Challenges for Biodiversity Science in the Era of Big Data

Jorge Soberón

Abstract

The extremely fast growing amounts of 'big data' made available to modern scientists, have consequences for particularly taxonomists, phylogeneticists, biogeographers and ecologists. Therefore, new machine learning algorithms designed for pattern recognition are now common, leading to an ultimately 'black box' model of science. While it is necessary to go ahead with this development in order to detect patterns in the increasing amount of 'big data', conventional theoretical analyses of the problems will still be indispensable for interpretation and to establish the limits of reliable forecasting – in other words, to enhance understanding.

Key Words: CONABIO, species distribution modelling, theoretical understanding

Jorge Soberón, Biodiversity Institute and Department of Ecology and Evolutionary Biology, University of Kansas, 1345 Jayhawk Boulevard, Lawrence, KS 66045, USA. E-mail: jsoberon@ku.edu

SCI. DAN. B. 6 · TROPICAL PLANT COLLECTIONS
THE ROYAL DANISH ACADEMY OF SCIENCES AND LETTERS · 2017

The term 'big data' has gained much popularity in recent years, as the extremely large and almost instantaneously updated bodies of data on banks, airlines, social networks, astronomy, elementary particles, and others are becoming accessible through the Internet, at the rate of Exabytes per day (McAfee *et al.* 2012). Big data is thus a label applied to bodies of knowledge that are digitized, so large that they require distributed or specialized storage facilities, are updated very frequently, and require special methods to be analysed. In biology, leaving aside the already vast holdings of the molecular communities (Stephens *et al.* 2015), the largest data repositories are the museum and herbarium holdings, that contain a few thousands of millions of records (Edwards 2004), and the holdings of observations recorded by amateurs (Dickinson *et al.* 2012). Although these data sources are 'big' in some sense, of the order of Terabytes, and growing reasonably fast [Megabytes per day; Soberón & Peterson 2004], they fall short of the spectacular volume and speed of change of other fields.

In view of the above, it is probably premature to assign the label of 'big data,' sensu stricto, to biodiversity data. However, it is undeniable that the organismic disciplines of biology, like taxonomy, systematics, ecology, biogeography and other similar ones, are experiencing an explosion on the amount of data digitally available (Kelling *et al.* 2009) and new technologies will increase this rate even more (Hampton *et al.* 2013). This creates challenges for our disciplines. In what follows I will explore two of them: the growing utilization of 'machine learning' methods that emphasize extracting patterns from large quantities of data, and the related question of whether that would imply that we are moving towards a different form of thinking and theorizing in organismic biology.

Machine Learning Methods in the Biodiversity Disciplines

The very large quantity of data available now, in the form of taxonomic authority files, occurrence data,

very large phylogenetic trees, sequence data, and soon image data, are difficult to organize, search, visualize and analyze without using very advanced computational methods. The sheer volume of data together with the complexity of the non-parametric algorithms that are now available (neural networks, maximum entropy, decision trees, genetic algorithms, etc.) indeed suggests that a black box approach to science, whereby patterns are found by software, and then applied to prediction without much human intervention will become the rule. For instance, simply by sequencing huge volumes of genetic materials found in seawater samples, Venter *et al.* (2004) found 148 unknown bacterial phylotypes as well as a very large amount of variation on rhodopsin receptors. Based on this type of research some have claimed that "the data deluge makes the scientific method obsolete" (Anderson 2008).

