Skills and Knowledge for Data-Intensive Environmental Research

STEPHANIE E. HAMPTON, MATTHEW B. JONES, LEAH A. WASSER, MARK P. SCHILDHAUER, SARAH R. SUPP, JULIEN BRUN, REBECCA R. HERNANDEZ, CARL BOETTIGER, SCOTT L. COLLINS, LOUIS J. GROSS, DENNY S. FERNÁNDEZ, AMBER BUDDEN, ETHAN P. WHITE, TRACY K. TEAL, STEPHANIE G. LABOU, AND JULIANN E. AUKEMA

The scale and magnitude of complex and pressing environmental issues lend urgency to the need for integrative and reproducible analysis and synthesis, facilitated by data-intensive research approaches. However, the recent pace of technological change has been such that appropriate skills to accomplish data-intensive research are lacking among environmental scientists, who more than ever need greater access to training and mentorship in computational skills. Here, we provide a roadmap for raising data competencies of current and next-generation environmental researchers by describing the concepts and skills needed for effectively engaging with the heterogeneous, distributed, and rapidly growing volumes of available data. We articulate five key skills: (1) data management and processing, (2) analysis, (3) software skills for science, (4) visualization, and (5) communication methods for collaboration and dissemination. We provide an overview of the current suite of training initiatives available to environmental scientists and models for closing the skill-transfer gap.

Keywords: ecology, informatics, data management, workforce development, computing

The practice of environmental science has changed dramatically over the past two decades as computational power, publicly available software, and Internet connectivity have continued to grow rapidly. At the same time, the volume and variety of data available for analyses continue to increase at a meteoric pace (Porter et al. 2009) because of the increased availability of data from long-term ecological research, environmental sensors, remote-sensing platforms, and genome sequencing, along with improved data-transfer capacity. The environmental research community is therefore faced with the exciting prospect of pursuing multidisciplinary scientific research at unprecedented resolution across multiple scales, making possible the synthetic research that can address pressing environmental problems (Green et al. 2005, Carpenter et al. 2009, Rüegg et al. 2014, Peters and Okin 2016). These exciting technological advances, however, have challenged the research community's capacity to rapidly learn and implement the concepts, techniques, and tools necessary to fully take advantage of this new era of big data and, more generally, data-intensive research (box 1). As a consequence, there is an urgent need to reevaluate how our training system can better prepare current and future generations of environmental researchers to thrive in this rapidly evolving digital landscape (Green et al. 2005, Hey et al. 2009, NERC 2010, 2012). Deep knowledge of ecological theory, ecosystem dynamics, and natural history prepares environmental researchers to ask the right questions within this data-rich landscape, minimizing the chances that spurious correlations will lead science down blind alleys, as it might for specialists trained primarily in computing and statistics. By proactively addressing the training challenge at a time when the field of data science is still young, environmental scientists will not only guide the environmental research questions but also guide the field toward a culture that is collaborative and inclusive.

Although the need for data skills is reflected across many if not all disciplines and sectors, the demand for training in the environmental workforce is particularly time sensitive given new flows of data from the National Ecological Observatory Network (NEON) in the United States, the Terrestrial Ecosystem Research Network (TERN) in Australia, and other large government investments in long-term research and observatories worldwide (Hampton et al. 2013). Environmental researchers must be prepared to use these data to address pressing environmental challenges. Furthermore, by developing training that can accommodate the exceptionally heterogeneous data that characterize environmental research (Jones et al. 2006)—from genes and "critter cam" videos to airborne and satellite sensors—training approaches will be readily adaptable to other fields.

BioScience 67: 546–557.
doi:10.1093/biosci/bix025 Advance Access publication 3 May 2017

Box 1. The current state of environmental sciences education in American universities.

The widespread lack of capacity among researchers in environmental biology for doing data-intensive science is a fundamental impediment to harnessing the potential power of big data and associated new technologies. The need for better preparation in these skills is increasingly acknowledged across diverse publications and forums (e.g., Jones et al. 2006, NERC 2010, 2012, Manyika et al. 2011, Joppa et al. 2013, Laney et al. 2015, Smith D 2015, Teal et al. 2015, Mokany et al. 2016, Peters and Okin 2016). A recent survey of graduate students in environmental sciences (Hernandez et al. 2012) is eye opening: Over 80% of students had received no formal training in computing or informatics at even the most basic level, and 74% stated that they had no skills in any programming language. Although 72% of the students said they understood the term *metadata*, about half had not created metadata for their dissertation data and had no plans to do so. Approximately one-third of the surveyed students were planning to use sensors in their research, which will lead them at least incidentally into learning some of these topics at some level, although likely not employing best practices. Why are these skills still so rare when the need for them is now widely recognized? Strasser and Hampton (2012) reported that when ecology instructors are asked why they do not train students in such foundational skills, they indicate the following eight obstacles: (1) limited time, (2) the topics were not appropriate at their course's level, (3) the topics were or should be covered in a lab section, (4) students in the course did not have the necessary quantitative or statistical skills to cover the topics, (5) lack of funding or resources, (6) the course was too large to cover these topics well, (7) the instructor was not knowledgeable in these topics, and (8) the topics were or should be covered in other courses. Essentially, we are attempting to fit more material into already-full courses and curricula, which are taught by people who do not feel prepared to address topics relevant to big data and data-intensive research. Clearly, the current situation is not satisfactory, but there is reason for optimism. Three decades ago, ecologists were ill prepared to use statistics in their research, and now statistics preparation is considered vital in ecology. It would be extremely difficult to publish a manuscript in ecology without any statistical testing. A similar revolution in computational proficiency must occur in order for environmental scientists to fully arrive in a digital age that requires data-intensive synthesis (Green et al. 2005).

