Environmental Modelling & Software 106 (2018) 4–12

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Environmental Data Science

Karina Gibert [a, *], Jeffery S. Horsburgh [b], Ioannis N. Athanasiadis [c], Geoff Holmes [d]

[a] Dep. Statistics and Operations Research, Knowledge Engineering and Machine Learning group at Intelligent Data Science and Artificial Intelligence Research Center (KEMLG at IDEAI), Research Institute on Science and Technology for Sustainability, Universitat Politècnica de Catalunya-BarcelonaTech, Spain

[b] Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, 84322-8200, USA

[c] Information Technology Group, Wageningen University, The Netherlands

[d] Department of Computer Science, University of Waikato, New Zealand

ARTICLE INFO

Article history:
Received 14 February 2018
Received in revised form
11 April 2018
Accepted 24 April 2018

Keywords:
Data Science
Environmental science
Data driven modelling

ABSTRACT

Environmental data are growing in complexity, size, and resolution. Addressing the types of large, multidisciplinary problems faced by today's environmental scientists requires the ability to leverage available data and information to inform decision making. Successfully synthesizing heterogeneous data from multiple sources to support holistic analyses and extraction of new knowledge requires application of Data Science. In this paper, we present the origins and a brief history of Data Science. We revisit prior efforts to define Data Science and provide a more modern, working definition. We describe the new professional profile of a data scientist and new and emerging applications of Data Science within Environmental Sciences. We conclude with a discussion of current challenges for Environmental Data Science and suggest a path forward.

1. Introduction

"*Data Science is the science of dealing with data …*" (Naur, 1974)

In recent years, we have observed an increasing popularity of Data Science methods that seem to be in the focus of many organizations, including those interested in a better comprehension or management of environmental systems. Data Science is already widely used in business to design successful strategies and policies, and the economic sector is facing a significant transformation as a result of the penetration of data-driven innovation in the business core. We believe that a similar transformation is underway within many scientific disciplines, among them those within the Environmental Sciences, to investigate the benefits that can be realized through use of appropriate Data Science approaches.

In this paper, we analyze the origins of Data Science as a new discipline that is diverse enough to be applied to any domain, including those within the Environmental Sciences. The potential of Data Science to advance our knowledge of the laws governing complex environmental phenomena is enormous. The technological development requisite for collecting the volume and resolution of data required to study these phenomena is mature, but classical data analysis methods are, in many cases, insufficient to cope with the size, speed and diversity of information sources providing evidence under the variety of forms (text, videos, audio recordings, numbers, images) that require global analysis and local tuning to elicit the hidden, relevant knowledge to support higher level decision making. Many investigators are already investigating how Data Science can address this deficiency.

We present the contributions of Data Science, together with an analysis of the new, specific skills associated with its inherent multidisciplinarity. As there is no common definition of Data Science, in the paper we present several definitions that have been used in the past and a propose a new conceptualization of what Data Science means. A discussion is also provided regarding its contact points with other emerging disciplines, such as Big Data Analytics. Emerging opportunities for new applications in Environmental Sciences are described. While not an exhaustive description of the opportunities for Data Science in Environmental Science applications, a wide perspective in the area is provided. Being an emergent field, a number of open issues envisage fertile areas for new research in the near future. The paper also provides some highlights, challenges, and trends with the aim to push the development of the Data Science field in general, and in

* Corresponding author.
E-mail address: karina.gibert@upc.edu (K. Gibert).

https://doi.org/10.1016/j.envsoft.2018.04.005
1364-8152/

Environmental Sciences in particular, where it can be of help.

The structure of the paper is as follows: In Section 2, the origins and a brief history of Data Science are provided. In Section 3, the added value of applying Data Science techniques to real problems using real data is discussed. Section 4 highlights the new skills required to become a qualified data scientist and the need to develop specific new curricula to provide appropriate training. Section 5 provides a more modern, contemporary view of Data Science, and Section 6 provides a general overview of how Data Science is being applied in Environmental Sciences. Section 7 identifies the main current challenges in the area. Section 8 provides a concluding discussion.

2. Origins and a brief history

Although Data Science is a relatively new discipline, the term Data Science is much older than might be expected. It is worth noting that there is no clear and agreed upon definition of the term Data Science. This lack of clarity appears in the first use of the term by Naur in 1960 (Sundaresan, 2017). Naur used the term to mean "data processing" in the computer science sense. However, it has also been used at times as a substitute name for the field of Statistics or, at the very least, Applied Statistics. Naur refined his earlier definition to: "Data science is the science of dealing with data, […] while the relation of data to what they represent is delegated to other fields and sciences […]" (Naur, 1974).

