1

Data Science and Human-Environment Systems

Transformation of the Earth's social and environmental systems is happening at an incredible pace. The global population has more than doubled over the last five decades, while food and water consumption has tripled and fossil-fuel use quadrupled. Attendant benefits such as longer lifespans and economic growth are increasingly joined by corresponding drawbacks, including mounting socioeconomic inequality, environmental degradation, and climate change. Over the past half-century, interregional differences in population growth rates, unprecedented urbanization, and international migration have led to profound shifts in the spatial distribution of the global population. Economic changes have been dramatic as well. The global per-capita gross domestic product doubled while economic disparities grew in many regions (Rosa et al. 2010).

These socioeconomic shifts have affected a host of natural systems and ecosystem services. Demographic shifts and economic development are distal causes of proximate drivers of environmental change, such as fossil-fuel emissions and land-cover change. These changes affect many natural systems, including land-cover composition, soil and water quality, climate regulation and temperature, and vegetation and animal communities. These environmental dynamics have profound implications for human well-being. Flooding, erosion of coastal areas, and drought already affect human societies in many ways, and these effects will grow sharply in coming decades. These shifts are all facets of interlinked human-environment systems that arise from complex interactions among individuals, society, and the environment (Ehrlich et al. 2012).

Can data science help address human-environment challenges? Scientific and policy bodies have called for more and better data and attendant analyses to support the research needed to meet the impacts of rapid human-environmental change (Millett & Estrin 2012). Socioeconomic, demographic, and other social data that can be closely integrated with Earth systems data are essential to describing the continuously unfolding transformation of human and ecological systems (Holm et al. 2013). Of particular interest is big data, or data sets that are larger and more difficult to handle than those typically used in most fields, and data science, the larger field concerned with big data and analysis.

Data science offers advances in processing and analysis for research and policy development. Special issues in leading journals like *Science* and *Nature* highlight the need for new data and methods to help answer a wide array of questions at the intersection of nature and society (Baraniuk 2011). National scientific bodies such as the US National Academy of Sciences, United Kingdom's Royal Society, European Science Foundation, and Chinese Academy of Sciences have issued high-profile calls to develop and use big data to understand and address scientific and policy challenges stemming from human-environment interactions. We also see the advent of specialized journals, such as *Big Data Earth* and the *International Journal of Digital Earth*, that focus on large human-environment data sets.

Researchers and policymakers see data science's promise and pitfalls for human-environment systems. The move toward analyzing vast new data sets redefines disciplines that range from physics to economics to Earth sciences. These data are gleaned from a host of new sensors, internet activities, and the merging of existing databases. At the same time, some of the initial hype around data science and big data has been tempered by how this work plays out in real-world contexts. The fast growth of some forms of data has highlighted the considerable gaps in other kinds. Humans have studied only a tiny part of the world's oceans or a fraction of the millions of species on the Earth's surface. There are also significant gaps in data on people and society over much of the globe. Human-environment data pose many significant unresolved methodological challenges because they represent complex social and environmental entities and relationships that span multiple organizational, spatial, and temporal levels (Kugler et al. 2015). Data science also faces many unsolved challenges around theory development and myriad policy dimensions. Even as vast databases become more readily accessible and tractable, many problems have yet to be addressed, and much of the promise of big data remains just that – a promise unfulfilled.

1.1 Data Science and Human-Environment Research

There is broad interest in using big data for understanding human-environment interactions and attendant issues – including climate change, natural hazards, ecosystem services, and sustainability. This volume brings together these various research streams while assessing the pros and cons of data science for human-environment scholarship. It draws on various sources but focuses almost exclusively on peer-reviewed research. The goal here is to bridge various camps of scholarly work on big data and data science for human-environment systems. Big data and data science are here to stay; maybe not in their current incarnation, but certainly in some form. Addressing the toughest human-environment issues requires scholars to work together across fields. This list includes (and is not limited to) data scientists, statisticians, and computer scientists; domain scientists working on social, environmental, and natural systems; and scholars in policy and law, and arts and humanities.

The nature of global environmental change and other human-environment topics is one of vast spatial and temporal scales in some ways and the hyperlocal in others. One need only look to action around climate change to see how global social and environmental systems are inextricably linked to individual behavior. These incredible scale shifts mean we deal with a vast range of data, methods, and theories across research domains. Scholars also deal with problems that do not neatly fall along human or environmental lines.

> Given these pervasive scale-related problems and the inherent complexity they create, it is not surprising that inter-disciplinary and trans-disciplinary research are both seen as necessary; the problems of global change transcend conventional disciplinary inquiry. Global change is often treated largely as an environmental problem, but the environment is not simply an ‘‘independent variable’’; indeed, global change is a consequence of social processes. *(Pahl-Wostl et al. 2013, p. 40)*

In simple terms, human-environment research is not the domain of any single research field. Doing this research well requires a deliberate commitment to boundary-crossing and integrated scholarship.

Data science is making deep inroads into many kinds of scholarship on human-environment topics, but the literature is splintered. Some of the most extensive work centers on big data (Section 1.2 dives into definitions of big data), primarily focused on providing wide-ranging and generic overviews. These are often trade books that cite primarily from the gray literature or nonpeer-reviewed blogs and web pages. Increasingly these works include research perspectives as data science has ramped up over the past decade. These resources often have an exuberant bent that is driven by just-so case studies that capture the attention of mass media. This large and general body of work directly (or often indirectly) reflects how big data is big business. Data science is vital to a growing array of economic sectors. This commercial success results from big data and data science, which means they are often couched with an optimistic viewpoint with a mercenary perspective at its core. Much of the early writing on big data was commercial, and the authors were understandably looking to sell their products (Wyly 2014).

Much of the early work in big data and data science relied on nonscholarly and nonpeer-reviewed sources. References to blog posts, web pages, and gray literature abound. Informal and nonpeer-reviewed sites will always be essential venues of information on rapidly emerging issues in technology since more deliberate and careful research and subsequent publications can require years. Apart from not being peer-reviewed, the major drawback of these sites is that they too often disappear. For example, the site www.bigdata-startups.com is cited by dozens of academic papers as a source of crucial information; however, it no longer exists beyond partial and fragmented backups in internet archives. Another example is the work of McKinsey & Company, a management consulting firm. This significant proponent of big data published well-cited work at the now-defunct website www.mckinseyonsociety.com, and its articles only live on as informal copies and references.

Scholarly work in data science and big data has proliferated over the past decade. This work falls into several camps and reflects the rapidity with which data science and big data worked their way into the arenas of science agenda setting, funding, and publication. Academia has always been as prone as any other human endeavor to embrace fads, fashions, and folderal (Dunnette 1966). The rapid embrace of all-things-data is driven in part by fashion, but it is also clear that data science approaches work well for many questions, even when there is room for improvement with others. As explored in later chapters, there are also deeper issues in how scholars can, or should, engage with these approaches. This book speaks to many communities in the hope of helping bring them together around a robust data science of human-environment systems.

Social scientists and humanities scholars have long been interested in nature and human-environment relationships. However, the recent increased visibility of human well-being, climate change, environmental justice, ecological resilience, and sustainability have rapidly expanded social science research on the environment. We are also seeing an increase in digital and environmental humanities, areas with an interest in data science as both a methodology and a subject of critical study. Social science and humanities scholarship comprises a large and growing body of perspectives on big data. The majority of this work critiques big data and its role in specific application areas, such as cities or policing, or from a specific perspective, especially in science and technology studies. There is also scholarship, still in the minority, that offers grounded accounts of the promise and drawbacks of big data for particular scientific and policy domains.

Earth, planetary, ecological, and natural scientists have embraced the study of the Earth as an integrated human-environment system. The physical, chemical, and biological impacts of human activities in the Anthropocene have taken on planetary import (Ruddiman 2013). The transdisciplinary field of Earth-system science focuses on ocean, land, and atmosphere processes, recognizing that changes in the Earth result from complex interactions among these Earth systems and human systems. Ecological, natural, and Earth sciences research with data science tends to center on fairly narrowly defined areas of interest, such as using remote sensing for climate change research or geospatial data to study animal movement. In keeping with environmental scientific publishing in general, this work is usually shared via articles, but a growing number of books, predominantly edited volumes, focus on specific research questions.

Information, data, and computer scientists perform much big data research. Many articles and editorials by these researchers call for greater engagement with domain experts to advance big data. One of the goals of this book is to offer these scholars an overview of significant challenges and opportunities in human-environment research. Information and computer science publications provide a mix of general overviews on the computational aspects of big data or advanced information on specific challenges. Articles and edited volumes also offer case studies within narrowly defined research topics. The vast majority of this work is in keeping with the general publishing model of computer sciences, which tends toward shorter pieces in conference proceedings that may or may not be peer-reviewed.

Debates over the potential and problems of data science can be uneven or narrowly defined. Hidalgo (2014) expresses frustration with these problems in his opinion piece “Saving big data from big mouths,” which argues that coverage of big data seems to oscillate between uncritical reports or even hyperbolic odes versus underinformed critiques or jeremiads about the big data strawman. Calls for greater collaboration among fields tend to revolve around linking core fields in data science, especially statistics, computer science, and domain fields in the social and natural sciences, and into the arts and humanities. One common complaint is that data science focuses too often on important yet narrow technical and computational considerations. It gives short shrift to many aspects of substantive domain knowledge. At the same time, domain scholars outside of data science run the risk of ham-handedly using data approaches or caricaturing the entire field based on limited engagement. As we explore later, there are many threads to this conversation. There are fundamental differences among fields and their conceptual and epistemological bases. There are marked disparities in funding and infrastructural support for some kinds of work over others that have far-reaching effects on the kinds of questions being asked and answered by scholars of all stripes.

Communication issues between data scientists and domain scholars are related to the need for better communication between human and environmental researchers. Three decades ago, Stern (1993) called for a *second environmental science* that highlighted the need for environmental science to embrace the human. While there have been positive developments in integration, there is much potential for greater collaboration. As Holm and others put it,

> various important disciplines, mainly social and human, are too often overlooked or neglected as a science, such as law, architecture, history, literature, communication, sociology, and psychology. These are important disciplines to fully understand Earth systems and human motivation and to guide decision-makers. However, they are not routinely seen as fundamental to giving policy advice. Proponents of interdisciplinary research at times relegate human and social science research to an auxiliary, advisory, and essentially nonscientific status. *(Holm et al. 2013, p. 26)*

Finally, while the focus on relationships between humans and nature anchors most discussion in this book, it is helpful to recognize that this division can be seen as an arbitrary. People have been looking at human-environment systems for thousands of years (Marsh 1864). At the same time, there is a long-standing body of work in *posthumanism* that questions human-centric explanations and correspondingly rejects the dual construction of nature and culture (Braun 2004). This scholarship rejects the concept that nonhuman beings lack agency and embraces the idea that human and nonhuman beings cocreate many spaces. These spaces range from our stomach microbiome to human relationships with animals to interactions with the Earth.

Posthumanism has critics. It can be seen as perpetuating Eurocentric forms of knowledge, as highlighted by Indigenous critiques of posthumanism that argue that the universalizing claims of ways of knowing and being are themselves problematic (Sundberg 2013). For example, there is an ongoing need for Euro-American scholarship to take more seriously Indigenous knowledge, and how the intellectual labor and activist work of Indigenous scholars and practitioners on the mutual interdependence of humans and the environment illustrates how this division may be illusory (Watts 2013). It is important to bear these issues in mind, even as this book primarily uses a human-environment framing as a helpful shorthand for a complex set of dynamics.

1.2 What Are Big Data?

Data science deals with data, unsurprisingly. Data science has subsumed many aspects of big data as a scholarly endeavor, but it is important to consider data and big data on their own. Most scholarly work relies on data harnessed to various methods and concepts. Most researchers can readily point to the kinds of data they use. The simple notion of data as measures of phenomena that we find interesting (e.g., temperature, population counts, or interviews) suffices for most conversations about data science. However, it is essential to dig a little deeper at times and recognize the long and fraught history of data in science. A note on terminology – we will use big data as a plural noun when speaking of the data as such (e.g., “big data are collected”) and as a singular noun when speaking of the larger field of big data (e.g., “big data offers perils and promise”).

People have collected data for millennia. People twenty thousand years ago were using *tally sticks*, where they would make notches in pieces of wood or bone to keep track of important things, which presumably came in handy for activities such as trading and keeping inventory of possessions (Mankiewicz 2000). Four thousand years ago, people used calculating devices such as the abacus and stored information in libraries. The rise of modern statistics and record-keeping originated in the 1600s and was codified by the 1800s. In the nineteenth century, people used data in ways nearly indistinguishable from how we employ data, statistics, and modeling today to design descriptive measures and find associations in data (Porter 1986). The scientific meaning of data, which underpins big data, came into being in the 1600s. The term *data* is the Latin plural for *datum*, or "what is given" from the verb *dare*, "to give," but it has a deep, contested, and varied history over the centuries for notions of facts or evidence (Rosenberg 2013). Data are not always simple!

The key to understanding data science is understanding that data are made or captured by an observer. Indeed, some scholars would argue that the term data should be better considered as the term *capta*, from the Latin verb *capere*, meaning "to take" (Checkland & Holwell 2006). This book uses "data" since capta is a technical term for what most people think of as data, but it is helpful to consider what the concept implies for big data. Since observers capture data, this information is biased from initial observations to subsequent data handling, interpretation, and analysis. Statisticians spend much time developing new ways to plumb the nuances of data. Social scientists debate endlessly about how data map onto complicated social phenomena like race or trust. Natural scientists are heavily invested in ensuring their instrumentation and observations are free of systemic bias. The humanities have led the charge against naïve realism, noting that data are not the same as related phenomena, despite how they are often treated as inseparable. Nonetheless, despite best efforts to reduce bias in data, it is inescapable (Section 2.3).

Despite (or perhaps because of) big data being a trendy topic, there is no single commonly shared definition. There is an ongoing scholarly conversation around the origins of big data. Diebold (2012) dives into its definition as one of the earlier users of the term during an academic presentation in 2000. He argues that for the field of econometrics, he is likely one of the originators of *big data* as a term that refers to data sets being too large to be used with existing approaches. However, he uncovers several instances of the term before 2001. Weiss and Indurkhya (1998) use the term repeatedly in their data mining textbook, and researchers with the firm Silicon Graphics used it as early as the mid-1990s. Big data is composed of two common words and associated ideas, so perhaps it is not surprising that there are multiple routes to current usage.

Despite being coined almost two decades ago, the definition of big data remains loose. Critical characteristics for many scholars define big data and data science. Among the most long-lived attributes are the "three v's" of big data: volume, velocity, and variety. We will not belabor these because there is a tremendous amount written on them already, but it helps frame the discussion. The v's of big data trace back to a four-page memo written by Laney (2001) in his role as an analyst for the now-defunct Meta Group. Volume refers to where there is much data, orders of magnitudes larger than is commonly used in most research fields. Velocity describes how data are collected and stored at speed or in real-time. Variety refers to how big data have varying degrees of organization and structure, from well-defined tables to text scraped from the web. Beyond these three basic characteristics, there are ongoing conversations on whether big data should have other v's, such as veracity (accuracy of data) and value (the usefulness of data to answer specific questions (Chen et al. 2014). Dozens of definitions relate to the three v's, additional v's, and other characteristics of big data that start with other letters besides "v."

Volume, or the raw amount of data, is central to any definition of big data and data science. Many fields have large volumes of data. Natural science disciplines, including particle physics, astronomy, and genomics, were early adopters of big data approaches. Genomics and astronomy are home to vast amounts of data. They will grow even more because research projects collect amounts of data that were unthinkable even a few decades ago – on the order of ~25 zettabytes per year. The volume of information generated globally doubles every three years, and this pace is increasing (Henke et al. 2016). Key challenges posed by these data are related to their acquisition, storage, distribution, and analysis. Outside of academia, platforms such as Twitter and Facebook collect and monetize large amounts of data, primarily by developing sophisticated analyses of their users to sell advertising.

A tremendous amount of ink has been dedicated to writing about the size of big data and attendant issues of measuring and defining what "big" means. There is not much value in rehashing those arguments here. Perhaps the easiest way to think about it is that context matters. "Big" is relative to the underlying technology and data format; video files are larger than tweets, but their use matters, such as trying to extract semantic understanding. Bigness tends to revolve around the inability of many existing computing systems or approaches to cope with data and the idea that the amount of data is increasing rapidly, exponentially in some cases. Bigness implies we are always moving toward the horizon and will never get there; in that what is big today, will someday be merely large, or just plain old data.

Most authors are careful to note that the term big is almost meaningless, given how increases in storage, processing speed, and analytical power almost always make the big data of yesterday into the small data of today. There are also debates over whether the bigness of data matters when data science in many fields goes well beyond the engineering and computing challenges that are the focus of so much work in big data (for a more in-depth take, see Chang & Grady 2015). As Jacobs (2009, p. 44) puts it, big data are those “whose size forces us to look beyond the tried-and-true methods that are prevalent at that time.” However, this definition is (necessarily) vague in order to apply to many specific problems. No matter the measure, the size of global data holdings is increasing (Figure 1.1).

People have attempted to measure how much information exists. The International Data Corporation is a maker of digital data storage and has attendant biases, but it predicts that the amount of data in the world (termed the Datasphere) will grow from 33 zettabytes in 2018 to 175 by 2025 (Reinsel et al. 2018, p. 3). The same study posits that over 75 percent of the world’s population will interact in some way with the data and, by definition, contribute to big data. Global data storage capacity is growing and increasingly moving to digital format. In 1986, 99.2 percent of all storage capacity was in analog forms such as paper volumes, and within two decades, 94 percent of storage capacity was digital (Hilbert & López 2011). Measuring data is an imprecise process and often relies on commercial interests using opaque methods, but it is safe to say there is a lot of data out there (more on data in Chapter 2).

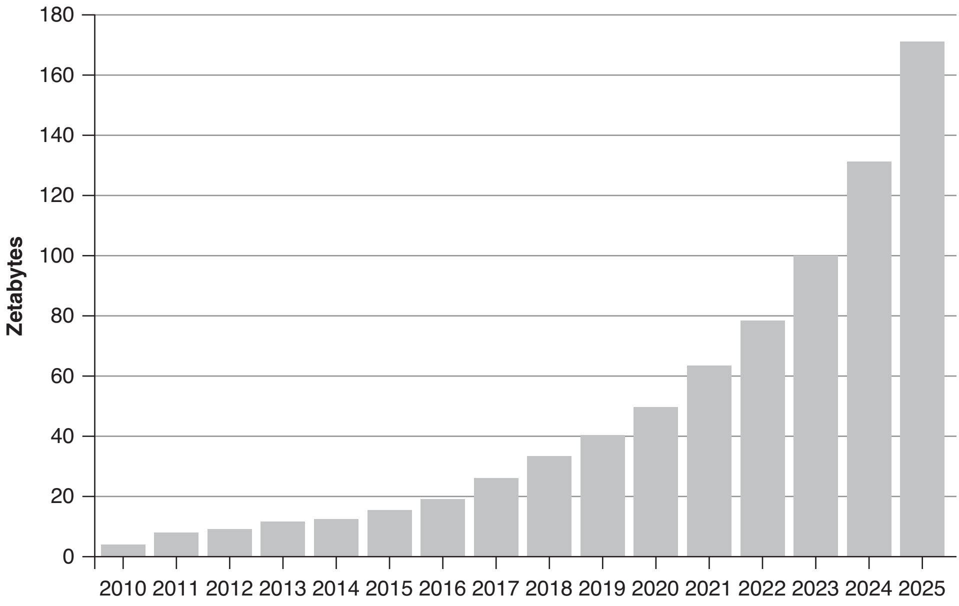

Figure 1.1 Size of global data holdings (2010–25) (Reinsel et al. 2018). Reprinted with permission from the International Data Corporation.

Cambridge University Press & Assessment
978-1-108-48628-6 — Data Science and Human-Environment Systems
Steven M. Manson
Excerpt
More Information

Velocity is another defining characteristic of big data, referring to the rate at which it is collected or moved. Most big data conversations center on how fast data are collected, but velocity also involves how quickly a given computer or processing system can perform calculations over these data. Human-environment data are derived and stored across time frames spanning from paper records and ship logs in the 1600s to real-time digital sensors operating today. The flow rate increases exponentially, especially when considering how scientists use computational modeling to generate simulated data alongside traditional sources (Overpeck et al. 2011). Many data sources are termed *streaming* because they are collected constantly. A significant challenge for human-environment research and data science is developing ways to analyze these data on the fly without assuming that they will be stored in their entirety for later use.

Standard units are used to measure data size. Computer performance has been typically measured by the number of floating-point arithmetic calculations a system can perform in a second (FLOPS). In contrast, data storage is usually measured in bits and bytes. The bit, a contraction of a *binary digit*, is the smallest unit of computer data storage and usually takes the binary value of 0 or 1. A byte is a collection of eight bits and is usually written in binary notation (i.e., 00000000 to 11111111). When using the FLOPS or bytes terminology, we use Greek prefixes to indicate speed or size (Table 1.1). More generically, the suffix *scale* denotes the

Table 1.1 *Size of big data in terms of speed and storage demands*

Prefix	Storage in bytes		Speed in FLOPS		Storage examples
	Byte (B)	1		10°	Single character
Kilo	Kilobyte (KB)	$1{,}024^1$	KiloFLOPS	10^3	Half a page of text
Mega	Megabyte (MB)	$1{,}024^2$	MegaFLOPS	10^6	Photograph
Giga	Gigabyte (GB)	$1{,}024^3$	GigaFLOPS	10^9	Hour-long video
Tera	Terabyte (TB)	$1{,}024^4$	TeraFLOPS	10^{12}	One day of Earth Observing System data in 2000 (Frew & Dozier 1997)
Peta	Petabyte (PB)	$1{,}024^5$	PetaFLOPS	10^{15}	One year of data collected by the United States National Aeronautics and Space Administration in 2015
Exa	Exabyte (EB)	$1{,}024^6$	ExaFLOPS	10^{18}	One day of data from the Square Kilometer Array (SKA) telescope (Farnes et al. 2018)
Zetta	Zettabyte (ZB)	$1{,}024^7$	ZettaFLOPS	10^{21}	One year of digital data in 2010 (Gantz & Reinsel 2010)
Yotta	Yottabyte (YB)	$1{,}024^8$	YottaFLOPS	10^{24}	One day of data generated globally in the mid-2020s (Parhami 2019)