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

stewardship seem very promising. Similarly, Caffo et al. (2016) warn that the Big Data/Data Science hype will flame out if it is only about "data" and not "science." They argue that Data Science is only useful when the data are used to answer a question, expressing views similar to those that have been expressed by others that there are limits to what can be accomplished using Data Science methods and tools and that a balance must be struck between newer and more traditional scientific methods. Other authors (Lauro et al., 2017) characterize Data Science as the process by which data are transformed into actionable knowledge to perform predictions as well to support and validate decisions. Lauro uses a metaphor such that "Computer Science represents the language of Data Science, Statistics the logics of Data Science, and domain expertise constitutes a catalytic element in the absence of which the transformation cannot be achieved".

3. The added value of Data Science

The development of Data Science promoted a new concept in decision-making in general (including business) where decisions are data-driven, and the added value to organizations (either institutions or companies) is not more technology, nor capital, but information, where data is considered to be a primary source of knowledge. This transformation is not restricted to business fields, and the value that Data Science processes can add to our understanding of complex phenomena is widely recognized.

Davenport claimed that, "*instead of competing on traditional factors, companies are beginning to employ statistical and quantitative analysis and predictive modeling as primary elements of competition*" (Davenport and Harris, 2007). The Climate Corporation, whose motto is "Data services for yield maximization," was acquired by Monsanto in 2013 for nearly $1 billion U.S. Monsanto, John Deere, and DuPont Pioneer are among many companies scrambling to help agricultural producers do more with their exploding volumes of data (Noyes, 2014), and the value of Data Science for agricultural applications is clear in the obvious market demand for these services.

In one specific example application that demonstrates this potential value, Elarab et al. (2015) coupled high resolution imagery from an unmanned aerial vehicle remote sensing platform with machine learning algorithms to estimate chlorophyll concentration in crops as an important biophysical parameter for use in precision agriculture. Their techniques enable farmers to assess the heterogeneity of the plants in their fields at fine resolution in space and time, aiding farmers in targeting management actions accordingly (e.g., effectively targeting application of fertilizers or water only where they are needed). The economic and environmental implications of enabling this type of precision agriculture could be significant.

Although the value of informed decision making was understood in the 1950s (Luhn, 1958), after the emergence of Knowledge Discovery in Databases (Fayyad et al., 1996), informed decision making became more popular (Brynjolfsson et al., 2011). Several authors discuss the relevance of data for decisions. For example, Craig Mundi, the head of research and strategy at Microsoft stated, "Managed well, data can unlock new sources of economic value. We are in front of a nascent data-centered economy" (Cukier, 2010).

However, despite the excitement about the explosion in available data and Data Science methods, the reality is that most available data are under-exploited, resulting in a loss of potential for decision making. In 2004, Hammond formulated what he called "the Fact Gap: the disconnect between data and decisions" (Hammond, 2004) to refer to this phenomenon of low consumption of available data at the decision making level. Data Science is an emergent discipline that focuses on the intensive consumption of available data to extract decisional knowledge relevant for informed decision making and to bridge Hammond's Fact Gap. In the environmental sciences, United Nations Agenda 21 has already since 1992 included a section on "bridging the data gap," highlighting that "the gap in the availability, quality, coherence, standardization and accessibility of data between the developed and the developing world has been increasing, seriously impairing the capacities of countries to make informed decisions concerning environment and development" (UN, 1992).

4. Data scientist: a new professional profile

It seems clear that the skills required to perform Data Science point to a new professional profile and claim that academia start to design new curricula to train this new type of professional. Hal Varian, Google's chief economist, when interviewed by McKinsey referred to the "scarce ability to extract wisdom from [data]" (Cukier, 2010). In many situations, data scientists are expected to have a broad set of skills, eloquently defined by Josh Wills (2012): "A Data Scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician." Acquisition of the appropriate skills for effectively applying Data Science is critical in order to ensure a solvent extraction of the right value contained in data. Some authors have reported how, in the absence of proper training, several data scientists have extracted contradictory conclusions from a single dataset by performing different analytical procedures (Baeza-Yates, 2017; Silberzahn et al., 2015).

Davenport and Patil in their 2012 article in the Harvard Business Review provocatively entitled *Data Scientist: The Sexiest Job of the 21st Century* argue that skills required for being an effective data scientist go beyond statistical or analytical capabilities and should include storytelling with data and excitement with potential for breakthroughs in the particular domain. Other authors have also described the new profile of a data scientist and the specific skills required to do Data Science properly (e.g., Sooraij Shah, 2013).

Contemporarily, the shortage of talent with required skills for capturing the whole potential of data was reported by McKinsey in 2011 by quantifying a *shortage of 140,000 to 190,000 data scientists by 2018* (Manyika et al., 2011). In his keynote at the Campus Party Europe in September of 2013, A. S. Pentland, head of MediaLab Entrepreneurship MIT, said, "There are too few data scientists in the world, and education needs to change in order to maximize the true potential of data science" (Palmer, 2013). At that time, 62% of executives realized that the lack of data scientists was causing a real problem (One Poll Survey, for soft supplier Teradata (McKenna, 2013)). The shortage is especially severe in the U.S. For example, 80% of new data scientist jobs were not filled within the year 2010–2011 (Harris et al., 2014). In the U.S., the demand for Data Science and analytics jobs is projected to grow by 15% between 2015 and 2020, with the highest rate of growth rate of 28% expected for the specific job title of Data Scientist (Markow et al., 2017). The lack of data scientists that can do high quality work with available data contributes to the Fact Gap (Burns, 2017), and this shortage still persists at the time of this writing (Business.com, 2017).

In spite of the shortage of skilled professionals, Data Science has become a relevant profession in the last few years. In 2013, Venture Beat reported Data Science as the second best new job in America. In a more recent report, they show data scientist as the number one best job in America (VentureBeat, 2017), and Piatetski-Shapiro (2017) reported on KDnuggets that "*Glassdoor again ranked Data Scientist as the no. 1 job in USA, and 5 of the top 10 US jobs are related to Analytics, Big-data, and Data Science*," a ranking that was repeated by Glassdoor in 2018.