In the same period, in the context of statistical sciences, there was also a process by which data became the center of interest of the discipline. Indeed, John W. Tukey (1962) had already envisaged the need for statistics to move its focus from inference to data analysis as an empirical science: "For a long time I thought I was a statistician, interested in inferences from the particular to the general. But […] I have come to feel that my central interest is in data analysis […] intrinsically an empirical science." The development of computer science near that time was opening an opportunity to this end. In the late 1970s, Tukey (1977) published Exploratory Data Analysis, promoting a new approach to statistics where "more emphasis needs to be placed on using data to suggest hypotheses to test […] Exploratory Data Analysis and Confirmatory Data Analysis can—and should—proceed side by side".

In 1977, the International Association for Statistical Computing (IASC, http://iasc-isi.org/about-iasc2/) was established as a section of the International Statistical Institute: "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge" (IASC, 1977; Rizzi and Vichi, 2006; Davenport and Dyche, 2013). In 1996, the International Federation of Classification Societies (IFCS) used, for the first time, the term Data Science in the title of their biennial conference ("Data science, classification, and related methods"). Aligned with this approach, Jeff Wu seems to have been the first to ask whether Statistics should change its name to Data Science in his talk entitled "Statistics = Data Science?," which was given first in November 1997 as the inaugural lecture for his appointment to the H. C. Carver Professorship at the University of Michigan. In 1998, this was his first P. C. Mahalanobis Memorial Lecture, in honor of Professor Mahalanobis, the founder of the Indian International Statistical Institute (IISI), and was archived by Wu (1999). In 2001, William S. Cleveland called for establishing Data Science as a field "to enlarge the major areas […] of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called 'data science.'" Cleveland put the proposed new discipline in the context of computer science and the contemporary work in data mining. One and two years later, the first journals in the area were launched: The Data Science Journal and The Journal of Data Science, respectively. These events are largely why the term Data Science is currently understood by many people to be closely related to Data Mining and Big Data Analytics rather than the original sense in which the term was used.

Data Science has also been approached from the perspectives of Artificial Intelligence (AI) and Machine Learning. In the early 1980s, there was a clear idea of the importance of using data as the main source of knowledge extraction. In 1985, Douglas Fisher and Bill Gale founded the Artificial Intelligence and Statistics society http://www.aistats.org/past.html with the aim of facilitating interactions between researchers in AI and Statistics. Nearly ten years later, Cheeseman and Oldford stated, "We feel that there is great potential for development at the intersection of Artificial Intelligence, Computational Science and Statistics" (Cheeseman and Oldford, 1994). In 1989, Gregory Piatetsky-Shapiro organized the first Knowledge Discovery in Databases (KDD) workshop as part of the International Joint Conferences on Artificial Intelligence (IJCAI) world conference. It soon (1995) became an independent series of conferences (ACM SIGKDD). Fayyad et al. (1996) edited the seminal book "Advances in Knowledge Discovery and Data Mining," introducing new techniques and tools for the discovery of knowledge from data as a response to the urgent need to address "data flooding." They defined the Knowledge Discovery from Databases (KDD) framework as the process of the non-trivial identifying of valid, novel, potentially useful, ultimately understandable patterns in data." In the KDD approach, Data Mining was considered a specific data exploitation step. One year later, the first international journal, Data mining and Knowledge Discovery, was launched.

Around the mid-1990s, Data Science started to be seen as a new business opportunity. At that time, most companies were aware of having large volumes of collected data that were not properly analyzed (Berry, 1994). In many current contexts, Data Science can be understood from a business perspective as the process of discovering what we do not know from data. It enables us to get predictive, actionable insight from data, creating data products with business impact, communicating relevant business from data, and building confidence in decisions that drive business value (Somohano, 2013).

More recently, data science has started to be seen as an enabler that has the potential to transform scientific inquiries. Mattmann (2013) identified algorithm integration and data stewardship as two components of data science that are essential for managing the data deluge in Earth and space sciences and other fields like physics and genomics. Mattmann described algorithm integration as including model integration in scientific workflows and interfacing with data repositories and infrastructures. Mattman also called for integrating data archival with data processing facilities and, in the same work, highlighted the diversity of science data that involve many formats, file types, and conventions. In fact, Data Science is often based on the analysis of datasets resulting from a previous conversion of videos, audio recordings, signals, data streams, or websites into sets of relevant and/or sufficient indicators by means of feature extraction techniques, thus finding the relationships between several sources of heterogeneous data together and identifying complex, hidden patterns useful for decision support.

In the last several years, data science has been challenged to make the next steps in science by enabling in-silico scientific discoveries from vast amounts of data, where computers are enabled to identify and prove hypotheses not constructed by scientists. For example, Agarwal and Dhar (2014) describe the explosion of opportunities for scientific inquiry with readily available, large, and complex datasets and suggest that computers are now powerful enough to not only verify hypotheses but also to suggest new theories. Though such claims may seem ambitious, advances in machine learning, artificial intelligence, data integration, and