between European parents' risk perception of online hazards (dependent variable), their level of 'digital skills' (independent variable) and the level of digital skills of their children. We tested the relevance of the covariates age (of child, coded in three groups), education level

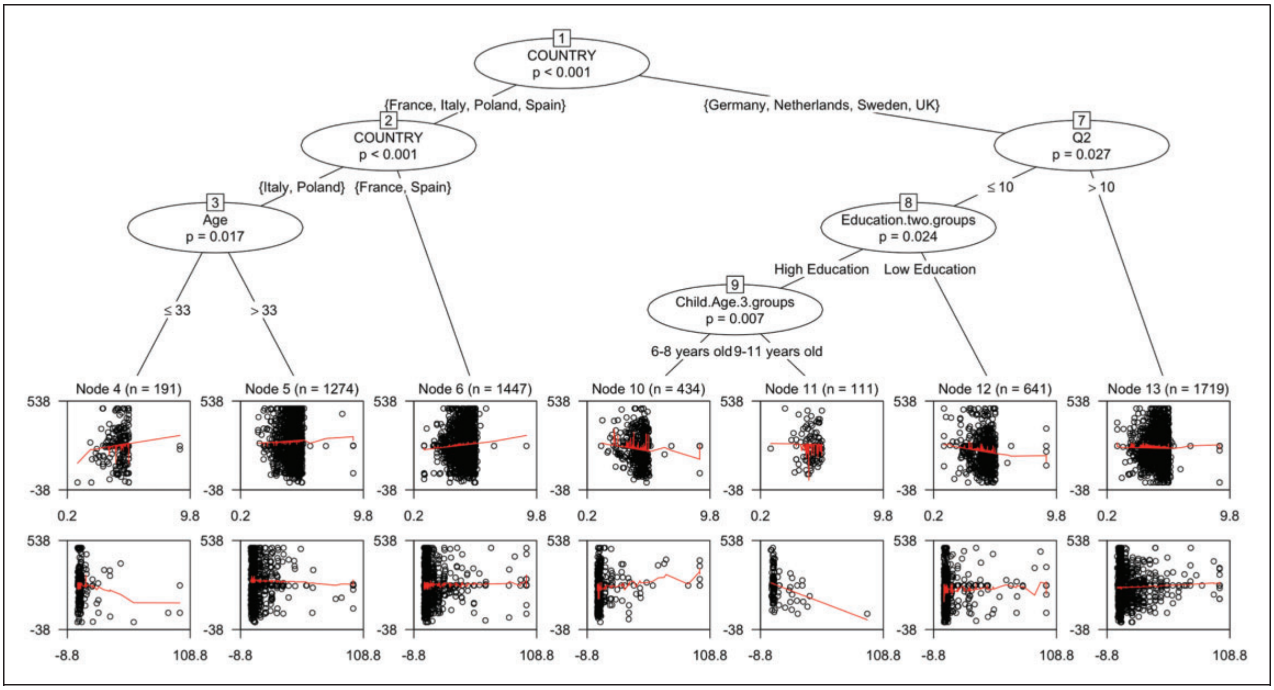

Figure 2. Example of model-based recursive partitioning of researcher's data.

(of parents), age of parents and country of residence. Such model is visualised using a regression tree in which the top part presents the splits according to different values of the covariates (for example, split 1 is done using the covariate 'country' resulting in France, Italy, Poland and Spain on one side and Germany, Netherlands, Sweden and UK on the other) while the bottom part represents the relationship dependent variable and two independent variables expressed in two scatterplots one after the other. The bottom part reports the 'local model' (local based on a subset of the original dataset) labelled as 'nodes': for example node 4 is the model expressed by parents from Italy and Poland with age below or equal to 33 years (this is the split value for the covariate age in this case). The model-based recursive partitioning finds different patterns of associations between the response variable and other covariates that have been pre-specified in the parametric model. In other words, it creates different versions of βs the model in terms of estimation, depending on different important values of covariates. Figure 2 shows how the relationship between the dependent variable and two independent variables does change considerably (in statistically significant matter, as indicated by the p values) in the different splits of the sample of respondents ($N = 6400$). Clearly, the initial model is insufficient to explain such relationship without taking some of these covariates into consideration. For example, if the relationship between the dependent variable and each independent variables reverse of sign for some partitions of the datasets represented by subgroups, a one-size-fits-all approach is highly problematic. Figure 2 shows how different covariates produce splits in the two groups of countries. Node 11 is an example of a different 'local' model of the relationship between parents' risk perception, their digital skills and those of their children for Northern Europeans parents with higher education and with 9–11 years old children compared to the rest.