Closer perhaps to organismic biology, the area of species distribution modelling makes full use of the hundreds of Gigabytes of occurrence data, as well as advanced and complex algorithms that extract patterns from the databases. For instance, the GARP (Genetic Algorithm for Rule Production) algorithm has been used as the engine of Lifemapper (Stockwell *et al.* 2006) to create a library of hundreds of thousands of unsupervised species distributions models. GARP is the ultimate 'black box' algorithm in the sense that it outputs long lists obtained by establishing a stochastic competition among different modelling algorithms (Stockwell 2007). The output is a set of rules that describes the pattern, but not all implementations of GARP provide access to the rules! Lifemapper later substituted Maxent (Phillips *et al.* 2006) for GARP, and although Maxent is well described theoretically, and calculates a relatively simple and well defined object (a Gibbs distribution, see Merow *et al.* 2013), it is clear that many, if not most of its users, simply "... fail to interpret the original algorithms, much less understand how they were implemented in the ... code" (Joppa *et al.* 2013).

Is the big data, 'machine learning' approach to science a new paradigm for the biodiversity disciplines? Kelling *et al.* (2009) believe this is the case, arguing that the complexity of ecological systems make difficult posing and testing hypothesis using parametric statistics. They describe an alternative methodology where big datasets are analyzed using sophisticated software, patterns are found, and then the patterns are tested in confirmatory analysis. Leaving aside problems with the uneven quality of big data data per se, and most relevantly, its biased (in time, space, and taxa) nature (Soberón *et al.* 2007; Engemann *et al.* 2015), and the need to deal with such biases using theoretical tools, the communities engaged in the biodiversity disciplines need to ponder in a serious way how it is that large quantities of digitally available data are changing the way we manage the relationship between the data provided by our instruments and senses, and the models, concepts, and theories that we use to describe, predict and understand the phenomena represented by the data. In the last section I will provide some reflections on this problem.

The Role of Theory in the Data-rich Disciplines

There is no doubt that science advances by a continuous interplay between observations of phenomena and our conceptual representations of them. This is well illustrated by a famous, and contradictory (Ayala 2009), pair of statements of Darwin: "I am turned into a kind of machine for grinding general laws out of large collections of facts" (Darwin, in Barlow 1958). This would be the 'machine learning' paradigm, whereby big data is grinded into patterns. However, the same Charles Darwin also stated: "How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service!" (Darwin, letter to H. Fawcett, 18 Sept 1861, in Darwin Correspondence Project, University of Cambridge https://www.darwinproject.ac.uk/), and this second statement corresponds to the conventional scientific paradigm, of contrasting hypotheses and models against data, for the purpose of understanding (Pigliucci 2009), as well as for the purpose of pattern discovering. Moreover, without understanding, in some theoretical sense (that goes beyond the mere

description of the patterns), there is no trust in prediction, another essential scientific objective.

This point is well illustrated by using an example of the Mexican Biodiversity Commission. When CONABIO started, its mandate was to create a database of the Mexican biodiversity. This lead to a large scale effort to digitize the data in national, and foreign, scientific collections. One of the first questions that the Mexican government wanted to address was how to identify suitable regions in which to create protected areas, and this in turn required them to be able to compile species' lists of arbitrary regions. Since large parts of the country are still unexplored, short of a full scale – and hugely expensive – field explorations program, the databases of specimens could be used to find patterns of high number of endemic species, and extrapolate them to unexplored regions. As soon as the specimen data started coming in large numbers (big data in relative terms), this question could be explored using Species Distribution Modeling (SDM). By 1994, CONABIO was resorting to the software called GARP, which at that time ran in the San Diego Supercomputing Centre, in California. One by one species were modeled, and although generally speaking the models made sense to the eyes of experts, they tended to overpredict, in the sense that the software often highlighted areas where the species had never been observed as part of the area of distribution. This problem was not due to lack of data. Even on a 'gedankenexperiment' with a perfect set of occurrence data, GARP (and many other algorithms) would still overpredict, but it was only later, thanks to theoretical understanding, that it became clear that the problem of overprediction was not really a problem. Theoretical understanding allowed scientists to realize that correlative species distribution algorithms (software quintessentially for pattern recognition) model something intermediate between an actual area of distribution and a potential area of distribution (Soberón 2010), and that the overprediction was actually very useful for the purpose of assessing potential impacts of invasive species, or any other species out of dispersal equilibrium (Peterson *et al.* 2011).

The morale that I would like to extract from this story is that, although the pattern looking exercise, based on large quantities of data and complicated software was indeed useful, and only possible in the era of large quantities of digitally available data and powerful software, the full comprehension, correct interpretation of the results, and an awareness of the possibilities and limits of extrapolation was the result of a conventional theoretical analysis of the problem.

Conclusion

In organismic biology we are now fully immersed in an era of exploding growth of digitally available data. This is a novel and exciting area for the biodiversity disciplines, one that will enable both fundamental discoveries and useful applications. However for the biodiversity disciplines it would be a serious mistake to accept the simple 'pattern discovery' paradigm that is so useful in commerce, banking, and other similar activities. Science is, at its core, about not only describing and predicting, but also about understanding. Leaving aside philosophical discussions, history shows that the most interesting and deep scientific advances are associated to an attempt to understand, in some of various senses, the patterns that are discovered, or why predictions are successful. This should not be left to machines or to algorithms. In the biodiversity disciplines, theoretical developments are more necessary than ever.

Acknowledgements

I am very grateful to Ib Friis and Henrik Balslev for their kindness and efficiency as organizers of this symposium, and for their patience with very late authors. I was partially supported by NSF grant 1208472.

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired* 16 -07. http://www.wired.com/2008/06/pb-theory/ (accessed April, 2016).

Ayala, F.J. (2009). Darwin and the Scientific Method. *Proceedings of the National Acadademy of Sciences USA* 106: 10033-10039.

Barlow, N. (ed.) (1958). *The Autobiography of Charles Darwin 1809-1882. With the Original Omissions Restored.* Collins, London.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment* 10: 291-297.

Edwards, J. (2004). Research and societal benefits of the Global Biodiversity Information Facility. *BioScience* 54: 485-486.

Engemann, K., Enquist, B.J., Sandel, B., Boyle, B., Jørgensen, P.M., Morueta-Holme, N., Peet, R.K., Violle, C. & Svenning, J.C. (2015). Limited sampling hampers 'big data' estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution* 5: 807-820.

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S. & Porter, J.H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11: 156-162.

Joppa, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D. & Emmott, S. (2013). Troubling trends in scientific software use. *Science* 340: 814-815.

Kelling, S., Hochachka, W.M., Fink, D., Riedewald, M., Caruana, R., Ballard, G. & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. *BioScience* 59: 613-620.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D. & Barton D. (2012). Big data. The management revolution. *Harvard Business Rev.* 90: 61-67.

Merow, C., Smith, M.J. & Silander, J. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography* 36: 1058-1069.

Peterson, A.T., Soberón, J., Pearson, R.G. Anderson, R.P. Enrique Martínez-Meyer, E., Miguel Nakamura, M. & Araújo, M.B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press, Princeton and Oxford.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190: 231-259.

Pigliucci, M. (2009). The end of theory in science? *EMBO Reports* 10: 534-534.

Soberón, J. (2010). Niche and area of distribution modeling: A population ecology perspective. *Ecography* 33: 1-9.

Soberón, J., Jiménez, R., Golubov, J. & Koleff, P. (2007). Assessing completeness of biodiversity databases at different spatial scales. *Ecography* 30: 152-160.

Soberón, J. & Peterson, A.T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B* 35: 689-698.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S. & Robinson, G.E. (2015). Big data: Astronomical or genomical? *PLoS Biology* 13: e1002195.

Stockwell, D.R. (2007). *Niche Modeling: Predictions from Statistical Distributions*. Chapman & Hall/CRC, London.

Stockwell, D.R.B., Beach, J.H., Stewart, A., Vorontsov, G., Vieglais, D. & Pereira, R.S. (2006). The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. *Ecological Modelling* 195: 139-145.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E. & Nelson, W. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304: 66-74.