Wildlife Society Bulletin; DOI: 10.1002/wsb.847

In My Opinion

Wildlife Biology, Big Data, and Reproducible Research

KEITH P. LEWIS,[1,2] *Department of Biology, Memorial University of Newfoundland, St. John's, NL A1B 3X9, Canada*
ERIC VANDER WAL, *Department of Biology, Memorial University of Newfoundland, St. John's, NL A1B 3X9, Canada*
DAVID A. FIFIELD, *Division of Wildlife Research, Environment and Climate Change Canada, 6 Bruce Street, Mount Pearl, NL A1N 4T3, Canada*

ABSTRACT Changes in technology have made it possible to gather vast amounts of data, often of high quality, that in turn can improve the quality of wildlife biology. However, with this growth in data, practices such as data management, exploratory data analysis, data-sharing, and reproducibility of an analysis have become increasingly complex. These practices often depend heavily on computer scripting languages, and are often hidden from the peer-review process despite their influence on the final results. Although these issues have been discussed in the literature, they are generally dealt with in a piecemeal fashion, preventing synthesis, and thereby slowing progress. We offer a conceptual framework to illustrate relationships among these practices, and show where wildlife biology as a field has embraced these changes, where awareness is growing, and where it lags behind other fields. We then present several case studies to emphasize the importance of adopting these practices. Any of these case studies could have been conducted with little attention to these practices or employing scripting languages, but there are many disadvantages to this approach including increased chance of errors, inefficiency, and lack of reproducibility. We suggest that a change in the culture of how wildlife biology is conducted is required and that this change will be fostered by integrating these practices into wildlife biology education, implementation, and embracing the idea of open data and open computer code.

KEY WORDS data management, data pipeline, exploratory data analysis, open science, reproducible research.

Changes in technology continually transform how science is conducted. Recently, technological changes have made available previously unimaginable amounts of, generally, very accurate data. This "data deluge" is generally a welcome trend that has brought aspects of wildlife biology into the realm of "big data" (Borgman et al. 2007; Michener and Jones 2012; Hampton et al. 2013, 2015; Kays et al. 2015; big data are defined here as unprecedented volumes and varieties of data that arrive at high velocities [Science International 2015]), or ecoinformatics (Michener and Jones 2012). Greater volumes of more accurate data enable studies to address questions at increasingly large spatial and temporal scales with stronger inference, as well as more accurate and predictive models, which, in turn, yield important biological insights (Baru et al. 2012, Michener and Jones 2012, Kays et al. 2015). However, as is often the case, new technologies can bring unintended problems. As data sets grow in size, it is increasingly clear that traditional means of data management, such as storing and manipulating data using spreadsheets, are no longer adequate (Urbano et al. 2010, Valle and Berdanier 2012, Chamanara and König-Ries 2014, Hampton et al. 2015, Kays et al. 2015), and new approaches are required for analyzing data (e.g., ecoinformatics, data science; Michener and Jones 2012, Valle and Berdanier 2012). Further, as analyses grow increasingly complicated and sophisticated, steps in the "data pipeline" that precede formal statistical tests, such as data processing and exploratory data analysis, increasingly influence the final outcome of the study (Fig. 1; Leek and Peng 2015*a*). Yet these steps have previously received little attention and are rarely scrutinized as part of the peer-review process, prompting calls for an increase in the openness and reproducibility of science (Peng 2009, 2011; Hampton et al. 2015). Open science or reproducible research can include data-sharing (i.e., open data, data archiving; Whitlock et al. 2010, Michener 2015), as well as sharing the code used to produce the analysis (Peng 2009, Hampton et al. 2015). Proponents of open science and reproducible research suggest that making data and code openly available (Fig. 1) will bring an array of benefits including creating a more productive and responsible scientific culture (Borgman et al. 2007, Peng 2009, Michener and Jones 2012, Hampton et al. 2015) and an ability to address larger and more

Received: 28 February 2017; Accepted: 2 October 2017

[1]*E-mail: keithl@mun.ca*
[2]*Present Address: Department of Fisheries and Oceans, NorthWest Atlantic Fisheries Centre, 80 East White Hills Road, PO Box 5667, St. John's, NL, Canada A1C 5X1*

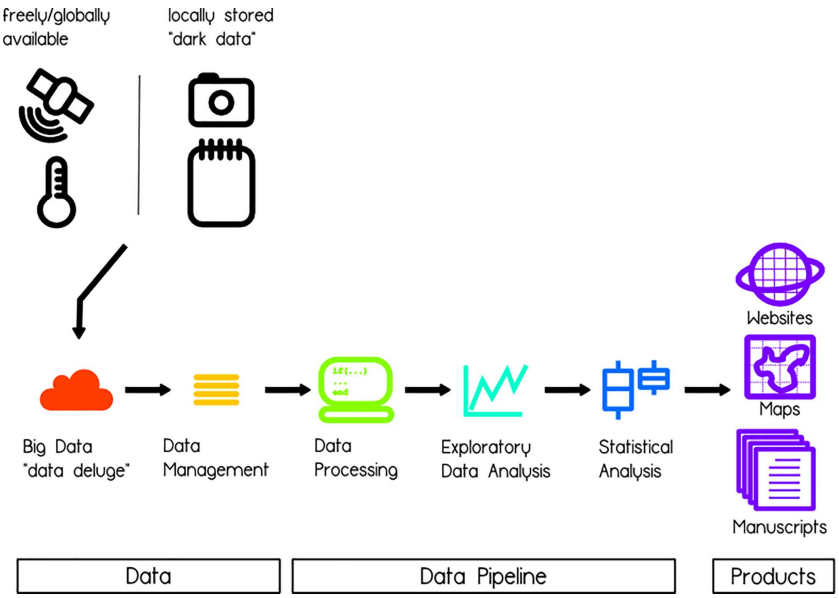

Figure 1. A conceptual framework for open science and reproducible research in wildlife biology. The "data pipeline" consists of a set of steps for processing, exploring, and statistically analyzing data and often depends on computer code. Big data can often be used to create a variety of products. Multiple documents and associated outputs such as appendices or maps can be managed and revised with a content management system. This framework represents an ideal but will often not be linear in practice. For example, exploratory data analysis may suggest further steps need to be taken during data processing or errors in the quality assurance/control process.