What is the advantage of having such information? The answer to this question refers to the initial distinction that was introduced about the two cultures of modelling. In the predominant (in social sciences) data modelling culture, the comparison between different models has always been difficult and a problematic point. The hybrid approach of model-based recursive partitioning modelling can help revise models that work for the full dataset and that do not neglect such information imposing on models, as 'global' strait jackets. Moreover, if the researcher in question values the working rule of Ockham's Razor (that a model should be no more complex than necessary but needs to be complex enough to describe the empirical data), model-based recursive partitioning can be used for evaluating different models. One more useful item of information, generated by this approach, is that the model-based recursive method allows the identification of particular segments of the sample under examination that might be worth further investigation.

Conclusion

Interestingly enough, Big Data has been criticised for being merely descriptive, often in the forms of correlational studies. While such critique is sometimes well founded, it does not take into consideration the difficulty in achieving 'causality' in social science. It has a long history to the point that some social scientists, for different reasons, call for abandoning the focus on causality to embrace an interest for descriptive and classification (Abbott, 2001; Bowker and Star, 2008; Latour, 2005; Pickstone, 2001). However, if we are not to endorse such a position, the critique of merely descriptive approaches is unfounded because it ignores the advances in analytical methods that are useful for theory development and model selections.

The contribution of Big Data to social science is not limited to data availability but includes the opening to analytical approaches that have been developed in computer science, and in particular in machine learning, which brings a new 'culture' of statistical modelling that bears considerable potential for the social scientist. This is illustrated in this brief discussion about model-based recursive partitioning. Computational methods, especially in the form of simulation, are not in conflict with theories' development but provide an opportunity to improve such process (Smith and Conrey, 2007). Rather than the availability of data, the family of innovations that are emerging under the umbrella term of Big Data should be considered in light of the current constraints and limitations in analytical methods, as much as in data collection issues.

For those interested in further readings, the reference book is 'Computer age statistical inference: algorithms, evidence and data science' by Efron and Hastie (2016). Both Alvarez (2016) and Cioffi-Revilla (2013) books are excellent introduction to computational social science. Last but not least, the edited volume by Gonçalves and Perra (2015) presents a wide range of computational methods applied to social phenomena.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Abbott AD (2001) *Time Matters: On Theory and Method*. Chicago: University of Chicago Press.

Alvarez RM (2016) *Computational Social Science*. Cambridge: Cambridge University Press.

Barabasi A-L and Posfai M (2016) *Network Science*. Cambridge: Cambridge University Press.

Berk RA (2006) An introduction to ensemble methods for data analysis. *Sociological Methods & Research* 34(3): 263–295.

Bowker GC and Star SL (2008) *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

Boyd D and Crawford K (2012) Critical questions for Big Data. *Information, Communication & Society* 15(5): 662–679.

Breiman L (2001) Statistical modeling: The two cultures. *Statistical Science* 16(3): 199–231.

Burrows R and Savage M (2014) After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society* 1(1): 1–6.

Cioffi-Revilla C (2013) *Introduction to Computational Social Science: Principles and Applications*. London: Springer.

Curti M, Iatus S, Porro G, et al. (2015) Measuring social well being in the Big Data era: Asking or listening? *arXiv.org*, cs.CY. Available at: http://arxiv.org/abs/1512.07271.

Efron B and Hastie T (2016) *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge: Cambridge University Press.

Gentle JE, Härdle WK and Mori Y (2012) *Handbook of Computational Statistics*, Heidelberg: Springer, p. 1192. Available at: http://books.google.co.uk/books?id=28S6jgEACAAJ&dq=intitle:Handbook+of+Computational+Statistics+concepts+and+methods+pdf&hl=&cd=1&source=gbs_api.

Gonçalves B and Perra N (2015) *Social Phenomena: From Data Analysis to Models*. Cham: Springer.

Hand DJ and Vinciotti V (2003) Local versus global models for classification problems. *The American Statistician* 57(2): 124–131.

Kahneman D and Tversky A (2000) *Choices, Values, and Frames*. Cambridge: Cambridge University Press.

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12.

Latour B (2005) *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: New York: Oxford University Press.

Lazer D, Pentland A, Adamic L, et al. (2009) Computational social science. *Science* 323(6): 721–723.

Pickstone JV (2001) *Ways of Knowing: A New History of Science, Technology, and Medicine*. Chicago: University of Chicago Press.

Podsakoff PM, MacKenzie SB, Lee JY, et al. (2003) Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology* 88(5): 879–903.

Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.

Schroeder R (2014) Big Data and the brave new world of social media research. *Big Data & Society* 1(2): 1–11.

Smith ER and Conrey FR (2007) Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review* 11(1): 87–104.

Tinati R et al. (2014) Big Data: Methodological challenges and approaches for sociological analysis. *Sociology* 48(4): 663–681. Available at: http://soc.sagepub.com/cgi/doi/10.1177/0038038513511561.

Zeileis A and Hornik K (2007) Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica* 61(4): 488–508. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2007.00371.x/full.

Zeileis A, Hothorn T and Hornik K (2008) Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17(2): 492–514.