geographical analysis

Geographical Analysis (2019) **0**, 1–15

Special Issue

Geographic Data Science

Alex Singleton, Daniel Arribas-Bel

Department of Geography and Planning, University of Liverpool, Liverpool, L69 7ZT, U.K.

It is widely acknowledged that the emergence of "Big Data" is having a profound and often controversial impact on the production of knowledge. In this context, Data Science has developed as an interdisciplinary approach that turns such "Big Data" into information. This article argues for the positive role that Geography can have on Data Science when being applied to spatially explicit problems; and inversely, makes the case that there is much that Geography and Geographical Analysis could learn from Data Science. We propose a deeper integration through an ambitious research agenda, including systems engineering, new methodological development, and work toward addressing some acute challenges around epistemology. We argue that such issues must be resolved in order to realize a Geographic Data Science, and that such goal would be a desirable one.

Introduction

There has never been a time in history with more abundant geographic data, offering great potential for the spatially enabled social sciences to advance understanding of a plethora of human and environmental problems (Elwood, Goodchild, and Sui 2012; Miller and Goodchild 2015). Such data are being generated by many sources including established and new earth observation technologies; the miniaturized and expanded mobile sensing platforms of smart phones (Batty 2013); wider sensor networks as part of a developing Internet of Things or other technologies related to the quantified self (Wilson 2015); and the warehousing, linkage and modeling of public and private sector consumer interactions (Miller 2015). The advance of such enabling instrumentation and those data that they generate have expanded both where and when points of computation and data collection can occur. Much of the resulting "data deluge" (Miller 2010; Kitchin 2014a, b) within this context have properties that can be argued as differentiating these new forms of data from those that have traditionally been the concern of the social sciences and geographers in particular (e.g., short- and long-form surveys or Censuses). Collectively, these new sources have been termed "Big Data," and although there is an array of different definitions (Kitchin 2014a), those properties that are most generally ascribed include being huge in *volume*, with high *velocity* (e.g., real time) and having diversity in *variety* (unstructured or structured) (Laney 2001).

Correspondence: Alex Singleton, Geography and Planning, University of Liverpool, Roxby Building, 74 Bedford St S, Liverpool, L69 7ZT, U.K.
e-mail: alex.singleton@liverpool.ac.uk

Submitted: 6 March 2018; Revised version accepted: 7 March 2019

doi: 10.1111/gean.12194

Often conflated into the discussion of "Big Data" are those processes and techniques involved in turning these resources into insight and understanding. However, we would argue that such approaches should be more accurately referred to as "Data Science" (Schutt and O'Neil 2013; Donoho 2015; Peng and Matsui 2015), and that this distinction is important beyond simple naming conventions. While a challenge of coping with "Big Data" is arguably technological, and there is good reason to believe innovations in this area will reduce this burden, how we process, analyze, and deploy insights from "Big Data" gives rise to a larger set of more enduring epistemological and ontological debates that are already taking place (e.g., Kitchin 2014a).

Our main thesis is that there are clear synergies and benefits to be realized from intensifying our interactions with Data Science, and that these should be bidirectional in nature and thus have positive collective impact. Although some proposed that "Big Data" will enable geographers to build better models of human relationships and activities over space and time (González-Bailón 2013), we argue that for the realization of this vision there needs to be intensified critical engagement of Data Science by geographers; while also ensuring better articulation, and embedding of knowledge concerning the unique properties of space. The long interdisciplinary tradition that exists within Geography makes it particularly well positioned to facilitate such engagement. At the same time, further interaction with Data Science will bring new methodological tools that can help Geography, and the Geographical Analysis community, to remain relevant in an increasingly data-driven and digital world (Miller and Goodchild 2015; Ash, Kitchin, and Leszczynski 2018). To realize such a vision and to foster interaction, we propose the term of Geographic Data Science, as a site for critique, collaboration, and co-creation. As it relates to the main theme of the present special issue, the next 50 years of Geographical Analysis, we see Geographic Data Science as a vehicle to maintain and intensify the relevance of this community in greater scientific and industrial arenas. We make the case for the use of this term as complementary rather than supplementary of related subfields or methodological approaches such as Geographic Information Science, Quantitative Geography or Geocomputation, which we discuss in the section "Towards a Geographic Data Science."

We advance our argument in three stages. First we contextualize Data Science, focusing on its origins to better understand some of its current day coverage (and gaps). We then review the role of Geographic information and knowledge in the context of Data Science to argue for a growing relevance for and to Geography. Together, these two sections serve as the foundation for our proposal of a Geographic Data Science, which we elaborate by suggesting three different phases of interaction that may contribute to its creation, and finally conclude with some future prospects for research synergy.

"Big Data" deluge and the emergence of Data Science

It is difficult to trace the exact emergence of the term Data Science given diversity of its intellectual lineage and its relative nascency. The term is simultaneously used to refer to a set of statistical, computational, and analytical techniques and workflows; the set of interconnected tools developed with such applications in mind; as well as the particular epistemological perspective that sustains these practices. Within the context of this article, we will refer mostly to the first understanding, a set of techniques which, although common in other areas of science have seen little adoption in Geography. The second dimension is touched upon briefly in relation to building technical bridges between Data Science and Geography, while the last conceptualization is used to call upon further examination of the challenges it poses in the context of Geography.