One symptom of the current curriculum's shortcomings is the recent emergence of a variety of extramural options for acquiring critical technological skills, including resources such as Software Carpentry, Data Carpentry, and other informatics and computational training workshops hosted at NEON, at environmental synthesis centers worldwide, or at meetings of professional societies such as the Ecological Society of America. Numerous self-guided online tutorials are also available, although such resources may vary widely in quality or are not tightly linked with topical environmental science domains. As these extramural opportunities proliferate, there is a paucity of systematic training within university programs to equip students with the computational skills they need to conduct data-intensive research. Lack of university-level training may reflect the sense among many environmental-science faculty that they themselves are not proficient in data management and the latest computational tools for data-intensive research (Strasser and Hampton 2012). In addition, environmental-science faculty may have difficulty redirecting students to high-quality instructional resources within universities, because mathematics, statistics, and computer-science departments are primarily focused on educating future practitioners in their respective fields. Therefore, within university courses and curricula, both faculty and students miss the opportunity to experience the pedagogical benefits of learning relevant technology concepts and skills while encountering the realistic data and analytical challenges associated with a particular domain science. It is possible that the pace of technological development will continue to demand that workshops and other resources thrive outside of university curricula, given the comparative flexibility of such activities to adapt materials rapidly and remain on the leading edge as it advances. Moreover, these workshops offer vital opportunities for technological advancement by a wide range of researchers working both inside and outside of academia.

Technical proficiency is necessary but not sufficient for modern scientific data management, processing, and synthesis challenges. Synthesis of heterogeneous environmental data usually requires collaboration skills as well as the ability to build on previous work (e.g., reuse of code). It is unreasonable to expect that every researcher can become an expert in domain science, statistics, informatics, data management, and software engineering, but researchers should at least be familiar with these concepts to foster effective collaborations. Collaboration between multiple researchers with diverse and complementary expertise is essential throughout all aspects of a project to define and integrate relevant concepts, data, models, and tools (Cheruvelil et al. 2014). Accordingly, collaborative data-intensive research should benefit greatly from community convergence on data structures, protocols for information exchange, and efficient reuse of code.

Community convergence on a common set of best-of-class software tools has become more apparent in recent years and may signal community maturation that facilitates more coordinated approaches to training in data-intensive research skills. For example, projects have grown out of the open-source software community using tools developed in R and Python. These free, open-source, cross-platform programming languages have decades of use and community building, and allow scaling up from desktop computers to powerful processing systems, such as HPC (high-performance computing) and cloud computing. When supported by the powerful and expanding information exchange standards set by the World Wide Web Consortium (W3C), these interoperable technologies

Table 1. Examples of existing training resources and events.

Type	Title	Organization	Topics					Target Audience	License	Web site
			Data	Analysis	Software	Visualization	Collaboration			
Lesson	Learn X and Y minutes, where X=json	Adam Bard/ Anna Harren			✓			Programmers	Open (CC-BY-SA)	*https://leamxinyminutes.com/docs/json/*
Lesson	R for reproducible scientific analysis	Software Carpentry		✓	✓			Researchers using R	Open (CC-BY)	*http://swcarpentry.github.io/r-novice-gapminder/*
Unit	NEON Data Skills	National Ecological obsevatory Network	✓	✓	✓	✓	✓	Researchers, Students, Instructors	Open (CC-BY)	*http://neondataskills.org*
Unit	DataONE Data Management Modules	DataONE	✓	✓			✓	Researchers, Instructors, Librarians	Open (CCO)	*https://www.dataone.org/education-modules*
Workshop	Data Cerpentry Workshops	Data Carpentry	✓	✓	✓	✓		Researchers	Open (CC-BY)	*http://www.datacarpentry.org/*
Workshop	Open Science for Synthesis	National Center for Ecological Analysis and Synthesis	✓	✓	✓	✓	✓	Researchers	Open (CC-BY)	*https://www.nceas.ucsb.edu/OSS*
Course	Data wrangling, exploration, and analysis with R	University of British Columbia	✓	✓	✓	✓		Graduate students	Open (CC-BY-NC)	*http://stat545.com/index.html*
Course	Programing for Biologists	Weecology Lab	✓	✓	✓	✓		Undergraduate students	Open (CC-BY)	*http://www.programmingforbiologists.org/programming/*
Program	Data Science Program (Coursera)	Johns Hopekins University	✓	✓		✓		Researchers, Students	Proprietary	*https://www.coursera.org/specializations/jhudatascience*
Program	Berkeley Data Science Education Program	University of California, Berkeley	✓	✓	✓	✓	✓	Undergraduate students	Open (CC-BY-NC)	*http://databears.berkeley.edu/*

afford opportunities for creating data, models, codes, and workflows that can be broadly shared, reviewed, and extended, resulting in transparent and reproducible synthesis (Hollister and Walker 2007, Rüegg et al. 2014, Hampton et al. 2015).

This article represents the consensus perspective of a group of environmental scientists, informatics experts, and computational scientists who have been building and delivering training that covers the concepts and skills environmental scientists need to stay on the leading edge (some examples are in table 1). The main challenge we address is how to significantly raise the data-science competencies of current and next-generation environmental researchers—that is, the concepts and skills needed to effectively engage the heterogeneous, distributed, and rapidly growing volumes of data available for addressing critical environmental questions. Here, we outline the skillset required by environmental scientists and many other scientific fields to succeed in the kind of data-intensive scientific collaboration that is increasingly valued. We also suggest the forms that such training could take now and in the future.

Key skills for the data-intensive environmental scientist

It is unrealistic for most individual researchers to master every aspect of data-intensive environmental research. Rather, we can identify the foundational knowledge and skills that are a gateway for researchers to engage in data science to the degree that best suits them. We emphasize that data-intensive environmental research is most likely to reach its full potential through collaboration among variously talented researchers and technologists. We distinguish five broad classes of skills (table 2): (1) data management and processing, (2) analysis, (3) software skills for science, (4) visualization, and (5) communication methods for collaboration and dissemination. The novice need not master all at once; in our experience, even basic familiarity with these skills and concepts has a positive impact on both research and collaboration capabilities.

Data management and processing. Data management has always been a challenge in research, and it continues to grow in magnitude and complexity, with the requisite skills a crucial

Table 2. A taxonomy of skills for data-intensive research.

