esa ECOSPHERE

CONCEPTS & THEORY

Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology

Debra P. C. Peters,[1,]† Kris M. Havstad,[1] Judy Cushing,[2] Craig Tweedie,[3]
Olac Fuentes,[4] and Natalia Villanueva-Rosales[4]

[1] *USDA ARS, Jornada Experimental Range and Jornada Basin Long Term Ecological Research Program, New Mexico State University, Las Cruces, New Mexico 88003 USA*
[2] *Computer Science Department, The Evergreen State College, Olympia, Washington 98005 USA*
[3] *Department of Biological Sciences, University of Texas, El Paso, Texas 79968 USA*
[4] *Department of Computer Science, University of Texas, El Paso, Texas 79968 USA*

Citation: Peters, D. P. C., K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales. 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5(6):67. http://dx.doi.org/10.1890/ES13-00359.1

Abstract. Most efforts to harness the power of big data for ecology and environmental sciences focus on data and metadata sharing, standardization, and accuracy. However, many scientists have not accepted the data deluge as an integral part of their research because the current scientific method is not scalable to large, complex datasets. Here, we explain how integrating a data-intensive, machine learning approach with a hypothesis-driven, mechanistic approach can lead to a novel knowledge, learning, analysis system (KLAS) for discovery and problem solving. Machine learning leads to more efficient, user-friendly analytics as the streams of data increase while hypothesis-driven decisions lead to the strategic design of experiments to fill knowledge gaps and to elucidate mechanisms. KLAS will transform ecology and environmental sciences by shortening the time lag between individual discoveries and leaps in knowledge by the scientific community, and will lead to paradigm shifts predicated on open access data and analytics in a machine learning environment.

Key words: analytics; data deluge; long-term data; machine learning; open data; paradigm shifts.

Received 6 November 2013; revised 28 April 2014; accepted 5 May 2014; **published** 5 June 2014. Corresponding Editor: J. Paruelo.

† **E-mail:** debpeter@nmsu.edu

Introduction

The data deluge—the large quantity of multifarious and validated data and information moving at faster rates—undoubtedly provides opportunities for the greatest scientific and technological advances of the early 21st Century (e.g., Ginsberg et al. 2009, Brumfiel 2011, King 2011, Manyika et al. 2011). These "big data" include both legacy data that are increasingly being rescued and captured digitally, and new data that are being acquired through autonomous or manual methods (Michener and Jones 2012, Peters et al. 2013). In ecology and related environmental sciences, datasets are growing in size, complexity, and type as a result of technological advances in sensor and sensor platform technologies (space-, air-, land-, aquatic-, marine-, and organismalbased), computational and analytical improvements in simulation models, and improved methodologies for probing samples, such as genome sequencing and the generation of 'omics' data (Drake et al. 2006, Hart and Martinez 2006, Cohen et al. 2009, Luo et al. 2011, Pfeifer et al. 2012, Porter et al. 2012). Research and development into using this data deluge have focused on cyber-infrastructure (CI) hardware and software constraints, the discovery and access to "dark data" and "deep web" information, and cultural concerns about sharing data (Price and Sherman 2001, Heidorn 2008, Trelles et al. 2011, Michener and Jones 2012, Parr et al. 2012, Peters et al. 2014*a*) that lead to calls for open science (Wolkovich et al. 2012, Hamilton et al. 2013). In spite of these advances in data acquisition and publishing, however, the use and re-use of data are not fully exploited.

Surprisingly, a key challenge has not been effectively addressed: big data are not readily accepted or utilized by most ecologists as an integral part of their research because the traditional scientific method is not scalable to large, complex datasets. In fact, only a small fraction of current data is actually reused by scientists (Reichman et al. 2011), and most data that are used (ca. 50%) are from relatively small, locally collected and stored datasets (Science Staff 2011). Even though there is overwhelming evidence of the importance of existing and emerging large datasets to fields as diverse as medicine, biology, and earth science (Garrett et al. 2006, Delaney and Barga 2009, Robinson et al. 2010, Hay et al. 2013), we believe that few ecologists will take advantage of these data *even if the technological and cultural challenges are met*.

Typically, the scientific method focuses on a small set of high quality data that are often collected, maintained, and analyzed locally by an individual investigator with a bias towards acquiring new data. Long time lags (i.e., years) often occur between individual discoveries and leaps in knowledge. These lags are associated with the time required for publication of results and for others to recreate the analyses and findings even when the data and metadata are readily accessible. Thus, serious consequences can result when a "small data" approach is used to address complex scientific problems (Hamilton et al. 2013).

Alternatively, data-intensive approaches using machine learning developed in other fields, for example to provide information retrieval, support business decision-making, and improve overall user experience on the Internet (Bryan and Leise 2006, Ginsberg et al. 2009), can explain patterns in ecological data. However, these correlation analyses of large quantities of mixed quality data, and numerous queries and analyses have limited direct application to scientific research where understanding underlying processes is paramount to knowledge discovery and problem solving (http://nyti.ms/1kgErs2). Thus, new and urgent solutions are needed to better exploit ecological data and to capacitate future generations of ecologists.

We contend that fundamental and urgent changes are needed in the way ecologists conceptualize and solve problems—changes that go beyond new tools, technologies, and infrastructure. In order to scale the scientific method to allow ecologists to take advantage of the data deluge, what is needed is a knowledge-driven, open access system that "learns" and becomes more efficient and easier to use as streams of data, and the number and types of user interactions, increase—similar to how internet searches and recommender systems work (Bryan and Leise 2006; http://microsoft.com*)*. Science is on the verge of a revolution, but for this to occur, scientists must radically change their way of thinking *and* their way of doing science to take advantage of the deluge of data and its global accessibility (Friedman 2005, Tolle et al. 2011). This paradigm shift is necessary for scientists to push the frontiers of knowledge discovery as well as to make important contributions towards solving the most pressing environmental problems facing society, today and in the future (NRC 2001, Sutherland et al. 2009, Peters 2010, Fleishman et al. 2011, Peters et al. 2014*a*).

Here, we present a novel soon-to-be automated Knowledge Learning and Analysis System (KLAS) that integrates a data-intensive, machine learning approach with a hypothesis-data-driven and process-based approach to take advantage of the relative strengths and offset the limitations of each approach when used in isolation (Fig. 1). This approach begins with a theory leading to hypotheses that are tested iteratively using data from a variety of sources. New experiments are

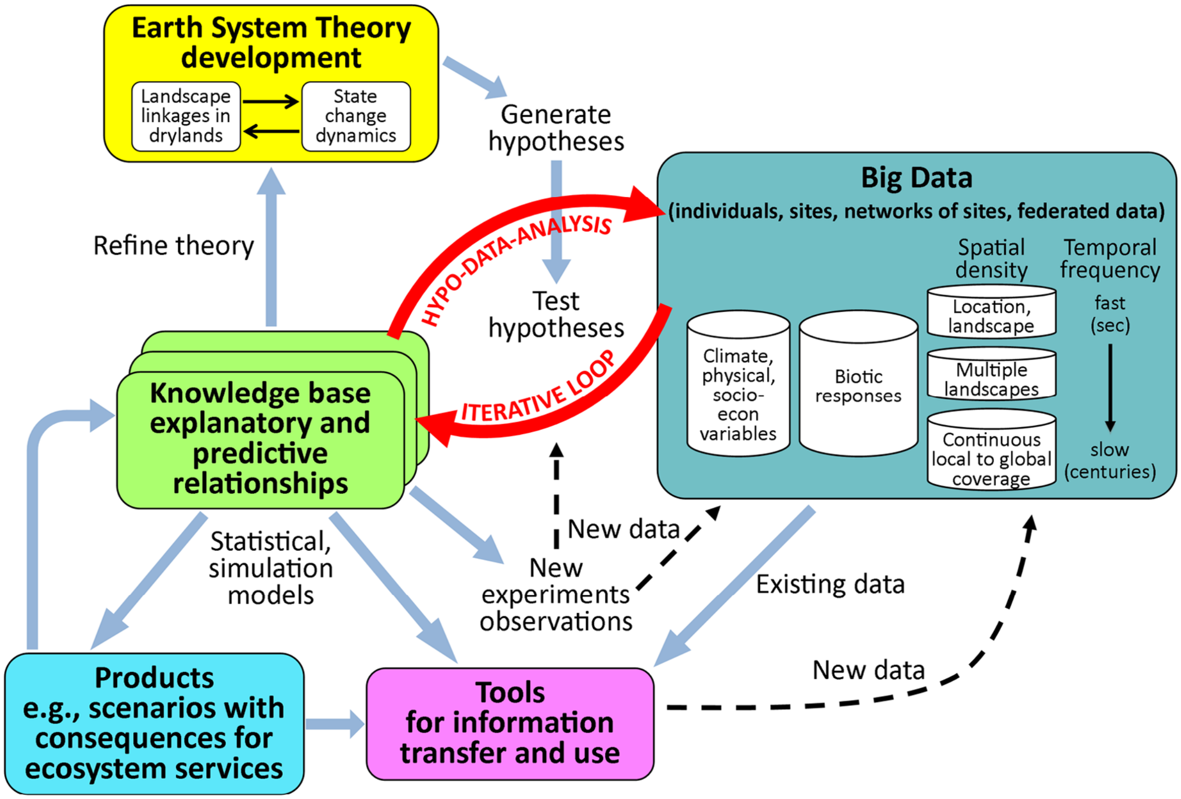
Earth System Theory development
Landscape linkages in drylands
State change dynamics
Generate hypotheses
Refine theory
Big Data
(individuals, sites, networks of sites, federated data)
HYPO-DATA-ANALYSIS
ITERATIVE LOOP
Test hypotheses
Spatial density
Temporal frequency
Location, landscape
Multiple landscapes
Continuous local to global coverage
fast (sec)
slow (centuries)
Climate, physical, socio-econ variables
Biotic responses
Knowledge base explanatory and predictive relationships
Statistical, simulation models
New data
New experiments observations
Existing data
New data
Products e.g., scenarios with consequences for ecosystem services
Tools for information transfer and use

Fig. 1. Iterative, process-based approach with incremental learning. This approach begins with a theory leading to hypotheses that are tested iteratively using data from a variety of sources. New experiments are conducted for the strategic collection of data based on knowledge gained from existing data. The knowledge base expands as more data are used and reused, and explanatory and predictive relationships are developed from statistical and simulation models. Products, including scenarios of future conditions with consequences for ecosystem services, and tools to transfer information to the public, resource managers, and decision-makers are developed. Infusing this scientific process with machine learning leads to more rapid refinements to theory and feedbacks to new data collection than possible by hypothesis-driven or data-intensive approaches used in isolation.

conducted for the strategic collection of data based on knowledge gained from existing data. The knowledge base expands as more data are used and reused, and explanatory and predictive relationships are developed from statistical and simulation models. Products, including scenarios of future conditions with consequences for ecosystem services, and tools to transfer information to the public, resource managers, and decision-makers are developed. Infusing this scientific process with machine learning will lead to more rapid refinements to theory and greater feedbacks to new data collection than possible using hypothesis-driven or data-intensive approaches used in isolation.

We first describe the traditional approaches, and how our approach to the scientific method integrates them. We then illustrate how existing long-term datasets can be reused in an iterative analysis to generate, test, and refine subsequent hypotheses, and how this leads to the strategic collection of new data. Finally, we explain how this manual process can be automated as a linked knowledge-learning-analytics system (KLAS) to test hypotheses in diverse ecosystems using a combination of mixed quality data, both individual investigator-generated and large federated. Based on its initial success, we believe KLAS has tremendous potential to transform the way ecologists think about big data and how, through the development of a new suite of open access software, ecologists can better take advantage of

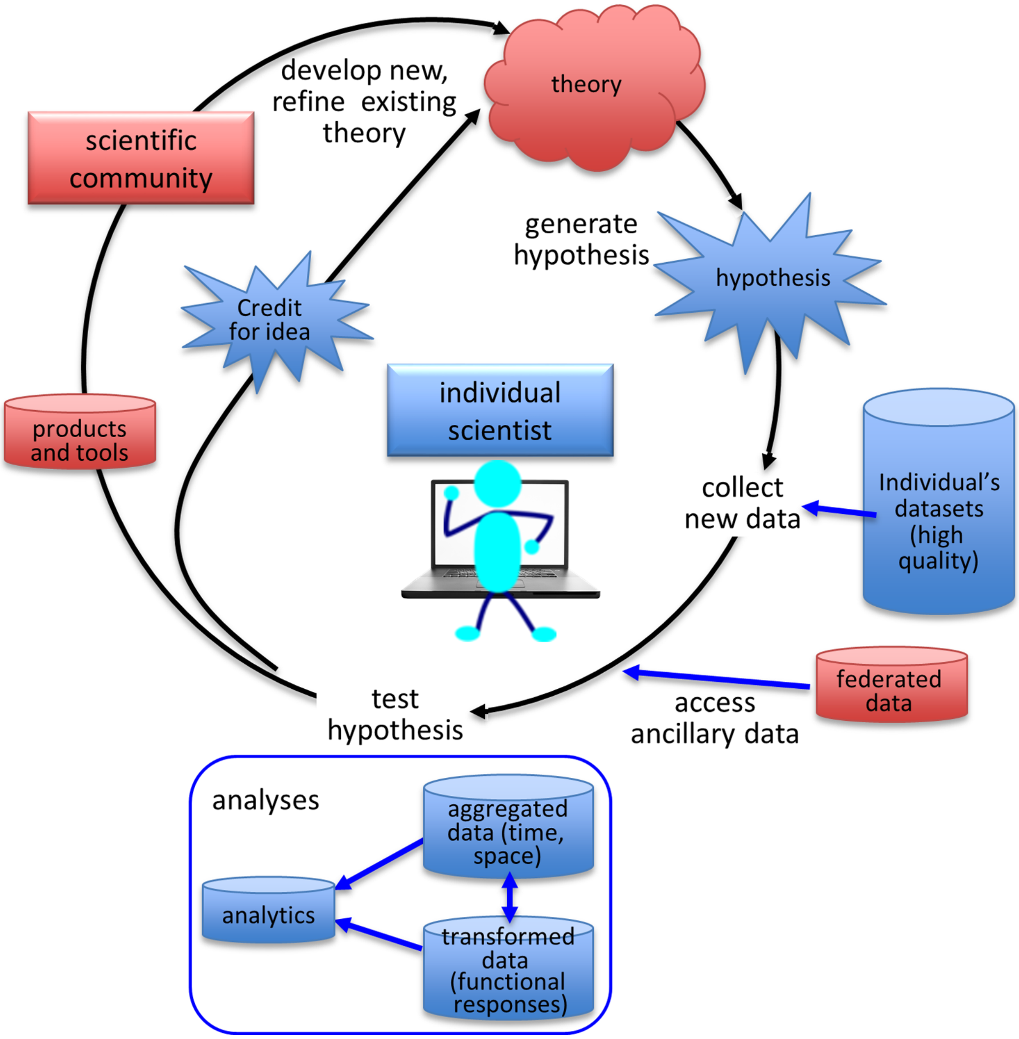

Fig. 2. The traditional hypothesis-driven scientific approach focuses on collecting new or reusing high quality data, primarily collected by the individual scientist, with selective use of ancillary data. Source data manipulations and analytics are maintained on the personal computer or workspace as part of the investigator's toolkit. Hypotheses are tested iteratively to understand mechanisms driving responses. Results and findings are made available to the community after a time lag through publications and internet sites. Credit for the idea remains with the individual, and the theory is refined through time with individual and community input.

the data deluge and move science rapidly forward.

Current Approaches to the Data Deluge

Although we present two current approaches (hypothesis-driven, data intensive) as a distinct dichotomy that has been studied by philosophers of science (Callebaut 2012), debated by some (Golub 2010, Weinberg 2010), and advocated for one approach or another by others (Kelling et al. 2009), we recognize that a gradient exists between the two approaches. Individual scientists may operate anywhere along the gradient, and the complementarity of the approaches was recently promoted conceptually for biodiversity studies (Nichols et al. 2012).

The hypothesis-data driven approach, as practiced most frequently by ecologists and environmental scientists, is a primarily sequential, yet iterative process that begins with a theory and leads to one or more hypotheses (Fig. 2). High quality source data collected or directed by a scientist in an experimental or observational setting are supplemented with ancillary data from data repositories or federated databases to test the hypotheses. The statistical analyses used for testing hypotheses often require: (1) aggregating source data to standard spatial and temporal units, and (2) transforming data to convert structural characteristics (i.e., measured, sensed, or collected data, such as plant biomass) to functional responses (e.g., plant growth). The theory is refined based on the outcome of the analyses, and new or modified hypotheses are developed and subsequently tested with additional data from new experiments or observations. Products (e.g., peer-reviewed publications), tools (e.g., websites, computational scripts), and open access to the data and metadata in public repositories allow the community of scientists to build on these results to refine the theory with additional experiments, to develop new theories, and to conduct new analyses. A primary driver of this approach is that the individual scientist receives credit for the original ideas, most often following publication of the papers.

The focus of this approach is twofold: (1) high quality data, either collected for a specific question or accessed from known databases, are analyzed locally on the individual's computer, and (2) an individual scientist's creativity and contributions to science are preserved. Patterns in data available from other sources can be used to support or help to refute hypotheses, but these phenomenological observations have relatively little use without an understanding of the underlying mechanisms elucidated by experimentation, most often conducted by the scientist asking the question. This approach is the most direct way to improve understanding; however, it has limitations related to: (1) an underutilization of potentially important data, that are not easily discovered or used, (2) the testing of only a small subset of alternative explanations defined by the observations or observer bias, (3) the sequential and manual testing of new hypotheses through time, and (4) the inaccessibility of the aggregated and transformed data with their analytical programs to the broader community. These limitations often lead to an inefficient use of resources and a delay in breakthroughs by the community that depend, at least in part, on the time required for findings to be published and released publicly. Additional time is then needed for other scientists to recreate the aggregated and transformed data from the source data, and to reprogram scripts that are seldom publicly available (Michener and Jones 2012). The result is a low probability of identifying rare, yet important processes or drivers of response, and a high probability of collecting data that already exist or new data that are not critical to testing the hypotheses. Thus, even when high quality data are used, large unexplained variance in the functional responses can arise that are attributed to stochastic processes or unmeasured interactions.

The data-intensive approach begins with the data, and uses statistical analyses and machine learning tools and techniques as the data increase in size and complexity, to examine correlations among variables of system response with potential drivers of that response (Fig. 3). No preconceived relationships are derived from a theory and many possible relationships are examined. Sensed, monitored, and measured environmental data of many types are analyzed from federated databases, data repositories or virtual databases (e.g., DataONE [www.dataone.org]; Pangaea [www.pangaea.de]; Group on Earth Observation System of Systems [www.geoportal.org]). The data may undergo aggregation to standard units in time and space, and transformations may be needed to create more meaningful variables. Data mining techniques and machine learning are often used to improve the selection of variables and to guide analysis methods. A benefit of this approach is that as additional data are included, the analyses become repetitive, more refined, and thus more efficient. For example, Google originally used PageRank to prune and order search queries; new search algorithms, however became even "smarter" as the number and types of searches increased over time, and more and better information was included in the query algorithms (Bryan and Leise 2006). Now, power users can direct the analyses and develop products and tools based on correlations that are then accessed by the user community for more specific applications (Garrett et al. 2006, Delaney and Barga 2009).

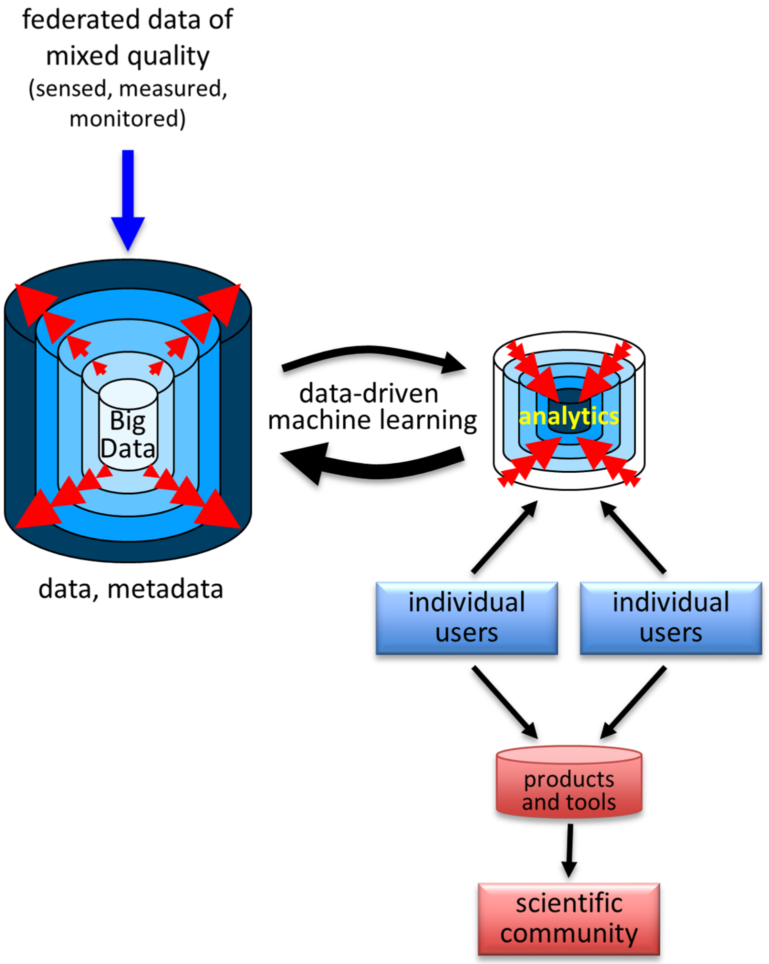

Fig. 3. A data-intensive, machine learning method focuses on analyzing massive amounts of data of mixed quality for correlations without a theory or guiding hypotheses. Machine learning leads to more efficient analytics (decreasing red arrows) through time as data increase in amount and type (increasing red arrows). Spurious correlations are possible, along with the identification of infrequent, yet important and novel relationships.

This approach can find infrequent, but meaningful observations, and can be used for short-term forecasting (i.e., now-casting) over the time period when correlations hold (Ginsberg et al. 2009). Ontologies (what entities exist, how such entities are related within a hierarchy and subdivided according to similarities and differences) can be used to identify, retrieve, and process dynamically multifarious and changing datasets (Callahan et al. 2011, Del Rio et al. 2013). Elucidating patterns and correlations can be an initial step in searching for mechanisms to explain these patterns. Limitations of this method in the sciences result primarily from an inability to apply a guiding theory in order to: (1) eliminate spurious results that appear when correlations among variables are not related to physical causation, (2) avoid using more data than are needed to extract knowledge when boundaries on questions are unknown, (3) determine when past relationships are poor predictors of future dynamics, and (4) identify outliers (Bollier 2010). Additional challenges, exacerbated without guiding theory, are that error propagation increases as the number of variables increases (Peters et al. 2004), and that the large quantities of data are often of mixed quality given the diversity of data sources. Identifying data with the largest meaning and filtering out low quality or misleading data are particularly challenging, but once identified, tagging data with this information has the potential to improve the efficiency of future applications. This approach, though not driven by a traditional hypothesis-based method, has become more readily accepted by scientists in some fields because of success associated with the Human Genome Project (Cohen et al. 2009).

Infusing the Scientific Method with Machine Learning

Our integrated approach begins with a knowledge base containing theories by which hypotheses could be generated, either by an individual or a group of individuals (gray box, Fig. 4). The hypotheses are iteratively tested and refined using a number of data sources (federated, big data; local, small data) and open access analytics (yellow cloud, Fig. 4). These analytics include the programming scripts used to create derived data products (aggregated and transformed data) from the source data as well as the models (e.g., conceptual, mathematical, simulation) needed to test the hypotheses against the theory or to make predictions.

As part of the linked human knowledge-analysis process (blue arrows, Fig. 4), the system learns and builds on previous analyses such that the analyses become more efficient, and easier to access and use as users create successful analyses (linking possibly new theories and hypotheses with data) and as the volume and types of data increase (red arrows in Fig. 4). The key machine learning components of KLAS that allow it to learn from usage patterns of data sets and analysis tools include: (1) *recommendations* of similar or complementary data sets and analytical tools based on previous user interactions with the system, references to relevant research work, and a history of user interactions with the system; (2) *caching* of intermediate results that may be useful for future users (e.g., data generated through time-consuming processes, such as model simulations, computations of derived and summary features for large datasets, and solutions to optimization problems); (3) *precomputing* data analysis tasks that could provide insight to the user, reduce discovery time, and broaden the user discovery experience. Precomputing is linked to the nature of the data, and includes linear and non-linear regression, decision tree analysis, aggregation, and feature selection; (4) *prefetching* data before it is requested based on previous users' interactions with the system; and (5) *filtering algorithms* for flagging or removing outliers and conducting quality assurance/quality control analyses. Importantly, this method provides an objective way to flag, correct, or delete poor quality (e.g., missing or out of range values, inadequate metadata) or unimportant data (e.g., extraneous or repetitive information) identified during the analysis. Because KLAS extends the knowledge base information with how specific datasets relate to certain theories or hypotheses as part of the learning process, higher quality, annotated data and associated metadata are available to future users, and poor quality or irrelevant data are flagged.

Making these intermediate and final data products openly available reduces the time lag for knowledge transfer from an individual to the community, and increases the speed of scientific progress, similar to the community-level sharing of information and derived data products in genomics and medicine (Hey and Trefethen 2005). New experiments can be strategically designed based on feedbacks from the hypothesis-data-analysis loop. Positive feedbacks to the scientific community using the data and to new scientific breakthroughs are generated in less time than currently possible under the traditional scientific approach. There are also positive feedbacks to technological advances in CI, and to the generation of more and different kinds of federated data. These feedbacks are expected to lead to future paradigm shifts that depend on, and lead to advances in, the collection, accessibility, and analysis of big data.

This integrated approach differs from the traditional scientific method where the analytics become more complicated and more difficult to use as data increase in quantity and decrease in quality (Fig. 2). Because current analytics are part of individual investigator's toolkits, the data aggregations and transformations, and statistical analyses need to be recalculated and repeated by

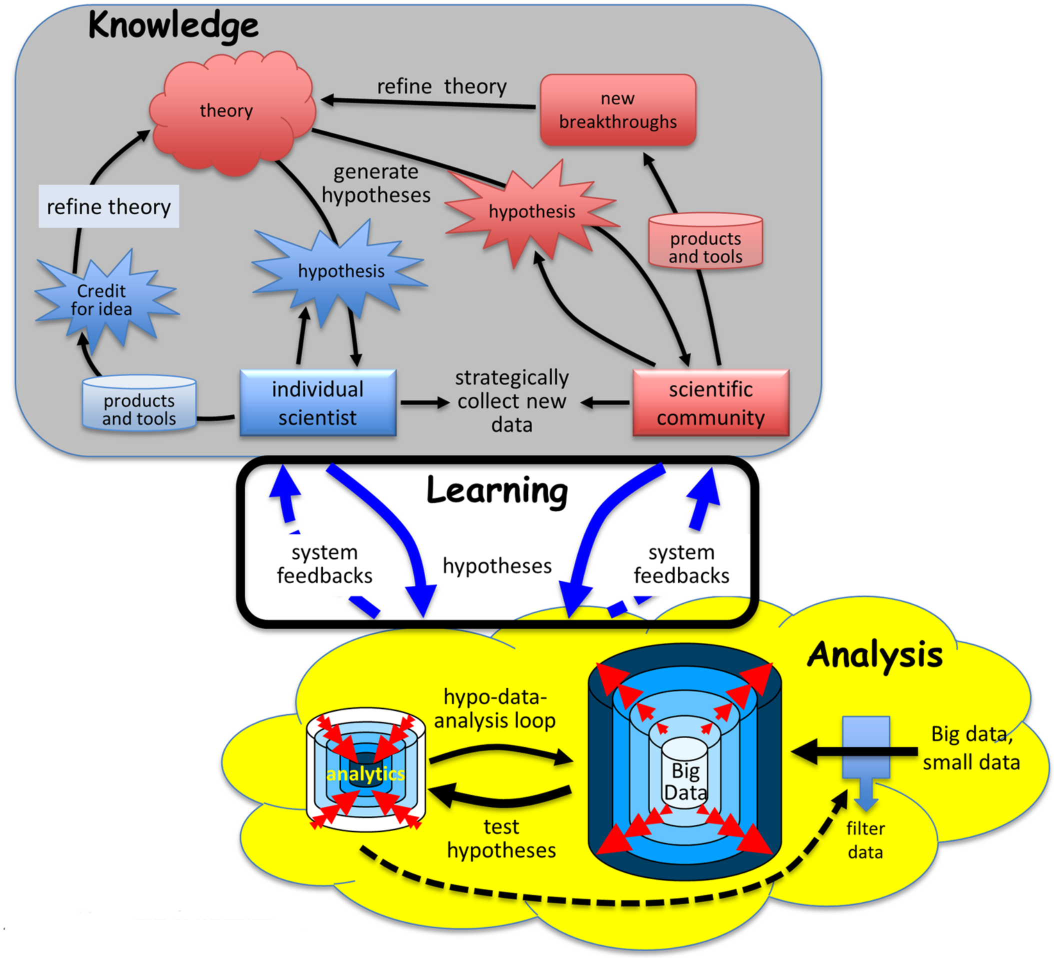

Fig. 4. The traditional modified scientific approach consists of a knowledge, learning, analytics system (KLAS) as a hypothesis-driven, yet data-intensive method for scientists to effectively and efficiently access and use big and little data. The approach begins with a theory to generate hypotheses as part of the knowledge base (gray rectangle) that are tested and refined using data and open access analytics (yellow cloud). Shared learning between humans and computers (blue arrows) make the data and analyses more efficient, and easier to access and use as the amount and types of data used by the community increase (red arrows). New experiments are strategically designed based on this iterative loop. The credit for new ideas remains with individual scientists (blue symbols in gray rectangle) whereas the analytics and data are developed in collaboration with the broader scientific community (red symbols in gray box). This approach provides an objective way to filter poor quality data with feedbacks from the analyses and learning by the system.

other members of the scientific community before they can be used or reused. As the data increase in volume, rate, quality, and type, only a small proportion of the scientific community currently has the technical skills or personal connections with colleagues needed to access and analyze the source data. Alternatively in KLAS, the derived data products and their analyses as well as the source data and metadata (as needed) are open access with continual change driven by user-interactions (red arrows, Fig. 3). Machine learning techniques allow new data to be incrementally refined, similar to the data-intensive approach, such that the process is scalable to large datasets. Thus, derivatives of big data will be easily accessible to the scientific community who may or may not be technologically adept. This integrated approach also differs from the data-intensive approach where machine learning and data mining depend on the characteristics of the data, and correlations are used to examine known and unknown patterns in the data. Although correlations can be used to investigate patterns, scientific understanding and prediction require knowledge about causation and underlying mechanisms governing the patterns (i.e., knowledge about theory). As a semi-automated system guided by frequent human input, KLAS maintains and learns from the history and use of source data, programming scripts, and derived data products, and uses this information to provide feedback to the user as to the next logical steps to be undertaken to guide the refinement of hypotheses.

An Ecological Example of KLAS

We demonstrate the utility of this integrated approach to catalyze timely knowledge discovery using an ecological example. We recently used this approach manually to provide new insights into controls on primary production (Peters et al. 2012; 2014*b*). Here we show how an automated KLAS that integrates human knowledge (i.e., the scientific approach) with machine learning can increase the speed and effectiveness of the process, and transform ecology through improved understanding (Fig. 5).

An important theory in ecology is that water drives dynamics in drylands. A common hypothesis is that Aboveground Net Primary Production (ANPP) is linearly related to annual precipitation (PPT). We tested this hypothesis for desertified shrublands of the Chihuahuan Desert where grass production is typically very low. Based on our experience, we developed two alternative hypotheses: (1) grass production increases linearly with increases in rainfall, as theory suggests (Huxman et al. 2004), or (2) grass production is not related to rainfall based on previous desertification studies (Huenneke et al. 2002). Grasses may be unable to respond to large rainfall years on desertified soils with low organic matter and low rates of infiltration.

The first step to test our hypotheses was to obtain relevant data for grass production in desertified shrublands. In our case, we used our knowledge to locate 20 years of grass production data in an open access database (hereafter referred to as: JRN LTER database) maintained by the Jornada Long Term Ecological Research Program (http://jornada.nmsu.edu) from southern New Mexico. In an automated KLAS, as more users test similar hypotheses and more data and findings are cached, the system would recommend to the user the datasets and variables that could be used, and would prefetch the data and analyses of highest priority or likelihood of being used based on these previous users' interactions. For datasets that were previously analyzed, these precomputed analyses would be available to the user. For example, both the data and the statistical relationship between ANPP and PPT based on >9000 data points in the Central Great Plains (published in Sala et al. [1988] and cited 777 times through 2013) would be available to users, after the data are part of KLAS. This improved accessibility to data and relationships from published papers would allow users to rapidly build on previous research without manually re-entering data and recreating analyses.

The user would then select the datasets to be analyzed (including user-collected data), and would perform exploratory analyses to view patterns in the data (Fig. 5A). Unusual values identified by the user or by KLAS would be examined further, and either flagged as outliers of unknown cause, corrected based on user knowledge of the data, or maintained as valid in the dataset. These exploratory analyses, including the sequence of steps, the findings, and identification of outliers, would be cached in KLAS to be used in: developing recommendations to future users, improving the quality of the data, and creating filtering algorithms to identify outliers in future datasets.

In the second step, we used a linear model to test our hypothesis about the relationship between grass ANPP and PPT (Fig. 5B). Our results showed that a linear relationship was significant for most years, but a cluster of points was clearly above the regression line (Peters et al. 2012). We then tried fitting lines to the points using different forms, such as an exponential curve, but were unsuccessful in improving the fit of the regression. In an automated KLAS, the process of

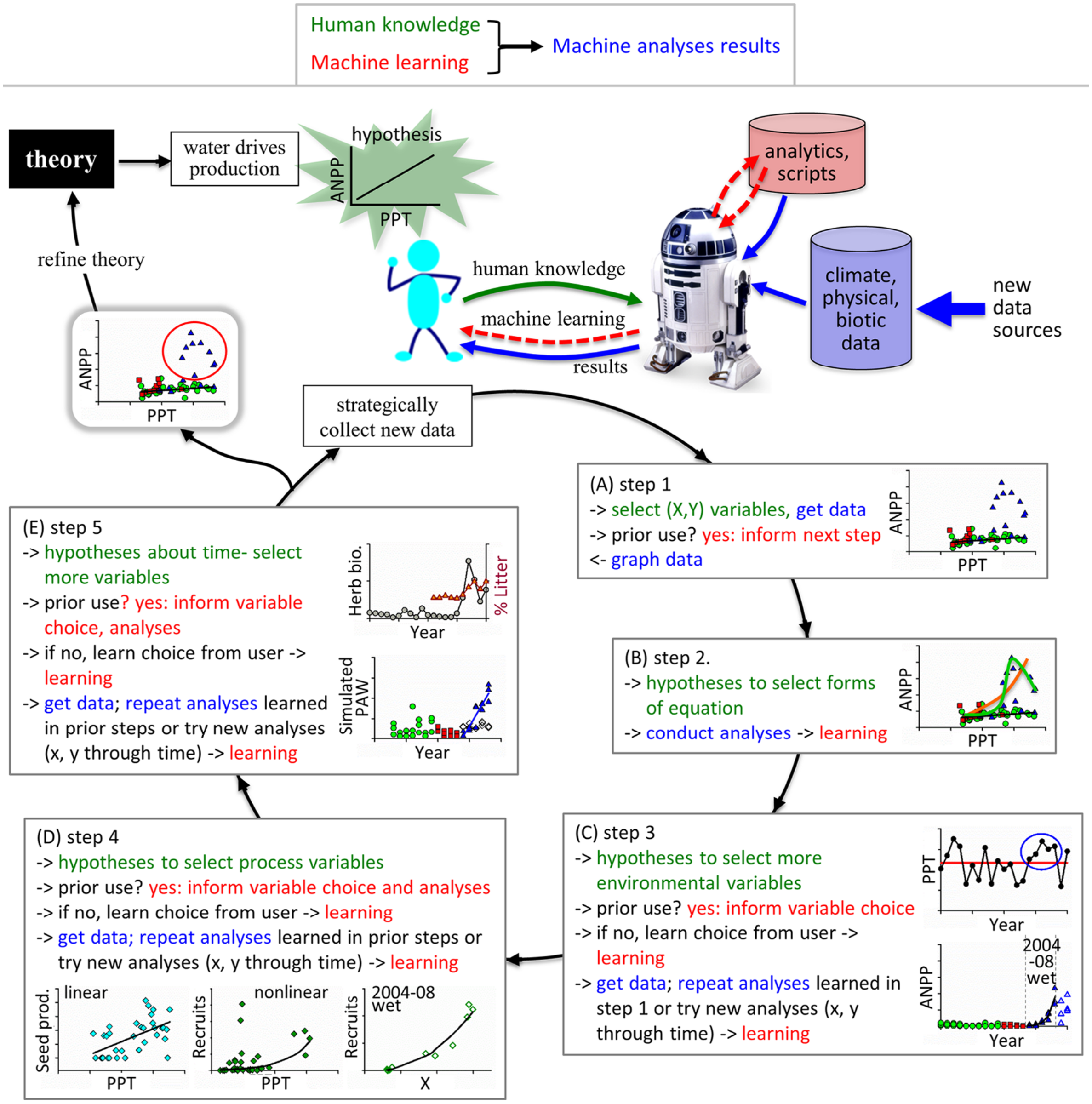
Human knowledge
Machine learning
Machine analyses results
theory
water drives production
hypothesis
ANPP
PPT
refine theory
human knowledge
machine learning
results
analytics, scripts
climate, physical, biotic data
new data sources
ANPP
PPT
strategically collect new data
(A) step 1
-> select (X,Y) variables, get data
-> prior use? yes: inform next step
<- graph data
ANPP
PPT
(B) step 2.
-> hypotheses to select forms of equation
-> conduct analyses -> learning
ANPP
PPT
(C) step 3
-> hypotheses to select more environmental variables
-> prior use? yes: inform variable choice
-> if no, learn choice from user -> learning
-> get data; repeat analyses learned in step 1 or try new analyses (x, y through time) -> learning
PPT
Year
2004 -08 wet
ANPP
Year
(D) step 4
-> hypotheses to select process variables
-> prior use? yes: inform variable choice and analyses
-> if no, learn choice from user -> learning
-> get data; repeat analyses learned in prior steps or try new analyses (x, y through time) -> learning
linear
Seed prod.
PPT
nonlinear
Recruits
PPT
2004-08 wet
Recruits
X
(E) step 5
-> hypotheses about time- select more variables
-> prior use? yes: inform variable choice, analyses
-> if no, learn choice from user -> learning
-> get data; repeat analyses learned in prior steps or try new analyses (x, y through time) -> learning
Herb bio.
% Litter
Year
Simulated PAW
Year

Fig. 5. A manual version of KLAS was used to identify the predictor variables associated with the processes leading to nonlinear dynamics in aboveground primary production (Peters et al. 2012, 2014*b*). Existing long-term datasets from the Jornada LTER were iteratively analyzed to generate and test hypotheses, and to inform the subsequent hypotheses and selection of variables and equations for analysis. The approach was also used to design a simulation model analysis that resulted in a specific hypothesis to be tested with targeted field experimentation in the future. Here, we show how this process could be automated and informed through machine learning in KLAS. The hypothesis generation steps will require human input (green text), the analyses will be conducted by the computer (blue text), and machine learning will inform future decisions and analyses (red text).

selecting and testing alternative forms of the equations would then be cached, and used for precomputing and prefetching data for future users.

In the third step, we had two options: we could either further explore characteristics of precipitation (e.g., seasonality, multi-year patterns), or search for relationships with additional explanatory variables, such as temperature (Fig. 5C). In examining patterns in precipitation, we discovered that these points were from the years 2004–2008, a sequence of wet years in southern New Mexico, USA. When we graphed grass ANPP through time, the increasing amount of ANPP through time was evident (blue points), and even though 2009 and 2010 were average rainfall years, ANPP remained higher than expected (Peters et al. 2012). This gave us confidence that we could explain the patterns in ANPP if we focused on processes occurring in the sequence of wet years. In an automated KLAS, the analysis used to classify individual years (dry, wet, average) and trends in years (drought, wet period, no trend) would be cached, and available to future users. This important distinction of a wet period in explaining patterns in ANPP is one example of an ecological insight that is typically contained only in published papers or through personal communication that may be challenging for the scientific community to find, in particular as information posted on the internet increases. Caching this information and making it readily available to the community through a centralized learning system, such as KLAS, is a paradigm shift that would likely lead to more rapid scientific advances than possible using current approaches.

In the fourth step, we had two options: *Option 1*: Use a traditional hypothesis-driven approach, and conduct an experiment to test a small set of hypotheses about the underlying mechanisms leading to patterns in ANPP during a multi-year wet period. A number of measurements would be needed in addition to ANPP, including soil water content, root growth, photosynthetic rates, etc. depending on the processes believed to be most important. Because the patterns in ANPP are non-specific, it is likely that either too many measurements or the wrong variables, time steps, and spatial resolutions would be obtained. In addition, given inter-annual variability in precipitation, our experiment would need to run for at least 5 years before we are likely to have gained an adequate understanding of the system. At that point, we could refine our hypothesis, and possibly conduct another experiment, add new measurements to the existing experiment or we may decide to move on to another question.

Instead of conducting an experiment, however, we used a data-intensive approach. *Option 2*: Use expert knowledge to focus on recruitment of grasses that need to occur before grass production can increase (Fig. 5D). Thus, we used additional long-term data sets in the JRN LTER database to examine the relationship between annual precipitation and number of seeds produced or number of recruits. Because the number of recruits was nonlinearly related to precipitation, we further examined this relationship to determine that recruitment is related to summer precipitation, seed production, and the number of consecutive wet years (Peters et al. 2014*b*). These results provided a partial explanation for our nonlinear relationship between ANPP and precipitation in the first step (above), but we still needed to identify the mechanism behind the consecutive wet year term. Similar to previous steps, the process of selecting datasets and variables in these analyses would be informed by previous users as part of an automated KLAS.

In the fifth step, we used additional long-term data to show that rain-use efficiency (RUE) by perennial grasses increases nonlinearly as the number of consecutive wet years increases (not shown). Thus, we refined our hypothesis further to focus on the accumulation of biomass and litter beneath individual grass plants that decrease evaporation and act as a positive feedback to fine-scale water availability to plants (i.e., Plant Available Water: PAW) (Fig. 5E). Additional long-term data and simulation model analyses supported the increase in litter and biomass during the period of wet years, and the increase in PAW as litter and biomass accumulated (Peters et al. 2014*b*). Under an automated KLAS, inputs and outputs of the model would be cached for use by future users.

Results from this step are leading to the design of a new field experiment to test the hypothesis that the accumulation of biomass and litter beneath individual grass plants increases PAW through time beyond the water available by rainfall alone. This is a very focused hypothesis that can be tested over a period of weeks instead of years. The measurements are restricted to soil water content at different depths for different amounts of biomass and litter. This targeted experiment requires far less time and effort than a multi-year rainfall manipulation experiment with many response variables. If the hypothesis is not supported by this experiment, then we could repeat the process by generating alternative hypotheses to be tested via simulation model analyses, recycling of existing data, or strategic collection of new data. Under an automated KLAS, the selection of variables and data would be recommended to the user based on previous experiences. Given that one-third of the world's land surface is arid and supports over one billion people, understanding ecological processes responsible for land restoration from research conducted in weeks or a few years rather than decades is essential to the development of land management policy, particular under a changing climate.

This example illustrates: (1) how our KLAS iterative process can reuse existing data to refine hypotheses and design new experiments, and (2) the logical steps required and decisions made by ecologists and environmental scientists that could be part of a machine learning environment to improve understanding and prediction. Selecting and analyzing independent and dependent variables as part of a correlation exercise leading to experimentation and causation is a general process conducted, at least in part, by many ecologists and environmental scientists manually on personal computers or workspaces. KLAS allows this process to be generalized, automated, and openly accessible to the scientific community, whose efforts would in turn feed back to the KLAS framework to synergistically and iteratively strengthen the system for future users.

Limitations of KLAS

The power of KLAS is a function of its use. As user interactions increase in number and type, the system will learn and provide more options for the reuse of data and analyses by future users. However, the number and type of options provided to users will be limited in the early stages of KLAS. In addition, there is the potential for KLAS to include inaccurate information if all data and analyses are added to the system. We envision a community of experts will be needed to determine which data and analyses are included in the system, similar to the way that open source community pages, simulation models, and software programs currently operate (e.g., http://en.wikipedia.org; https://www2.cesm.ucar.edu). As KLAS grows and expands in the types of data and analyses included, a peer-review system may be needed to evaluate the accuracy and usefulness of data, information, and analyses to be included. Finally, KLAS will need to allow for information that is protected by privacy rights when local sources of data are combined with federated databases.

Implementing the Vision: Addressing Conceptual and Technological Challenges

Big data are rapidly making inroads in some disciplines (e.g., particle physics, genomics) where research centers or groups of scientists have joint or open access to the CI required for shared source and manipulated data, and analysis tools (e.g., Hey and Trefethen 2005, Green et al. 2011). However, in many other disciplines, there is a clear lack of interest, capacities, or, in some cases disdain, by individual scientists for the data deluge. These individual views reflect, in part, the cultural, sociological, and technological challenges of sharing, archiving, and managing federated data for use in research, and a frustration with the media hype and unfulfilled promises of these data (The Economist 2010, Reichman et al. 2011, Hamilton et al. 2013). But, these views also reflect conceptual challenges associated with datasets that are much larger in size, scope, and complexity than previously measured or even imagined. For scientists in disciplines where experimentation has been the primary mode of hypothesis testing, a shift from small, highly controlled, high quality datasets to extremely large, federated datasets of mixed quality is an uncomfortable one.

To implement KLAS will require two key shifts in thinking with associated changes in technology. *First,* the full suite of analytics needs to be publicly available and part of the iterative learning process. Much of the current focus is on open access source data and metadata as part of federated databases (Michener and Jones 2012, Hamilton et al. 2013). However, a more efficient use of resources will occur if the derived data products and analyses are also in the public domain and continually modified as more scientists use and learn from the data. This conceptual shift will require a CI that: (1) incorporates machine learning techniques that become more efficient as the number of user interactions and data sources increase, (2) develops linkages between the knowledge and the analysis components that allow the system to learn through time to guide the analyses and feedback to the hypotheses, (3) maintains a community-level history of the data sources, procedures, and findings to allow users to quickly and easily build on previous studies, and (4) accesses, checks, and potentially modifies streams of data of mixed quality.

Second, there needs to be a shift towards the use of existing and federated data before new data are collected by individual researchers. As shown in our example, more focused experiments with targeted response variables and treatments are possible after using powerful insights obtained from big data. Many disciplines have accumulated vast amounts of historic data that can be integrated with the large datasets being collected by new technologies and the smaller datasets collected by individuals. Importantly, our approach is able to identify and filter data of mixed quality from disparate sources. KLAS provides a framework for taking advantage of these data sources for knowledge discovery and problem solving rather than being overwhelmed by them.

Conclusions

Ecology and environmental sciences must be more broadly informed by lessons of genomics where it is recognized that large-scale studies alone are insufficient, yet most data analysis and interpretation come from individual researchers (Green et al. 2011). Hypothesis-driven research by individuals and research groups requires access to data catalogues and technological tools (Frew and Dozier 2012). However, analytical tools and derived data products also need to be in the public domain to encourage multi-disciplinary, collaborative science (Hey and Trefethen 2005). Our knowledge-learning-analysis system (KLAS) adapts the scientific method to accommodate vast amounts of data, and make them accessible to a broad range of users via an open access, iterative learning process. Use of this hypothesis-driven, data-intensive scientific method will require a shift from individual efforts at experimentation and analysis on personal workspaces to: (1) the reuse of historic data integrated with new data streams followed by strategic experimentation, (2) open access analyses that become increasingly efficient as the data increase in type, volume, and rate, and (3) an automated machine learning approach that builds on past experience of the broader community to guide hypothesis testing and refinement by individuals. Positive feedbacks to both intellectual capacity and technological developments resulting from this modernized scientific method will lead to rapid leaps in knowledge and future paradigm shifts that depend on, and lead to advances in, big data.

Acknowledgments

Reported data reported are presented, archived, and openly available (http://jornada.nmsu.edu). Funding was provided by the National Science Foundation to New Mexico State University (EF-1065699; DEB-1440166) and the Jornada Basin Long Term Ecological Research Program (DEB-1235828), by the Sevilleta Long Term Ecological Research Program at the University of New Mexico (DEB-0832652), and by the US Department of Agriculture—Agricultural Research Service to the Jornada Experimental Range. This work used resources from Cyber-ShARE Center of Excellence at the University of Texas at El Paso, which is supported by the National Science Foundation (HRD-1242122). Cushing's contribution was funded in part by the National Science Foundation (BIO/DBI 1062572). We thank Jin Yao and Stacey Scroggs for assistance with the figures. Scott Collins, Susan Moran, and Osvaldo Sala provided helpful comments on an earlier version. We thank William B. Batista and an anonymous reviewer for comments on the manuscript.

Literature Cited

Bollier, D. 2010. The promise and peril of big data. The Aspen Institute, Washington, D.C., USA.

Brumfiel, G. 2011. Down the petabyte highway. Nature 469:282–283.

Bryan, K. and T. Leise. 2006. The $25,000,000,000 eigenvector: the linear algebra behind Google. SIAM Review 18:569–581.

Callahan, A., M. Dumontier, and N. Shah. 2011. HyQue: evaluating hypotheses using Semantic Web technologies. Journal of Biomedical Semantics 2 (Supplement 2):S3.

Callebaut, W. 2012. Scientific perspectivism: a philosopher of science's response to the challenge of big data biology. Studies in History and Philosophy of Biology and Biomedical Sciences 43:69–80.

Cohen, I. R., H. Atlan, and S. Efroni. 2009. Genetics as explanation: limits to the Human Genome Project. Encyclopedia of Life Sciences. doi: 10.1002/9780470015902.a0005881.pub2

Delaney, J. R., and R. S. Barga. 2009. A 2020 vision for ocean science. Pages 27–38 *in* T. Hey, S. Tansley, and K. Tolle, editors. The fourth paradigm, data-intensive scientific discovery. Microsoft Research, Redmond, Washington, USA.

Del Rio, N., N. Villanueva-Rosales, D. Pennington, K. Benedict, A. Stewart, and C. J. Grady. 2013. ELSEWeb meets SADI: Supporting data-to-model integration for biodiversity forecasting. AAAI Fall Symposium on Discovery Informatics: AI Takes a Science-Centered View on Big Data. American Association for the Advancement of Artificial Intelligence Technical Report FS-13-01.

Drake, J. M., C. Randin, and A. Guisan. 2006. Modelling ecological niches with support vector machines. Journal Applied Ecology 43:424–432.

Fleishman, E., et al. 2011. Top 40 priorities for science to inform US conservation and management policy. BioScience 61:290–300.

Frew, J., and J. Dozier. 2012. Environmental informatics. Annual Review of Environment and Resources 37:449–472.

Friedman, T. L. 2005. The world is flat: a brief history of the twenty-first century. Farrar, Straus and Giroux, New York, New York, USA.

Garrett, K. A., S. P. Dendy, E. E. Frank, M. N. Rouse, and S. E. Travers. 2006. Climate change effects on plant disease: genomes to ecosystems. Annual Review of Phytopathology 44:489–509.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. Nature 457:1012–1014.

Golub, T. 2010. Counterpoint: data first. Nature 464:679.

Green, E. D., and M. S. Guyer. and National Human Genome Research Institute. 2011. Charting a course for genomic medicine from base pairs to bedside. Nature. 470:204–213.

Hamilton, 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.

Hart, J. K., and K. Martinez. 2006. Environmental sensor networks: a revolution in the earth system science? Earth Science Reviews 78:177–191.

Hay, S. I., D. B. George, C. L. Moyes, and J. S. Brownstein. 2013. Big data opportunities for global infectious disease surveillance. PLoS Medicine 10:e1001413.

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

Hey, T., and A. E. Trefethen. 2005. Cyberinfrastructure for e-Science. Science 308:817–821.

Huenneke, L. F., J. P. Anderson, M. Remmenga, and W. H. Schlesinger. 2002. Desertification alters patterns of aboveground net primary production in Chihuahuan ecosystems. Global Change Biology 8:247–264.

Huxman, T. E., et al. 2004. Convergence across biomes to a common rain-use efficiency. Nature 429:651–654.

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

King, G. 2011. Ensuring the data-rich future of the social sciences. Science 331:719–721.

Luo, Y., K. Ogle, C. Tucker, S. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. S. Schimel. 2011. Ecological forecasting and data assimilation in a data-rich era. Ecological Applications 21:1429–1442.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. 2011. Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, McKinsey & Company. www.mckinsey.com/mgi

Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27:85–93.

Nichols, J. D., E. G. Cooch, J. M. Nichols, and J. R. Sauer. 2012. Studying biodiversity: is a new paradigm really needed? BioScience 62:497–502.

NRC [National Research Council]. 2001. Grand challenges in environmental sciences. National Academies Press, Washington, D.C., USA.

Parr, C. S., R. Guralnick, N. Cellinese, and R. D. M. Page. 2012. Evolutionary informatics: unifying knowledge about the diversity of life. Trends in Ecology & Evolution 27:94–103.

Peters, D. P. C. 2010. Accessible ecology: synthesis of the long, deep, and broad. Trends in Ecology & Evolution 25:592–601.

Peters, D. P. C., H. W. Loescher, M. D. SanClements, and K. M. Havstad. 2014*a*. Taking the pulse of a continent: expanding site-based research infrastructure for regional- to continental-scale ecology. Ecosphere 5:art29. http://dx.doi.org/10.1890/

ES13-00295.1

Peters, D. P. C., D. L. Urban, R. H. Gardner, D. D. Breshears, and J. E. Herrick. 2004. Strategies for ecological extrapolation. Oikos 106:627–636.

Peters, D. P. C., J. Yao, D. Browning, and A. Rango. 2014*b*. Mechanisms of grass response in grasslands and shrublands during dry or wet periods. Oecologia 174:1323–1334.

Peters, D. P. C., J. Yao, O. E. Sala, and J. P. Anderson. 2012. Directional climate change and potential reversal of desertification in arid and semiarid ecosystems. Global Change Biology 18:151–163.

Peters, D. P. C., et al. 2013. Long-term trends in ecological systems: a basis for understanding responses to global change. Technical Bulletin No. 1931, U.S. Department of Agriculture, Washington D.C., USA.

Pfeifer, M., M. Disney, T. Quaife, and R. Marchant. 2012. Terrestrial ecosystems from space: a review of earth observation products for macroecology applications. Global Ecology and Biogeography 21:603–624.

Porter, J. H., P. C. Hanson, and C.-C. Lin. 2012. Staying afloat in the sensor data deluge. Trends in Ecology & Evolution 27:121–129.

Price, G., and C. Sherman. 2001. The invisible web: uncovering information sources search engines can't see. Information Today.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.

Robinson, G. E., et al. 2010. Empowering 21st century biology. BioScience 60:923–930.

Sala, O. E., W. J. Parton, L. A. Joyce, and W. K. Lauenroth. 1988. Primary production of the central grassland region of the United States. Ecology 69:40–45.

Science Staff. 2011. Challenges and opportunities. Science 331:692.

Sutherland, W. J., et al. 2009. One hundred questions of importance to the conservation of global biological diversity. Conservation Biology 23:557–567.

The Economist. 2010. Clicking for gold: how internet companies profit from data on the web. 25 February. The Economist Newpaper Limited, London, UK.

Tolle, K. M., D. S. W. Tansley, and A. J. G. Hey. 2011. The fourth paradigm: data-intensive scientific discovery. Proceedings of the IEEE 99:1334–1337.

Trelles, O., P. Prins, M. Snir, and R. C. Jansen. 2011. Big data, but are we ready? Nature Reviews Genetics 12. doi: 10.1038/nrg2857-c1

Weinberg, R. 2010. Point: hypotheses first. Nature 464:678.

Wolkovich, E. M., J. Regetz, and M. I. O'Connor. 2012. Advances in global change research require open science by individual researchers. Global Change Biology 18:2102–2110.