complex questions (Michener and Jones 2012). A more immediate and practical benefit of reproducible research may simply be that the author has developed a more efficient workflow, has properly documented their own analysis and can therefore, rapidly reproduce or update their own work (Fomel and Claerbout 2009, Wilson et al. 2016, Lowndes et al. 2017).

The practices discussed above, or the extent to which they can be employed, may not be familiar to all wildlife ecologists because they have generally been pioneered in other fields (e.g., bioinformatics [Peng 2009], climatology [Steenweg et al. 2016], oceanography, biodiversity [Michener and Jones 2012], and genetics [Hampton et al. 2013]). As such, these practices have only recently been discussed in the wider ecological literature, and have not yet become a part of the lexicon or culture of our field. Furthermore, they are usually dealt with piecemeal in the literature, slowing the synthesis of these ideas and making it more difficult for scientists to readily apply them in practice or education. Therefore, we suggest a conceptual framework to illustrate relationships among these practices and suggest this framework as a starting point to further a change in the culture of how wildlife biology is taught and implemented (Fig. 1). We argue that wildlife biology, like other subdisciplines of ecology, would benefit by adopting this conceptual framework (i.e., fully documenting data management, data processing, exploratory data analysis, as well as the formal statistical analysis in an open and reproducible manner). First, we illustrate what we view as the current state of wildlife biology and show where it has embraced these practices, where awareness is growing, and where it lags behind other fields (Table 1). We then present several case studies to emphasize the importance of adopting this framework and conclude with suggestions for furthering their adoption and empowering the necessary cultural changes.

THE STATE OF PRACTICE IN WILDLIFE BIOLOGY

The current state of wildlife biology contrasts sharply with the past when wildlife biology was viewed as a data-poor field (Kays et al. 2015). Consider animal movement studies. Beginning in the 1960s, Very-High-Frequency (VHF) telemetry tags (collars) provided relatively small amounts of data; relocation rates were low; and resulting sample sizes limited the rigor of many studies. Indeed, the study of animal movement was traditionally not part of mainstream ecology because of sparse data (Kays et al. 2015). The advent of technological changes in tracking devices, such as Geographical Positioning System (GPS) tags, have revolutionized the field enabling collection of large quantities of accurate location data on animal movement; this technology has been rapidly adopted by wildlife biologists. This shift from sparse to large volumes of data is certainly not limited to animal movement studies. Other technological innovations such as bio-loggers and autonomously triggered cameras

Table 1. A summary of the general state of practice in wildlife biology in response to recent changes in technology.

Risen to the challenge	Growing awareness	Lag behind
Large data sets	Data management	Reproducible research
Statistical sophistication	Data processing	Exploratory data analysis
		Data sharing

(often known as camera traps) have created floods of data, as have incorporation of remote sensing data and genetics into wildlife studies providing insights into a wide range of ecological fields (Urbano et al. 2010, Reichman et al. 2011, Kays et al. 2015, Wilmers et al. 2015, Steenweg et al. 2016).

Wildlife biology has also adopted increasingly sophisticated statistical techniques. Consider habitat selection studies. Prior to the early 2000s, data from such studies were typically analyzed with univariate use/availability selection ratios, often with dozens of points per animal and relatively rudimentary land cover maps (Jones 2001, Kays et al. 2015). With the advent of GPS tags and readily available remote sensing data, thousands of data points per animal are the norm and detailed land cover maps are available (Borgman et al. 2007, Wilmers et al. 2015). These data are routinely used in resource selection analyses, such as resource selection functions (McLoughlin et al. 2010), step (Fortin et al. 2005), and integrated step selection analyses (Avgar et al. 2016) as part of a habitat selection studies. The study design associated with resource selection analyses must be carefully considered so that proper inferences can be made (Gillies et al. 2006, Thomas and Taylor 2006, Northrup et al. 2013). Similarly, resource selection analyses are routinely analyzed with Generalized Linear Mixed Models or Generalized Estimating Equations necessitating a solid statistical education and experience (Gillies et al. 2006, Koper and Manseau 2009). Increased statistical sophistication has also occurred with other methodologies such as camera traps where spatially explicit capture–recapture models are routinely employed (Chandler and Royle 2013).

The emergence of wildlife biology as a big-data field has also created a growing awareness of new challenges such as a vast increase in the time and resources devoted to data management and data processing (Kays et al. 2015, Leek and Peng 2015*a*). It has recently been recognized that failure to properly manage large data sets results in inefficient analyses, loss of data, and a lack of reproducibility (Urbano et al. 2010, Hampton et al. 2013, Kays et al. 2015). Big data have also created new challenges in terms of data processing. For example, data from GPS collars are often fraught with errors and require extensive processing or "cleaning" (Bjørneraas et al. 2010). Although VHF data could be manipulated with a spreadsheet, this becomes impractical when hundreds of thousands or millions of records are involved. The reality of modern-day analytical workflows with big data is that data processing, even with good data management structures in place, requires a substantial amount of time, and decisions made at this stage can have a profound effect on the outcome of a study.