Data management and processing	Software skills for science	Analysis	Visualization	Communication for collaboration and results dissemination
Fundamentals of data management	Software development practices and engineering mindset	Basic statistical inference	Visual literacy and graphical principles	Reproducible open science
Modeling structure and organization of data	Version control	Exploratory analysis	Visualization services and libraries	Collaboration workflows for groups
Database management systems and queries (e.g., SQL)	Software testing for reliability	Geospatial information handling	Visualization tools	Collaborative online tools
Metadata concepts, standards, and authoring	Software workflows	Spatial analysis	Interactive visualizations	Conflict resolution
Data versioning, identification, and citation	Scripted programming (e.g., R and Python)	Time-series analysis	2D and 3D visualization	Establishing collaboration policies
Archiving data in community repositories	Command-line programming	Advanced linear modeling	Web visualization tools and techniques	Composition of collaborative teams
Moving large data	Software design for reusability	Nonlinear modeling		Interdisciplinary thinking
Data-preservation best practices	Algorithm design and development	Bayesian techniques		Discussion facilitation
Units and dimensional analysis	Data structures and algorithms	Uncertainty propagation		Documentation
Data transformation	Concepts of cloud and high-performance computing	Meta-analysis and systematic reviews		Website development
Integrating heterogeneous, messy data	Practical cloud computing	Scientific workflows		Licensing
Quality assessment	Code parallelization	Scientific algorithms		Message development for diverse audiences
Quantifying data uncertainty	Numerical stability	Simulation modeling		Social media
Data provenance and reproducibility	Algorithms for handling large data	Analytical modeling		
Data semantics and ontologies		Machine learning		

Note: Many if not most of these elements apply across multiple categories. This taxonomy was initially created in a workshop involving natural and physical scientists, information scientists, and computer scientists (*isees.nceas.ucsb.edu*), with modest refinements by the authors.

component of environmental work in the coming decade (e.g., NERC 2010, 2012). Many classical ecological studies are based on data that were collected and stored in personal notebooks. Today, there is an expectation that data will be stored digitally, backed up, and available for future analysis (Heidorn 2008, Hampton et al. 2013). A new set of data-management skills (table 2) is required to ensure that data storage and sharing are not prohibitively burdensome to investigators and that scientists are prepared to articulate and adhere to a well-structured data-management plan from beginning to end (e.g., Michener and Jones 2012).

Metadata, or data about the data, provide the descriptions and documentation that enable one to understand the content, format, and context of a data set (Michener 2006, Michener and Jones 2012). Clear metadata are essential for a researcher to understand how a data set was collected and processed, by whom, its format and structure, and its associated uncertainties (Jones et al. 2006, Edwards et al. 2011, White et al. 2013). At the very least, scientists must learn to routinely generate metadata in easily accessed machine-readable formats. Even better, metadata standards such as Ecological Metadata Language (EML; Fegraus et al. 2005) can greatly facilitate data sharing and reuse. Data storage formats that tightly package metadata with data are becoming more common (e.g., netCDF and HDF5); however, few environmental scientists understand and can work with these formats. Furthermore, documentation of the data set itself is often not sufficient in cases of large ecological syntheses: Process metadata, which documents the alterations made to produce a final data set, are needed for research to be truly repeatable and reproducible (Ellison 2010).

There is broad variation in the kinds of data that are collected and used in environmental research, such that users are challenged not only to understand many data types and formats, from text to raster and video (Jones et al. 2006, Michener and Jones 2012), but also to integrate them in

order to accomplish meaningful synthetic analyses. A large-scale study may call for the integration of many different types of data, creating philosophical, logistical, and analytical challenges (Jones et al. 2006, Soranno et al. 2015).

Although good data management can facilitate data integration, for the efficient synthesis of diverse data, scientists may need to dig deeper in the toolbox and learn about formalized semantics and ontologies. The *semantics* of a data set (e.g., the context and compatibility of similarly labeled attributes across studies) necessary for full integration may still be missing or incomplete (Madin et al. 2007). From spatially explicit data (e.g., Al-Bakri and Fairbairn 2012) to species-level observations (e.g., Kennedy et al. 2005), semantic dissimilarity can hinder integration. For example, in a synthesis of stream restoration effectiveness, Barnas and Katz (2010) found that minor differences in how stream restoration projects were characterized in metadata resulted in major qualitative differences in overall evaluation of restoration actions' efficacy. Using formalized ontologies has benefited other fields, such as molecular biology and urban planning (Bada et al. 2004, Michalowski et al. 2004). Within a research domain, an *ontology* represents knowledge in both standardized terminology and by characterizing the relationships among domain objects (Madin et al. 2007). Broader use of ontologies in ecology would facilitate syntheses by streamlining and simplifying decisions about whether and how diverse data sets are integrated (Madin et al. 2008).

Whether focal data sets are well organized or messy, scientists must have the tools to work with varied data formats and types in a reproducible workflow. Informally, researchers may describe this stage of data processing as *data wrangling*, the process of manipulating data sets into consistent formats appropriate for analysis and synthesis. For example, a plant ecologist may want to aggregate and summarize sensor data collected at various temporal frequencies and merge these data with point or regional values extracted from raster files. Adding observational and experimental data on traits such as chemical composition or growth rates would add another layer of complexity. Although essential, this process is rarely taught in courses or described thoroughly in publications. Popular scripting languages (e.g., R and Python) provide a large set of dedicated tools for researchers to perform the necessary data wrangling steps in a transparent and reproducible manner.

Analysis

Just as ecological data have become richer, more complex, and more challenging to navigate, so, too, have statistical methods (Green et al. 2005). The breadth and complexity of methods now employed in ecological research are overwhelming. Rather than attempt to run ever faster in the hamster wheel of statistical methods, data-intensive training programs should focus on the general skills that will best enable researchers to survive and thrive in this rapidly changing environment (table 2). Specific statistical methods frequently are determined by the researcher's field, and a continued emphasis on rigor in these statistical methods will be synergistic with learning fundamental computing skills. Such skills are needed to facilitate not only the creation and use of efficient code for diverse statistical analyses but also the critical evaluation of its implementation, including peer review (Joppa et al. 2013).

Computational building blocks for statistics. First, we recommend a computational approach to statistics training. Whereas calculus and a basic statistics course might have been sufficient background for classical ecological statistics, some basic computational training is essential to understand today's algorithms (Wilson 2006). A computational approach to statistics training offers an opportunity to avoid the overload of highly specialized methods contingent on a narrow set of assumptions in favor of a more general approach that emphasizes basic concepts such as simulation, sampling, visualization, and summary statistics (table 2).

