Received: 23 December 2019 | Revised: 24 December 2019 | Accepted: 24 December 2019

DOI: 10.1002/asmb.2506

DISCUSSION PAPER

WILEY

Discussion of "A review of data science in business and industry and a future view," by Grazia Vicario and Shirley Coleman

Piercesare Secchi[1,2]

[1]MOX—Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

[2]Center for Analysis Decisions and Society, Human Technopole, Milan, Italy

Correspondence
Piercesare Secchi, Dipartimento di Matematica, Politecnico di Milano, Via Bonardi 9, 20133 Milan, Italy.
Email: piercesare.secchi@polimi.it

First, I want to congratulate the authors for this very informative review on the state of affairs between the newly emerging Data Science and its applications in business and industry. Stimulated by it, in the following lines, I will put in my two cents and comment on a very few issues raised by the article.

The origin of the name notwithstanding, when Data Science enters a narrative today, the focal point is invariably the first term: data. Data are abstract—and now digital—representations of reality. Whether they are vectors, curves, operators, networks, images, texts, and so on, or any combination of the former, data are the final atoms of the analysis and they are eventually treated as points of an abstract mathematical space. It is the cloud of points representing data in this space that forms the landscape explored by the data scientist in search for informative patterns. Contrary to the cited Chris Anderson's opinion,[1,2] even when the cloud is massive and conforms to the many V's required today to be worth of interest, I would suggest the data scientist starting the mighty endeavor of its exploration to be armed with something more than correlation. If nothing else, the data scientist will soon discover that correlation is not a property of data, independent of the geometry of the mathematical space where the scientist embeds them, or invariant with respect to the coordinate system set by the scientist to frame the cloud and move around in it.

Stronger and more reliable tools than correlation are offered by the second, and neglected, term in Data Science: science, the systematic and organized knowledge held by the data scientist. This is especially true in the applications of Data Science to business and industry, where the moving force driving the analysis is the need to solve a practical problem and not necessarily to expand our knowledge of the universe. The problem comes first, and the first responsibility of the data scientist is to reformulate it in terms of formal questions apt to be explored and answered leveraging the accessible data and the conceivable models, methods, and algorithms offered by science, sub specie mathematicae. Following the object-oriented data analysis approach, first introduced by Wang and Marron[3] and more recently overviewed by Marron and Alonso,[4] the atoms of the statistical analysis (vectors, curves, operators, networks, etc.) are indivisible objects, which should be modeled as points of a mathematical space whose dimensionality, topology, and geometrical properties must not neglect the data complexity in the face of the goals of the analysis. This modeling effort is directed by the competent scientific knowledge of the specific domain that frames the problem and not driven exclusively by the data. Hence, I am in favor of the terminology Data Sciences advanced by Paolo Giudici[5] since, by articulating the second term as a plural, it clearly indicates that Data Science is not a singular new scientific discipline but a novel approach for the exploitation of the massive, heterogeneous, and complex data—made available by the digital revolution that generated the Information Age—within different scientific domains, each one "having different objectives and different languages" as aptly reported by Vicario and Coleman.[1]

Hence, my vision of a data scientist is not that of a know-it-all transposable nerd armed with computer power and passepartout data-driven algorithms—in the best of cases, the T-shaped version of the data scientist—but a picture of a host of data scientists, each one possessing diverse skills and competences and acting in a different business or industry domain—the Π-shaped data scientists. Here the horizontal edge of the Π indicates collaborative ability combined with two domains of deep vertical expertise; the first leg, common to all data scientists, being the mathematics of learning and the digital technology to implement it, and the second leg deeply expanding in the specific area where data science is applied, being it within the natural or the social sciences, or within specific science-based technical disciplines, like engineering or medicine. Moving the metaphor from Π to a wicket should be pretty obvious to the casual reader.