As a response to this reality, the last few years have seen

recommendations and guidance for integration of Data Science into degree programs (e.g., Association for Computing Machinery, 2017). Many universities have created qualifications in Data Science or in some of its related areas such as Data Analytics, Business Intelligence, and so on. These new programs tend to be associated with the Computer Science or Mathematics/Statistics disciplines. While Earth, Environmental, and Life science curricula have not caught up to those of Computer Science and Mathematics/Statistics in the area of Data Science, it is becoming more common to find specific courses or discipline specific focus areas on Data Science methods in additional scientific disciplines.

5. A modern view of Data Science

So far we have seen that there is no clear, agreed definition of the term Data Science, but the intrinsic multidisciplinary nature of the field seems to be clear. Conway's Data Science Venn Diagram (Fig. 1, Conway, 2013) provides a useful conceptualization for how coding skills (Conway calls them hacking skills), math and statistics knowledge, and domain science expertise (Conway calls this substantive expertise) come together to enable Data Science. Domain expertise combined with math and statistics knowledge is where most traditional research is conducted. Coding with math and statistics knowledge may lead to insight through machine learning using data, but without driving scientific questions and hypotheses that come through domain expertise, mechanistic or process understanding may be limited. Coding skills combined only with domain expertise may lead to incorrect interpretation of results without knowledge of math and statistics (the danger zone in Conway's diagram). It is only at the intersection between the three elements that Data Science can be most effective. Indeed, multidisciplinarity has become a main pillar of Data Science.

In recent applications, it is still possible to see the relationship between Data Science and Computer Science, Statistics, Data Mining, and Big Data Analytics. Essentially, Data Science is broader than these fields, but can make use of all of them. For example, Data Science does not necessarily concern itself with the size of the data being processed, but rather with transforming data into added value for the end user and extracting relevant knowledge from data coming from complex phenomena. Where data are prolific, Data Science uses techniques from Big Data Analytics to perform parts of its workflow. Similarly, the Data Mining process works well for the overall development of some Data Science applications. Robust statistical methods are needed in most Data Science applications, but the potential heterogeneity, complexity, and size of data can require innovative solutions from Computer Science.

Trying to precisely define the relationship between these fields is difficult, as their definitions and the boundaries between them are not altogether clear. Even the terms used to describe Data Science methods and the methods themselves are often confusing, as discussed in Section 2. From a practical perspective, we are very much in agreement with Lauro et al. (2017) that the novelty of Data Science is identified to be the role played by knowledge, which is definitely integrated into the process, including interpretation issues, with the main purpose to give *meaning* to data.

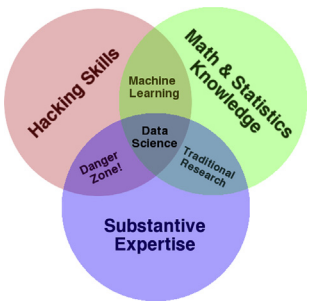

Fig. 1. The data science venn diagram (Conway, 2013).

In the current context, we *consider Data Science as the multidisciplinary field that combines data analysis with data processing methods and domain expertise, transforming data into understandable and actionable knowledge relevant for informed decision making*, thus contributing to bridge Hammond's Fact Gap. This modern view and its associated techniques and applications are enabling Data Scientists to synthesize value-added products from diverse datasets that would be otherwise impossible to obtain. Data Science often requires the ability to overcome data complexity and the limitations of classical statistics and machine learning techniques - for example dealing simultaneously with heterogeneous data sources (e.g., videos, text, or streams) or coping with non-independencies, non-normalities, and few technical hypothesis on variable's distributions, when required.

6. Data Science in Environmental Sciences

Much of our discussion up to this point has involved mathematics, statistics, computer science, and applications of Data Science in general, including impact in business. However, there is no shortage of opportunity for applications in Environmental Sciences (Gibert et al., 2008). Indeed, as the size and complexity of environmental datasets continue to grow, the demand for Data Science in environmental applications is increasing. Data Science is able to add value to Environmental Sciences in many different ways.

Data Science is at work in examining, interpreting, and deriving useful and actionable information from environmental sensor data streams (Athanasiadis and Mitkas, 2007; Reis et al., 2015; Bifet et al., 2017). These datasets, produced by *in situ* and remote sensors of many different types, span the environmental sciences from climate and weather observations from satellite and ground-based sensors (e.g., Hill et al., 2011) to air quality sensors (e.g., Wiemann et al., 2016), hydrologic and water quality sensors installed in aquatic environments (e.g., Wong and Kerkez, 2016), water quality sensors in wastewater treatment processes (e.g., Corominas et al., 2017), sensors used to track the movement and behaviors of biological organisms (e.g., Kranstauber et al., 2011), ground-based sensors for detecting and quantifying the magnitude of earthquakes and geological events, and many other applications. The size of these monitoring networks and the datasets they produce is growing as the cost of sensors and related systems is falling, producing a greater need for analysts capable of managing the datasets produced and assimilating them with simulation models and other applications. Sometimes, evolutionary and population-based algorithms are used to extract relevant parameters of a reference phenomenon based on these sensor data Afshar et al. (2015) describe an application to water resource management, Levasseur et al. (2008) uses genetic algorithms in soil applications and Nagesh Kumar et al. (2006) describe an application for optimal reservoir operation for irrigation of multiple crops.

Another quickly growing application is that of interpreting data collected from aerial drones, sailing or aquatic drones, and satellite remote sensing (Roelofsen et al., 2014; Elarab et al., 2015; Gauci et al., 2018). Remote sensing data of these types are becoming more and more vital for accurate and broad environmental monitoring. The scope of these applications is simply too broad for individual sensors or ground-based stations to be effective, but the volume of data produced by remote sensing drones and satellites can dwarf most other environmental datasets, requiring specialized tools, techniques, and training for data analysts.

Many new applications at the nexus between water and energy are generating high spatial and temporal resolution datasets using smart metering technologies that collect observations of water or power usage at frequencies on the order of seconds (or even more frequent) to support applications such as end use disaggregation, demand estimation, and demand management (Cominola et al., 2015; Gurung et al., 2015). Sophisticated algorithms and data management techniques are required for these applications, and without them, high resolution data can be an unnecessary barrier for water and power system managers who have traditionally used metering data for monthly billing purposes (Horsburgh et al., 2017).

Preparing datasets designed to support high-performance, large-scale (e.g., continental scale) modeling and analyses is another area Data Science is being used within the environmental sciences. These datasets are typically derived from existing geospatial data, but are organized over large spatial scales using sophisticated data modeling to provide the backbone for environmental models. Examples include the National Hydrography Dataset Plus (NHDPlus) being used as the hydrologic network underlying the continental-scale National Water Model in the U.S. Another example includes ongoing efforts by the United States Geological Survey to link aquatic monitoring sites and features to the network of streams in the U.S. (ELFIE - https://opengeospatial.github.io/ELFIE/json-ld/) using concepts from the Open Geospatial Consortium's HyFeatures specification (Dornblut and Atkinson, 2014). Data Science is needed in the design and preparation of these types of datasets to enable high resolution and high performance use.

Model outputs from high resolution, long temporal period climate simulations represent another class of environmental data requiring specialized skills for analysis. Analysts and modelers working with global or continental scale models are often faced with hundreds of terabytes of model-generated results that must be reduced to useful products that can be used for downstream analyses (Ashraf Vaghefi et al., 2017). The volume of data involved is challenging from not only the data use perspective, but also from the perspective of provisioning basic storage hardware and software required for short and longer term uses of these massive datasets. In many contexts, integrating data and domain knowledge for obtaining better models requires application of Data Science methods. In this sense, Uusitalo (2007) discusses the use of Bayesian Networks in environmental modelling. Blattenberger and Fowles (2017) have also used them to evaluate avalanche danger, and Gibert et al. (2010a) used prior knowledge to bias a clustering process in waste water treatment plant applications.

Combining disparate datasets from multiple scientific domains for synthesis studies and to create new, derived datasets can also be enabled using Data Science. Vitolo et al. (2015) provide a broad review of many of the techniques and technologies that have been applied in enabling data integration, particularly via the Internet. This is a problem of data fusion, where complexity and heterogeneity in data can be more challenging than size (Nativi et al., 2015). Often, combining datasets from different sources in a single analyses requires specialized skills in data management, programming, and visualization that are central to Data Science (as in Porter et al., 2014). Geospatial machine learning methods (e.g., Kanevski et al., 2008) are also helping in this area (McCord et al., 2017; Hengl et al., 2017).

Another emerging area in environmental Data Science involves linking human health to environmental conditions, especially in the cases of natural disasters such as hurricanes, flooding, or earthquakes (e.g., Klise et al., 2017). Major hurricane events seriously impact the availability of power, potable water, and other basic human needs. Relatively little is currently understood about the short and longer-term effects events like these have on human health because it is only in the past few years that more detailed datasets have been available for characterizing environmental conditions during and following events along with health impacts that can be attributed to these conditions. Exposure (Gómez-Losada et al., 2014; Schlink et al., 2016) is another new issue that requires attention. The impact of air pollution on public health is currently a focus of interest, and modelling air quality on the basis of data streams provided by smart sensors currently available in many cities is crucial to understand how exposure affects public health, but also to perform real time air pollution forecasting to run preventive and protective healthcare plans for citizens' health. Data Science linking environmental and human health data raises multiple challenges. First, human health data are sensitive, restricted, and must be anonymized. Second, the format, vocabularies, and syntax of these two types of data are very different and require the combination of different resources to deal with them simultaneously. Properly linking environmental and human health data requires unique and careful approaches for both data management and analysis.

7. Current challenges and trends in Environmental Data Science

While the use of Data Science techniques to enhance research in the Environmental Sciences is rapidly growing, it is not without significant challenges that must be overcome. In this section, we describe some of these challenges, and, while likely not exhaustive, the list we provide here clearly illustrates that there is much room for improvement.

Challenge 1: Shortage of trained Data Science experts to cover real demand: As mentioned, many authors identify a relevant shortage of properly trained data scientists to cover the real demand. New training programs are being deployed at different levels of academy to increase the number of formally trained data scientists. However, it is unlikely that the number of personnel formally trained as data scientists will be able to keep up with the growing demand from so many different disciplines in which they are now employed. From the Environmental Science perspective, it will remain difficult to attract the best and brightest to work on environmental problems when salaries for data scientists are so much higher elsewhere. This is common with Environmental Informatics, as already pointed out by Swayne (2003).

Challenge 2: Lack of Data Science skills within Environmental Science curricula: Applying Data Science approaches to Environmental Systems and data requires all three skills described by Conway (2013) - data programming and data modelling skills (in general, including statistical and machine learning approaches), and domain knowledge. It will continue to be difficult to find individuals that can effectively do all three with sufficient expertise. Collaborations between environmental scientists and computer scientists can potentially address this need, but are difficult to foster because the needs, interests, and expertise of collaborators are not always aligned. Even though curricula in Environmental Engineering and the Environmental Sciences are evolving to address this, they have not totally caught up with the need yet, and sustained efforts are required to resolve this deficiency. Developing resources within open environmental data repositories accompanied with problem descriptions can provide materials for introducing realistic practicums in education that helps in better training new scientists.

Challenge 3: Methodological gaps for designing Data Science processes in real applications: The number and variety of environmental problems suitable for application of Data Science approaches is high. It is currently the responsibility of the data scientist to translate the environmental problem into a Data

Science workflow that encompasses both the goals of the problem and available data. This includes identifying the proper preprocessing, data mining, and knowledge production methods and their proper sequencing and interactions to create a Data Science workflow that will advance the project. Gibert et al. (2016) provides some guidelines to design preprocessing steps. There are currently no established guidelines or standards for how to design Data Science workflows, leaving subjectivity in their design and difficulty in comparing results where different workflows may have been used to address the same or similar problems. Research is needed in this respect to build a conceptual framework with standard Data Science Processes providing answers to a certain kind of problems.

Challenge 4: Guidelines to map families of environmental problems with prototypical Data Science processes that help in environmental application: From a structural point of view, there are some commonalities in some families of environmental problems that fit well with a similar kind of Data Science process. As an example, analyzing the effect of pollutants may require similar analytics for air pollution or water pollution. Analogously, predicting the survival of certain protected species might require similar methods for forest fauna and freshwater fish. Even if the related environmental systems are radically different, the first example may entail in both cases modelling continuous multiresponse variables (several coexisting pollutants) that are never independent and develop in a spatiotemporal space, whereas the latter example is, in both cases, about discrete prediction methods with a single counting response variable. Very little work has been done on finding families of environmental problems that share structure and building guidelines to map them into standard Environmental Data Science workflows. A deep analysis of the different environmental problems suitable for Data Science is required to produce insights about them and how they map to standard Data Science solutions. Providing environmental scientists with tools to help them identify the structure of an environmental problem would be of great benefit to do these mappings and opens the door for more effective application of Data Science techniques within the Environmental Sciences.

Challenge 5: Data quality and dealing with uncertainty in data: Data quality is currently one of the hotpoints of the Data Science process. As environmental data always includes mistakes or biases, the validity of Data Science processes that use these data becomes limited at best, or even dangerous under worst case scenarios, enabling the production of potentially incorrect conclusions that may lead to decisions with dramatic consequences. As an example, in 2012, the secretary of environmental protection in Pennsylvania told Congress that there was no evidence the state's water quality had been affected by fracking. In August 2014, the same department published a list of 248 incidents of damage to well water due to gas development. This discrepancy was caused because six regions in the state had missing data in 2012, and data collection was not the same among regions (Barrett and Greene, 2015).

With the diversity of information sources combined in Data Science projects, often including videos, audio recordings, images and real time data streams, guaranteeing the quality of data requires development of stronger methodologies that go beyond the current catalogs of unconnected preprocessing operations provided by existing software. Along with the need for more research in data preprocessing approaches (Gibert et al., 2016), there is also a need for efforts to improve technologies behind smart sensors to reduce failures in measurements, eliminate noise, and increase the quality of data transmissions. In spite of the need for improvement in these areas, it has to be noted that Data Science offers a certain robustness with regard to noise that other approaches cannot achieve. More recent approaches like deep-learning have shown reasonable performance with noisy data, pointing that Data Science might create useful results even from noisy or lower quality data. The potentials in this direction is still under-explored.

Challenge 6: Privacy and security: Guaranteeing privacy of personal data circulating in the Internet, from sources ranging from smart sensors to communication networks, data centers, or the cloud is another critical issue that requires attention and limits the scope of applying Data Science to all available environmental data. For example, federal laws in the U.S. and EU directives in Europe govern the collection, management, and disclosure of personally identifiable information, including health and medical records or farmer's data. Yet, there are many potential opportunities for combining this type of data with those from environmental sensors or samples to learn more about the effects of exposure to environmental conditions (e.g., Reis et al., 2015). Another example is privacy preservation in creating crowdsourced noise maps (Drosatos et al., 2014).

Challenge 7: Methods to choose pertinent, correct, sufficient, and non-superfluous data for analysis: As the volume and heterogeneity of available data continues to grow, no clear criteria have been established to assess which out of all available data is required for an analysis, or whether available data is representative enough of the whole target population. “Big” is not necessarily linked with “sufficient” or “unbiased,” and more work is required to provide guidelines for making these decisions. This necessarily requires clarification of which possible biases different data sources might include by construction. Furthermore, as the volume, speed and diversity of collected data is growing, it is hardly possible for scientists to keep on manually preprocessing data - i.e., performing tasks such as data linking, cleaning, and integration. Data sharing and discovery need to be performed using methods that make it possible for machines to tackle these tasks with semantic interoperability, rather than human experts. Additionally, clear policies on use of big or small data for specific environmental applications is needed.

Challenge 8: Need for development of integral data mining methods: There is a clear need for data mining methods able to cope with heterogeneous data that might include traditional databases, data derived from environmental samples, smart sensor data or data from supervisory control and data acquisition (SCADA) systems, data with intrinsic uncertainty (like georadar data), data streams, geospatial datasets, images, sounds, and free text to provide an integrated overview of a complex system. These methods must also address any structural complexities involved (e.g., high order interactions, multigranularity, spatiotemporality, etc.).

Challenge 9: Guidelines to choose the right analytics method for a given problem:

Challenges 5 to 7 are related to the data used in the Data Science process. Next step in the process is analyzing data with suitable methods. Selection of the proper methods to effectively process data can be equally difficult. Not much work has been done to establish consensus about which analytics methods are effective and appropriate for specific applications (Gibert et al., 2010b). As such, there are few clear guidelines for analyzing a certain kind of data when addressing a certain kind of question. We have even seen how analyses of the same dataset can provide contradictory conclusions when analyzed by two independent data scientists without a common set of guidelines for conducting the analysis in the proper way.

Challenge 10: Clear policies on long-term data storage and computational costs in terms of both sustainability and information availability: When data has been consumed for a primary analysis, it is useful to store it in a long-term data repository for future exploitations, including combining it with other information

sources for reuse or comparative analysis, as well as for independent verification of results. However, all of these data consume storage space and require energy for both storage and processing. The costs of storage space and energy consumption associated with data and processing envisaged in the near future present critical sustainability challenges. In 2013, data centers in the U.S. consumed 91 TWh of power (this is approximately 2.2% of the total power generated in the U.S.) (CGO, 2014). It is expected that in 2020 this consumption will increase to 3.5%. Communication networks consumed approximately 5% of total energy generated in 2012, and it is expected that this will increase to 10% by 2020. As another example, the penetration and buildout of the Internet of Things (IoT) is expected to raise data consumption around 27% between 2015 and 2020 (CGCI, 2015). TREND (2013) estimates that approximately 4.6% of world energy consumption (~9000 TWh per year) is currently devoted to information and communications (ICT) systems, with yearly increase of 7%. Given these estimates of current consumption and projected growth, finding ways to become more efficient with computation and long-term storage, along with criteria for which data need to be stored versus which data can be discarded will become critical. This also includes considerations about the long-term effective life of digital objects that may suffer from degeneration along time (Conway, 2010). For example, images and audio recordings (Corrado and Moulaison Sandy, 2017) require specific methodologies to guarantee they will still be usable in the long future, in spite of software upgrades (de la Rosa et al., 2010).

Challenge 11: Reproducibility and interoperability: Some scientific communities have aligned themselves around standard formats for data and standard software for particular analyses. Examples include the use of the Network Common Data Form (netCDF) file format in the weather and climate community or the use of the Quantitative Insights into Microbial Ecology (QIIME - http://qiime.org/) software used by many scientists and bioinformatics experts to perform microbiome analysis from raw DNA sequencing data. Such standardization can promote the reusability of data and the reproducibility of analyses. However, other scientific domains have not achieved this level of standardization, and heterogeneity in both data and analysis techniques is still a barrier to rapid scientific progress using Data Science. The sheer volume of data used in some scientific analyses can also be a barrier to reproducibility because it is not always practical or possible to store and maintain the large volumes of data used for a particular analysis (see previous challenge). Also, reproducibility is not just about preserving data in an accessible form in the long-term, but also about retaining the exact conditions of the analysis itself, i.e. the algorithms used, the input parameters, software versions involved, the intermediate data preprocessing and transformation steps, and documentation with enough detail and precision to allow the reproduction of exact results using the same data. Often, inadequate details are recorded about the complete Data Science workflow, and some random elements are involved in the middle of the process (like initial random class seeds in K-means, for example) making reproducibility difficult, and, as a consequence, the scope of conclusions may also become limited because generality of conclusions is impacted by lack of reproducibility. Given that it is unlikely that all scientific disciplines will settle on standards for data collection, management, and analysis, better methods are needed for capturing scientific workflows to enhance the reproducibility of data-intensive analyses.

Some of the challenges described above may be addressed through the results of future, interesting research trends from the side of Data Science that may provide new techniques and methodologies for using environmental datasets to provide answers to certain environmental problems. On the other hand, the future will also require research to create new environmental datasets and to make existing environmental data more suitable for use with Data Science approaches. This will likely rely more on future research from Environmental Science experts who better understand techniques for making measurements, formulating domain specific models of environmental phenomena, and generating data. In our opinion, the most productive approach for addressing the problems listed above is to enhance collaborations in multidisciplinary teams where both environmental and data scientists work together to contribute to overcoming of these challenges in the near future.

8. Conclusions and the path forward

In this paper we have provided insight into the origins of the Data Science field, the intrinsic nature of Data Science, and how it can contribute to improving understanding and management of Environmental Systems. The new view of the field we have provided stresses the multidisciplinary nature of Data Science as a combination of data analytics, data processing, and knowledge management in order to provide added value for decision making. The term is still controverted as it is a high-level umbrella encompassing methods and techniques from many interrelated areas with a common ambition of providing global understanding of complex phenomena.

Indeed, “the magic” involved in Data Science (i.e. extracting wisdom from data as Hal Varian expected in 2009) requires specific skills that are not frequent yet in professionals Corporations realized the potential impact from highly qualified data scientists and underwent a deep transformation in recent years towards the new “data-centered” economy announced by Cukier in 2010. The phenomenon is not restricted to corporations and we are encouraged to see that academia has started to deploy new Data Science curricula. However, current efforts are still far from covering the existing shortage. One approach to addressing the lack of Data Science professionals (Challenges 1 and 2 above) is to promote a new culture where the keystone for Data Science is no longer a single multifaceted professional, but rather performing Data Science within a multidisciplinary team. Such teams could be composed of statisticians, machine learners, software engineers, knowledge engineers, and domain experts like environmental scientists for environmental applications that guarantee the highest levels of expertise in all the skills involved in real Data Science projects to achieve the “extraction of wisdom from data” already mentioned. In that case, a common language among such working teams is still required and needs to be part of specific, postgraduate curricula to be urgently developed by academia.

Our current ideas about what are necessary components of engineering and science curricula may also need to change to create a next generation of engineers and scientists who are better trained in Data Science and who are more capable of working in collaborative teams. This requires development of specific training aimed at transferring the basics of Data Science to environmental scientists and the basics of Environmental Sciences to data scientists. We anticipate working toward a new generation of professionals with the necessary skills to not only understand the scientific concepts within a domain, but who also have the data processing and computer science expertise to be able to work in computationally complex and data intensive fields. Stronger links may also be needed between academia and corporations to enable students to be actively engaged with corporations as part of their educational experience (without leaving their degree program entirely), making education more of a partnership between educators and employers.

The Environmental Sciences cannot elude the transformation produced by the penetration of Data Science (and data scientists or

data scientist teams) that has been experienced in many other application fields. In fact, we have already seen how much potential Data Science has for analyzing environmental systems, providing a broad perspective of the complexity involved in these systems, and, as a consequence, a nice support for deeper understanding of environmental phenomena and enhanced information for decision-making. However, significant challenges remain. We envisage and encourage new research trends that contribute to integration between Data Science and Environmental Sciences.

At a high level, new tools that understand the structure of environmental problems and can refer analysts to a standardized family of reference environmental problems would be a major help in mapping environmental problems to Data Science methods and workflows that can provide an appropriate solution. Standardization of Data Science methods would also be a major benefit in this scenario, addressing Challenge 3 above, and would contribute to reproducibility and interoperability (Challenge 11). This is a high-level activity very much related to the design of Data Science processes.

On the other hand, specific criteria have to be developed to properly manage each of the internal steps within a Data Science workflow, including measurement, data transmission, data storage, analysis, and sharing (Challenges 6 to 11). Many users who have collected heterogeneous, noisy, non-linear, multigranular, spatio-temporal environmental data face the prospect of not knowing which methods to use, how to evaluate their effectiveness, or what constitutes an acceptable result (Challenges 8 and 9). Thus, further methodological development followed by software development supporting new methods and guidelines for appropriate usage is encouraged. To do this, a bespoke repository of environmental problems with their associated Data Science workflows could provide a basis to support research development, benchmarking, and a complementary platform of typical pre-processing methods, modelling tools (either predictive of descriptive), and post-processing methods. It could provide a focal point in the area of Environmental Data Science to centralize methodological achievements with their corresponding guidelines. These tools could be used with a level of intelligent guidance and explanation adapted to the level of expertise of the user and would be useful to train a new generation of data scientists as well.

The paper elicits that Environmental Data Science is a fruitful research area providing strategic added value to both corporations and environmental systems and merits attention for further developments in the short, mid, and long-term.

References

Afshar, Abbas, et al., 2015. State of the art review of ant colony optimization applications in water resource management. Water Resour. Manag. 29 (11), 3891–3904.

Agarwal, R., Dhar, V., 2014. Big data, data science, and analytics: the opportunity and challenge for IS research. Inf. Syst. Res. 25 (3), 443–448. https://doi.org/10.1287/isre.2014.0546.

Ashraf Vaghefi, S., Abbaspour, N., Kamali, B., Abbaspour, K.C., 2017. A toolkit for climate change analysis and pattern recognition for extreme weather conditions – case study: California-Baja California Peninsula. Environ. Model. Software 96, 181–198. https://doi.org/10.1016/j.envsoft.2017.06.033.

Association for Computing Machinery, 2017. Information technology curricula 2017 IT2017: curriculum guidelines for baccalaureate degree programs in information technology. In: Association for Computing Machinery (ACM) and IEEE Computer Society (IEEE-CS). https://doi.org/10.1145/3173161.

Athanasiadis, I.N., Mitkas, P.A., 2007. Knowledge discovery for operational decision support in air quality management. Journal of Environmental Informatics 9 (2), 100–107. https://doi.org/10.3808/jei.200700091.

Baeza-Yates, R., 2017. Big-data or Small Data? the Correct Answer Is Both. Inside BIG-data Editorial. July 13th 2017.

Barrett, K., Greene, R., 2015. The causes, costs and consequences of bad government data, Governing the States and localities. June 21, 2015. http://www.governing.com/topics/mgmt/gov-bad-data.html.

Berry, J., 1994. Database Marketing, Business Week, September 5, pp. 56–62.

Bifet, A., Gavaldà, R., Holmes, G., Pfahringer, B., 2017. Machine Learning for Data Streams. MIT Press, Massachusset, USA.

Blattenberger, G., Fowles, R., 2017. Treed avalanche forecasting: mitigating avalanche danger utilizing bayesian additive regression trees. J. Forecast. 36 (2), 165–180. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2421.

Burns, E., 2017. Lack of skills remains one of the biggest data science challenges. In: Search Business Analytics Jan 11th, 2017. https://searchbusinessanalytics.techtarget.com/news/450410819/Lack-of-skills-remains-one-of-the-biggest-data-science-challenges.

Brynjolfsson, E., Hitt, L.M., Kim, H., 2011. Strength in Numbers: how does data-driven decision-making affect firm performances?. In: International Conference on Information Systems. December 2011, Shanghai, China.

Businesscom Editorial Staff, 2017. Big Data, Big Problem: Coping with Shortage of Talent in Data Analysis. Business. com, February 22, 2017.

Caffo, B., Peng, R.D., Leek, J., 2016. Executive Data Science. Lean Publishing.

CGCI, 2015. Cisco global cloud index: forecast and methodology, 2015–2020 white paper. http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/whitepaper-c11-738085.pdf.

CGO, 2014. G20 energy action plan, 16 november 2014. https://ipeec.org/upload/publication_related_language/pdf/11.pdf.

Cheeseman, P., Oldford, R.W. (Eds.), 1994. Selecting Models from Data. LNStats 89. Springer.

Cleveland, W.S., 2001. Data science: an action plan for expanding the technical areas of the field of statistics. Int. Stat. Rev. 69 (1), 21–26.

Cominola, A., Giuliani, M., Piga, D., Castelletti, A., Rizzoli, A.E., 2015. Benefits and challenges of using smart meters for advancing residential water demand modeling and management: a review. Environ. Model. Software 72, 198–214. https://doi.org/10.1016/j.envsoft.2015.07.012.

Conway, P., 2010. Preservation in the age of Google: digitization, digital preservation, and dilemmas. Libr. Q. 80.1, 61–79.

Conway, 2013. The DataScience Venn Diagram. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram.

Corominas, Ll, Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M., 2017. Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. Environ. Model. Software. https://doi.org/10.1016/j.envsoft.2017.11.023.

Corrado, E.M., Moulaison Sandy, H., 2017. Digital Preservation for Libraries, Archives, and Museums. Rowman & Littlefield.

Cukier, K., 2010. Data, Data Everywhere. In Special Report on Managing Information. The Economist. February 27th 2010.

Davenport, T.H., Dyche, J., 2013. Big Data in Big Companies. Tech Report. SAS institute Press, Cary NC.

Davenport, T.H., Harris, J.G., 2007. Competing on Analytics: the New Science of Winning. Harvard Business Press.

Davenport, T.H., Patil, D., 2012. Data scientist. Harv. Bus. Rev. 90 (5), 70–76.

de la Rosa, Lluis, Josep, et al., 2010. Agents for social search in long-term digital preservation, Semantics Knowledge and Grid (SKG). In: Sixth International Conference on. IEEE.

Dornblut, I., Atkinson, R., 2014. OGC HY_Features: a common hydrologic feature model, open geospatial Consortium discussion paper, OGC 11-039r3. https://portal.opengeospatial.org/files/?artifact_id=55157.

Drosatos, G., Efraimidis, P., Athanasiadis, I., Stevens, M., DHondt, E., 2014. Privacy-preserving computation of participatory noise maps in the cloud. J. Syst. Software 92, 170–183. https://doi.org/10.1016/j.jss.2014.01.035.

Elarab, M., Ticlavilca, A.M., Torres-Rua, A.F., Maslova, I., McKee, M., 2015. Estimating cholorophyll with thermal and broadband multispectral high resolution imagery from an unmanned aerial system using relevance vector machines for precision agriculture. Int. J. Appl. Earth Obs. Geoinf. 43, 32–42. https://doi.org/10.1016/j.jag.2015.03.017.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press/ MIT Press.

Gauci, A., Abela, J., Austad, M., Cassar, L.F., Zarb Adami, K., 2018. A Machine Learning approach for automatic land cover mapping from DSLR images over the Maltese Islands. Environ. Model. Software 99, 1–10. https://doi.org/10.1016/j.envsoft.2017.09.014.

Gibert, K., Spate, J., Sánchez-Marré, M., Athanasiadis, I.N., Comas, J., 2008. Data mining for environmental systems. In: Environmental Modelling, Software and Decision Support: State of the Art and New Perspective. Elsevier, pp. 205–228. https://doi.org/10.1016/S1574-101X(08)00612-1.

Gibert, K., Rodríguez-Silva, G., Rodríguez-Roda, I., 2010a. Knowledge discovery with clustering based on rules by states: a water treatment application. Environ. Model. Software 25 (6), 712–723.

Gibert, K., Sànchez-Marrè, M., Codina, V., 2010b. Choosing the right data mining technique: classification of methods and intelligent recommenders. In: Proc. Of the IEMSs'10, 5th Biennial Meeting (III DMTES Workshop). S23.03.1–S23.03.9, 2010.

Gibert, K., Sànchez-Marrè, M., Izquierdo, J., 2016. A Survey on pre-processing techniques in the context of environmental data mining. Artificial Intelligence in Communications 29 (6), 627–663. https://doi.org/10.3233/AIC-160710. IOSPress.

Gómez-Losada, Álvaro, et al., 2014. Finite mixture models to characterize and refine air quality monitoring networks. Sci. Total Environ. 485, 292–299.

Gurung, T.R., Stewart, R.A., Beal, C.D., Sharma, A.K., 2015. Smart meter enabled water end-use demand data: platform for the enhanced infrastructure planning of contemporary urban water supply networks. J. Clean. Prod. 87, 642–654.

https://doi.org/10.1016/j.jclepro.2014.09.054.
Hammond, M., 2004. The Fact Gap: the Disconnect between Data and Decisions, Business Objects.
Harris, J.G., Shetterley, N., Alter, A.E., Schnell, K., 2014. It takes teams to solve the data science shortage. Wall St. J. https://blogs.wsj.com/cio/2014/02/14/it-takes-teams-to-solve-the-data-scientist-shortage/ Feb 14, 2014.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., et al., 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One 12 (2), e0169748. https://doi.org/10.1371/journal.pone.0169748.
Hill, D.J., Liu, Y., Marini, L., Kooper, R., Rodriguez, A., Futrelle, J., Minsker, B., Myers, J., McLaren, T., 2011. A virtual sensor system for user-generated, real-time environmental data products. Environ. Model. Software 26 (12), 1710–1724. https://doi.org/10.1016/j.envsoft.2011.09.001.
Horsburgh, J.S., Leonardo, M.E., Abdallah, A.M., Rosenberg, D.E., 2017. Measuring water use, conservation, and differences by gender using an inexpensive, high frequency metering system. Environ. Model. Software 96, 83–94. https://doi.org/10.1016/j.envsoft.2017.06.035.
IASC, 1977. Statutes of the IASC, 1977-2015.
Kanevski, M., Pozdnukhov, A., Timonin, V., 2008. Machine learning algorithms for geospatial data. Applications and software tools. In: Proceedings of the 4th International Congress on Environmental Modelling & Software, Barcelona, Catalonia, Spain, July 2018. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2710&context=iemssconference.
Klise, K.A., Bynum, M., Moriarty, D., Murray, R., 2017. A software framework for assessing the resilience of drinking water systems to disasters with an example earthquake case study. Environ. Model. Software 95, 420–431. https://doi.org/10.1016/j.envsoft.2017.06.022.
Kranstauber, B., Cameron, A., Weinzerl, R., Fountain, T., Tilak, S., Wikelski, M., Kays, R., 2011. The Movebank data model for animal tracking. Environ. Model. Software 26 (6), 834–835. https://doi.org/10.1016/j.envsoft.2010.12.005.
Lauro, N.C., Amaturo, E., Grassia, M.G., Aragona, B., Marino, M. (Eds.), 2017. Data Science and Social Research, Epistemology, Methods, Technology and Applications. Studies in Classification, Data Analysis and Knowledge Organization V. 1564. ISBN: 978-3-319-55477-8, Springer Int'l 2017.
Levasseur, Séverine, et al., 2008. Soil parameter identification using a genetic algorithm. Int. J. Numer. Anal. Meth. GeoMech. 32 (2), 189–213.
Luhn, H.P., 1958. A Business Intelligence System. IBM Journal of Research and Development.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A., 2011. Big Data: the Next Frontier for Innovation, Competition, and Productivity. Technical Report May 2011. McKinsey Global Institute. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.
Markow, W., Braganza, S., Taska, B., Miller, S.M., Hughes, D., 2017. The quant crunch: how the demand for data science skills is disrupting the job market, burning glass technologies. https://public.dhe.ibm.com/common/ssi/ecm/im/en/iml14576usen/analytics-analytics-platform-im-analyst-paper-or-report-iml14576usen-20171229.pdf.
Mattmann, C.A., 2013. Computing: a vision for data science. Nature 493 (7433), 473–475. https://doi.org/10.1038/493473a.
McCord, S.E., Buenemann, M., Karl, J.W., Browning, D.M., Hadley, B.C., 2017. Integrating remotely sensed imagery and existing multiscale field data to derive rangeland indicators: application of bayesian additive regression trees. Rangel. Ecol. Manag. 70 (5), 644–655. http://www.bioone.org/doi/abs/10.1016/j.rama.2017.02.004. www.bioone.org.
McKenna, B., 2013. C-level Executives Cry Out for Data Scientists, Computer Weekly. Sep 24, 2013.
Kumar, D.N., Srinivasa Raju, K., Ashok, B., 2006. Optimal reservoir operation for irrigation of multiple crops using genetic algorithms. J. Irrigat. Drain. Eng. 132 (2), 123–129.
Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., Ochiai, O., 2015. Big data challenges in building the global Earth observation system of systems. Environ. Model. Software 68, 1–26. https://doi.org/10.1016/j.envsoft.2015.01.017.
Noyes, K., 2014. Cropping up on Every Farm: Big Data Technology, Fortune. May 30.
Palmer, D., 2013. Not Enough Data Scientists, MIT Expert Tells Computing, Computing. Published online. https://www.computing.co.uk/ctg/news/2292485/not-enough-data-scientists-mit-expert-tells-computing.
Piatetski-Shapiro, G., 2017. Data Scientist – best job in America, again, KDNuggets blog post. Available online: https://www.kdnuggets.com/2017/01/glassdoor-data-scientist-best-job-america.html.
Porter, C.H., Villalobos, C., Holzworth, D., Nelson, R., White, J.W., Athanasiadis, I.N., Janssen, S., Ripoche, D., Cufi, J., Raes, D., Zhang, M., Knapen, R., Sahajpal, R., Boote, K., J.Jones, J.W., 2014. Harmonization and translation of crop modeling data to ensure interoperability. Environ. Model. Software 62, 495–508. https://doi.org/10.1016/j.envsoft.2014.09.004.
Reis, S., Seto, E., Northcross, A., Quinn, N.W.T., Convertino, M., Jones, R.L., Maier, H.R., Schlink, U., Steinle, S., Vieno, M., Wimberly, M.C., 2015. Integrating modelling and smart sensors for environmental and human health. Environ. Model. Software 74, 238–246. https://doi.org/10.1016/j.envsoft.2015.06.003.
Rizzi, A., Vichi, M. (Eds.), 2006. Compstat 2006—Proceedings in Computational Statistics. Preface. Physica Verlag, Heidelberg, Germany.
Roelofsen, H.D., Kooistra, L., van Bodegom, M., Verrelst, J, P., Krol, J., Witte, J.M.P., 2014. Mapping a priori defined plant associations using remotely sensed vegetation characteristics. Rem. Sens. Environ. 140, 639–651.
Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E.C., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.A., Rosa, A.D., Dam, L., Evans, M.H., Cervantes, I.F., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T., Eriksson, K.H., Heene, M., Mohr, A.H., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D.M., Lei, R., Lindsay, T.A., Dr, S.L., Madan, C.R., Molden, D.C., Molleman, E., Morey, R.D., Mulder, L.B., Nijstad, B.A., Pope, B., Pope, N.G., Prenoveau, J.M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlueter, E., F., S., Sherman, M.F., Sommer, S.A., Sotak, K.L., Spain, S.M., Spörlein, C., Stafford, T., Stefanutti, L., Täuber, S., Ullrich, J., Vianello, M., Wagenmakers, E.-J., Witkowiak, M., Yoon, S., Nosek, B.A., 2017. Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results PsyArXiv. http://doi.org/10.17605/OSF.IO/QKWST.
Schlink, U., Röder, S., Kohajda, T., Wissenbach, D.K., Franck, U., Lehmann, I., 2016. A framework to interpret passively sampled indoor-air VOC concentrations in health studies. Build. Environ. 105, 198–209.
Shah, S., 2013. Analysis: it Takes Skills to Explore Big-data, Computing. Jan 24, 2013.
Somohano, C., 2013. What does a data scientist Do? Data science london. Jan 2013 http://www.datasciencecentral.com/profiles/blogs/what-does-a-data-scientist-do. .
Sundaresan, N., 2017. The History of Data Science, Huffpost 2017. May 25th, Quora.
Swayne, D.A., 2003. Applying computer research to environmental problems. Environ. Model. Software 18, 485–486. https://doi.org/10.1016/S1364-8152(03)00022-7.
TREND, 2013. TREND workshop. http://www.fp7-trend.eu/system/files/content-public/502-final-trendworkshop-brussels-24-october-2013-presentations/energyconsumptionincentives-energy-efficientnetworks.pdf.
Tukey, J.W., 1962. The future of data analysis. Ann. Math. Stat. 33 (1), 1–67.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley.
United Nations, 1992. Earth Summit: Agenda 21: programme of action for sustainable development. Rio declaration on environment and development. Department of public information. United Nations, Rio de Janeiro, Brazil. Available online: https://sustainabledevelopment.un.org/content/documents/Agenda21.pdf.
Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203 (3–4), 312–318.
VentureBeat, 2017. Glassdoor: 14 of the top 50 U.S. jobs for 2017 are in tech. https://venturebeat.com/2017/01/23/glassdoor-14-of-the-top-50-u-s-jobs-for-2017-are-in-tech/.
Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C.J.A., Buytaert, W., 2015. Web technologies for environmental big data. Environ. Model. Software 63, 185–198. https://doi.org/10.1016/j.envsoft.2014.10.007.
Wiemann, S., Brauner, J., Karrasch, P., Henzen, D., Bernard, L., 2016. Design and prototype of an interoperable online air quality information system. Environ. Model. Software 79, 354–366. https://doi.org/10.1016/j.envsoft.2015.10.028.
Wills, J. (2012). Posted online: https://twitter.com/josh_wills/status/198093512149958656.
Wong, B.P., Kerkez, B., 2016. Real-time environmental sensor data: an application to water quality using web services. Environ. Model. Software 84, 505–517. https://doi.org/10.1016/j.envsoft.2016.07.020.
Wu, C.F.J., 1999. Statistics = data Science? (Talk) P.C. Mahalanobis memorial lectures, 7th series, indian statistical Institute. https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf.