Received: 28 May 2020 | Revised: 14 October 2020 | Accepted: 6 November 2020

DOI: 10.1111/ejss.13071

REVIEW

European Journal of Soil Science WILEY

Perspectives on data-driven soil research

Alexandre M. J.-C. Wadoux | **Mercedes Román-Dobarco** | **Alex B. McBratney**

Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia

Correspondence
Alexandre M. J.-C. Wadoux, Sydney Institute of Agriculture & School of Life and Environmental Sciences, The University of Sydney, Sydney, Australia.
Email: alexandre.wadoux@sydney.edu.au

Abstract

Soil is a complex system in which biological, chemical and physical interactions take place. The behaviour of these interactions changes in spatial scale from the atomic to the global, and in time. To understand how this system works, soil scientists usually rely on incremental improvements in the knowledge by refinement of theories through hypothesis testing and development using carefully designed experiments. In the last two decades, the primacy of this knowledge construction process has been challenged by the development of large soil databases and algorithms such as machine learning. The data-driven research approach to soil science, the inference of soil knowledge directly from data by using computational tools and modelling techniques, is becoming more popular. Despite the wide adoption of a data-driven research approach to soil science, there has been little discussion on how a research driven by data instead of hypotheses affects scientific progress. In this paper, we provide an introductory perspective on data-driven soil research by discussing some of the issues and opportunities of knowledge discovery from soil data. We show that while data-driven soil research may seem revolutionary for some, soil science has a long history of exploratory efforts to generate knowledge from data. Empirical and factual soil classifications, for example, were data driven. We further discuss, with examples, (i) data, databases and the logic of data storage for data-driven soil research, (ii) the issues of extreme empiricist claims that arise corollary to the increase in the use of computational tools, and (iii) the challenge of formulating a scientific explanation based on patterns observed in the data and data analysis tools. By considering the epistemic challenges of the data-driven scientific research in the light of the historical literature, we found that there is a continuity of practices, some being certainly amplified by recent technological changes, but that the core methods of scientific enquiry from data remain essentially unchanged.

Highlights:

- Historical account of data-driven soil science research.
- Describe data to be used for data-driven soil science.
- Discuss conceptual issues and opportunities for data-driven soil science.
- Investigate the challenge of formulating an explanation from soil data.

KEYWORDS

data science, epistemology, knowledge discovery, machine learning, pedology, pedometrics, soil science

 © 2020 British Society of Soil Science

1 | INTRODUCTION

The last decade has witnessed a considerable increase in electronic digital information and information technologies available for academic research. This increase is questioning how the sciences are approached from a methodological perspective. Criticisms are being made of research driven by data and computational tools, thus reinvigorating debates on the scientific method and scientific practices in many fields of science. Soil science is no exception. Despite consensus that soil science fundamentally relies on experts and extensive domain knowledge, the development of sensing techniques, analytical methods of soil analysis, and the ease of storing and processing these data are changing the practice of soil science (Roudier, Ritchie, Hedley, & Medyckyj-Scott, 2015; Rossiter, 2018). The indubitable challenge for soil science is that of extracting knowledge and relevant information from increasingly large, diverse and complex soil datasets.

A great deal of attention has thus recently been paid to data-intensive or data-driven research in both the scientific (e.g. Bui, 2016) or popular (e.g. Minasny & McBratney, 2013) soil science literature. Paraphrasing Kelling et al. (2009), data-intensive research takes an approach where progress is compelled by data, as opposed to the "knowledge-driven" or "expert-centred" approaches in which a hypothesis is developed on or corroborated by data. Data-intensive research is emerging through the combination of several timely factors, which are 1. the ease of data generation, processing and storing, 2. the development of computer, computational power and software resources, and 3. the popularization of complex statistical and algorithmic tools, which increasingly engage machine learning calculi, to explore these repositories of data.

The use of statistics for exploring databases and finding patterns in data is not new in soil science. Yaalon (1975), for example, suggested statistical search procedures to explore the functional relationships in a natural soil system. When more is known about the soil structure and process, the statistical model can gradually be replaced by a mechanistic one. Others (e.g. Jenny and Leonard, 1934; Webster, 1997) have used statistical modelling for correlation analysis, data dimensionality reduction or regression. Much of statistical modelling in the twentieth century was model-based: it requires the scientist to specify the candidate independent variables that could enter the model, the functional form of the relationships (linear, quadratic) between independent and dependent variables, and the assumptions of the underlying nature of the soil data (e.g. deterministic or random; Webster, 2000) (Hochachka et al., 2007). Lately, soil scientists have witnessed an increase in the use of flexible data-driven models and algorithmic strategies, in particular machine learning, to tackle this data-rich environment. Neural networks or random forests are primary examples of these models, whereas regressions with a specified form of the functional relationship (e.g. linear) are examples of parametric model-based statistics (Breiman, 2001). In machine learning, no explicit assumption is made on the functional form of the relationship between independent and dependent variables. Instead, machine learning models seek an estimate of the form of the relationship which best detects and describes patterns in data, given some accuracy indices, thus avoiding many of the assumptions of parametric statistical modelling outlined previously. The major disadvantage of most machine learning models is their lack of interpretability, i.e. the model structure is very complex and cannot be readily visualized or perceived.

In several subdisciplines of soil science, research driven by database and computational tools have flourished over the past 20 years. For example, Bui, Henderson, and Viergever (2009) used a large (i.e. containing more than 10,000 soil samples) soil organic carbon database and environmental covariates to infer the organic content of agricultural soils of Australia. Morellos et al. (2016) compiled laboratory-derived measurements and infrared soil spectral data containing several thousands of wavelength values, and used machine learning models to estimate soil properties using their molecular vibration in the spectra. In soil hydrology, Kornelsen and Coulibaly (2014) estimated soil moisture in the root-zone by artificial neural networks at various local study sites, using a dataset generated by the physically constrained HYDRUS-1D hydrological model. These example studies, far from being an exhaustive summary of the current literature, illustrate that data-driven research has a real impact on the current production of knowledge in soil science.

Data-driven research has generated much enthusiasm in soil science, in particular in the sub-fields of soil survey and pedometrics, where research has always relied more on field and observational data than on manipulated experiments. The abundance of data and their use as a primary driver of knowledge in subdisciplines of soil science have several methodological and epistemological implications that have not been documented so far. This raises several questions. Is data-driven soil research unprecedented in its history? Are all data equally valid for use in data-driven science? What are the risks, challenges and extreme claims of data-driven soil research? If knowledge is to be found in the data, should we invest all our efforts in generating more data? This paper discusses data-driven soil science, and attempts to provide some contexts and an introductory perspective to the conceptual challenges pertaining to the research strategies driven by data instead of hypotheses. Notably, this paper aims to serve as a starting point for further discussions

on the new epistemological challenges facing soil science research in the information era.

The perspective developed in this study stems from challenges encountered in pedometrics, digital soil mapping, numerical soil surveys or pedotransfer functions, fields with which the authors are most familiar. The general comments made in the paper are also applicable to a large number of applications in soil science where the development of large soil databases and their exploration with machine learning is emerging (e.g. in soil genotype studies). We leave it to the reader's discretion to apply the concepts discussed here to various disciplines and subdisciplines.