Finally, coming closer to my expertise in applications of data science to business and industry, I see two themes emerging, which are worth to mention as complements to those illustrated by Vicario and Coleman.[1]

The first concerns new methods and algorithms for the analysis of dependence when data are complex and high dimensional. Industries and public agencies capture and store data with a definite reference in space or time. These data may come from sensors tracking objects moving over a urban network, like mobile phones, or recording the change over time of a vector of indicators observed in different locations of a textured and topologically complex space domain, like the composition of chemical species measured in an estuarine system or in a large-scale ground-water body. Additive manufacturing, which is feeding the fourth industrial revolution, requires real-time monitoring of parts represented as shapes and identified as space-dependent data points of a manifold. Data observed over time on complex manifolds, as the cerebral cortex of the brain, are becoming more common in medical research due to the advent of new diagnostic devices. The analysis of these space-time complex data cannot be conducted leveraging global models that assume some sort of stationarity, a Euclidean spatial domain, and a simple linear space for data embedding. New methods and algorithms are required to address these GeoData-driven problems, in areas of applied sciences well beyond the classical fields of application of geostatistics. Object-oriented spatial statistics[6,7] is a recent system of ideas, which provides a solid framework where these new and revolutionary challenges can be faced by grounding the analysis on a powerful geometrical and topological approach.

The second theme I want to mention is that of data fusion and integration. Personalized medicine requires the integration of large-scale genomic data with lifestyle and medical data captured and recorded by the health system. Unstructured granular data from administrative records and private sources can be integrated and managed to study the structure of the economy and the impact of policy decisions. Natural risks prevention, and the design of precision policies for mitigation and containment of vulnerability, demand the integrated analysis of a host of heterogeneous data, generated by different public agencies, having different space and time references and recorded with different degrees of uncertainty: census and demographic data, seismic hazard estimates based on historical data, microzonation data, flood hazard assessments based on mathematical models of basins, fragility curves and indicators for buildings based on their dimensions, age and construction material, indicators of social and material vulnerability, and so on. Developing new statistical models and algorithms for integrating different sources of data, exploring their joint and individual variability, and quantifying the uncertainty of predictions and inferences generated by their analysis are part of the mission of the Center for Analysis, Decisions and Society of the Human Technopole (https://www.humantechnopole.it/en/home), the large-scale research infrastructure financed by the Italian Government and under development at the former Expo site in Milan, now called MIND. Human Technopole will advance personalized approaches, both in the medical and the nutritional fields, aimed to fight cancer and neurodegenerative diseases by integrating large-scale genomics with the analysis of complex data systems and the development of new diagnostics techniques.

ORCID

Piercesare Secchi https://orcid.org/0000-0003-4048-9552

REFERENCES

1. Vicario G, Coleman S. A review of data science in business and industry and a future view. *App Stochastic Models Bus Ind*. 2019. doi:https://doi.org/10.1002asmb.2488.
2. Anderson C. The end of theory: the data deluge makes the scientific method obsolete. WIRED Magazine. 2008. https://www.wired.com/2008/06/pb-theory/. Accessed December 22, 2019.
3. Wang H, Marron JS. Object oriented data analysis: sets of trees. *Ann Stat*. 2007;35(5):1849-1873.
4. Marron JS, Alonso AM. Overview of object oriented data analysis. *Biom J*. 2014;56(5):732-753.
5. Giudici P. Financial data science. *Stat. Probab. Lett.* 2018;136:160-164.

6. Menafoglio A, Secchi P. Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *Eur J Oper Res*. 2017;258(2):401-410.

7. Menafoglio A, Secchi P. O2S2: A new venue for computational geostatistics. *Appl Comput Geosci*. 2019;2. https://doi.org/10.1016/j.acags.2019.100007.

How to cite this article: Secchi P. Discussion of "A review of data science in business and industry and a future view," by Grazia Vicario and Shirley Coleman. *Appl Stochastic Models Bus Ind*. 2020;1–3. https://doi.org/10.1002/asmb.2506