Despite the increasing scale, sophistication, and complexity of analytical workflows, too often these decisions and the code used to execute them are opaque and not presented within the body or appendices of peer-reviewed manuscripts (i.e., they are not reproducible). In addition, exploratory data analysis, although often critical to the outcome of the study, typically receives little attention in most methods sections, statistical text books, or as part of general statistical training despite champions of such distinction as John Tukey (see Tukey 1977, Zuur et al. 2010, Leek and Peng 2015*a*). Finally, data-sharing is common in other fields such as oceanography, remote sensing, biodiversity, and genetics but less so in wildlife biology (Michener and Jones 2012, Hampton et al. 2013). We suggest that wildlife biology lags behind some other fields in these areas and we should collectively consider benefits of adopting improved documentation of our workflows, exploratory data analysis, and data-sharing.

CASE STUDIES

The following case studies illustrate that the practices synthesized in Figure 1 are critical for conducting modern ecology, and intellectual as well as human resources must be devoted to these areas. In all cases, it rapidly became clear as we conducted these studies that the traditional practices were either impractical or completely infeasible, and availability of big data presented new and unique challenges. We surmise that our experiences conducting these studies are not unique. On the contrary, we suspect that these experiences are becoming increasingly common place, not only for the types of data that we describe below, but for many others as well (Lowndes et al. 2017). In addition to the problems created by big data, we offer our experiences as one potential solution in the areas of data management, data processing, and exploratory data analysis because these have received less treatment in the literature than other aspects of the data pipeline (Fig. 1).

Data Management and Processing With Big Data

Filtering biotelemetry data for spatial statistical analyses.—Biologists have used GPS tags to better understand, among other things, caribou (*Rangifer tarandus*) habitat selection, survival, and management (Bastille-Rousseau et al. 2015, Lewis et al. 2017). Data from GPS tags have traditionally been stored locally in flat files (e.g., spreadsheets) and examined using a variety of software by a single analyst. Such workflows are unlikely to ever be reproducible, error prone, and tedious to manage. An example of how the traditional system became unworkable is animal tagging in Newfoundland and Labrador, Canada. Since 1979, 2,744 animals—including caribou, black bears (*Ursus americanus*), eastern coyote (*Canis latrans*), and lynx (*Lynx canadensis*)—have been fitted with VHF, ARGOS platform transmitter terminals (PTTs), and GPS tags in Newfoundland producing >2 million fixes. In addition to the volume of these data, tags were purchased from a variety of vendors over the 30+ years, and, depending on the vendor, data files from the PTTs and GPS tags came in a variety of formats (column meanings and order varied, and date fields were displayed in multiple ways) and a variety of means (e-mail, downloads, etc.) with varying schedules (2 hr, 4 hr, days) and delivery times (velocity; Campbell et al. 2016). Furthermore, data from these tags are often fraught with errors such as records with no data, erroneous fixes, duplicate fixes, and bursts of 5–20 fixes being taken in a short time period (e.g., 5 min). These data were originally processed and manipulated in spreadsheets, but it became clear that this approach was inefficient and inappropriate.

To remedy these many problems, a custom-built tool, written in the Python programming language and integrated into the ArcGIS environment, was created to automatically load data from the many different file and format types directly into a Microsoft ACCESS database (Microsoft Corp., Redmond, WA, USA). Code was written to remove the many types of errors and produce a "clean" data set depending on the question at hand. A simple, but powerful, menu was created in ACCESS to perform queries by animal, species, sex, collar type, location, a set geographic area, and the type of errors to be automatically removed could be specified.

Live capture and management of seabird abundance data at sea.—Environment Canada's Eastern Canada Seabirds at Sea program monitors the pelagic distribution of birds at sea in eastern Canadian waters using distance sampling line-transect surveys from ships and aircraft (Buckland et al. 2001, Fifield et al. 2009). Since 1965, >450,000 observations have been collected during >600 surveys. Traditionally, seabird observations were recorded on paper (which required observers to divert their attention away from observing) or on tape, for later, error-prone transcription. Additionally, ancillary vessel data including ship position, speed, and heading, were read from vessel instruments on the bridge with attendant transcription errors.

To improve the workflow for observers, data managers, and analysts, we created a Microsoft ACCESS database featuring voice recognition technology that allows observers to dictate observations while concentrating on observing the birds at sea (Robertson et al. 2012). Automated quality checks occur during live data entry to ensure adherence to the survey protocol, and the database automatically records ancillary vessel data through an electronic connection to the vessel (Gjerdrum et al. 2012). After a survey trip, the database administrator imports data from the observers' computer into the master database using a simple menu-driven interface that includes a suite of data quality checks. The database provides a feature-rich data query form to allow flexible data extraction in several formats suitable for mapping and direct import to downstream analysis steps such as Distance Sampling (Thomas et al. 2010). With this approach, we have essentially eliminated transcription errors, and post-observation data-processing time is greatly reduced facilitating a more rapid formal analysis.

Geolocator data loggers and data processing for spatial statistical analyses.—Using geolocators, Fifield et al. (2014) identified important migratory and overwintering hotspots for northern gannets (*Morus bassanus*) in North America ($n = 65$ migrations) and showed that a few even cross the Atlantic Ocean and winter off northern Africa. Geolocators capture raw light-level data, which are converted to latitude and longitude based on day length and time of local noon, respectively (Wilson et al. 1992, Hill 1994). Prior to testing of ecological hypotheses (e.g., colony- or sex-based differences in duration of migration or size of wintering range), these data required a considerable amount of processing. Data processing included converting raw light data ($n > 50.5$ million light measurements) to geographic locations, excluding positions at the colony before or after migration, filtering of consecutive positions requiring unrealistic speeds, interpolation of missing positions, smoothing of the interpolated track with a boxcar filter, extraction of behavioral modes (e.g., migrating, at stopover site, at winter grounds) with a set of decision rules resulting in 22,871 estimated locations, and finally calculating the research quantities of interest including migration timing-duration, winter home-range size, site fidelity, etc.

Many of these data processing steps are mathematically complex and involve a variety of input parameters. This data processing workflow potentially needed to be repeated multiple times for each of the 65 geolocator data sets as improved methods became available to produce the initial geographic coordinates (e.g., sea surface temp correction; Teo et al. 2004), or to investigate the effect of changes in parameter values. Given these constraints, it was clear that the data processing workflow would need to be efficient and reproducible. An integrated set of programs were written in the Program R statistical programming language (R Development Core Team 2015) and Matlab (Mathworks, Natick, MA, USA) that applied all processing steps while managing all intermediate files to produce final migration tracks for each bird. This approach allowed the entire data processing workflow to be repeated any number of times, drastically reducing processing time and avoiding mistakes. Subsequently, as new geolocator data, or as new approaches and tools for generating the initial geographical locations from the raw light data have become available, the entire data-processing workflow (as well as the formal statistical analyses) could be redone simply by rerunning the code (e.g., GeoLight, Lisovski and Hahn 2012, Rakhimberdiev et al. 2015).

Exploratory Data Analysis with Big Data

Spatial ecology of eastern coyote.—Fifield et al. (2013) conducted an examination of coyote spatial ecology in Newfoundland including home range size, daily movement rate, and site fidelity by season and sex using GPS and PTTs tags ($n = 79$ with 68,286 positions; see also Ellington 2015). As elsewhere, Newfoundland coyote exhibit ≥ 2 broad types of space use behavior: residency and transience (Boisjoly et al. 2010). Resident coyotes have home ranges of approximately 250 km^2, whereas, transient coyotes range widely and can cross Newfoundland in approximately 20 days, covering a distance of >1,000 km. Before proceeding to further analyses, animals needed to be categorized in 1 of these 2 behavioral modes. This categorization was critical to produce category-specific estimates of spatial ecology parameters including home range size, movement rate, and site fidelity.

After preprocessing these data using the aforementioned tool for cleaning biotelemetry data, the maximum difference between any 2 locations was calculated for each animal ($n = 79$). A histogram of these distances was constructed for each animal and animals with distances greater than the 75th percentile of distances classified as transient. This threshold was subjectively chosen, but was further assessed by mapping each individual. There was good correspondence between

this threshold and the maps. Other approaches to behavior classification were also attempted, although the one described here was the most robust (Fifield et al. 2013). This process was made reproducible by conducting the workflow in the R statistical programming language (R Development Core Team 2015), which allowed for the assessment of repeated runs of the software each with different parameters to arrive at the optimal characterization approach. In addition, the same code can be reused to identically process new data from other animals or years as it becomes available.

Lessons From Case Studies

We acknowledge that all of the above case studies could have been carried out using flat data files (e.g., spread sheets) and manual data manipulation (e.g., manually down loading data files from e-mails or from the World Wide Web, manually removing erroneous or duplicate points in a spreadsheet, cutting–pasting of data, summarizing data using pivot tables, making maps, and calculating metrics individually); adopting such an approach could have likely yielded similar results. Yet in all cases, this course of action would present several serious disadvantages. First, mistakes in manual data processing are difficult to detect and harder to rectify. Usually, one must begin the whole process over from the point at which the mistake was found. Subsequent changes are then often forgotten due to lack of documentation resulting in further errors. Second, discussions with colleagues, as well as reading the literature or online forums, can suggest adjustments to ways to analyze a given data set. It is simply not practical to start an analysis over multiple times using more traditional approaches. Finally, big data are often used for multiple projects including government reports, student theses, and academic papers. Querying and manipulating these data for efficient execution of these various projects was only possible with relational databases and scripts given the available human resources and time frames.

We also stress that simply building data management tools is not enough. Data management requires planning and forethought just as any other part of a scientific endeavor (Baru et al. 2012, Michener and Jones 2012, Sutter et al. 2015). Data management also requires dedicated human resources, often in the form of a data architect or curator, to successfully manage large and error-prone data sets and provide quality assurance and control (Sergeant et al. 2012, Kolb et al. 2013, Hunt et al. 2015, Science International 2015, Mislan et al. 2016).

NEXT STEPS

This is an exciting time in wildlife biology with technology creating many positive changes. Some authors have characterized this time as a "golden age" (Wilmers et al. 2015), a threshold (Hampton et al. 2015), or transition between old ways of doing things and new (Borgman et al. 2007), and a quiet revolution (Sutter et al. 2015, Borregaard and Hart 2016). Sutter et al. (2015) have even called for a cultural change in science to better take advantage of these developments. We echo these statements and have integrated a variety of practices that have arisen in recent years and presented them in the context of wildlife biology (Fig. 1). In many ways, wildlife biology has admirably risen to these challenges as we illustrate in the Case Studies, although there are areas where awareness is growing, and where it still lags behind other fields such as bioinformatics, oceanography, and genetics.

However, to fully take advantage of these practices, we need a change in how we conduct wildlife biology and educate the next generation of wildlife biologists. What needs to occur for these practices to become standard? Do we want to actively embrace the change in culture or let it proceed in an *ad hoc* manner? In the fourth part of this paper, we offer 3 suggestions for fostering change as well as some caveats that may hinder it.

Next Steps—Changing Wildlife Biology

First, one of the most important ways in which we can change the current culture is through education. Leek and Peng (2015*b*) emphasized the importance of education in improving reproducible research. Data management, for example, used to be a minor task in many studies and was rarely discussed or taught a generation ago. We agree with Rüegg et al. (2014) and Sutter et al. (2015) that most environmental scientists have little expertise in data management practices and personal experience suggests that data management is usually dealt with in a *post hoc* and *ad hoc* manner rather than as a planned exercise. Indeed, Hernandez et al. (2012) found that only 26% of ecology graduate students ever created metadata and 28% did not know what it meant to create metadata. We also suspect that the majority of ecologists continue to manage and process data using spreadsheets although these are poor substitutes for relational databases (Michener and Jones 2012, Kolb et al. 2013, Michener 2015).

There is also a pressing need for teaching computer programming skills that can expedite all of the above processes and make them more reproducible (Fig. 1; Valle and Berdanier 2012, Baumer et al. 2014, Hunt et al. 2015). However, the vast majority of scientists are largely self-taught as computer programmers (Wilson et al. 2014). We suggest that the most expedient means to incorporate these concepts into standard wildlife ecology practice is by incorporating them into standard ecology and wildlife biology educational curricula. Alternatives to traditional academic education are nonprofit organizations that teach best practices in software development as well as data management and analysis (e.g., Software Carpentry [https://software-carpentry.org/] and Data Carpentry [http://www.datacarpentry.org/]). Regardless of where computer programming skills are learned, relying on trial and error or "learn-as-you-go" is not an efficient or responsible way to conduct research (especially when publicly funded) or rapidly respond to pressing conservation and management concerns.

Second, in addition to improving education, traditional peer review of manuscripts is no longer an adequate approach to determine whether research is of an acceptable caliber because many of the steps in the analysis are not reviewed or

even reproducible (Ioannidis 2013, Baumer et al. 2014). Lack of reproducibility can be remedied if the final product of the research is not viewed as a manuscript, but as the full computational environment, which includes data, its management, and code used in the analysis (Peng 2011, Xie 2013). Fortunately, there are many tools available for conducting reproducible research such as Standard Query Language that can record information critical to reproducing a database query, and script-based statistical languages such as R (R Development Core Team 2015) that record and perform data management, data processing, and many analytical functions. Indeed, simply using a script-based analysis is a major step toward reproducible research. Once the formal analysis has been completed, the results of interest (i.e., computational results, figures, or tables) can be easily integrated into dynamic documents, which are constructed by embedding the analysis code into the document text (Xie 2013). This can be accomplished using fairly simple facilities such as Program R Markdown (Baumer et al. 2014) or in more publication-quality style with LaTeX (Xie 2013); both approaches can be implemented conveniently in RStudio (Gandrud 2014), a popular R Graphical User Interface. Subsequently, each time such a dynamic document is "reprocessed" all the embedded analytic code is rerun, which automatically updates any embedded tables, figures, etc.; thereby, keeping the document perennially up to date as code and input data change.

Augmenting this facility are readily available management tools that range from the relatively simple content management systems (e.g., GitHub—https://github.com/, Bitbucket) that simplify version control, foster collaboration, and make research more transparent, to complex scientific workflow systems such as Kepler (Curcin and Ghanem 2008, Reichman et al. 2011: for an expansive list of tools, see Hampton et al. 2013, Michener 2015). Finally, a content management system for documents (e.g., WordPress—https://wordpress.com/) can manage multiple code and manuscript revisions, distributed authorship, as well as production, display, and distribution of multiple outputs (e.g., manuscripts, appendices, etc.). Clearly, time and expertise must be invested in learning and adopting tools of reproducible research, and this investment can be substantial. However, once the investment has been made, it brings the benefits of open and reproducible research as well as improvements in work habits, teamwork, and potentially, greater impact research as well as increased public trust (Gandrud 2014, Leek and Peng 2015*b*).

Last, if science is to be truly reproducible, then we must foster a cultural shift in the sharing of data and code. Failure to share and archive data results in "dark data" (i.e., potentially valuable data that is essentially inaccessible; Whitlock et al. [2010], Hampton et al. [2013]). Data-sharing is common in other fields such as astronomy, oceanography, and genomics to the benefit of these fields as a whole (Reichman et al. 2011). The shift to more open data seems to have begun in ecology as evidenced by a number of authors who have discussed data-sharing (Whitlock et al. 2010, Baru et al. 2012, Michener 2015, Mills et al. 2015, Science International 2015) and the increasing number of ecologically related data repositories such as DataONE (https://www.dataone.org/), MoveBank (www.movebank.org), and Wildlife Insights (https://www.wildlifeinsights.org/WMS/#/). Likewise, the sharing of code is equally important because nowhere else are the details and intricacies of the data processing pipeline so evident. We concur with many authors that sharing code would benefit science by 1) increasing analytical efficiency by allowing the reuse of relevant data and code, 2) identifying problems with existing code and speed the fixing of these problems, 3) fostering the ability to ask new questions by combining data and analyses from multiple sources, and 4) making the scientific process more transparent and reproducible (Michener 2015). This is perhaps best expressed by Sutter et al. (2015:461) "What is called for is a culture shift away from viewing data as a single-purpose, 'consumable' item, toward that of developing a valuable, irreplaceable resource that may even increase in value over time." We suggest this sentiment applies to computer code as well.

Caveats

We acknowledge that completely open and reproducible research will not always be easy or in some cases possible. In addition to the aforementioned investment in time and expertise, ecological studies were traditionally conducted over short spatial and temporal scales and ecological data are often not standardized because of differences in experimental protocols and field methods that vary by organisms, ecosystems, and questions (Reichman et al. 2011, Hampton et al. 2013). Heterogeneity among data sets makes archiving and sharing data difficult and hinders subsequent, synthetic analyses. Improving metadata practices and standardizing field protocols will partially ameliorate this issue (Baru et al. 2012, Steenweg et al. 2016), but cannot fully for the reasons listed above.

Further, even if all code and data are provided, it may not be feasible in practice for the research to be completely scrutinized. How realistic is it for editors and reviewers, people who are usually vastly overcommitted, to check the code for all but the most cursory analysis (Banks 2011)? Even coauthors are unlikely to be able to properly vet an analysis if they lack the analytical background or are engaged in multidisciplinary research where it is not possible to deeply scrutinize work outside of one's field of expertise. In addition, it may simply not be practical to reproduce analyses that involve terabytes of data without access to large data storage and very fast computers. Time constraints, computer speed, and limits to one's personal expertise are very real issues that must be acknowledged if a shift toward open and reproducible research is not to be a token gesture.

Finally, there is the issue of data ownership. There is still understandable apprehension among some researchers of granting unfettered access to hard-earned data (Mills et al. 2015). In addition, researchers often do not own the data they are analyzing, or ownership is unclear, involving multiple partners, especially in wildlife ecology. Open data can be particularly problematic for government scientists.

Governments are often the owners of large data sets and, as curators of these data, government scientists frequently contribute to the academic literature (e.g., The Wildlife Society journals). However, some governments may not be willing to release data gathered at public expense to academic journals. Although there are many obvious exceptions, and views on data provenance are changing, journals should be cognizant that government scientists may not have a free hand in how their data are handled (Obama 2013, Hunt et al. 2015, Sutter et al. 2015).

CONCLUSION

In conclusion, we think that this is a transitional time in wildlife ecology; with marked advances in technology and abundance of data comes an imperative for matching transparency and reproducibility. Tremendous advances in the scale and complexity of studies are possible, but we must change our culture (i.e., our research and educational practices) if we are to take full advantage of all the benefits of the new age (Fig. 1). Following Fomel and Claerbout (2009), we urge researchers to ask themselves the following question before they publish: "have I done enough to allow the readers of my paper to verify and reproduce my ... experiment?"

ACKNOWLEDGMENTS

We thank F. Norman for his help and his role in inspiring this paper as well as the many members of the Wildlife Evolutionary Ecology Lab at Memorial University for considering previous versions of this manuscript. We thank 2 anonymous reviewers for their helpful comments as well as Roger Applegate, the Associate Editor.

LITERATURE CITED

Avgar, T., J. R. Potts, M. A. Lewis, and M. S. Boyce. 2016. Integrated step selection analysis: bridging the gap between resource selection and animal movement. Methods in Ecology and Evolution 7:619–630.

Banks, D. 2011. Reproducible research: a range of response. Statistics, Politics, and Policy 2:4.

Baru, C., E. H. Fegraus, S. J. Andelman, S. Chandra, K. Kaya, K. Lin, and C. Youn. 2012. Cyberinfrastructure for observatory and monitoring networks: a case study from the TEAM network. Bioscience 62:667–675.

Bastille-Rousseau, G., J. R. Potts, J. A. Schaefer, M. A. Lewis, E. H. Ellington, N. D. Rayl, S. P. Mahoney, and D. L. Murray. 2015. Unveiling trade-offs in resource selection of migratory caribou using a mechanistic movement model of availability. Ecography 38:1049–1059.

Baumer, B., M. Cetinkaya-Rundel, A. Bray, L. Loi, and N. J. Horton. 2014. R Markdown: integrating a reproducible analysis tool into introductory statistics. Technology Innovations in Statistics Education 8:1–29.

Bjørneraas, K., B. Van Moorter, C. M. Rolandsen, and I. Herfindal. 2010. Screening global positioning system location data for errors using animal movement characteristics. Journal of Wildlife Management 74:1361–1366.

Boisjoly, D., J.-P. Ouellet, and R. Courtois. 2010. Coyote habitat selection and management implications for the Gaspésie caribou. Journal of Wildlife Management 74:3–11.

Borgman, C. L., J. C. Wallis, and N. Enyedy. 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries 7:17–30.

Borregaard, M. K., and E. M. Hart. 2016. Towards a more reproducible ecology. Ecography 39:349–353.

Buckland, S. T., D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas. 2001. Introduction to distance sampling estimating abundance of biological populations. Oxford University Press, Oxford, England, United Kingdom.

Campbell, H. A., F. Urbano, S. Davidson, H. Dettki, and F. Cagnacci. 2016. A plea for standards in reporting data collected by animal-borne electronic devices. Animal Biotelemetry 4:1.

Chamanara, J., and B. König-Ries. 2014. A conceptual model for data management in the field of ecology. Ecological Informatics 24:261–272.

Chandler, R. B., and J. A. Royle. 2013. Spatially explicit models for inference about density in unmarked or partially marked populations. Annals of Applied Statistics 7:936–954.

Curcin, V., and M. Ghanem. 2008. Scientific workflow systems-can one size fit all? Pages 1–9 *in* 2008 Cairo International Biomedical Engineering Conference. IEEE, Cairo, Egypt. DOI: 10.1109/CIBEC.2008.4786077

Ellington, E. H. 2015. Beyond habitat: individual and population-level drivers of coyote space use. Dissertation, Trent University, Peterborough, Ontario, Canada.

Fifield, D. A., K. P. Lewis, C. Gjerdrum, G. J. Robertson, and R. Wells. 2009. Offshore seabird monitoring program. Environment Studies Research Funds Report 183. Environment Canada – Atlantic Region, St. John's, Newfoundland and Labrador, Canada.

Fifield, D. A., W. A. Montevecchi, S. Garthe, G. J. Robertson, U. Kubetzki, and J.-F. Rail. 2014. Migratory tactics and wintering areas of northern gannets (*Morus bassanus*) breeding in North America. Ornithological Monographs 79:1–63.

Fifield, D. A., K. Unger, K. P. Lewis, S. E. Gullage, and S. P. Mahoney. 2013. Spatial ecology of black bear (*Ursus americanus*), coyote (*Canis latrans*) and lynx (*Lynx canadensis*) in Newfoundland. Technical Bulletin No. 007, Sustainable Development and Strategic Science, Department of Environment and Conservation, Government of Newfoundland and Labrador, St. John's, Canada.

Fomel, S., and J. F. Claerbout. 2009. Reproducible research. Computing in Science & Engineering 11:5–7.

Fortin, D., H. L. Beyer, M. S. Boyce, D. W. Smith, T. Duchesne, and J. S. Mao. 2005. Wolves influence elk movements: behavior shapes a trophic cascade in Yellowstone National Park. Ecology 86:1320–1330.

Gandrud, C. 2014. Reproducible research with R and RStudio. Second edition. CRC Press, Taylor & Francis Group, Boca Raton, Florida, USA.

Gillies, C. S., M. Hebblewhite, S. E. Nielsen, M. A. Krawchuk, C. L. Aldridge, J. L. Frair, D. J. Saher, C. E. Stevens, and C. L. Jerde. 2006. Application of random effects to the study of resource selection by animals. Journal of Animal Ecology 75:887–898.

Gjerdrum, C., D. A. Fifield, and S. I. Wilhelm. 2012. Eastern Canada Seabirds at Sea (ECSAS) standardized protocol for pelagic seabird surveys from moving and stationary platforms. Canadian Wildlife Service Technical Report, Canadian Wildlife Service, Atlantic Region, Sackville, New Brunswick, Canada.

Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W. C. Lenhardt, A. MacDonald, W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P. Schildhauer, K. H. Woo, and N. Zimmerman. 2015. The tao of open science for ecology. Ecosphere 6(7):120.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment 11:156–162.

Hernandez, R. R., M. S. Mayernik, M. L. Murphy-Mariscal, and M. F. Allen. 2012. Advanced technologies and data management practices in environmental science: lessons from academia. BioScience 62:1067–1076.

Hill, R. D. 1994. Theory of geolocation by light levels. Pages 227–236 *in* B. J. Le Boeuff and R. M. Laws, editors. Elephant seals: population ecology, behavior, and physiology. University of California Press, Berkeley, USA.

Hunt, V. M., S. K. Jacobi, M. G. Knutson, E. V. Lonsdorf, S. Papon, and J. Zorn. 2015. A data management system for long-term natural resource monitoring and management projects with multiple cooperators. Wildlife Society Bulletin 39:464–471.

Ioannidis, J. P. A. 2013. This I believe in genetics: discovery can be a nuisance, replication is science, implementation matters. Frontiers in Genetics 4:33. DOI: 10.3389/fgene.2013.00033

Jones, J. 2001. Habitat selection studies in avian ecology: a critical review. Auk 118:557–562.

Kays, R., M. C. Crofoot, W. Jetz, and M. Wikelski. 2015. Terrestrial animal tracking as an eye on life and planet. Science 348:1–9.

Kolb, T. L., E. A. Blukacz-Richards, A. M. Muir, R. M. Claramunt, M. A. Koops, W. W. Taylor, T. M. Sutton, M. T. Arts, and E. Bissel. 2013. How to manage data to enhance their potential for synthesis, preservation, sharing, and reuse—a Great Lakes case study. Fisheries 38:52–64.

Koper, N., and M. Manseau. 2009. Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. Journal of Applied Ecology 46:590–599.
Leek, J. T., and R. D. Peng. 2015*a*. Statistics: P values are just the tip of the iceberg. Nature 520:612.
Leek, J. T., and R. D. Peng. 2015*b*. Reproducible research can still be wrong: adopting a prevention approach. Proceedings of the National Academy of Sciences of the United States of America 112:1645–1646.
Lewis, K. P., S. E. Gullage, D. A. Fifield, D. H. Jennings, and S. P. Mahoney. 2017. Manipulations of black bear and coyote affect caribou calf survival. Journal of Wildlife Management 81:122–132.
Lisovski, S., and S. Hahn. 2012. GeoLight-processing and analysing light-based geolocator data in R. Methods in Ecology and Evolution 3:1055–1059.
Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang, and B. S. Halpern. 2017. Our path to better science in less time using open data science tools. Nature Ecology & Evolution 1:160.
McLoughlin, P. D., D. W. Morris, D. Fortin, E. Vander Wal, and A. L. Contasti. 2010. Considering ecological dynamics in resource selection functions. Journal of Animal Ecology 79:4–12.
Michener, W. K. 2015. Ecological data sharing. Ecological Informatics 29:33–44.
Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27:85–93.
Mills, J. A., C. Teplitsky, B. Arroyo, A. Charmantier, P. H. Becker, T. R. Birkhead, P. Bize, D. T. Blumstein, C. Bonenfant, and S. Boutin. 2015. Archiving primary data: solutions for long-term studies. Trends in Ecology & Evolution 30:581–589.
Mislan, K. A. S., J. M. Heer, and E. P. White. 2016. Elevating the status of code in ecology. Trends in Ecology & Evolution 31:4–7.
Northrup, J. M., M. B. Hooten, C. R. Anderson, and G. Wittemyer. 2013. Practical guidance on characterizing availability in resource selection functions under a use-availability design. Ecology 94:1456–1463.
Obama, B. 2013. Executive Order—making open and machine readable the new default for government information. whitehouse.gov. <https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-> Accessed 18 Jul 2016.
Peng, R. D. 2009. Reproducible research and biostatistics. Biostatistics 10:405–408.
Peng, R. D. 2011. Reproducible research in computational science. Science 334:1226–1227.
R Development Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rakhimberdiev, E., D. W. Winkler, E. Bridge, N. E. Seavy, D. Sheldon, T. Piersma, and A. Saveliev. 2015. A hidden Markov model for reconstructing animal paths from solar geolocation loggers using templates for light intensity. Movement Ecology 3:25.
Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.
Robertson, G. J., D. A. Fifield, W. A. Montevecchi, A. J. Gaston, C. M. Burke, R. Byrne, K. H. Elliott, C. Gjerdrum, H. G. Gilchrist, and A. Hedd. 2012. Miniaturized data loggers and computer programming improve seabird risk and damage assessments for marine oil spills in Atlantic Canada. Journal of Ocean Technology 7:42–58.
Rüegg, J., C. Gries, B. Bond-Lamberty, G. J. Bowen, B. S. Felzer, N. E. McIntyre, P. A. Soranno, K. L. Vanderbilt, and K. C. Weathers. 2014. Completing the data life cycle: using information management in macrosystems ecology research. Frontiers in Ecology and the Environment 12:24–30.
Science International. 2015. Open data in a big data world. Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP). <https://twas.org/sites/default/files/open-data-in-big-data-world_short_en.pdf>. Accessed 18 Feb 2016.
Sergeant, C. J., B. J. Moynahan, and W. F. Johnson. 2012. Practical advice for implementing long-term ecosystem monitoring. Journal of Applied Ecology 49:969–973.
Steenweg, R., M. Hebblewhite, R. Kays, J. Ahumada, J. T. Fisher, C. Burton, S. E. Townsend, C. Carbone, J. M. Rowcliffe, and J. Whittington. 2016. Scaling-up camera traps: monitoring the planet's biodiversity with networks of remote sensors. Frontiers in Ecology and the Environment 15:26–34.
Sutter, R. D., S. B. Wainscott, J. R. Boetsch, C. J. Palmer, and D. J. Rugg. 2015. Practical guidance for integrating data management into long-term ecological monitoring projects. Wildlife Society Bulletin 39:451–463.
Teo, S. L. H., A. Boustany, S. B. Blackwell, A. Walli, K. C. Weng, and B. A. Block. 2004. Validation of geolocation estimates based on light level and sea surface temperature from electronic tags. Marine Ecology Progress Series 283:81–98.
Thomas, D. L., and E. J. Taylor. 2006. Study designs and tests for comparing resource use and availability II. Journal of Wildlife Management 70:324–336.
Thomas, L., S. T. Buckland, E. A. Rexstad, J. L. Laake, S. Strindberg, S. L. Hedley, J. R. Bishop, T. A. Marques, and K. P. Burnham. 2010. Distance software: design and analysis of distance sampling surveys for estimating population size. Journal of Applied Ecology 47:5–14.
Tukey, J. W. 1977. Exploratory data analysis. Pearson, London, England, United Kingdom.
Urbano, F., F. Cagnacci, C. Calenge, H. Dettki, A. Cameron, and M. Neteler. 2010. Wildlife tracking data management: a new vision. Philosophical Transactions of the Royal Society B: Biological Sciences 365:2177–2185.
Valle, D., and A. Berdanier. 2012. Computer programming skills for environmental sciences. Bulletin of the Ecological Society of America 93:373–389.
Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data archiving. American Naturalist 175:145–146.
Wilmers, C. C., B. Nickel, C. M. Bryce, J. A. Smith, R. E. Wheat, and V. Yovovich. 2015. The golden age of bio-logging: how animal-borne sensors are advancing the frontiers of ecology. Ecology 96:1741–1753.
Wilson, G., D. A. Aruliah, C. T. Brown, N. P. C. Hong, M. Davis, R. T. Guy, S. H. Haddock, K. D. Huff, I. M. Mitchell, and M. D. Plumbley. 2014. Best practices for scientific computing. PLoS Biology 12:e1001745.
Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. 2016. Good enough practices in scientific computing. PLoS Computational Biology 13:e1005510.
Wilson, R. P., J. J. Ducamp, W. G. Rees, B. M. Culik, and K. Niekamp. 1992. Estimation of location: global coverage using light intensity. Pages 131–134 *in* I. G. Priede and S. M. Swift, editors. Wildlife telemetry: remote monitoring and tracking of animals. Ellis Horwood, New York, New York, USA.
Xie, Y. 2013. Dynamic documents with R and knitr. Volume 29. CRC Press, New York, New York, USA.
Zuur, A. F., E. N. Ieno, and C. S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1:3–14.

Associate Editor: Applegate.