2 | KNOWLEDGE-BASED VERSUS DATA-DRIVEN RESEARCH

2.1 | Scientific methods in soil science

Much of today's epistemology acknowledges that scientific progress is either goal-driven or anomaly-driven. In soil science, the role of hypotheses is crucial but is not given the same importance in each of the cases. When the research is goal-driven, viz. the soil scientist wants to solve something, the hypothesis defines areas of focus for the research. The hypothesis is tested through an experiment on generated data tailored to answer a question. From the results of this test, a deduction is made on whether the hypothesis is corroborated. The explanation can also be of the inductive-type, by assigning probabilities to the soil phenomenon to be explained using statistical laws (inductive statistical explanation). Much of soil science is goal-driven: in the past when soil science was associated with agronomy and had to fulfill the promise of agricultural production (McDonald, 1994), and still today in several sub-fields of soil science, for example in digital soil mapping when producing accurate quantitative soil information. In an experimental science, the hypothesis is formulated at the beginning of the research and drives data collection. Alternatively, progress may be anomaly driven when a phenomenon or observation conflicts with existing knowledge. Think of a field pedologist noticing in a soil profile an unexpected soil feature (colour or aggregate), which cannot be readily explained using existing knowledge. The hypotheses that follow with the aim of explaining this unexpected soil feature may be as numerous as possible tracks to investigate. The hypotheses are elaborated within existing theories, scientific laws, tacit knowledge of the soil scientist and a scientific context. This scientific development based on extensive expert knowledge is different from scientific development driven by the exploration of (large) soil datasets and databases by statistical and algorithmic tools, in particular machine learning. In today's data-driven research, large stores of soil data are explored using algorithms which do not rely on explicit hypotheses, but aim to find patterns, correlation and order in data. This is fundamentally different from scientific progress based on carefully designed experiments and hypothesis testing.

2.2 | Data-driven soil science is not new

One may yet rightfully claim that progress based on data accumulation and ordering is nothing new to soil science (Philip, 1991). In the nineteenth century, agricultural stations in western Europe were collecting daily data from several experiments on, for example, fertilizer applications, climate or organic matter degradation. The amount of data was such that in 1919 the Rothamsted experimental station hired Ronald Fisher, young statistician, to extract information from the large amount of soil and crop data amassed over 70 years (Johnston, 1994). Similarly, many of Vasilii Dokuchaev's findings were based on massive amounts of data on soil properties, geological and geomorphological features, climate and land use, collected by him and a network of students in large areas of the chernozem belt (Moon, 2005). Many aspects of past and present soil science, as with other natural sciences (Leonelli, 2014; Strasser, 2012b), are not about testing hypotheses, but concern creating order in what is observed in nature by classifying a (large) number of soil data. Accumulation of soil data is never an end in itself, it is meant to classify and reduce complexity of the soil system and to understand relationships among soils (Hartemink, 2015; Isbell, 1992). American or Australian early classifications, for example, were built on empirical topsoil parameters such as colour, texture and organic matter content, free of "scientific basis" (Krasilnikov, Arnold, & Ibanez, 2010). This view of science based on "facts" and observations (i.e. data-driven or data-intensive research) belongs to a nominalist epistemology. Nominalism states that scientific laws and theories are derived using logic and reason from empirically derived facts and observations. Nominalism is thus based on empiricism and opposes rationalism: rationalists recognize ideas and theorizing. This duality is visible in the fundamental differences between soil classifications.

The seventh approximation is one example of nominalism in soil science, so are the "factual classifications" of Northcote (1971) or the numerical classifications (e.g. Hughes, McBratney, Minasny, & Campbell, 2014), which are alleged to be free of subjective judgement: "This is to say that the soil scientist should use soil genesis in the

form of the empirical geographic correlations [...] but should not make them dependent upon hypotheses of soil-forming processes and should not translate them into theories" (Cline, 1963). French and Russian early soil classifications, based on morphogenetic soil characteristics (e.g. that of Duchaufour, 1963), were examples of rationalism in soil science. What is striking is how closely 1960s nominalist views on soil classifications are similar to contemporary definitions of data-driven science. For example, Moore, Isbell, and Northcote (1983), in a discussion of the two main Australian soil classifications existing in the 1960s, relate classification to pattern recognition. Isbell (1992) explains the two immutable aspects of classification since the 1920s in Australia, the first being the grouping of similar soils into classes, and the second being the assignment of a new entity to one of the classes, and that the classification needs to be updated when new information becomes available. In each of the examples, subjectivity is avoided in the treatment of the soil data, should the classifier be the soil scientist in the past or the computerized statistical or algorithmic tools in the present.

2.3 | What is the difference with current data-driven science?

The above suggests that some early developments in soil science were guided by data alone, but this was most likely not the case because soil scientists are never free from ontological assumptions about the arrangement of the natural processes (Strasser, 2012b). The soil is not an isolated body, but a continuum. In any classification, however, each soil individual is assigned to a class, classes that are artificial, bounded or fuzzy (McBratney & Odeh, 1997) entities. For example, the soil surveyor assumes the existence of a definite number of soil classes as there exist no "raw" observations defining what an exact soil class might be, thereby forcing existing soil data into ontological categories. Soil classifications are revised periodically as knowledge increases. Each new scheme was based on a different assumption about the number of classes "useful" for a wide range of purposes (Krasilnikov, Ibanez Marti, Arnold, & Shoba, 2009). Similarly, soil texture classes are another example of ontological categories because the soil texture is continuous in nature. Clearly, this shows that many aspects of soil science are never driven by data alone, but by a combination of knowledge and observations/experiments.

What stands as new in the present-day data-driven soil research is thus not immediately apparent. As Strasser (2012b) puts it, "natural history has been 'data-driven' for many centuries". We have outlined previously that the practices of data collection, storage, ordering and classification, and the methods of analysis of these data, arrived long before the advent of computers and electronic databases. Indeed, soil scientists in the past, like contemporary ones, were not less exposed to vast amounts of data. They stored data in soil archives, created order by means of classifications, and modelled trend and pattern using statistical modelling. The practice of data-driven enquiry is not so new in soil science and seems not to be a characteristic of twenty-first century soil science research. Most of the components of the present-day data-driven soil research are perhaps just a reflection of a change in magnitude of the amount of data collected, their storage and in the capacity to analyse them with complex statistical tools aided by computational power. We will refrain from providing definite dissimilarities between past and today's data-driven science, but two aspects appear as potential differences. They are also found elsewhere (e.g. in Strasser, 2012a, Strasser, 2012b; Sepkoski, 2018) in the literature on natural sciences. The first is that the link to the physical object of the soil material becomes smaller in today's data-driven soil research. Practitioners in the past had a close link to the soil material. At the Rothamsted station in the early twentieth century, statistician Fisher had on-site scientific exchanges with chemists and agronomists. Similarly, large-scale soil classification schemes, such at that from the USA from around the same period, were derived by a soil scientist with significant field practice. Today's data-driven soil science can be performed entirely from the office desk, on soil data stored in electronic databases, without the fundamental need for field experience (regrettably, some argued already in the 1990s, for example in Philip, 1991). The second is the omnipresence of statistical methods (Strasser, 2012b). We stressed previously that past soil scientists were also finding patterns and correlation in data aided by statistics, such as Fisher at Rothamsted, but statistical modelling has increased significantly in the past 30 years. Interestingly, past statistical models were intimately related to and designed for the problem to be solved, whereas current models are more complex, were not initially developed for nor thought to be applied on soil or environmental data, and their structure disregards the underlying mechanisms that created the data.

2.4 | Data-driven science for hypothesis generation

There are many opportunities in the scientific literature (e.g. Elragal & Klischewski, 2017; Leonelli & Ankeny, 2012; Miller, 2010; Sepkoski, 2018) to learn about the elements of contemporary data-driven research. Some are also discussed later in this article. In

short, data-driven soil science is essentially a hypothesis-generation process (Bui, 2016; Hochachka et al., 2007; Kitchin, 2014a). The argument is that, with large volumes of data describing a situation and aided by statistical and algorithmic tools, in particular machine learning, it is possible to discover patterns and correlations in the soil data (Bui, 2016; Pennock, 2004). These correlations, when interpreted by the soil scientist, may trigger new hypotheses that can ultimately be tested using the hypothetico-deductive approach to corroborate that a correlation found in the data indicates an actual mechanism occurring in the soil system. Data-driven science is thus likely to be useful at an early stage of the discovery process in proposing ideas to the researcher that would otherwise remain unseen.

In this perspective, data-driven science is a form of abductive reasoning (Kitchin, 2014b; Miller, 2010). Abductive reasoning builds on data describing a situation and ends with a hypothesis possibly explaining the data. This is a weaker form of inference compared to the well-known deductive and inductive reasoning (Miller & Goodchild, 2015). Deduction is a syllogistic logic, from the general to the specific. A hypothesis is formulated and tested by an experiment. Induction, conversely, builds on observations (the specific) to make a generalization (a hypothesis or a theory) and a prediction, which can be validated by data. For Gohau (1992), the boundary between deductive and inductive reasoning in the research process is not strict because knowledge is often acquired by going back and forth between data, hypothesis and theory. The author questions the origin of the hypothesis in the inductive reasoning and shows that this is often the output of a previous induction. Gohau (1992) thereby calls "invention" the process of generating a hypothesis from the data. In data-driven science, the "invention" can be proposed by an algorithm by mining a complex multivariate database. Abductive reasoning, proposing hypotheses, is thus better used at an early stage of the discovery process, and precedes inductive and deductive reasoning.

3 | WHICH DATA?

3.1 | Observational versus experimental data

On the basis of this data-driven science are the computable electronic digital data (Strasser & Edwards, 2017), referred to as digital data hereafter. The soil data arise either from a controlled experiment or from an observation of the (uncontrolled) natural environment (Dijkerman, 1974). Observational data are, for example, remote sensing images (e.g. Landsat, SMOS, hyperion), spectroscopic information (infrared, nuclear magnetic or electron spin resonance) of a soil sample, soil DNA, laboratory soil analysis, text from existing literature on soil science (Furey, Davis, & Seiter-Moser, 2019; Wang et al., 2019), qualitative soil information or farmers expert saying. In this perspective, observational data can be either a measurement, a recorded instrument reading or a human observation. In soil science, where the level of physical understanding is deep, it is unlikely to have a scientist collecting blindly a large amount of data with the objective of finding a potential pattern when, ultimately, these data are analysed. We rely instead on the opportunity to access and store legacy data from multiple sources. There is a direct link between the data collection method and the quality of inference that can be made (Kelling et al., 2009). Data from carefully designed experiments are the most suitable to corroborate a hypothesized causal relationship between variables. These data, however, should be used in data-driven scientific research with caution because the environment of their collection is highly manipulated and controlled. The difficulty to re-use data from a controlled experiment has long been noted in soil science. This was the case, for example, in the 1970s when applications of experimental (laboratory) results began in hydropedology. Uniform wetting recorded in an experimental setup had to give way to preferential flow in the uncontrolled soil environment (Warkentin, 1994). This was also noted by Dijkerman (1974): when the experimental models of the underlying soil processes are too simplified, either because our knowledge is incomplete or the technology precludes us from doing better, the application of experimental knowledge on soil is limited. Another limitation for the re-use of experimental data is that experiments are mostly made at a specific scale, generally local for field experiments (e.g. pedon scale in the field for soil warming experiments; Ettinger et al., 2019), or at horizon scale in the laboratory, which restricts their re-use for larger-scale data-driven modelling and the comparison of results between studies. Hypothesis-driven experiments are thus the most adequate to generate data that can corroborate causal relationships and explain a phenomenon, but use of these data in data-driven research to investigate soil processes should be made with more discretion than when using purely observational data, because only few of the components found in the natural system were allowed to vary in an experimental setup.

3.2 | Challenges for multi-source observational data collection

The use of observational data from different sources logically leads to some concerns, notably whether sufficient metadata are supplied together with the soil data.

Typical soil metadata are the soil depth or the pedogenetic horizon at or from which the soil observation has been made. Some important metadata in a data-driven context are the origin, description of the laboratory analysis protocol, identification code, ownership or even uncertainty (Heuvelink & Brown, 2006) of the measurement or analysis method. The metadata helps the scientist to evaluate the value of the information conveyed in the data, and to understand the assumptions and decisions that were made during data curation. In this sense, they provide useful information for exploratory analysis, to select a subset of the database for analysis or, perhaps more importantly, to account for possible bias that may have occurred during data collection or synthesis (Kelling et al., 2009). The use of guidelines and persistent digital object identifiers (Brase, 2009) enables a worthwhile picture of observational data available, associated gaps, and prepares the way for data-driven modelling. Several efforts for data collection have been made. Soil databases such as the Africa Soil Profiles Database (Leenaars, 2013) and the International Soil Carbon Network (ISCN; Harden et al., 2018), for example, have strict quality control procedures and require explicit information on the provenance and data sharing policy. The African database contains over 18,500 georeferenced legacy soil profile records for 40 sub-Saharan African countries, from 54 different sources. Users can access the data online by specific query. Similarly, the soil carbon database is a community-based data sharing effort. The data are required to be geo-referenced and go through three phases before being released. The data must be formatted by a template and pass a quality test where the provenance and methodology are checked. The data are released after a validation step made by the network members.

Historically, for example in the nineteenth century, the soil data producers were also the users. Data sharing was not essential for knowledge production because the collection of the soil samples in the field, and their analysis in the laboratory were performed by the same person in an institution. Both the soil material or the descriptions (drawing and pictures) and measurements of this material were amassed at a single geographic location. The soil samples as a physical (material) entity were stored in soil archives and transformed into a multitude of more convenient representations, such as images or summary digits of their composition. The centralization of the data enabled cataloguing, comparison and classification of soil into different types, as was done in a natural history museum for animal or plant species (Strasser & Edwards, 2017).

Crucially, present-day data users are most often different from those who produce them. In the last few decades, more divide appeared between the soil scientists who produce the data and those who analyse them, between the database producer and the data analyst. Present-day soil data are decomposed into multiple fragments of digital bits. The electronic nature of the database means that these bits of soil data can be analysed simultaneously at multiple locations, but also readily combined with different data sources, for example a satellite data stream.

To be sure, soil data have been shared for centuries through publications and conferences, re-used in different research works or combined through meta-analyses, but the increased circulation of electronic databases and the specialization of researchers who analyse them have made subsequent data-sharing problems more acute. Resources for data sharing exist but are limited by governmental policies and ethical concerns on the re-use of data by private companies, or inversely companies are reluctant to share data for which they made investments. This increase in database circulation also questions the centralized logic of data and knowledge production, inherited from the past when data were supplied and analysed in research institutions. Strasser and Edwards (2017) argued that this centralized logic is still prevailing today, but in a different form. Electronic databases, like soil data on pieces of paper in the past, have a physical reality in large computers able to store petabytes of data. These computers require specialized infrastructures found in large research institutions. The same research institutions are the ones possessing the largest soil data generation tools, such as spectrometers of all kinds, large-scale observational data collection (e.g. national-scale soil monitoring) or remote-sensing imagery. Similarly, most specialized researchers able to analyse the databases work at large private companies and a few research institutions. In this sense, the present-day logic of centralized soil databases is in substance the same as that of the past.

Combining a large amount of soil data increases the risk of joining data that have major incompatibilities or that altered from their initial use. Soil observations (e.g. pH) are obtained with specific spatial and temporal characteristics. The information they carry, and the statistical properties thereof, are relative to these specific spatial and temporal characteristics. Each soil datum provides a snapshot of the soil in space and time. Without sufficient metadata on the characteristics of the soil observation (e.g. confounding factors such as climate), the site-by-site comparison of soil observations or the use of the soil data for purposes other than those they were initially generated for, is to be made with caution. In large datasets, the problem of temporal replication is also important, i.e. do we have sufficient information in time to model a

dynamic soil property? An attempt was made by Heuvelink et al. (2020) on 5,000 legacy soil organic carbon data collected in Argentina between 1982 and 2007. Using a machine learning ensemble tree model, the organic carbon at different soil depths was modelled over the 36-year period. The study found that model accuracy was limited because of both the lack of appropriate environmental covariates and poor temporal and spatial coverage of the soil data. To fit user-demand monitoring of soil organic carbon change, the solution is found in the development of standardized soil sampling and monitoring schemes, such as that proposed by Batjes, Ribeiro, and Oostrum (2020).

However, while the creation of national and global datasets of all kinds of soil properties is now a pursuit for many communities of sub-disciplinary soil science, data accumulation and constitution of harmonized databases is never an end in itself. In fact, databases are useful only when the user makes them speak *qua* representation of the soil system. But who makes them speak? Undoubtedly, the ones who produced them are the main users: individual researchers and academics in research institutions and indirect "soil science practitioners" such as farmers or private companies. As Strasser and Edwards (2017) put it, the social landscape of science is changing. For several reasons, databases are increasingly made publicly available, which opens up avenues to citizen-assisted science. Undeniably, soil science has on several occasions benefited from a network of non-specialist volunteers. Rossiter, Liu, Carlisle, and Zhu (2015) detail some examples of such initiatives, such as mySoil for collecting soil observations using smartphones or the OPAL Soil and Earthworm Survey (Bone et al., 2012) to identify earthworm species and record earthworm density. In these examples, non-specialists contribute to building the soil database with various levels of engagement and expertise. Citizens not only amplify the work of scientists by adding their observations to the pool of data, but the novelty is that they increasingly act to build soil knowledge by performing non-routine soil data analysis. Perhaps another novelty in this contemporary data-driven soil science, corollary to the elements defined in Section 2, is the production of knowledge outside, both geographically and in terms of actors, of academic research institutions (Strasser & Edwards, 2017).

4 | THE RISK OF RADICAL EMPIRICISM

As a consequence of the increase in the use of computational tools and information technologies in soil research, discourse made in other fields, such as in data science, is being introduced into soil science. Most of the arguments are rooted in a radical form of nominalism called empiricism. Kitchin (2014b) describes three main propositions associated with empiricism in data-driven sciences: 1. data analysis is free of any theory, 2. the data speak for themselves, free of human bias, and 3. no domain knowledge is needed. In fact, the propositions of empiricism made explicit by Kitchin (2014b) can be challenged and represent extreme claims made by advocates of data-driven sciences. We discuss each of these propositions in the context of soil science in the following subsections.

4.1 | Theory-free analysis of data

The argument here is that analysis of data can be made free of any theory, model, hypothesis and scientific method. This was provocatively endorsed by Anderson (2008), who argued that knowledge production directly from data will lead to "the end of theory". The illusion that pattern discovery in data is free of any subjacent theory originates from marketing and retail (Kitchin, 2014a) but also appears in some empirical soil science studies. Automated soil mapping (Hengl et al., 2014) is an example of an attempt at theory-free discovery of insights from a large soil database. Automated soil mapping is the process of generating soil spatial information from a highly automated and flexible machine learning model, information which can be updated when new input data become available. The soil scientist does not need to build a reasoning or hypothesize a pattern in relation to environmental covariates; the algorithm based on machine learning searches for the optimal soil pattern using all the information provided by the scientist in an error-minimization procedure. The soil pattern is not constrained to any scientific law or tested against existing theories, but directly returned to the end user in a geographic information system.

Yet, the argument of theory-free data analysis from empirical epistemology does not hold for long in soil science. Any pattern discovery or data mining is guided by scientific reasoning or ontological assumptions. Data are never self-explanatory. The validity of analytics, algorithms or machine learning models has previously been approved and refined, and any model expresses a vision of how the pattern exists within a specific scientific approach. In this sense, machine learning models categorized under connectionist (e.g. neural network) or evolutionary (e.g. genetic programming) each represent a different viewpoint on the solution space partitioning (Domingos, 2015). In the earlier example on automated soil mapping, using a different machine learning algorithm would lead to discovering a different relationship

between covariates and point soil observations, and ultimately a different soil pattern. In short, discovery of patterns from data does not occur in a scientific vacuum (Kitchin, 2014a), it is always contextualized within a scientific domain. Discovering patterns relates to the "invention" of hypotheses from data (abductive reasoning; see Section 2) and is useful only if framed within the existing theories to infer knowledge by deduction. For example, Rasmussen et al. (2018) investigated which soil physico-chemical properties, other than clay-sized particles, may predict soil organic matter (SOM) stabilization. To do so, they built a large dataset of 5,500 soil profiles spanning different ecoregions, climate gradients and soil taxa, and used a regression model on independent soil variables. Their findings suggest that the relative importance of SOM stabilization mechanisms varies with climate and soil pH, and urge updating the set of variables used as predictors of SOM stability by most biogeochemical models. This "guided knowledge discovery" process, that is, the generation of a hypothesis to update existing theories, is common and has been later adopted by others (e.g. by Vos et al., 2019).

4.2 | The data speak for themselves, free of human bias

Data collection is a somewhat objective, but selective process (Dijkerman, 1974). The soil colour, moisture or texture, among other pieces of information, can be observed by a pedologist in a soil profile, but only the most significant information is used for the problem under study. To decide what is significant, the soil scientist uses a reservoir of pedological and tacit knowledge, that is, information and techniques learned through experience (Hudson, 1992). Similarly, when a computer model simulates soil organic carbon dynamics, the model is governed by some *a priori* knowledge because a number of environmental variables are supplied to the model. Data "cleaning", described in Section 3.2, is another step involving many assumptions and decisions. The data always "speak" within a context. In empirical epistemology, however, the pattern found in the data is sufficient evidence of the phenomenon under study, without the need for human contextualization. Kitchin (2014a) showed that this notion holds if two assumptions are met. The first is that data are neutral and generated without bias. The second is that associations between data are necessarily meaningful, which makes the human interpretation of these correlations irrelevant. Both of these claims are problematic in soil science. Indeed, data do not pre-exist by themselves and are always generated from a particular view and context, which is the sensor (including the human) or the observer's sense. As mentioned previously, data pass through subjective filters, called data "cleaning", before they are used in any process. The second assumption also does not hold because random association in environmental data is the rule rather than the exception. Interpreting random associations as meaningful may lead to flawed conclusions about the causes and determinism of a soil process. Soil scientists are well aware of this matter, as shown by the attempts to be alert to the risk of interpreting correlation as causation in the pattern found in soil data (Wadoux, Samuel-Rosa, Poggio, & Mulder, 2020). When seen through this lens, it is clear that the way data are selected, assembled and interpreted transmits rather than mitigates human bias, and determines the answers that the soil scientist obtains from the data.

4.3 | No pedological knowledge is needed

Soil science has a long tradition of research carried out by scientists whose first affiliation was not soil science but forestry, biology, chemistry or engineering (McDonald, 1994). These scientists have conceived soil science from their viewpoint and expert-specific contextual knowledge. In recent years, computer scientists, machine learning experts and data scientists became active in soil science, in particular in subdisciplines where data availability is important and the use of statistics is prevailing. Philip (1991), for example, noted in the 1990s the supplementation of laboratory and field experiments by computer modelling of data by "computer jockeys", which the author argued is inadequate for serious soil science research. The increase in computational studies in soil science was also indirectly noted, 20 years ago, by Hartemink, McBratney, and Cattle (2001). In 2001, the field of pedometrics and information systems represented about 20% of published papers in the *Geoderma* journal, compared to less than 5% in 1971. This increase came at the expense of qualitative soil genesis and morphological studies, which in parallel decreased rapidly in the same period. The same study showed that modelling/simulation was the second most important subject in 2001. No doubt that this trend is confirmed today with the increase in computer facilities that we have witnessed over the last two decades. The increase in soil modelling using databases and statistical or algorithmic tools has led to a number of models describing a phenomenon using soil data, which contain little or no soil science knowledge, and which may well be conducted in the absence of practising pedologists. In this regard, the increased availability of soil data poses some challenges,

as data can be accessed and analysed outside the context in which they have been produced.

A prototypical example is found in Padarian, Minasny, and McBratney (2019), where a global soil pattern is predicted using online machine learning tools and soil organic carbon data. As a means to avoid constraining data-sharing policies, an online platform enables soil modelling without accessing the data. The user is connected to the calibrated model, without information on the data from neighbouring parties involved in the modelling. In this vision, no pedological knowledge is required for modelling, which precludes any contextualization of the prediction or data. This raises the question: can anyone with a reasonable understanding of computational techniques contribute to soil science, without the need to contextualize the data, model and results with pedological knowledge? Hudson (1992) argued against this: it takes 2 to 3 years for a field soil scientist to internalize the soil landscape relationships. The possibility of having soil science studies carried out by non-soil scientists raised concerns, echoed for example by Basher (1997) or more recently by Walter, Lagacherie, and Follain (2006): relationships found in soil data can be misinterpreted or, perhaps more worryingly, not flagged as deserving more attention. In many ways, recent soil modelling studies, in particular those on digital soil mapping using machine learning, often bypass a significant body of earlier literature and represent a narrow contribution to the understanding of the processes that take place in the soil. The rhetoric of soil knowledge production by computational scientists is blurred by missing expertise that would ensure that data or models are contextually interpreted. A compromise, perhaps obvious but not less true, is found in the integration of soil expertise and computational scientists in multidisciplinary research projects.

5 | THE CHALLENGE OF FORMULATING EXPLANATIONS

By acknowledging that data-driven soil science harmonizes with the longstanding nominalist view of soil science, and that its pitfalls are rooted in empiricism, the question that logically follows is how do soil scientists obtain a scientific explanation from data? This section focuses on the use of modelling techniques (i.e. statistical and mathematical tools and models, data mining) to detect patterns (e.g. Jorda, Bechtold, Jarvis, & Koestel, 2015; Vos et al., 2019) in complex multivariate databases, or to provide a valid and generalizable representation of the reality, as a basis for making predictions (e.g. in digital soil mapping by Behrens et al., 2014). Soil scientists calibrate these models on data and use them to formulate explanations in light of the existing knowledge. With the increase in dataset size and complexity, there has been a parallel (seemingly related) increase in modelling complexity to mine these stores of data. This poses some challenges to the soil scientist. Usually, the objectives of analysing the data and modelling are 1. to obtain information and 2. to make predictions (Breiman, 2001). Highly complex models, insofar as they are not overfitted, are beneficial to detect a pattern and to make predictions, but the increase in modelling complexity has made it difficult for the soil scientist to obtain information from the model and thus explanations about the underlying structure of the soil system and process.

5.1 | Model complexity and the principle of parsimony

On a global meta-database of tension infiltrometer measurements, Jorda et al. (2015) used boosted regression trees, a machine learning algorithm, to identify the key environmental variables that determine saturated and near-saturated hydraulic conductivity in undisturbed soils. The authors found two different models, one with seven parameters and the other with five parameters, but almost equivalent in terms of accuracy. Using the principle of parsimony, they discarded the more complex model. The principle of parsimony recommends that from several competing models, one should select the simplest. In other words, the representation of the reality should be made as simple as possible. Parsimony is generally expressed in terms of number of adjustable parameters, but can also entail other criteria such as coherence, or the possibility to obtain insights from the model. In statistical modelling, parsimony is reached by a balance between model complexity and accuracy (how well the model agrees with the data). If the accuracy was the only criterion for selecting a model, the best model would reach each data point, thus containing many adjustable parameters, and be highly complex. Soil scientists usually refer to the parsimony principle when deciding which model to choose (e.g. Jorda et al., 2015; Lark, 2001; McBratney, Santos, & Minasny, 2003) using criteria such as the Akaike information criterion or the number of adjustable parameters. The simplicity of a model is seen as a desirable feature. Should it be?

A simple model has three main advantages (Gauch Jr, 2003). First, the model is easily interpretable and the relationships among variables can be understood by the user. Second, the number of simple models is generally much smaller than the number of complex models. This is related to Breiman's (2001) "Rashomon effect": there

exists an infinite number of complex models with similar solutions, but often few simple models or a single simple model. Third, a simple model is less flexible and hence more vulnerable when compared to new observations; it can be falsified or, perhaps as is often done in practice, refined when confronted with the natural system under study.

Faced with the desire to obtain explanations, soil scientists may refrain from building complex models. When using models on complex and large soil databases, one may claim that because the soil system is complex, then we also need complex models. We stress here that there are no logical or epistemological reasons to use parsimony as a guiding principle, nor can we affirm that a model is plausible if it is simple (Sober, 1990). The parsimony of models, and the ontological representation of simple soil structures do not foreshadow the complexity of the soil in nature (Gauch Jr, 2003). One should find the right balance between parsimony and how the model approaches the reality for the objective in hand (explanation of the underlying soil process or pattern detection/prediction). Parsimony only connects with model plausibility within a context and assumptions. There is thus no explicit demand for parsimony when exploring large datasets, other perhaps than when choosing a more likely and understandable model between several competing models, ideally all being consistent with the existing pedological knowledge.

5.2 | Correlation and causation

In their attempt to formulate explanations of the complexity of the soil system, scientists are interested in discovering mechanistic links between variables. Several models have been developed to investigate potential causal effects driving variation in soil properties. For instance, Angelini (2018) used structural equation modelling to understand the causes of variation in soil properties and to test hypotheses of these relationships built on previous knowledge. However, as the authors rightfully noted, there is discussion on whether a statistical model can truly reveal cause-effect relationships from observational data. In addition, causality in soil science is difficult to prove, especially as experimental confirmations of causality can never be fully established in a natural soil system. The first step in attempting to establish causation is to find an association between variables through correlation. The trained soil scientist is aware that association does not imply that a variable is mechanistically related to the change in another. Faced with the increase in dataset size, false correlation is yet more likely to occur. Calude and Longo (2017) have shown, for example, that the ratio between correlation and causation is a function of the sample size. In other words, as the sample size increases, the likelihood of finding a correlation among variables increases. The difference between correlation and causation becomes more difficult to discern.

In practice, however, soil scientists search for empirical correlations among data and use them as a heuristic to guide research and to develop models. Two types of models exist, the ones based purely on correlation among data, and the ones based on the current theories and known mechanistic links between variables. This is the fundamental distinction made by Jenny (1941) between state factor models (empirical, based on correlation) and process models (based on mechanisms). The calibrated models based on correlation are tested by comparing predictions and observations. If the model agrees with the observations it is said that the model is validated. Otherwise its structure can be refined or the model rejected. This is often done in practice, for example when building spectroscopic, mapping or hydrological models empirically from data, but a model is never entirely confirmed by observations (Oreskes, Shrader-Frechette, & Belitz, 1994); it agrees to a certain degree and can be partially validated using, for example, some quality of fit indices to measure the predictive accuracy. The higher the accuracy, the more faith is put in the representation of the reality made by the model, or in the scientific explanations that we formulate from it. This is what Hempel (1965) called the statistical-inductive explanation. Hempel's principles are close to the logic of abduction (Section 2), but instead of explaining a phenomenon with laws, it is explained by probabilities under statistical laws. The explanation of a phenomenon is successful if the statistical law confers a high probability to the description of the phenomenon to be explained. For example, if soil carbon (the phenomenon to be described) is predicted with high accuracy by a statistical model (the statistical law) of the soil organic carbon variation (the description of the phenomenon to be explained), then the model can be used to provide an inductive-statistical explanation of the soil organic carbon variation. Since a scientific explanation can be formulated from a model built on correlations among data, as is often done in practice, this means that the soil scientist can use the model to explain, model which can be refined (e.g. by retaining plausible correlations) as more knowledge of the system under study becomes available.

5.3 | Interpreting and explaining data-driven models

The previous section on model complexity suggests that to formulate an explanation from data, the soil scientist

is guided by the dilemma between accuracy and interpretability. A highly complex model often best emulates the underlying structure of the soil process but does not reveal readily how the prediction has been made. The exact relationship between input and output is obscure. This is the case, for example, in machine learning models composed of a large ensemble of decision trees (random forest), or artificial neural network models composed of potentially millions of adjustable parameters. These models are referred to as black box since the complexity of the inner relationships is beyond human understanding. Faced with this dilemma, nearly all soil scientists choose intepretability, for example to ensure the validity of the relationships found among data with existing scientific theories and laws (e.g. as in Häring, Dietz, Osenstetter, Koschitzki, & Schröder, 2012) or as in Behrens et al. (2014) to extract and reveal new knowledge on soil formation. Resuming the example on soil hydraulic conductivity from Section 5.1, Jorda et al. (2015) selected the parsimonious model composed of five parameters, and hence dropped two variables that had little effect on prediction performance. The two variables explained little, but were part of the soil hydraulic property variation, and there was a rational reason, perhaps mechanistic in nature, to keep them in the model. In this dilemma between accuracy and interpretability, soil scientists go for intepretability at the expense of model accuracy, but also at the expense of obtaining, perhaps unexpected, new information from the data.

Breiman (2001) argued that posing the objective of modelling as the dilemma between accuracy and interpretability is framing the wrong question. We also argued in this sense previously: there is no epistemological reason to formulate parsimonious models. Recall that the objective of modelling is to obtain information and to make predictions (see Section 5). Interpretability is only a means to obtain the information. Paraphrasing Breiman (2001), a model does not need to be simple to provide reliable information. In fact, complex models are often more accurate than simple models, and hence carry a better representation of the natural system under study; that is, we can put more faith in accurate models and use them to provide an inductive-statistical explanation of the soil system under study, according to Hempel (1965).

There is apparent contradiction here: complex models are more accurate than simple models and hence provide more reliable information, but if the information cannot be extracted (models are black boxes), do we really have an explanation? We have seemingly accepted the exaggeration that complex models, in particular machine learning models, are black boxes. These models do not evidently provide the same level of insight as simple models. A single decision tree is intuitive for the human but a stack of 500 trees is beyond human comprehension. It is not the scope of this study to give an account of all existing methods for model interpretation, but at the higher level, one might distinguish model-specific or model-agnostic methods. Both provide users with a set of techniques to interpret highly complex models. Another way to increase model interpretability is to move away from connectionist models (e.g. neural networks) and go towards evolutionist concepts where the emphasis is not only on prediction but also on model structure search and prediction (Beriro, Abrahart, & Diplock, 2014), for example gene expression programming (Ferreira, 2001), which reveals the model structure in the form of equations.

The following theoretical considerations in pedology illustrate this discussion. Pedologists usually use the two-term soil scheme initially presented by Dokuchaev (1883):

$$soil\ forming\ factors \rightarrow soils,$$

and explored later by Jenny (1941) (among others) in the form $S = f(clorpt)$, where S is the soil and the acronym *clorpt* stands for climate, organisms, relief, parent material and time, respectively. McBratney et al. (2003) and Grunwald (2009) considered soil properties or classes to be functions of external environmental factors such as elevation or age. McBratney et al. (2003) stressed that the approach is largely empirical and not theoretical. Causality of the factors on the soil property is not a prerequisite. The form of the empirical quantitative function f is flexible and can accommodate non-linear relationships. In the scheme *soil forming factors→soils* the only unknown is the form of the function f, which can be a complex model. In fact, the form of the function f is often unknown, such as in most machine learning black-box models. In this scheme, the pedologist usually attempts to select either an interpretable (simple) model (e.g. as in Behrens et al., 2014) or to devote all efforts to obtaining high accuracy (as in Hengl et al., 2014). We previously argued that the choice between accuracy and interpretability is not judicious. When the model is accurate, it is sensible to obtain information on the model that enables interpretation of the underlying soil structure and process. This opens avenues for obtaining new knowledge, for example by using the scheme later proposed by Gerasimov (1984) instead, in which soil-forming factors impact processes that are linked to soil:

$$soil\ forming\ factors \rightarrow process \rightarrow soils.$$

While the difference between the two schemes seems modest because both admit that soil is dependent on soil factors, the notion of process introduces evolutionary

(genetic) concepts and theoretical considerations. Characterizing the processes is possible when enough is known about the functional relationship of the factor to the soil, i.e. when information is obtained from an accurate model by means of interpretation. When more is known about the process, as per the solution found in the model and by the refinement of it, the complex model can gradually be replaced by a physical model of the soil consistent with the body of scientific laws. This type of approach has been proposed previously in the soil science literature (e.g. by Yaalon, 1975).

6 | CONCLUSIONS AND OUTLOOK

A series of incremental but rapid changes in electronic digital information and information technologies available for academic research have given the impression that research engaging large and diverse amounts of data is new to soil science. In this paper we have argued that while it may seem revolutionary for some, soil science has a long history of data-intensive and exploratory efforts to generate knowledge from soil data. This is precisely because soil science, like other natural sciences, started with an inventory of the various properties representing the diversity and complexity of the soil. Working in a data-rich environment is thus not new to soil scientists, but perhaps less so in subdisciplines of soil sciences (e.g. pedology/pedometrics) with a longstanding history of data collected from an observation of the uncontrolled environment rather than from a controlled and manipulated experiment.

Just as soil scientists were storing and classifying soil data on pieces of paper and in physical archives in the past, the available tools for present-day soil data storage are computers and electronic databases. These electronic databases are analysed simultaneously and remotely by multiple users, but for all that the logic of data storage at one location has not disrupted. The databases are still geographically centralized in large research institutions combining computer power, storage and large projects bringing a constant stream of new data. Perhaps the social landscape shows a sign of change. While citizens have much contributed to the collection of soil data in past years, the development of publicly available soil databases and the accessibility of software provided by personal computers increasingly encourage citizens to analyse data and be involved in knowledge production.

What is typically seen as revolutionary in contemporary data-driven science is to a large extent a change of magnitude in the amount of data collected and the capacity to analyse them with computational power. The soil data are currently generated rapidly, in large amounts and from multiple sources, leading to concerns on whether they can effectively be combined. The methods of analysing these data come to a large extent from the use of computer power and complex statistical and algorithmic solutions. In this "new" data-driven science, computer experts are playing a role in the production of soil knowledge by the analysis of data. It is no coincidence that corollary to this increase, discourse in other fields, such as in data science, is being made in soil science. These claims, which do not hold for long in soil science, are rooted in a radical form of empiricism expressing that soil data analysis can be made free of any theory, hypotheses or pedological knowledge.

In substance, there appears to be no major recent change in the way soil scientists obtain a scientific explanation from data. Finding correlations in soil data is at best a starting point, useful in an exploratory phase to generate hypotheses and to be used as heuristic to develop more realistic, mechanistic models based on causation when knowledge increases. In the quest for explanations on soil processes, soil scientists thus select simple models at the expense of accuracy. Perhaps by framing the problem differently, we may be tempted to use complex models, often more accurate than simple models. Complex models are more accurate and thus may provide a better representation of the natural system under study. This is a desirable feature when the data analysis technique aims to provide an inductive statistical explanation of a soil process.

The recent use (or abuse) of data-driven scientific research aided by formidable computer power has generated concerns to as a possible lack of new production of knowledge or, perhaps more worryingly, a lack of field and laboratory experience by young scientists. By analysing the epistemic challenges of the data-driven scientific research in the light of the historical literature, we found that there is a continuity of practices, some being certainly amplified by recent technological changes, but that the core methods of scientific enquiry from data, i.e. the scientific methods for knowledge production, remain largely unchanged.

Finally, we argue that this paper is an introductory and thus necessarily incomplete analysis of the current epistemological challenges of data-driven soil science. More research is needed, in particular to analyse whether data quantity makes a difference to the usual tools of scientific inquiry, such as sampling. Usually, the soil scientist selects a carefully-designed fragment of the environment, called a sample. When the dataset is large enough, or that nearly all possible data about a phenomenon have been assembled, do we still need to do sampling? This brings many questions, such as that of bias in data collection, or that of the value of the information

contained in data. Another area for future research is to analyse the nature of understanding. In which extent the users of data-intensive tools for data analysis (e.g. machine learning) obtain a scientific explanation from data. When a model returns information from data to the user, how to discriminate the phenomenological feeling of understanding, and the "epistemic" understanding (McCain, 2016), which forms the basis of scientific explanation? Assessing this difference would certainly make a useful contribution to future soil science studies.

ACKNOWLEDGEMENTS

We thank two anonymous referees for their helpful comments on an earlier draft of the manuscript. We are particularly grateful to one of them, who put incredible efforts into providing clear and insightful advice which was of great help in improving the manuscript.

AUTHOR CONTRIBUTIONS

Alexandre Wadoux: Conceptualization; writing-original draft; writing-review and editing. **Mercedes Román Dobarco:** Conceptualization; writing-review and editing. **Alex McBratney:** Conceptualization; writing-review and editing.

DATA AVAILABILITY STATEMENT

No data have been used in this study.

ORCID

Alexandre M. J.-C. Wadoux https://orcid.org/0000-0001-7325-9716
Mercedes Román-Dobarco https://orcid.org/0000-0001-8078-8616
Alex B. McBratney https://orcid.org/0000-0003-0913-2643

REFERENCES

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, *23*, 7–16.

Angelini, M. E. (2018) *Structural equation modelling for digital soil mapping*. PhD thesis, Wageningen University & Research, Wageningen.

Basher, L. R. (1997). Is pedology dead and buried? *Soil Research*, *35*, 979–994.

Batjes, N. H., Ribeiro, E., & Van Oostrum, A. (2020). Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth System Science Data*, *12*, 299–320.

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., & Scholten, T. (2014). Hyper-scale digital soil mapping and soil formation analysis. *Geoderma*, *213*, 578–588.

Beriro, D. J., Abrahart, R. J., & Diplock, G. (2014). Genetic programming: Magic bullet, poisoned chalice or two-headed monster?. In R. J. Abrahart & L. M See (Eds), *GeoComputation* (2nd ed., pp. 188–221). Boca Raton, FL: CRC Press.

Bone, J., Archer, M., Barraclough, D., Eggleton, P., Flight, D., Head, M., … Voulvoulis, N. (2012). Public participation in soil surveys: Lessons from a pilot study in England. *Environmental Science & Technology*, *46*, 3687–3696.

Brase, J. (2009) DataCite-A global registration agency for research data. In *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, pp. 257–261. IEEE.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*, 199–231.

Bui, E. N. (2016). Data-driven critical zone science: A new paradigm. *Science of the Total Environment*, *568*, 587–593.

Bui, E. N., Henderson, B., & Viergever, K. (2009). Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. *Global Biogeochemical Cycles*, *23*, GB4033.

Calude, C. S., & Longo, G. (2017). The deluge of spurious correlations in big data. *Foundations of Science*, *22*, 595–612.

Cline, M. G. (1963). Logic of the new system of soil classification. *Soil Science*, *96*, 17–22.

Dijkerman, J. C. (1974). Pedology as a science: The role of data, models and theories in the study of natural soil systems. *Geoderma*, *11*, 73–93.

Dokuchaev, V. V. (1883). *The Russian Chernozem. Report to the free economic society*. St. Petersburg: Imperial University of Saint Petersburg.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York, NY: Basic Books.

Duchaufour, P. (1963). Soil classification: A comparison of the American and the French systems 1. *Journal of Soil Science*, *14*, 149–155.

Elragal, A., & Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data*, *4*, 1–20.

Ettinger, A. K., Chuine, I., Cook, B. I., Dukes, J. S., Ellison, A. M., Johnston, M. R., … Wolkovich, E. M. (2019). How do climate change experiments alter plot-scale climate? *Ecology Letters*, *22*, 748–763.

Ferreira, C. (2001). Gene expression programming: A new adaptive algorithm for solving problems. *Complex Systems*, *13*, 87–129.

Furey, J., Davis, A., & Seiter-Moser, J. (2019). Natural language indexing for pedoinformatics. *Geoderma*, *334*, 49–54.

Gauch, H. G., Jr. (2003). *Scientific method in practice*. Cambridge: Cambridge University Press.

Gerasimov, I. P. (1984). The system of basic genetic concepts that should be included in modern dokuchayevian soil science. *Soviet Geography*, *25*, 1–14.

Gohau, G. (1992). Esprit déductif versus esprit inductif. *Aster*, *14*, 9–19.

Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. *Geoderma*, *152*, 195–207.

Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., et al. (2018). Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. *Global Change Biology*, *24*, e705–e718.

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T., & Schröder, B. (2012). Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. *Geoderma*, *185*, 37–47.

Hartemink, A. E. (2015). The use of soil classification in journal papers between 1975 and 2014. *Geoderma Regional*, *5*, 127–139.

Hartemink, A. E., McBratney, A. B., & Cattle, J. A. (2001). Developments and trends in soil science: 100 volumes of Geoderma (1967–2001). *Geoderma*, *100*, 217–268.

Hempel, C. G. (1965). *Aspects of scientific explanation*. New York, NY: Free Press.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., ... Walsh, M. G. (2014). SoilGrids1km—Global soil information based on automated mapping. *PLoS One*, *9*, e105992.

Heuvelink, G. B. M., Angelini, M. E., Poggio, L., Bai, Z., Batjes, N. H., van den Bosch, H., ... Olmedo, G. F. (2020). Machine learning in space and time for modelling soil organic carbon change. *European Journal of Soil Science*. https://onlinelibrary.wiley.com/doi/full/10.1111/ejss.12998.

Heuvelink, G. B. M., & Brown, J. D. (2006). Towards a soil information system for uncertain soil data. In P. Lagacherie A. B. McBratney & M. Voltz (Eds), *Developments in soil science* (Vol. 31, pp. 97–106). Amsterdam: Elsevier.

Hochachka, W. M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., & Kelling, S. (2007). Data-mining discovery of pattern and process in ecological systems. *The Journal of Wildlife Management*, *71*, 2427–2437.

Hudson, B. D. (1992). The soil survey as paradigm-based science. *Soil Science Society of America Journal*, *56*, 836–841.

Hughes, P. A., McBratney, A. B., Minasny, B., & Campbell, S. (2014). End members, end points and extragrades in numerical soil classification. *Geoderma*, *226*, 365–375.

Isbell, R. F. (1992). A brief history of national soil classification in Australia since the 1920s. *Soil Research*, *30*, 825–842.

Jenny, H. (1941). *Factors of soil formation: A system of quantitative pedology*. New York, NY: McGrawHill.

Jenny, H., & Leonard, C. D. (1934). Functional relationships between soil properties and rainfall. *Soil Science*, *38*, 363–382.

Johnston, A. E. (1994). The Rothamsted classical experiments. In A. E. Johnston & R. A. Leigh (Eds.), *Long-term experiments in agricultural and ecological sciences Wallingford, Oxon* (pp. 9–37). Wallingford: CABI International.

Jorda, H., Bechtold, M., Jarvis, N., & Koestel, J. (2015). Using boosted regression trees to explore key factors controlling saturated and near-saturated hydraulic conductivity. *European Journal of Soil Science*, *66*, 744–756.

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

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

Kitchin, R. (2014b). *The data revolution: Big data, open data, data infrastructures and their consequences*. Thousand Oaks, CA: SAGE.

Kornelsen, K. C., & Coulibaly, P. (2014). Root-zone soil moisture estimation using data-driven methods. *Water Resources Research*, *50*, 2946–2962.

Krasilnikov, P., Arnold, R. W. and Ibanez, J. J. (2010) Soil classifications: Their origin, the state-of-the-art and perspectives. In Gilkes, R. J., & Prakongkep, N. (Eds), *Proceedings of the 19th World Congress of Soil Science: Soil Solutions for a Changing World*, *Brisbane*, *Australia*, *1–6 August 2010. Symposium 1.4. 2 Soil Classification Benefits and Constraints to Pedology*, 19–22. International Union of Soil Sciences (IUSS).

Krasilnikov, P., Ibanez Marti, J. J., Arnold, R., & Shoba, S. (2009). *A handbook of soil terminology, correlation and classification*. London: Routledge.

Lark, R. M. (2001). Some tools for parsimonious modelling and interpretation of within-field variation of soil and crop systems. *Soil and Tillage Research*, *58*, 99–111.

Leenaars, J. G. B. (2013) Africa soil profiles database, version 1.1. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). *Tech. rep.*, Africa Soil Information Service (AfSIS), Wageningen, the Netherlands: ISRIC-World Soil Information.

Leonelli, S. (2014). What difference does quantity make? On the epistemology of big data in biology. *Big Data & Society*, *1*, 2053951714534395.

Leonelli, S., & Ankeny, R. A. (2012). Re-thinking organisms: The impact of databases on model organism biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*, 29–36.

McBratney, A. B., & Odeh, I. O. A. (1997). Application of fuzzy sets in soil science: Fuzzy logic, fuzzy measurements and fuzzy decisions. *Geoderma*, *77*, 85–113.

McBratney, A. B., Santos, M. M., & Minasny, B. (2003). On digital soil mapping. *Geoderma*, *117*, 3–52.

McCain, K. (2016). *The nature of scientific knowledge*. Berlin: Springer.

McDonald, P. (1994). *The literature of soil science*. Ithaca, NY: Cornell University Press.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, *50*, 181–201.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*, 449–461.

Minasny, B., & McBratney, A. B. (2013). Jenny, pca and random forests. *Pedometron*, *33*, 10–13.

Moon, D. (2005). The environmental history of the Russian steppes: Vasilii Dokuchaev and the harvest failure of 1891. *Transactions of the Royal Historical Society*, *15*, 149–174.

Moore, A. W., Isbell, R. F., & Northcote, K. H. (1983). Classification of Australian soils. In *Soils: An Australian viewpoint*. Glen Osmond: CSIRO Division of Soils.

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., ... Mouazen, A. M. (2016). Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosystems Engineering*, *152*, 104–116.

Northcote, K. H. (1971). *Factual key for the recognition of Australian soils*. Glenside: Rellim Technical Publications.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*, 641–646.

Padarian, J., Minasny, B., & McBratney, A. B. (2019). Online machine learning for collaborative biophysical modelling. *Environmental Modelling & Software*, *122*, 104548.

Pennock, D. J. (2004). Designing field studies in soil science. *Canadian Journal of Soil Science*, *84*, 1–10.

Philip, J. R. (1991). Soils, natural science, and models. *Soil Science*, *151*, 91–98.

Rasmussen, C., Heckman, K., Wieder, W. R., Keiluweit, M., Lawrence, C. R., Berhe, A. A., ... Wagai, R. (2018). Beyond clay: Towards an improved set of variables for predicting soil organic matter content. *Biogeochemistry*, *137*, 297–306.

Rossiter, D. G. (2018). Past, present & future of information technology in pedometrics. *Geoderma*, *324*, 131–137.

Rossiter, D. G., Liu, J., Carlisle, S., & Zhu, A.-X. (2015). Can citizen science assist digital soil mapping? *Geoderma*, *259*, 71–80.

Roudier, P., Ritchie, A., Hedley, C. and Medyckyj-Scott, D. (2015) The rise of information science: A changing landscape for soil science. In *IOP Conference Series: Earth and Environmental Science*, vol. 25, 012023. IOP Publishing.

Sepkoski, D. (2018). Data in time: Statistics, natural history, and the visualization of temporal data. *Historical Studies in the Natural Sciences*, *48*, 581–593.

Sober, E. (1990). Explanation in biology: Let's razor ockham's razor. *Royal Institute of Philosophy Supplements*, *27*, 73–93.

Strasser, B. J. (2012a). Collecting nature: Practices, styles, and narratives. *Osiris*, *27*, 303–340.

Strasser, B. J. (2012b). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*, 85–87.

Strasser, B. J., & Edwards, P. N. (2017). Big data is the answer... but what is the question? *Osiris*, *32*, 328–345.

Vos, C., Don, A., Hobley, E. U., Prietz, R., Heidkamp, A., & Freibauer, A. (2019). Factors controlling the variation in organic carbon stocks in agricultural soils of Germany. *European Journal of Soil Science*, *70*, 550–564.

Wadoux, A. M. J.-C., Samuel-Rosa, A., Poggio, L., & Mulder, V. L. (2020). A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science*, *71*, 133–136.

Walter, C., Lagacherie, P., & Follain, S. (2006). Integrating pedological knowledge into digital soil mapping. In P. Lagacherie, A. B. McBratney, & M. Voltz (Eds.), *Digital soil mapping: An introductory perspective* (pp. 281–615). Amsterdam: Elsevier.

Wang, D.-S., Liu, J.-Z., Zhu, A.-X., Shu, W., Zeng, C.-Y., et al. (2019). Automatic extraction and structuration of soil–environment relationship information from soil survey reports. *Journal of Integrative Agriculture*, *18*, 328–339.

Warkentin, B. P. (1994). Trend and developments in soil science. In P. McDonald (Ed.), *The Literature of Soil Science* (pp. 1–19). Ithaca, NY: Cornell University Press.

Webster, R. (1997). Regression and functional relations. *European Journal of Soil Science*, *48*, 557–566.

Webster, R. (2000). Is soil variation random? *Geoderma*, *97*, 149–163.

Yaalon, D. (1975). Conceptual models in pedogenesis: Can soil-forming functions be solved? *Geoderma*, *14*, 189–205.

How to cite this article: Wadoux AMJ-C, Román-Dobarco M, McBratney AB. Perspectives on data-driven soil research. *Eur J Soil Sci*. 2021;72: 1675–1689. https://doi.org/10.1111/ejss.13071