What is clear however is that several disciplines claim ownership, with early references within both Computer Science (Naur 1974) and Statistics (Wu 1997; Cleveland 2001; Provost and Fawcett 2013). Data Science is also promoted widely by industry as *the solution* to the problem of making sense of and monetizing the increasing volumes of "Big Data" produced by computer-mediated systems (Kitchin 2014b; Varian 2014). Although an agreed definition does not exist to date, Loukides (2011) considers Data Science as "gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others." Extrapolating from such an all-encompassing definition implies a foundation in statistics and computer science, but also a firm grasp on software and database engineering, data/information visualization, and communication skills (Schutt and O'Neil 2013; Patil and Mason 2015). From an industry perspective, narratives associated with Data Science place clear emphasis on predictive modeling and building "data products," those whose very existence depends crucially on data (Loukides 2011). From a methodological standpoint, the statistical areas most stressed relate to techniques that, instead of imposing structure on the data ex ante (as was traditionally customary), rely on the amount of data to identify ("learn") such structure and more flexibly adapt to it, often providing better predictive performance. Such methods by themselves, however, are not what makes Data Science distinctive; in fact, many of the purported new Data Science techniques have lengthy history. It is their combination with "Big Data" that is reshaping landscapes both within industry and academia, and producing results that only a few years ago seemed within the realm of science fiction, from self-driving cars to personalized health applications. Every methodological turn is marked by a key distinctive characteristic. Within science, many of these shifts are linked explicitly to novel and differentiating technological advances that allow fields of research to evolve into distinctively new phases. In the case of Data Science, this is undoubtedly those methods and tools that make it possible to take full advantage of "Big Data."

Data Science presents a set of interconnected practices that have gained significant traction in the commercial sector. Although such developments are not limited to a single industry, and there are certainly numerous examples with a long history of generating large volumes of data (e.g., earth observation, finance, precision manufacturing, aerospatial engineering, logistics, etc), a significant contributor to the platforms and techniques of contemporary Data Science came from the activity of information technology companies. The Internet was one of the first platforms where the explosion of automated data production took place. Contrasting with traditional companies up to the late 1990s and early 2000s, the majority of their business, operations, and interactions with customers were almost entirely mediated through the web, thus enabling hitherto unseen potential to track rich details about activities of individual users. At some point during their development, such companies realized that the storage and creative use of these data held great value (Weinberger 2011); as both an asset for the business that was useful for enhancing valuations when seeking venture capital (Cassidy 2002), or as an operational resource (Vise and Malseed 2008) that enabled the streamlining or customization of services, consumer targeting (Petrison, Blattberg, and Wang 1997), or even the creation of entirely data-based products (Van Dijck 2013). Many of these developments were kept in-house, as they were (rightly) deemed as giving a competitive advantage, but early examples of such activities could be observed at firms including Google, Facebook, or Amazon (Rao and Scaruffi 2013). This revolutionary discourse of "Big Data" and Data Science has however been challenged (e.g., Barnes and Wilson 2014); for example, being argued by Dalton, Taylor, and Thatcher (2016) and Dalton and Thatcher (2015) as strategically beneficial to industry and often "black box."

Data Science and the production of geographic knowledge

Contemporary Data Science, as described in the previous section, emerged in a significant part as a collection of methods, tools, and supporting infrastructure to make sense of mostly non-geographical data derived through Internet activity. If considered at all, Geography would most typically be coarsely coded (country or city) through a connected device's unique IP address, and not necessarily at the forefront or integral to analysis. However, as documented in Arribas-Bel (2014) and Kitchin (2014b), many contemporary "Big Data" are generated by companies whose activities are also mediated digitally, but often have clear spatial and geographical dimensions to their operations. Furthermore, in many instances, the warehousing of such data has made it possible to link individuals to their associated attributes or events through historic records, thus creating not only highly detailed spatial but also temporal profiles (Miller 2015).

However, for Geography, there are two important considerations that emerge as Data Science is applied to geographic questions: firstly, of what or where are the underlying data representative; and secondly, how divergent is the extraction of knowledge within this context from more widely accepted epistemologies such as those emerging from Quantitative Geography, Geographic Information Science, or Geocomputation? Both pose important challenges for Data Science within this context, and learnings from Geography are well positioned to play a significant role in their resolution.

There are substantial issues related to the provenance of "Big Data" (Goodchild 2013) and the associated implications for computation, methodology, and interpretation (Gorman 2013). For example, "Big Data" are rarely raw (Gitelman 2013; Dalton and Thatcher 2014), given the extent to which such data (or, indeed, any data) can be considered as socially constructed (boyd and Crawford 2012). There are further issues related to how geographic features are encoded within "Big Data," with their geographic ontology being particularly vague (Goodchild and Li 2012). There are therefore a range of significant challenges around how more sophisticated understanding of Geography can be computed (Goodchild and Li 2012; Crampton et al. 2013; Miller 2015; Leszczynski and Crampton 2016), which have been a focus of GIScience since its inception (i.e., Goodchild 1991).

However, given that much Data Science is situated outside of Geography, there is increased risk within such contexts that location continues to be rationalized only as a supplementary column within a database, no more or less important than any other attribute. We argue that such effects are clearly counterproductive, and will be worrying for geographers; specifically given the large body of knowledge associated with the many unique properties of spatial data and the additional functionality they unlock that necessitate particular considerations in their analysis (e.g., Anselin 1989).

At the same time, those ways in which "Big Data" are turned into information challenge established epistemologies within the social sciences (Kitchin 2014a). This relates to an emphasis particularly within commercial Data Science on the "Fourth Paradigm" (Hey, Tansley, and Tolle 2009) which, taken to its extreme, is a held view that data in themselves are enough to extract knowledge and thus detached from theory (Miller 2015) or the consideration of process (O'Sullivan 2018). Within this context, the analysis of "Big Data" represents a shift away from carefully designed experiments with known sample sizes (Brunsdon 2014), the traditional approach of hypothesis testing, and the confirmation of exogenously stated theories through models that are carefully specified with relevant and rationalized attributes. To some extent, these issues reflect long-held tensions within Geography and social science more generally—between

idiographic (specification of the unique properties) and nomothetic (generalization and derivation of laws) forms of knowledge production (Schaefer 1953; Miller 2015). Although data-driven knowledge can be considered as idiographic (Miller 2015), Data Science does not represent a purely ideographic form of knowledge production; and often "Big Data" provide a rich yet incomplete representation of reality. Beyond such input, and core to many Data Science methods are various forms of explanatory models that can account for the characteristics of the input data that they are fed. As discussed earlier, such methods will typically seek to find rules and associations on the basis of input data, however, unlike many traditional mathematical or statistical frameworks, the exact specification of such rules are often determined endogenously by the technique (Gould 1981). This aspect also underlies one of the main methodological critiques of Data Science, in that models can become very sensitive to the original input data used for their specification, and this may not correspond to subsequent realizations.

The interplay between data, code, and the production of knowledge are typically integral to the curriculums of Geography programs that teach GIS, and it would be expected that most students would have a grounding of these fundamental issues by culmination of their studies (Johnston et al. 2014). However, this is not necessarily the case for the interdisciplinary area of Data Science, where many researchers and practitioners are drawn from a wider constituency of disciplines, and often outside of the social sciences. Given that much "Big Data" have locational attribution, our argument here is that Data Science should introduce critical geographical notions and reflection in a more fundamental way for these methods to build credibility within the social sciences. Indeed, Kitchin (2013: 264) notes there is a significant role for Geography within this context to "push back against naïve forms of predatory science," which is echoed by O'Sullivan and Manson (2015); and is a good example of what Sui and DeLyser (2012) call a "boundary project": the integrating of practices thought to be incompatible.

Geographers and other areas of cognate social science have historically had limited access to transactional (commercial/administrative) and more recently "Big Data" (Manovich 2011). Unsurprisingly, the necessary special considerations for their analysis have therefore had limited curriculum integration (Kitchin 2013; Johnston et al. 2014). This gives rise to the significant risk that Data Science applications become the preserve of the non-social sciences where there is technical training, but perhaps not the embedding of an epistemology that emphasizes the social and ethical considerations necessary for the analysis of socio-spatial problems (Ruppert 2013).

At the same time, GISc might be argued as having parallel tensions. From an instrumentation perspective, many curriculums have historically produced a body of GIS professionals where the focus of work processes are bound by specific GIS software platforms, including data creation, management, and representation. Gorman (2013) discusses how the rise of many new forms of (geographic) data gathered through social, mobile, and location applications have occurred external to GIS, and such software tools were not built to manage such large volumes of externally generated data. As a result, much of the GIS ecosystem has fragmented into multiple distributed but connected components that demand a wider set of skills than may traditionally have been acquired. Such issues are not only of concern to the spatial sciences (Hardin et al. 2015), and although progress is being made within this context (Bowlick, Goldberg, and Bednarz 2017), in order to stay relevant in a rapidly changing data economy, Geography must continue to embrace this shifting context, widen the base of skills taught, and encompass some of those contemporary approaches being developed within Data Science. Conversely, as Data Scientists move onto questions framed by location, space, and other geographical considerations, they will therefore run into similar issues as those that geographers have been dealing with (and proposing solutions

to) for decades. Unless an explicit action is taken, there is a clear risk of "reinventing the wheel," which would be counterproductive. Geography has the potential to help Data Science avoid this situation by bringing, literally and epistemologically speaking, the role of context and decades of experience with these questions. However, to realize this contribution to the Data Science community, Geography needs to be able to establish a common field where interaction and exchange with the disciplines and industries of Data Science and "Big Data" are encouraged and fostered.

Toward a Geographic Data Science

Geographic Information Science takes a critically reflective view on the application of computational methods to locational problems (Elwood 2008, 2010) and, in doing so, GISc is enriched by the breadth and depth of debates long held in Geography about competing perspectives, epistemological and ontological paradigms, and ethical considerations. In his seminal contribution, Goodchild (1991) defined the domain of GISc as a research agenda consisting of five distinct topics: spatial analysis and spatial statistics; theories of spatial relations; artificial intelligence and expert systems; visualization; and social, institutional, and economic issues. In later reflection on 20 years on the subdiscipline, Goodchild (2010) points out that, because it was considered more engineering than science, and despite earlier engagement (e.g., Couclelis 1986), the theme of artificial intelligence and expert systems were underrepresented within the ongoing NCGIA research at the time, and as such were removed from the more elaborate definition proposed in Goodchild (1992). To some extent, a related line of inquiry was taken up by academics in the Geocomputation sister field (e.g., Openshaw and Abrahart 1996; Openshaw and Openshaw 1997; Longley et al. 1998; Gahegan 1999), which bridged the "spatial analysis and spatial statistics" component of GISc with greater emphasis on the computational dimension (Fotheringham 1998; Brunsdon and Singleton 2015; Harris et al. 2017). These areas are, in a sense, all cognate of Quantitative Geography, a term with less traction today perhaps but which underpins several of the advances described in this context and we see also as one of the potential links to bridge Geography with Data Science.

It is interesting to consider such developments within a broader historical context of AI research. The so-called "AI winter" of the mid-80s (Hendler 2008), a period of discontent and disinvestment in artificial intelligence research, was at its peak when the GISc agenda was being formed. However, in the following two decades, the field has made significant leaps that have delivered progress in a wide range of fields of everyday life (Kitchin and Dodge 2011; Tenney and Sieber 2016) and academic research, from language translation, to autonomous transportation. Many core Data Science methods, and particularly those that have emerged from Computer Science, are in essence AI: they perform machine learning tasks that allow computers to make individual predictions and, in cases, decisions based upon them. This process can happen in an entirely automated way (subject to calibration), without human intervention and sometimes even in real time. Since these techniques rely heavily on the amount of input data fed into the model, one of the key factors responsible for this renaissance in AI has to do with the advent of "Big Data," which has made it possible to use similar techniques, yet obtain significantly superior results.

Geography has, for the most part, remained disconnected from many of these developments. While elements of the discipline (e.g., remote sensing) have engaged with several components of what is considered Data Science (e.g., image analysis), such interactions have taken place in a fragmented and indirect fashion. We argue that there should be a more orchestrated

cross-pollination between the two. We envisage that a productive way forward in this direction is to foster common spaces of interaction in what we could call a Geographic Data Science that effectively combines the long-standing tradition and epistemologies of Geographic Information Science and Geography with many of the recent advances that have given Data Science its relevance in an emerging "datafied" world.

There are various ways in which such integration might occur. In this context, we will sketch a process that moves from simple coupling of tools through assimilation of methods into a fully integrated Geographic Data Science. Coupling of tools refers to the linking of functionality from one platform into another, and within open source GIS and statistical platforms, this has become common practice. Through similar mechanisms, coupling of Data Science technologies with GIS features represents a productive start to expose both communities to the advantages that may emerge from engaging with each other. Indeed, this process has already been set in motion. For example, the two start-up companies Carto (mapping and cartography) and Plot.ly (statistical visualization) offer interfaces that allow the integration of their analytics tools into other common platforms. Alternatively, ESRI (www.esri.com) have developed tools that enable the ArcGIS platform to interact with Hadoop clusters (hadoop.apache.org), one of the industry standard platforms to store and process "Big Data."

Assimilation represents a further level of embeddedness of not only functionality but also those practices and methods surrounding the analytical process. Much the same way in which GIS approaches to data storage and query are now found within many other classes of software (e.g., spatially enabled databases), the tools of Data Science are also starting to assimilate elements that go beyond simple coupling and engage with both GIS and spatial analysis principles. For example, the Spatial Hadoop (spatialhadoop.cs.umn.edu) project integrates spatial analysis functionality into Hadoop, thus enabling data to be queried using spatial operators (e.g., distance or topology-based queries). Although such developments show great promise, it is important to highlight that more advanced insights and components of the GISc literature, such as spatial uncertainty, statistics or modeling, have received much less attention in this context so far. Both coupling and assimilation represent examples of bidirectional dissemination between Geography and Data Science. Such interaction offers tangible benefit, but we would argue it should only be the starting point for a more ambitious agenda where Geography as a discipline can influence the representation, analysis, and use of spatial "Big Data." In this context, the effects go beyond simply sharing best practice or exploring the utility of new tools from other fields. Geography has a long history of attracting scholars and their associated methodologies/epistemologies from over multiple disciplines (Agnew and Livingston 2011); and indeed outwardly contributing new methods and approaches (Warf and Arias 2009; Brunsdon and Singleton 2015). This provides an enviable meeting point for discussion and deeper integration, drawing on decades of interdisciplinary experience.

Progressively, we also argue that there is potential for the development of a new set of Geographic Data Science methods and tools, as well as their associated epistemological frameworks. Designing these with direct contributions from the Geography/GISc tradition and modern Data Science approaches would aspire to realize the full potential of spatial "Big Data" (Gorman 2013). In order to foster the debate, the remainder of this section presents a research agenda that suggests how and where integration could occur, and in particular those areas where challenges may emerge. We focus specifically on systems, methods, and established epistemology that can or do directly connect and extend nonspatial approaches that are current in Data Science, yet may currently be implemented to explore geographic phenomena without specific consideration

of those unique properties of space. However, this should be taken as a starting rather than end point for discussion and debate.

Systems engineering

The first component of this research agenda relates to core systems engineering, and includes the development of spatial databases and file formats that are explicitly designed to store, retrieve, and manipulate spatial "Big Data"; and secondly, how such spatial "Big Data" might be translated into information from these systems through visual display. The nature of spatial "Big Data" gives rise to specific challenges that warrant focused research on data structures. For example, efficiently integrating space and time at scale (Cheng 2012; Miller 2014; Rey 2014) or nonplanar representations of space such as spatial networks (Goodchild 2006; Barthélemy 2011; Okabe and Sugihara 2012). These all require a flexible ontology that is able to deal with a host of different types of geographic features and their conceptualizations. Much work within this area has progressed under the umbrella of Cyber GIS (Wang 2010, 2016; Wang et al. 2013; Evans et al. 2019); with some specific examples including alternative storage and transfer mechanisms (Lv, Rehman, and Chen 2013) or the development of new routing platforms (Shekhar et al. 2012). Developing core systems from first principles we argue holds the greatest potential, where explicit design can be embedded to account for those unique properties of spatial "Big Data." In some sense, we would expect this to follow similar advantages to those leveraged recently in other contexts, such as the creation of databases specifically designed to store, manage, and manipulate graph or network data. As the earlier presented definition described, there is a particularly strong focus within Data Science on the visual display of information, which also has parallel to the role that cartography plays in GIS. In both contexts, modern approaches have leveraged the advantages of computer-driven representations (e.g., Cheshire and Uberti 2014; Kirk 2016). As argued by Andrienko, Andrienko, and Weibel (2017), closer integration between Geography and Data Science could also infuse new developments in the area of infrastructure to support Exploratory Spatial Data Analysis (ESDA; Haining, Wise, and Ma 1998; Anselin 1999). This direction ensues a range of challenges around how spatial relationships (associations, significant clusters, etc.) can be identified and represented that go beyond the efficiency of applying techniques to large data sets (Andrienko, Andrienko, and Weibel 2017). For example, how to account for greater uncertainty in the underlying spatial data (Kinkeldey et al. 2015); to what extent traditional significance testing, for example, as it relates to spatial autocorrelation, is relevant in the context of very large samples; or how such approaches can be implemented in a real-time environment, where georeferenced data are conceptualized as a continuous flow, rather than as a large batch.

Modeling

Beyond the storage and visual representation of spatial "Big Data," there are clear opportunities to integrate various aspects of modeling as applied within Geography/GISc and Data Science. Many of the techniques widely used in Data Science come from a branch of computational statistics called machine learning (ML). ML is usually split into supervised and unsupervised methods. The former aim at identifying structure in the data without any form of previous instruction. There are clear precedents of unsupervised applications within Geography through, for example, geodemographic analysis (Singleton and Spielman 2014) or even explicitly spatial through regionalization and zone design (Openshaw 1977; Martin 1998; Duque, Ramos, and Suriñach 2007) which, in addition to statistical similarity, imposes geographic constraints to obtain the

resulting groupings. There are also a range of applications where established analysis techniques within Quantitative Geography have been reconfigured within the context of new infrastructure such as graphics processing unit architecture (Zhang, You, and Gruenwald 2014; Liang et al. 2015; Zhou et al. 2016; Tang and Feng 2017) or utilization of machine learning frameworks (Sun et al. 2015). All are a good example of areas of preexisting collaboration, however, there is potential and need to expand these interactions.

Data Science methods usually neglect location in their estimation, even when it is an important element of the problem at hand, trading apparent simplicity for potentially suboptimal outcomes. At the same time, explicitly spatial unsupervised learning, although promising, is very much in its infancy in terms of scalability to a point where it is a feasible option with "Big Data." Geographic Data Science would enhance advances at this intersection and enable innovative perspectives on long-standing questions and themes within Geography, such as the modifiable areal unit problem (MAUP; Openshaw 1984). Supervised learning, on the other hand, aims at building models and representations of phenomena that allow a machine to generate predictions in an automated fashion when new input data are presented to the model. The parallel with Geography in this context is less direct, although well-established approaches to integrate space in a regression context, such as spatial econometrics (Anselin and Rey 2014) or geographically weighted regression (Brunsdon, Fotheringham, and Charlton 1998) come closest. Although the main interest usually differs between Data Science (prediction), and Geography (explanation), also here there is scope for fruitful and productive interaction. The explicit inclusion of space in modeling contexts where it plays an important role improves predictive performance. To the extent that this is an almost unexplored field in Data Science, there are clear benefits to be realized in that respect. At the same time, some applications in Geography/GISc either require (e.g., small-area estimation) or could benefit from better predictive performance, which supervised learning is likely to ensue when combined with a formal representation of space. We would argue this is one of the most fruitful methodological areas where Geographic Data Science could comprehensively rework some of those core techniques of Data Science when considering problems associated with recorded attributes within spatial "Big Data."

Data-driven epistmology

Finally, we support the view that the practice of Data Science needs to be more effectively embedded within what Kitchin (2014a) terms a data-driven epistemology or Hey, Tansley, and Tolle (2009) describe as the "fourth paradigm" in Science. This is an approach that, grounded in scientific theories, extends their traditional approaches, adopting data and computation as an additional tool not only to test existing theories but also to develop new ones. In this respect, disregard of past scientific and academic practice, or a blind move into complete empiricism devoid of theory is undesired. Conversely, Geography has been argued as ill prepared theoretically for an era of "Big Data" (Kitchin 2013; Ruppert 2013).

Epistemological challenges that emerge are related to differences between some of the practice of Data Science vis-à-vis traditional social science. An example of this are modeling approaches implemented to predict an outcome effectively, but which use techniques whose inner predictive mechanisms are opaque and difficult to interpret; or cases where predictive analytics are deployed in real-world situations devoid of context or the social consequences of the decisions made by those models (O'Neil 2016). Such exercises are usually described as "black boxes," are less open to scrutiny or reproducibility (Singleton, Spielman, and Brunsdon 2016), and risk making poor decisions in terms of social justice and fairness. In the context of

a commercial production system which only requires a good prediction, this is not necessarily a source of concern; indeed Wyly (2014: 681) notes "[t]he capitalist correlation imperative is clear: spurious correlation is fine, so long as it is profitable spurious correlation." However, more acutely in the context of scientific inquiry, where process is as relevant, if not more, as the outcome, this can produce a significant and understandable backlash.

These debates are however not new (Shmueli 2010) and, as it is the case in the other elements we have highlighted, there is already important work taking place in this respect. In this particular area, there is much and interesting work being carried out in the nascent field of critical data studies, where several geographers are making active contributions (e.g., Leszczynksi and Crampton 2016; Zook 2017). We would argue that Geography, as a "discipline of disciplines" where different and often-confronting paradigms coexist, is well prepared to take an active role in advancing them toward more socially desirable outcomes. In this context, Geographic Data Science would closely align with core critical and ethical principles in this regard that have been advanced within Geography and, in particular, the subdisciplinary field of GISc. Furthermore, a Geographic Data Science would also act as a platform where the outcomes of these debates are more effectively disseminated across Data Science researchers and practitioners who, as covered above, are not necessarily aware of developments in the various fields of Geography. Such developments will be necessary to unlock all the potential in spatial "Big Data," without repetition of where Geographic research has already spent considerable effort (Schwanen and Kwan 2009; Barnes 2010).

Conclusions

This article considers the emergence of the interdisciplinary field of Data Science and critically examines the role that Geography and subdisciplinary approaches such as GISc can play in the development of new methodological and epistemological frameworks. The rapid expansion of instrumentation generating spatial "Big Data" generates clear research opportunities, but also significant challenges. We discuss how "Big Data" has spawned Data Science and how the field has evolved to consider ever more inherently geographic problems. However, this expansion has not been accompanied by an extension of the original methodological approaches and epistemological frameworks, potentially making its application to problems where location is key suboptimal. Given such disconnect, we make a case for closer and careful coupling and assimilation of the connected fields of Geography with Data Science, and provide some evidence that such practices are already taking place.

We argue strongly that there is substantial potential for the establishment of a Geographic Data Science within Geography, which provides a historical lineage of interdisciplinary working, and which we see as an important component of the next 50 years of the Geographical Analysis community. In this context, Data Science can benefit from the critically reflective perspective that Geography takes on new computational approaches to locational problems, as well as methodological contributions that better account for some of the key challenges in building models with spatial data. Such a relationship is and should be bidirectional in nature, since the discipline of Geography also has much to gain from Data Science, particularly in the methodological and technical aspects of working with "Big Data." We recognize the lineage of a Geographic Data Science would be closely related to Geocomputation, Geographic Information Systems and, in a broadest sense, Quantitative Geography and Geographical Analysis. But we also stress the need for a distinct Geographic Data Science, given the interdisciplinarity of this

endeavor; and, furthermore, the step change that the technological innovation of new forms of "Big Data" implies and requires methodologically to take full advantage. We conclude with a research agenda toward a Geographic Data Science that will emerge through deeper integration of the discipline of Geography and Data Science around three areas that include aspects of systems engineering, new methodological development, and work toward addressing some acute challenges of epistemology.

It is clear to us that there are benefits for this integration, both in practical terms of being able to implement more effective, ethical, and epistemologically robust analytics; but also, and importantly, in sustaining the relevance of Geography and subdisciplinary approaches within a rapidly changing socio-technological landscape. We concur with Graham and Shelton (2013: 259) when they state that "the futures of geography and big data are still to be made," and that there is still much exciting work to be done for a range of scholars with differing interests. To this end, we are firmly convinced there can only be positive outcomes from stronger interaction and cross-fertilization between the Geography and Data Science, and that this will strengthen our discipline and reaffirm its future relevance.

References

Agnew, J. A., and D. N. Livingston. (2011). *Geographical Knowledge*. London: Sage.

Andrienko, G., N. Andrienko, and R. Weibel. (2017). "Geographic Data Science." *IEEE Computer Graphics and Applications* 37(5), 15–7.

Anselin, L. (1989). "What is Special about Spatial Data? Alternative Perspectives on Spatial Data Analysis." In *Symposium on Spatial Statistics, Past, Present and Future*, 89-4. New York: Syracuse University.

Anselin, L. (1999). "Interactive Techniques and Exploratory Spatial Data Analysis." In *Geographical Information Systems: Principles, Techniques, Management and Applications*, 251–64, edited by P. Longley, M. Goodchild, D. Maguire and D. Rhind. New York: Wiley.

Anselin, L., and S. J. Rey. (2014). *Modern Spatial Econometrics in Practice: A Guide to GeoDa GeoDaSpace and PySAL*. AZ: GeoDa Press LLC.

Arribas-Bel, D. (2014). "Accidental Open and Everywhere: Emerging Data Sources for the Understanding of Cities." *Applied Geography* 49, 45–53.

Ash, J., R. Kitchin, and A. Leszczynski. (2018). "Digital Turn, Digital Geographies?" *Progress in Human Geography* 42(1), 25–43.

Barnes, T. J. (2010). "Taking the Pulse of the Dead: History and Philosophy of Geography, 2008–2009." *Progress in Human Geography* 34(5), 668–77.

Barnes, T. J., and M. W. Wilson. (2014). "Big Data, Social Physics, and Spatial Analysis: The Early Years." *Big Data & Society* 1(1), 1–14.

Barthélemy, M. (2011). "Spatial Networks." *Physics Reports* 499(1), 1–101.

Batty, M. (2013). "Big Data, Smart Cities and City Planning." *Dialogues in Human Geography* 3(3), 274–9.

Bowlick, F. J., D. W. Goldberg, and S. W. Bednarz. (2017). "Computer Science and Programming Courses in Geography Departments in the United States." *The Professional Geographer* 69(1), 138–50.

boyd, d., and K. Crawford. (2012). "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication, & Society* 15(5), 662–79.

Brunsdon, C. (2014). "Spatial Science – Looking Forward." *Dialogues in Human Geography* 4(1), 45–9.

Brunsdon, C., and A. D. Singleton. (2015). *Geocomputation: A Practical Primer*. London: Sage.

Brunsdon, C., S. Fotheringham, and M. Charlton. (1998). "Geographically Weighted Regression." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3), 431–43.

Cassidy, J. (2002). *Dot Con: The Greatest Story Ever Sold*. London: Harper Collins.

Cheng, T. (2012). "Guest Editorial: Integrated Spatio-Temporal Analysis and Data Mining." *GeoInformatica* 16(4), 623–4.

Cheshire, J., and O. Uberti. (2014). *London-the Information Capital*. London: Particular Books.

Cleveland, W. S. (2001). "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics." *International Statistical Review* 69(1), 21–6.

Couclelis, H. (1986). "Artificial Intelligence in Geography: Conjectures on the Shape of Things to Come." *Professional Geographer* 38(1), 1–11.

Crampton, J. W., M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, and M. Zook. (2013). "Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb." *Cartography and Geographic Information Science* 40(2), 130–9.

Dalton, C. M., L. Taylor, and J. Thatcher. (2016). "Critical Data Studies: A Dialog on Data and Space." *Big Data & Society* 3(1), 1–9.

Dalton, C. M., and J. Thatcher. (2014). "What Does A Critical Data Studies Look Like, and Why Do We Care?" *Society and Space*. Retrieved 5 July 2017 from http://societyandspace.org/2014/05/12/what-does-a-critical-data-studies-look-like-and-why-do-we-care-craig-dalton-and-jim-thatcher/

Dalton, C. M., and J. Thatcher. (2015). "Inflated Granularity: Spatial "Big Data" and Geodemographics." *Big Data & Society* 2, 1–15.

Donoho, D. (2015). *50 Years of Data Science Tukey Centennial Workshop*. Princeton, NJ. Retrieved 5 July 2017 from http://coursescsailmitedu/18337/2015/docs/50YearsDataSciencepdf

Duque, J. C., R. Ramos, and J. Suriñach. (2007). "Supervised Regionalization Methods: A Survey." *International Regional Science Review* 30(3), 195–220.

Elwood, S. (2008). "Volunteered Geographic Information: Future Research Directions Motivated by Critical Participatory and Feminist GIS." *GeoJournal* 72(3), 173–83.

Elwood, S. (2010). "Geographic Information Science: Emerging Research on the Societal Implications of the Geospatial Web." *Progress in Human Geography* 34(3), 349–57.

Elwood, S., M. F. Goodchild, and D. Z. Sui. (2012). "Researching Volunteered Geographic Information: Spatial Data Geographic Research and New Social Practice." *Annals of the Association of American Geographers* 102(3), 571–90.

Evans, M. R., D. Oliver, K. Yang, X. Zhou, R. Y. Ali, and S. Shekhar. (2019). "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities." In *CyberGIS for Geospatial Discovery and Innovation*, vol. 118, 143–70, edited by S. Wang and M. Goodchild. Dordrecht: GeoJournal Library, Springer.

Fotheringham, A. S. (1998). "Trends in Quantitative Methods II: Stressing the Computational." *Progress in Human Geography* 22(2), 283–92.

Gahegan, M. (1999). "Guest Editorial: What is Geocomputation?" *Transactions in GIS* 3(3), 203–6.

Gitelman, L. (2013). *Raw Data is an Oxymoron*. Boston, MA: MIT Press.

González-Bailón, S. (2013). "Big Data and the Fabric of Human Geography." *Dialogues in Human Geography* 3(3), 292–6.

Goodchild, M. F., and L. Li. (2012). "Formalizing Space and Place." In *Fonder les Sciences du Territoire*, 83–94, edited by P. Beckouche, C. Grasland, F. Gue´rin-Pace and J.-Y. Moisseron. Paris: Editions Karthala.

Goodchild, M. F. (1991). "Geographic Information Systems." *Progress in Human Geography* 15(2), 194–200.

Goodchild, M. F. (1992). "Geographical Information Science." *International Journal of Geographical Information Systems* 6(1), 31–45.

Goodchild, M. F. (2006). "GIScience Ten Years after Ground Truth." *Transactions in GIS* 10(5), 687–92.

Goodchild, M. F. (2010). "Twenty Years of Progress: GIScience in 2010." *Journal of Spatial Information Science* 1, 2–20.

Goodchild, M. F. (2013). "The Quality of Big (Geo)Data." *Dialogues in Human Geography* 3(3), 280–4.

Gorman, S. P. (2013). "The Danger of a Big Data Episteme and the Need to Evolve Geographic Information Systems." *Dialogues in Human Geography* 3(3), 285–91.

Gould, P. (1981). "Letting the Data Speak for Themselves." *Annals of the Association of American Geographers* 71(2), 166–76.

Graham, M., and T. Shelton. (2013). "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography* 3(3), 255–61.

Haining, R., S. Wise, and J. Ma. (1998). "Exploratory Spatial Data Analysis." *Journal of the Royal Statistical Society: Series D (The Statistician)* 47(3), 457–69.

Hardin, J., R. Hoerl, N. J. Horton, D. Nolan, B. Baumer, O. Hall-Holt, P. Murrell, R. Peng, P. Roback, D. Temple Lang, and M. D. Ward. (2015). "Data Science in Statistics Curricula: Preparing Students to 'Think with Data'." *The American Statistician* 69(4), 343–53.

Harris, R., D. O'Sullivan, M. Gahegan, M. Charlton, L. Comber, P. Longley, C. Brunsdon, N. Malleson, A. Heppenstall, A. Singleton, D. Arribas-Bel, and A. Evans. (2017). "More Bark than Bytes? Reflections on 21+ Years of Geocomputation." *Environment and Planning B* 44(4), 598–617.

Hendler, J. (2008). "Avoiding Another AI Winter." *IEEE Intelligent Systems* 23(2), 2–4.

Hey, T., S. Tansley, and K. M. Tolle. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research.

Johnston, R., R. Harris, K. Jones, D. Manley, C. E. Sabel, and W. W. Wang. (2014). "Mutual Misunderstanding and Avoidance Misrepresentations and Disciplinary Politics: Spatial Science and Quantitative Analysis in (United Kingdom) Geographical Curricula." *Dialogues in Human Geography* 4(1), 3–25.

Kinkeldey, C., A. M. MacEachren, M. Riveiro, and J. Schiewe. (2015). "Evaluating the Effect of Visually Represented Geodata Uncertainty on Decision Making: Systematic Review, Lessons Learned and Recommendations." *Cartography & Geographic Information Science* 44(1), 1–21.

Kirk, A. (2016). *Data Visualisation: A Handbook for Data Driven Design*. London: Sage.

Kitchin, R. (2013). "Big Data and Human Geography: Opportunities, Challenges and Risks." *Dialogues in Human Geography* 3(3), 262–7.

Kitchin, R. (2014a). "Big Data New Epistemologies and Paradigm Shifts." *Big Data & Society* 1(1), 1–12.

Kitchin, R. (2014b). *The Data Revolution: Big Data Open Data Data Infrastructures and Their Consequences*. London: Sage.

Kitchin, R., and M. Dodge. (2011). *Code/Space: Software and Everyday Life*. Boston, MA: MIT Press.

Laney, D. (2001). 3D Data Management: Controlling Data Volume Velocity and Variety. META Group Research Note 670.

Leszczynski, A., and J. Crampton. (2016). "Introduction: Spatial Big Data and Everyday Life." *Big Data & Society* 3(2), 1–6.

Liang, Y., H. P. Huynh, K. Rupnow, R. S. M. Goh, and D. Chen. (2015). "Efficient GPU Spatial-Temporal Multitasking." *IEEE Transactions on Parallel and Distributed Systems* 26(3), 748–60.

Longley, P. A., S. M. Brooks, R. McDonnell, and W. D. Macmillan. (1998). *Geocomputation: A Primer*. Chichester, U.K.: John Wiley & Sons.

Loukides, M. (2011). *What is Data Science?* Sebastopol, CA: O'Reilly Media Inc.

Lv, Z., S. U. Réhman, and G. Chen. (2013, November). "Webvrgis: A p2p Network Engine for VR Data and GIS Analysis." In International Conference on Neural Information Processing, 503–10. Berlin, Heidelberg: Springer.

Manovich, L. (2011). *Trending: The Promises and the Challenges of Big Social Data*. Retrieved 5 July 2017 from http://www.manovich.net/DOCS/Manovich_trending_paper.pdf

Martin, D. (1998). "Optimizing Census Geography: The Separation of Collection and Output Geographies." *International Journal of Geographical Information Science* 12(7), 673–85.

Miller, H. J. (2010). "The Data Avalanche is Here. Shouldn't We Be Digging?" *Journal of Regional Science* 50(1), 181–201.

Miller, H. J. (2014). "Space-Time Data Science for a Speedy World." *Journal of Law and Policy for the Information Society* 10(3), 10705–20.

Miller, H. J. (2015). "Spatio-Temporal Knowledge Discovery." In *Geocomputation: A Practical Primer*, 97–109, edited by C. Brunsdon and A. D. Singleton. London: Sage Publications.

Miller, H. J., and M. F. Goodchild. (2015). "Data-Driven Geography." *GeoJournal* 80(4), 449–61.

Naur, P. (1974). *Concise Survey of Computer Methods*. Sweden: Studentlitteratur.

Okabe, A., and K. Sugihara. (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. Chichester, U.K.: John Wiley & Sons.

Openshaw, S., and R. J. Abrahart. (1996). "Geocomputation." In Proceedings of the 1st International Conference on GeoComputation, 665–6, edited by R. J. Abrahart. Leeds, U.K.: University of Leeds.

Openshaw, S. (1977). "Optimal Zoning Systems for Spatial Interaction Models." *Environment and Planning A* 9(2), 169–84.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich: Geobooks.

Openshaw, S., and C. Openshaw. (1997). *Artificial Intelligence in Geography*. Chichester: Wiley.

O'Neil, C. (2016). *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Penguin Random House LLC.

O'Sullivan, D., and S. M. Manson. (2015). "Do Physicists have 'Geography Envy'? And What Can Geographers Learn From It?" *Annals of the Association of American Geographers* 105(4), 704–22.

O'Sullivan, D. J.. (2018). "Big Data: Why (oh Why?) this Computational Social Science?" In *Thinking Big Data in Geography: New regimes, New Research*, 21–38, edited by J. Thatcher, J. Eckert and A. Shears. Nebraska, NE: University of Nebraska Press.

Patil, D. J., and H. Mason. (2015). *Data Driven: Creating a Data Culture*. Sebastopol, CA: O'Reilly Media Inc.

Peng, R., and E. Matsui. (2015). *The Art of Data Science*. Victoria: Leanpub.

Petrison, L. A., R. C. Blattberg, and P. Wang. (1997). "Database Marketing: Past Present and Future." *Journal of Interactive Marketing* 11(4), 109–25.

Provost, F., and T. Fawcett. (2013). "Data Science and its Relationship to Big Data and Data-Driven Decision Making." *Big Data* 1(1), 51–9.

Rao, A., and P. Scaruffi. (2013). *A History of Silicon Valley: The Largest Creation of Wealth in the History of the Planet; 1900–2013*. Palo Alto: Omniware Group.

Rey, S. J. (2014). "Spatial Dynamics and Space-Time Data Analysis." In *Handbook of Regional Science*, 1365–83, edited by P. Nijkamp and M. Fischer. Berlin, Heidelberg: Springer.

Ruppert, E. (2013). "Rethinking Empirical Social Sciences." *Dialogues in Human Geography* 3(3), 268–73.

Schaefer, F. K. (1953). "Exceptionalism in Geography: A Methodological Examination." *Annals of the Association of American Geographers* 43(3), 226–49.

Schutt, R., and C. O'Neil. (2013). *Doing Data Science: Straight Talk from the Frontline*. Sebastopol: O'Reilly Media Inc.

Schwanen, T., and M. Kwan. (2009). "'Doing' Critical Geographies with Numbers." *The Professional Geographer* 61(4), 459–64.

Shekhar, S., V. Gunturi, M. R. Evans, and K. Yang. (2012). "May Spatial Big-Data Challenges Intersecting Mobility and Cloud Computing." In Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access, 1–6. ACM.

Shmueli, G. (2010). "To Explain or Predict?" *Statistical Science* 25(3), 289–310.

Singleton, A. D., and S. E. Spielman. (2014). "The Past Present and Future of Geodemographic Research in the United States and United Kingdom." *The Professional Geographer* 66(4), 558–67.

Singleton, A. D., S. Spielman, and C. Brunsdon. (2016). "Establishing a Framework for Open Geographic Information Science." *International Journal of Geographical Information Science* 30(9), 1507–21.

Sui, D., and D. DeLyser. (2012). "Crossing the Qualitative-Quantitative Chasm I. Hybrid Geographies, the Spatial Turn, and Volunteered Geographic Information (VGI)." *Progress in Human Geography* 36(1), 111–24.

Sun, Z., F. Chen, M. Chi, and Y. Zhu. (2015, August). "A Spark-Based Big Data Platform for Massive Remote Sensing Data Processing." In International Conference on Data Science, 120–6. Cham: Springer.

Tang, W., and W. Feng. (2017). "Parallel Map Projection of Vector-Based Big Spatial Data: Coupling Cloud Computing with Graphics Processing Units." *Computers, Environment and Urban Systems* 61, 61187–97.

Tenney, M., and R. Sieber. (2016). "Data-Driven Participation: Algorithms, Cities, Citizens, and Corporate Control." *Urban Planning* 1(2), 101–13.

Van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford: Oxford University Press.

Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." *The Journal of Economic Perspectives* 28(2), 3–27.

Vise, D. A., and M. Malseed. (2008). *The Google Story: For Google's 10th Birthday*. London: Random House Digital Inc.

Wang, S. (2016). "CyberGIS and Spatial Data Science." *GeoJournal*, 81(6), 965–68.

Warf, B., and S. Arias. (2009). *The Spatial Turn*. Abingdon: Routledge.

Wang, S. (2010). "A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis." *Annals of the Association of American Geographers* 100(3), 535–57.

Wang, S., L. Anselin, B. Bhaduri, C. Crosby, M. F. Goodchild, Y. Liu, and T. L. Nyerges. (2013). "CyberGIS Software: A Synthetic Review and Integration Roadmap." *International Journal of Geographical Information Science* 27(11), 2122–45.

Weinberger, D. (2011). *Too Big to Know*. Nova Iorque: Basic Books.

Wilson, M. (2015). "Flashing Lights in the Quantified Self-City Nation." *Regional Studies Regional Science* 2(1), 39–42.

Wu, J. (1997). *Statistics = Data Science?* Retrieved 5 July 2017 from http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf

Wyly, E. (2014). "Automated (Post)Positivism." *Urban Geography* 35(5), 669–90.

Zhang, J., S. You, and L. Gruenwald. (2014). "High-Performance Spatial Query Processing on Big Taxi Trip Data Using Gpgpus." In 2014 IEEE International Congress on Big Data (BigData Congress), June, 72–79. IEEE.

Zhou, E., S. Mao, M. Li, and Z. Sun. (2016). "PAM Spatial Clustering Algorithm Research Based on CUDA." In 2016 24th International Conference on Geoinformatics,1–7. IEEE.

Zook, M. (2017). "Crowd-sourcing the Smart City: Using Big Geosocial Media Metrics in Urban Governance." *Big Data & Society* 4(1), 2053951717694384.