Scripting for efficient, reproducible, and transparent analysis. Second, scientists who execute their analyses in a scripting language have a tremendous advantage in synthetic integrative work, enjoying greater flexibility and efficiency, and with the important benefit of creating transparency and reproducibility for collaborators and colleagues (White et al. 2013). Compared with spreadsheet tools that allow users to mix the data processing with the data set itself, scripting approaches help to clearly separate data processing from the data, paving the way toward capturing the scientific workflow for a specific analysis. Increasing transparency and direct reproducibility, by sharing scripted analyses, is crucial as data-intensive analyses become more complex and varied (Ellison 2010). Providing well-documented code and data to accompany manuscripts helps reviewers and readers to understand both familiar and unfamiliar analyses (Wilson et al. 2014). Although the code itself can assist transparency, skills in the appropriate documentation of codes are perhaps just as important for reuse and reproducibility. The novice will make great strides by becoming comfortable with fundamental computational approaches to statistics, in a scripted environment. And as analyses become more challenging, scientists are sometimes faced with the surprising idea that they are not just doing analysis but also actually developing software.

Software skills for science

Any scientist who writes data-processing and -analysis code is functionally a software developer, but few have been trained in best practices of software development (Wilson et al. 2014). Researchers in the vanguard of data science have suggested that scientists adopt software-development best practices: version control, literate programming documentation, unit testing, continuous integration, software development and release patterns, and code peer review (table 2).

Although these techniques are valuable, they are likely too advanced to serve as a starting point for most domain scientists. We suggest the following starting points for every researcher to learn.

Learning a computing language and its "ecosystem." Like the scientific process itself, software stands on the shoulders of giants. Learning to discover, assess, and manage dependencies within software is an important part of becoming proficient in a computing language (Wilson et al. 2014). Scripts that reuse existing proven and tested methods are faster to write, simpler to understand, and easier to trust than those that reinvent the wheel. Learning how to find software that already provides the required functionality is often just as important as knowing how to write that functionality from scratch. However, not all software is created equal, and buggy, unstable, or untested dependencies are the Achilles heel of many scripts. Telling the good from the bad is a skill that scientists need to acquire; Wilson and colleagues (2014) have provided more detailed advice on best practices in software development.

Code organization. Like most aspects of research, good software practice requires good organization. Following existing practices and recommendations for a software language or field will help an individual researcher and others who read the code to find the correct lines and scripts for a particular result. Good organization goes beyond files to how code itself is written. A fundamental concept of clean, well-organized code is the don't repeat yourself (DRY) principle (Wilson et al. 2014). Although heavy use of copy–paste is a common strategy, researchers should learn to identify and reorganize common tasks or subroutines into separate scripts or functions. Like any other writing, good code requires frequent revision and rewriting, which saves time and reduces errors.

Data visualization

Historically, scientific data visualizations have been static, two dimensional, and created as a scientific "end product," often designed for publication. However, as data streams continue to evolve and expand and as analyses that integrate these data become more complex, it is crucial for data visualization to be included throughout the scientific workflow (Fox and Hendler 2011), tightly connected to the original and derived data to support current results and reproducibility, and effective as a communication tool that disseminates developing research to the scientific and other communities.

Integrating data visualization throughout the scientific workflow. Visualization products created early in the data-exploration and -analysis stages are tools to understand trends and relationships that inform analysis methods, constraints, and interpretations (Kelling et al. 2009). Furthermore, in an era of high-frequency streaming data, static two-dimensional output can quickly become outdated. It is increasingly important that visualizations maintain a close connection to the original data (Fox and Hendler 2011) to support dynamic outputs that can adapt to methodological and data updates and to maintain reproducibility by connecting the community more directly to the original data.

Interactive visualization as a compelling communication tool. Interactive visualization can allow researchers to more readily explore data with each other and also foster dialog with stakeholders. A recent explosion of application program interface access to powerful, interactive data-intensive visualization (e.g., plot.ly and Shiny) and mapping (e.g., leaflet and mapbox) tools accessed through widely used programming environments (e.g., R and Python) empowers environmental scientists to place analysis results in the hands of a broader audience (Zastrow 2015).

Despite advances in visualization tools and approaches, harnessing the full value of these resources requires proficiency in scripting and knowledge of an ever-expanding array of tools. It also requires an understanding of graphic or cartographic principles that support readability (Brewer 1996, Tufte 2001, Brewer and Buttenfield 2007). These are skills that not all scientists can be expected to have. As for all the skills discussed here, scientists who lack specific skills would benefit from collaboration with relevant experts.

Communication for collaboration and dissemination of results

Communication is the cornerstone both for collaborative team science that produces high-impact scientific products and for effective dissemination that ensures these products are useful for society. Collaboration is now the norm for successful scientific endeavors (Wuchty et al. 2007), particularly for data-intensive environmental research, which implicitly requires a broader suite of cross-disciplinary data, skills, and knowledge. Successful science communicators integrate technologies that augment but still cannot replace good people skills, the soft skills of interacting productively with other human beings in a professional endeavor.

Communication tools. A growing suite of tools facilitates dynamic collaboration across geographic and disciplinary boundaries. Version control tools such as Git and the user-friendly GitHub interface support cross-team development of processing algorithms, documentation, and code that integrates and synthesizes heterogeneous data, in addition to issue assignment and tracking. Collaborative writing tools such as Google Docs and Mozilla Etherpad support remote meeting participation and community-developed documentation of methods, protocols, and scientific workflows. Data repositories relevant to environmental sciences are on the rise, supporting data sharing, discovery, and documentation (Michener and Jones 2012). Furthermore, a diversity of new

platforms is emerging to integrate the tools for collaboration and data sharing, such as the Open Science Framework (*osf. io*) and Jupyter/iPython Notebooks (*jupyter.org*). Of course, these tools do not create a vibrant exchange of ideas by themselves; they only make it easier for researchers to communicate with each other and a broader audience, and this skill set is not one that can be as easily coded.

Communication skills. Effective communication (table 2) both within a collaborative team and to the broader scientific and nonscientific community is a critical skill in science. In both cases, researchers who are effective communicators learn to invest time and energy in understanding their audience—whether it is a research team or a policy-setting organization—and honing their skills to engage in a meaningful, respectful and productive dialogue (Baron 2010, Pace et al. 2010, Cheruvelil et al. 2014).

Early in a collaboration, scientists from different disciplines often spend substantial effort in assuring that they are using a common language in their work, defining terms in the same way, and working toward the same objective (Eigenbrode et al. 2007, Hackett et al. 2008). Even with initial hurdles cleared, successful teams must continue to expend considerable energy communicating with each other clearly to ensure that individual as well as collective expectations for research productivity are met and that sources of conflict are addressed (Cheruvelil et al. 2014). High-performing teams excel in communication, achieve results beyond what any could have realized alone, and thus richly return on the investment they make upfront on human interactions (Smith and Imbrie 2007).

Many scientists assume that communication skills are innate; in our experience, they are like any other skill in that some people are more predisposed than others, but most if not all researchers can improve their communication skills. Some useful exercises are those based on the "message box" in Baron (2010) and development of collaboration policies in Cheruvelil and colleagues (2014).

Changes in mindset

As the research and training landscapes change, the need for new skills will be accompanied by a need for changes in mindset to make data-intensive training effective. These changes in mindset must occur among administrators, instructors, and individual learners who together shape the capabilities of the workforce in environmental science.

Administrators and faculty in higher education will need to recognize that data-intensive research skills are core skills that need to be widely introduced into departmental courses and curricula. Both faculty and students need these changes. Funding organizations with finite resources and large commitments to environmental sensor networks (e.g., in the form of national observatories and satellite-based sensors) expect a return on these investments, which requires researchers to acquire the capacity to use these data effectively. Furthermore, students with data skills clearly are more marketable across sectors, a trend that is expected to grow (NERC 2010, 2012, Manyika et al. 2011, Smith 2015). To better prepare the next-generation of scientists for modern data-intensive research, skills should be taught both as stand-alone courses and incorporated as integral learning objectives of existing science courses. Incorporating data-intensive skills into university programs will raise the baseline for data literacy (box 2).

Bringing data into the classroom requires recognition of ongoing changes in data availability and variety, as well as the speed with which data are now generated, and how these shifts affect approaches to data management, integration, and analysis. In introducing students to data-intensive research in undergraduate ecology, Langen and colleagues (2014) additionally found that students had very diverse perceptions about whether public data were more or less "authoritative" than those they generated themselves and whether these activities were really "doing science." Given that addressing environmental questions at appropriately broad scales will likely require the use of large-scale public data (e.g., NASA, EPA, and NEON), Langen and colleagues' (2014) findings suggest a need to address students' (and instructors') questions about how data-intensive research fits into the scientific endeavor overall.

Changing learning objectives for data-intensive training will require educators to restructure existing courses and develop new teaching materials, but collaborating in course design and sharing materials can ease the burden on individual instructors. A variety of initiatives provide freely available data sets to be slotted into existing courses for specific learning objectives (e.g., the Portal Project Teaching Database, Ernest et al. 2015; NEON Teaching Data Subsets, *https://dx.doi.org/10.6084/m9.figshare.2009586.v9*). It is also becoming more common for instructors to openly share their full course materials. Community sharing of course materials allows educators to teach "field-tested" courses broadly, discuss best practices, share experiences and perspectives, and, ultimately, to improve and refine training to be higher quality and more effective (Teal et al. 2015). Software Carpentry and Data Carpentry have been leading examples of collaborative course development for the workshop model (Teal et al. 2015), but other models exist, ranging from single units (*www.dataone.org/education-modules*) and lesson sets (*http://neondataskills.org/tutorial-series*) to full-semester courses (*www.programmingforbiologists.org*). Unfortunately, the growth of learning management systems at many institutions has acted to limit the transferability of course materials, because access is typically limited to members of the institution.

The training landscape for data-intensive research skills

Currently, the resources for training in data-intensive research skills are both broad and scattered (table 1), complicating navigation for novices and experts alike. On the

Box 2. Building the next-generation workforce.

Several opportunities are presented by integrating data science into university curriculum. First, the skills for data-intensive research are largely high-demand, transferable skills that will benefit students across sectors and disciplines (Manyika et al. 2011). The marketability of these skills therefore argues for their early introduction in university curricula. Second, data-science initiatives can be positioned to foster diversity in high-demand research areas. Berman and Bourne (2015) made a powerful argument that data science should build gender balance into its foundations, and we suggest here that data-intensive environmental research has a special opportunity in this regard. The life sciences typically are gender balanced from undergraduate through postdoctoral stages, whereas women represent only 23% of engineering and 25% of computer-sciences graduate students (*www.nsf.gov/statistics/seind14/index.cfm/chapter-2*). As these fields meet at the intersection of data-intensive environmental research and education, women from biological sciences may find employment in data science such that the field can become both more gender balanced and representative of society at large in terms of ethnicity and other demographics.

Figure 1. A Software Carpentry Boot Camp for Women at Lawrence Berkeley National Lab. Photo credit: The Regents of the University of California, through the Lawrence Berkeley National Laboratory.

bright side, as "big data" and "data science" become household terms, these diverse resources are rapidly accumulating, presenting a tremendous opportunity for advancing widespread training. A variety of resources and events support self-paced and facilitated learning, ranging in format from single stand-alone lessons to domain-themed degree programs in data science. The delivery format for instructional resources includes blog posts, websites, videos, text documents, documented code, and training events (box 3).

Improving training approaches: The ideal and the practical

Ideally, the skills for data-intensive science will be incorporated into existing curricula at university or preceding levels. The need to integrate the skills for data-intensive research into higher education has been highlighted repeatedly. For example, the National Science Foundation's *Vision and Change in Undergraduate Biology* stated, "students should be competent in communication and collaboration, as well as have a certain level of quantitative competency and a basic ability to understand and interpret data. Furthermore, to be current in biology, students should have experience with modeling, simulation, and computational and systems-level approaches to biological discovery and analysis, and should be familiar with using large databases" (Smith 2015).

Data across the curriculum. The potential for data literacy to transform approaches across disciplines and sectors argues strongly for an initiative similar to that of writing across the curriculum in the 1970s, which persists in various forms throughout higher education (McLeod 1989). In that case, the recommendation has been for writing to be integrated into many required university courses so that students use writing as a tool to understand disciplinary material and also understand good writing to be fundamental to professional success, no matter what career trajectory one takes. This approach has many benefits. Students will acquire skills

Box 3. Types of resources and events currently available to support training in data-intensive environmental research and the diverse manners by which training can build on modular components.

In table 1, we show illustrative examples of these initiatives, and figure 2 below describes how they can be productively coordinated, spanning the following: (a) *lesson*, an atomic module containing material that covers specific learning outcomes; (b) *unit*, a group of lessons that have been designed to collectively cover a broader data concept or set of concepts and tasks; (c) *data*, a data set useful for one or more particular teaching goals; (d) *code*, a set of machine-readable instructions that constitute or contribute to a computer program; (e) *seminar*, a single or series of presentations that may be presented in person or online; (f) *workshop*, a facilitated collection of lessons or units that address a specific concept or set of skills, offered across a short duration; (g) *course*, a facilitated longer collection of units, usually part of a university degree program or a course of study that may or may not be recognized for credit toward the granting of an approved degree; and (h) *program*, a facilitated degree- or certificate-bearing suite of courses that follow a particular domain.

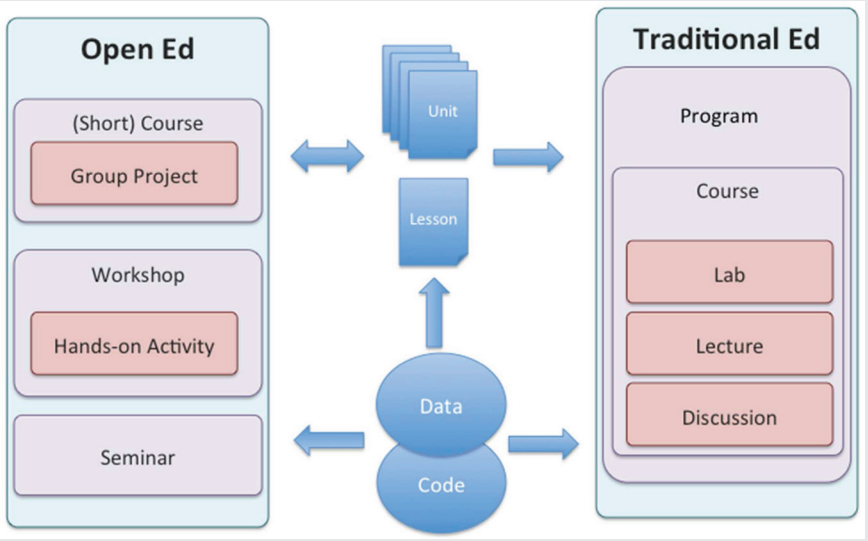

Figure 2. Resources that promote data-intensive research skills, emerging from open education initiatives, can be incorporated into traditional education programs and coordinated either inside or outside of academia to form the basis of a data-intensive curriculum.

that can be applied broadly and reinforced through application throughout their academic programs. Furthermore, it allows equal access to these skills rather than creating the "haves" and "have nots" of students with this critical skill set. A drawback, of course, is the difficulty of asking instructors to add material to existing courses, with their potentially limited expertise (Strasser and Hampton 2012), but this transition can be aided by access to appropriate short courses and online resources, including data-discovery tools, example data sets, code, and instructional materials. In a manner similar to the writing centers that are in place at many institutions and that reach out to assist diverse units to enhance quality writing, quantitative learning centers that act as outreach centers for the data-science skills discussed here may be a highly effective strategy to broadly enhance quantitative skills across disciplines. Such centers might support learners through both structured and *ad hoc* tutorials tailored to localized curricula and resident researchers.

Stand-alone university courses. A useful interim step—until such a curriculum-wide approach is adopted—is providing specialized courses akin to scientific writing or statistics. At the undergraduate level, a course on data skills for environmental science early in the curriculum has the advantages of early exposure and common knowledge that helps to encourage continued learning through peer networks. Drawbacks include the fact that many curricula are already quite full and demanding, uncertainty about where such a course would fit within the university structure (e.g., biology, mathematics, or computer science), and whether such an approach might be too generic for specific disciplinary needs.

Coordinating workshop resources and events. In the absence of full integration within university curricula, there are multiple effective mechanisms by which current modes of data-intensive training can continue to have positive influence. The diversity of training opportunities available

outside of universities (e.g., table 1) can help to build skills for individuals that seek them out, train more trainers, produce educational materials used by others, and build a like-minded community that is independent of institutional affiliation. Such programs have many benefits, including raising awareness of data-intensive skills among more established researchers, for whom stand-alone university courses are unlikely. Once a basic understanding of coding has been achieved, students can readily gain additional advanced skills. Many online programs are user paced, such that knowledge can advance rapidly. However, these programs also have drawbacks. There are initial barriers to entry; for example, people with minimal introduction to data-intensive research may not be aware of these opportunities or motivated to enroll. We suggest that the benefits of data-intensive training will be realized more rapidly through sustained coordination that enables discovery of training opportunities and the sharing of materials, lessons learned, and convergence on standards such as training-effectiveness assessment instruments.

Assessment and evaluation. Just as training in the use of data-intensive skills for science has not yet matured, neither has the development of best practices in its implementation. A great need for assessment and evaluation of training approaches exists for data-intensive research skills in science. However, this endeavor will build on a solid foundation of existing knowledge and instruments from related disciplines.

There are well-developed tools available to assist in assessing skill development in certain areas that relate to data science, examples being instruments such as concept inventories including those in statistics (Allen 2006) and computer science (Taylor et al. 2014). These tools have been instrumental in fostering an appreciation for the benefits of active learning methods (Epstein 2013). The development of such instruments in interdisciplinary areas of quantitative science has not yet occurred, although there are instruments to assess comprehension of the nature of science. The variety of topics, concepts, and skills in the burgeoning area of data-intensive science has not yet fostered the development of an inventory-type assessment tool. Such a tool could motivate the diverse stakeholders in data-intensive science education to prioritize effort across the range of potential topics. Differences in such prioritization are to be expected, and we propose that the challenges of data-intensive science in environmental fields, due to the heterogeneity of data types and analysis methods, make this a particularly appropriate area for development of such an assessment tool. If available, it could serve as a model for assessments in other areas of data-intensive research. A useful aspect of such an assessment is that once evaluated in a few settings, it could more readily meet Institutional Review Board (IRB) approval in other settings and enhance the capacity for the community of data-oriented science educators to compare and contrast the benefits of the variety of instruction methods.

Conclusions

The availability of information about the environment is unprecedented and growing at an overwhelming rate as automated sensors, satellite products, and large-scale environmental observatories come online (Jones et al. 2006, Hampton et al. 2015, Peters and Okin 2016). In addition, many more funding organizations are requiring that grant recipients make their independent data sets well documented and publicly available, such that a large pool of heterogeneous environmental data is rapidly becoming easily accessible. It is exciting to contemplate the advances that will be enabled by syntheses of the rich data resources now at hand and how these prospects will steadily grow in the near future. Conducting robust, synthetic analyses of environmental issues in today's rapidly changing world—and indeed performing environmental synthesis in general—requires a broad set of skills and concepts in data integration, analysis, and fundamental computing that currently are not accessible to most researchers, regardless of career stage. And the pace of change in quantitative methods and technology makes it extremely difficult for environmental scientists to stay on the leading edge as society and science move deeper into the information age. Surveys indicate that the scale of the problem is massive and not yet addressed by existing education, training, and mentorship (box 1).

Ideally, universities ultimately will integrate data-intensive skills training into their curricula in a widespread manner that emulates the writing-across-the-curriculum movement that began in the 1970s; however, few institutions have moved in this direction (but see UC Berkeley; *http://databears.berkeley.edu*). In the meantime, across the sciences, a diversity of workshops and online materials have proliferated, demonstrated high demand, and will benefit from systemic coordination that increases their efficiency and users' ability to navigate them. This stage of development in data-intensive research and education provides fertile ground for improving its trajectory in several dimensions. First, sharing and coordinating resources and events across environmental sciences will lower barriers for those seeking training and for instructors seeking support for delivering content. Second, coordinating the evaluation of training effectiveness will improve the future quality of training delivered at various scales. Third, understanding that skills for data-intensive science are core to being a professional scientist will facilitate the integration of training into the university, where existing resources can provide a foundation on which to build activities, courses, and curricula. Finally, targeting training opportunities toward women and underrepresented groups will motivate the creation of a more diverse workforce at the ground floor of this exciting movement (box 2). The rapid growth in data availability and technologies creates not only unprecedented research potential but also the timely opportunity for researchers to establish the standards of scientific

rigor and inclusive community that will define the field for decades.

Acknowledgments

We thank Mike Smorul, Mary Shelley, and members of the Institute for Sustainable Earth and Environmental Software's Workforce Development Working Group for key conversations during idea development, as well as Ben Bolker, Jim Regetz, Steve Katz, Gavin Simpson, and the two anonymous reviewers for suggestions that improved this manuscript.

Funding statement

This work was supported by National Science Foundation award nos. EF-1358900 and ACI-1216894, and the Gordon and Betty Moore Foundation grant nos. 4563 and 4855.

References cited

Al-Bakri M, Fairbairn D. 2012. Assessing similarity matching for possible integration of feature classifications of geospatial data from official and informal sources. International Journal of Geographical Information Science 26: 1437–1456.

Allen K. 2006. The Statistics Concept Inventory: Development and Analysis of a Cognitive Assessment Instrument in Statistics. Social Science Research Network (SSRN). SSRN Scholarly Paper no. 2130143.

Bada M, Stevens R, Goble C, Gil Y, Ashburner M, Blake JA, Cherry JM, Harris M, Lewis S. 2004. A short study on the success of the Gene Ontology. Web Semantics: Science, Services and Agents on the World Wide Web 1: 235–240.

Barnas K, Katz SL. 2010. The challenges of tracking habitat restoration at various spatial scales. Fisheries 35: 232–241.

Baron N. 2010. Escape from the Ivory Tower: A Guide to Making Your Science Matter, 1st ed. Island Press.

Berman FD, Bourne PE. 2015. Let's make gender diversity in data science a priority right from the start. PLOS Biology 13 (art. e1002206).

Brewer CA. 1996. Guidelines for selecting colors for diverging schemes on maps. Cartographic Journal 33: 79–86.

Brewer CA, Buttenfield BP. 2007. Framing guidelines for multi-scale map design using databases at multiple resolutions. Cartography and Geographic Information Science 34: 3–15.

Carpenter SR, et al. 2009. Accelerate synthesis in ecology and environmental sciences. BioScience 59: 699–701.

Cheruvelil KS, Soranno PA, Weathers KC, Hanson PC, Goring SJ, Filstrup CT, Read EK. 2014. Creating and maintaining high-performing collaborative research teams: The importance of diversity and interpersonal skills. Frontiers in Ecology and the Environment 12: 31–38.

Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL. 2011. Science friction: Data, metadata, and collaboration. Social Studies of Science 41: 667–690.

Eigenbrode SD, et al. 2007. Employing philosophical dialogue in collaborative science. BioScience 57: 55–64.

Ellison AM. 2010. Repeatability and transparency in ecological research. Ecology 91: 2536–2539.

Epstein J. 2013. The calculus concept inventory: Measurement of the effect of teaching methodology in mathematics. Notices of the American Mathematical Society 60: 1018–1026.

Ernest M, Brown J, Valone T, White EP. 2015. Portal Project Teaching Database. Figshare. (28 February 2017; *https://figshare.com/articles/Portal_Project_Teaching_Database/1314459*)

Fegraus EH, Andelman S, Jones MB, Schildhauer M. 2005. Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation. Bulletin of the Ecological Society of America 86:158–168.

Fox P, Hendler J. 2011. Changing the equation on scientific data visualization. Science 331: 705–708.

Green JL, et al. 2005. Complexity in ecology and conservation: Mathematical, statistical, and computational challenges. BioScience 55: 501–510.

Hackett EJ, Parker JN, Conz D, Rhoten D, Parker A. 2008. Ecology transformed: The National Center for Ecological Analysis and Synthesis and the changing patterns of ecological research. Pages 277–296 in Olson GM, Zimmerman A, Bos N, eds. Scientific Collaboration on the Internet. MIT Press.

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH. 2013. Big data and the future for ecology. Frontiers in Ecology and the Environment 11: 156–162.

Hampton SE, et al. 2015. The Tao of open science for ecology. Ecosphere 6: 1–13.

Heidorn PB. 2008. Shedding light on the dark data in the long tail of science. Library Trends 57: 280–299.

Hernandez RR, Mayernik MS, Murphy-Mariscal ML, Allen MF. 2012. Advanced technologies and data management practices in environmental science: Lessons from academia. BioScience 62: 1067–1076.

Hey T, Tansley S, Tolle K. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

Hollister JW, Walker HA. 2007. Beyond data: Reproducible research in ecology and environmental sciences. Frontiers in Ecology and the Environment 5: 11–12.

Jones MB, Schildhauer MP, Reichman OJ, Bowers S. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics 37: 519–544.

Joppa LN, et al. 2013. Troubling trends in scientific software use. Science 340: 814–815.

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

Kennedy JB, Kukla R, Paterson T. 2005. Scientific names are ambiguous as identifiers for biological taxa: Their context and definition are required for accurate data integration. Pages 80–95 in Ludäscher B, Raschid L, eds. Data Integration in the Life Sciences. Springer.

Laney CM, Pennington DD, Tweedie CE. 2015. Filling the gaps: Sensor network use and data-sharing practices in ecological research. Frontiers in Ecology and the Environment 13: 363–368.

Langen TA, et al. 2014. Using large public datasets in the undergraduate ecology classroom. Frontiers in Ecology and the Environment 12: 362–363.

Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F. 2007. An ontology for describing and synthesizing ecological observation data. Ecological Informatics 2: 279–296.

Madin JS, Bowers S, Schildhauer MP, Jones MB. 2008. Advancing ecological research with ontologies. Trends in Ecology and Evolution 23: 159–168.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A. 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

McLeod SH. 1989. Writing across the curriculum: The second stage, and beyond. College Composition and Communication 40: 337–343.

Michalowski M, Ambite JL, Thakkar S, Tuchinda R, Knoblock CA, Minton S. 2004. Retrieving and semantically integrating heterogeneous data from the Web. IEEE Intelligent Systems 19: 72–79.

Michener WK. 2006. Meta-information concepts for ecological data management. Ecological Informatics 1: 3–7.

Michener WK, Jones MB. 2012. Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology and Evolution 27: 85–93.

Mokany K, et al. 2016. Integrating modelling of biodiversity composition and ecosystem function. Oikos 125: 10–19.

[NERC] Natural Environment Research Council. 2010. Most Wanted: Postgraduate Skills Needs in the Environment Sector. NERC.

———. 2012. Most Wanted II: Postgraduate and Professional Skills Needs in the Environment Sector. NERC.

Pace ML, et al. 2010. Communicating with the public: Opportunities and rewards for individual ecologists. Frontiers in Ecology and the Environment 8: 292–298.

Peters DPC, Okin GS. 2016. A toolkit for ecosystem ecologists in the time of Big Science. Ecosystems. doi:10.1007/s10021-016-0072-1

Porter JH, Nagy E, Kratz TK, Hanson P, Collins SL, Arzberger P. 2009. New eyes on the world: Advanced sensors for ecology. BioScience 59: 385–397.

Rüegg J, et al. 2014. Completing the data life cycle: Using information management in macrosystems ecology research. Frontiers in Ecology and the Environment 12: 24–30.

Smith D. 2015. Vision and Change in Undergraduate Biology Education: Chronicling Change, Inspiring the Future. American Association for the Advancement of Science.

Smith KA, Imbrie PK. 2007. Teamwork and Project Management. McGraw-Hill.

Soranno PA, et al. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. GigaScience 4 (art. 28).

Strasser CA, Hampton SE. 2012. The fractured lab notebook: Undergraduates and ecological data management training in the United States. Ecosphere 3: 1–18.

Taylor C, Zingaro D, Porter L, Webb KC, Lee CB, Clancy M. 2014. Computer science concept inventories: Past and future. Computer Science Education 24: 253–276.

Teal TK, Cranston KA, Lapp H, White E, Wilson G, Ram K, Pawlik A. 2015. Data Carpentry: Workshops to increase data literacy for researchers. International Journal of Digital Curation 10: 135–143.

Tufte ER. 2001. The Visual Display of Quantitative Information, 2nd ed. Graphics Press.

White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp SR. 2013. Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution 6: 1–10.

Wilson G 2006. Software Carpentry: Getting scientists to write better code by making them more productive. Computing in Science and Engineering 8: 66–69.

Wilson G, et al. 2014. Best practices for scientific computing. PLOS Biology 12 (art. e1001745).

Wuchty S, Jones BF, Uzzi B. 2007. The increasing dominance of teams in production of knowledge. Science 316: 1036–1039.

Zastrow M. 2015. Data visualization: Science on the map. Nature 519: 119–120.

Stephanie E. Hampton (s.hampton@wsu.edu) is affiliated with the Center for Environmental Research, Education and Outreach at Washington State University, in Pullman. Matthew B. Jones is affiliated with the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara. Leah A. Wasser is affiliated with EarthLab at the University of Colorado, in Boulder. Mark P. Schildhauer is with the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara. Sarah R. Supp is affiliated with the University of Maine's School of Biology and Ecology, in Orono. Julien Brun is with the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara. Rebecca R. Hernandez is affiliated with the Land, Air, and Water Resources Department at the University of California, Davis; with the Energy and Resources Group at the University of California, Berkeley; and with the Climate and Carbon Science Program at the Lawrence Berkeley National Lab, in Berkeley, California. Carl Boettiger is affiliated with the Department of Environmental Science, Policy, and Management at the University of California, Berkeley. Scott L. Collins is with the Department of Biology at the University of New Mexico, in Albuquerque. Louis J. Gross is affiliated with the Departments of Ecology and Evolutionary Biology and Mathematics at the University of Tennessee, in Knoxville. Denny S. Fernández is with the Department of Biology at the University of Puerto Rico at Humacao. Amber Budden is affiliated with DataONE at the University of New Mexico, in Albuquerque. Ethan P. White is with the Department of Wildlife Ecology and Conservation and The Informatics Institute at the University of Florida, in Gainesville. Tracy K. Teal is affiliated with Data Carpentry, in Davis, California. Stephanie G. Labou is with the Center for Environmental Research, Education and Outreach, at Washington State University, in Pullman. Juliann E. Aukema is affiliated with the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara.