

Infusing the Scientific Method with Machine Learning in Ecology

Debra P.C. Peters, Judith B. Cushing, Kris M. Havstad, Craig Tweedie, Olac Fuentes, Natalia Villanueva-Rosales

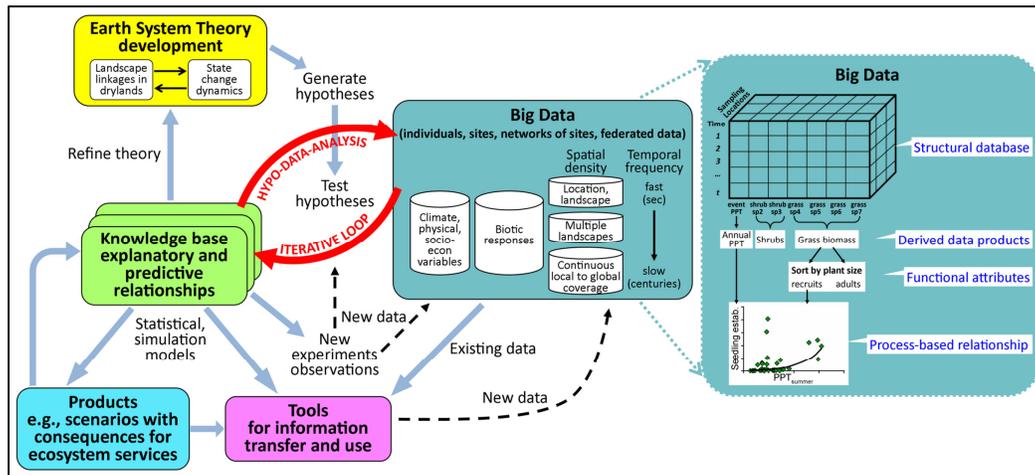


Fig. 1. Iterative, process-based approach to generating and testing ecology hypotheses with incremental learning

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 could 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. This novel, hybrid approach would likely prove more acceptable to scientists than machine learning alone, and could significantly shorten the considerable time lag between individual discoveries and leaps in knowledge by the community, as is currently the case with the traditional hypothesis-driven only approach. We believe it could lead to paradigm shifts predicated on open access data and analytics in a machine learning environment.

Index Terms—Analytics, long term data, machine learning, scientific method, earth system theory, visualization and ecology

1 INTRODUCTION

While not always recognized as “big data” by the computer science community, ecologists are faced with increasingly large quantities of multifarious and validated data and information moving at faster rates. While this *ecological data deluge* (including both legacy and new data) undoubtedly provides opportunities for scientific and technological advances [2, 14, 23, 27]. 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 organismal-based), computational and analytical improvements in simulation models, and improved methodologies for probing samples, such as genome sequencing and the generation of ‘omics’ data [6, 9, 18]. Research and development into using these data (cyber-infrastructure, discovery and access, and data sharing

concerns [9, 24, 26, 31, 37] lead to calls for open science [9], but in spite of such advances, however, the use and re-use of data are not fully exploited, and a key challenge remains: long term data stores are not readily accepted or utilized by most ecologists as an integral part of their research. In fact, only a small fraction of current data is actually reused by scientists [32], and most data that are used (ca. 50%) are from relatively small, locally collected and stored datasets [34]. We believe this is because the traditional scientific method is not scalable to large, complex datasets.

Typically, the scientific method focuses on a small set of high quality data that are often collected, maintained, and analysed locally by an individual investigator with a bias towards acquiring new data. Due to the time required for publication of results and for others to recreate the analyses and findings, long time lags occur between individual discoveries and leaps in knowledge. Alternatively, data-intensive approaches using machine learning developed in other fields [3, 14], could explain patterns in ecological data; unfortunately, however, such work typically uses mixed quality data, and have limited direct application to scientific research where understanding underlying processes is critical to knowledge discovery (<http://nyti.ms/1kgErs2>).

New solutions are needed to better exploit ecological data and to capacitate future generations of ecologists—fundamental changes in how ecologists conceptualize and solve problems, changes beyond new tools, technologies, and infrastructure. To scale the scientific method, ecologists need a knowledge-driven, open access system

- Judy Cushing is with The Evergreen State College, Olympia WA, judyc@evergreen.edu.
- Deb Peters, USDA ARS, Jornada Experimental Range and Jornada Basin Long Term Ecological Research Program, New Mexico State University, Las Cruces, NM.
- Craig Tweedie, Olac Fuentes, Natalia Villanueva-Rosales are all on the faculty at University of Texas, El Paso, TX..

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 internet searches and recommender systems. To effectively use such systems, scientists will need to change their way of doing science [12, 36]. We envision a data-intensive, machine learning approach integrated 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, and are developing KLAS—Knowledge Learning and Analysis System (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 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.

2 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 [5], debated by some [15, 38], and advocated for (or not) by others, 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 [25].

The *hypothesis-data driven approach*, as practiced most frequently by ecologists and environmental scientists, is primarily sequential and begins with a theory leading to one or more hypotheses. 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 using statistical analyses. The focus of this approach is twofold: (1) high quality data, collected for a specific question or accessed from known databases, are analyzed locally, 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.

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. 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

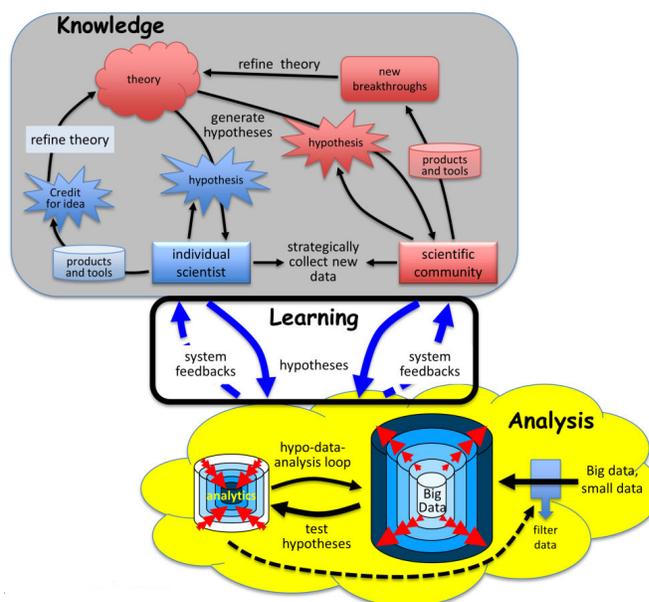


Fig. 2 A modified scientific approach incorporating machine learning

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. 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 [8, 13].

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 [14]. Ontologies can be used to identify, retrieve, and process dynamically multifarious and changing datasets [4, 7]. 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 to: (1) eliminate spurious results when correlations among variables are not related to physical causation, (2) avoid using more data than needed to extract knowledge when a question's boundaries are unknown, (3) determine when past relationships are poor predictors of future dynamics, and (4) identify outliers [1]. Additional challenges, exacerbated without guiding theory, are that error propagation increases as the number of variables increases [28], 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 [6, 10].

3 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. 2). 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. 2). These analytics include the programming scripts for creating 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. 2), the system learns and builds on previous analyses;

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. 2). Key machine learning components will allow KLAS to learn from usage patterns of data sets and analysis tools: (1) *recommendations* of similar or complementary data sets and analytical tools, 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 model simulations, derived data from large datasets, and solutions to optimization problems); (3) *precomputing* data analysis tasks (e.g., linear and non-linear regression, decision tree analysis, aggregation, and feature selection) that could provide insight, reduce discovery time, or broaden user discovery; (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/ control.

4 AN ECOLOGICAL EXAMPLE

We describe here our recent manual use of the above approach to provide new insights into controls on primary production [30, 29]; that manual use is extrapolated to an automated KLAS that integrates human knowledge (i.e., the scientific approach) with machine learning (Fig. 3). 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.

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 [22], or (2) grass production is not related to rainfall based on previous desertification studies [21]. Grasses may be unable to respond to large rainfall years on desertified soils with low organic matter and low rates of infiltration. The following five steps below and in Figure 5 illustrate how KLAS might be used for ecology research.

4.1 Step 1

To test our hypotheses we first obtained relevant data for grass production in desertified shrublands, using 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 datasets and variables, and would prefetch the data and analyses of highest priority based on these previous users' interactions. For previously analyzed datasets, 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. This 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 perform exploratory analyses to view patterns in the data. 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.

4.2 Step 2

We used a linear model to test our hypothesis about the relationship between grass ANPP and PPT. 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 selecting and testing alternative forms of the equations would then be cached, and used for precomputing and prefetching data for future users.

4.3 Step 3

We considered 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. 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 [30]. 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

4.4 Step 4

Instead of conducting a traditional hypothesis-drive experiment to test a small set of hypotheses, we used a data-intensive approach—expert knowledge to focus on recruitment of grasses that need to occur before grass production can increase. 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 [29]. 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.

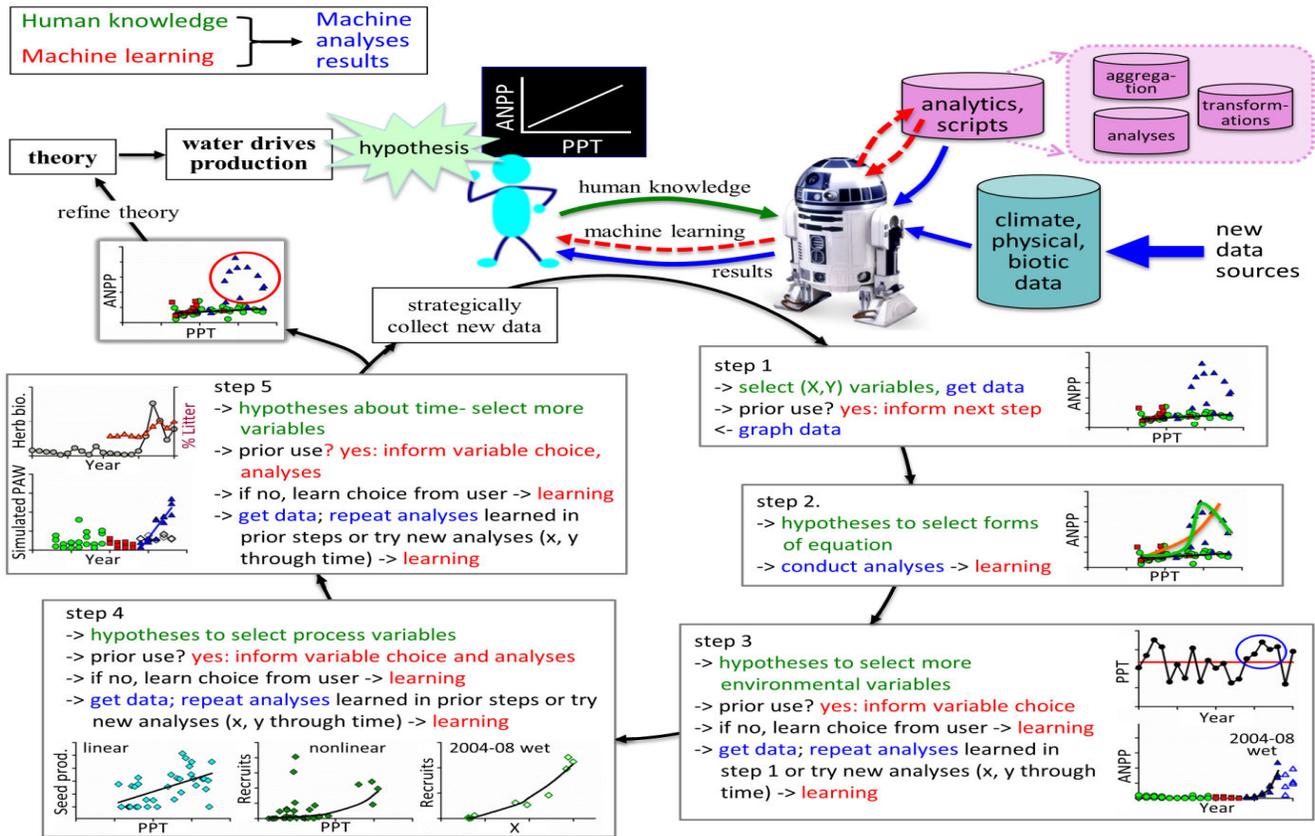


Fig. 3. A manual version of KLAS

5 STEP 5

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). 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 [29]. 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.

6 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. 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.

7 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., 16, 20]. However, in many other disciplines, including ecology, 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 [35, 17, 32].

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.

Implementing 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 [17, 24]. 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.

8 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 [16]. Hypothesis-driven research by individuals and research groups requires access to data catalogues and technological tools [11]. However, analytical tools and derived data products also need to be in the public domain to encourage multi-disciplinary, collaborative science [20]. 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 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) 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. We also note that this position paper for VisWeek 2015's *1st Workshop on Data Systems for Interactive Analysis* has been extracted specifically from a longer publication by the same authors: Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology, *ESA Ecosphere*, 5 June 2014, <http://www.esajournals.org/doi/pdf/10.1890/ES13-00359.1>.

REFERENCES

- [1] Bollier, D. 2010. The promise and peril of big data. The Aspen Institute, Washington, DC, USA.
- [2] Brumfiel, G. 2011. Down the petabyte highway. *Nature* 469:282-283
- [3] Bryan, K. and T. Leise. 2006. The \$25,000,000,000 eigenvector: the linear algebra behind Google. *SIAM Review* 18:569-581
- [4] Callahan, A., M. Dumontier, and N. Shah. 2011. HyQue: evaluating hypotheses using Semantic Web technologies. *Journal of biomedical semantics* 2 (Suppl 2):S3
- [5] 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
- [6] 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
- [7] 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*.
- [8] Delaney, J. R., and R. S. Barga. 2009. A 2020 vision for ocean science. Pages 27-38 *in* Hey, T., S. Tansley, and K. Tolle, editors. The fourth paradigm, data-intensive scientific discovery. Microsoft Research, Redmond, Washington, USA
- [9] Drake, J. M., C. Randin, and A. Guisan. 2006. Modelling ecological niches with support vector machines. *Journal Applied Ecology* 43:424-432.
- [10] Farnsworth, A., Sheldon, D., Geevarghese, et al. (2014). Reconstructing velocities of migrating birds from weather radar - A case study in computational sustainability. *AI Magazine*, 35(2), 31-48. doi:10.1609/aimag.v35i2.2527.
- [11] Frew, J., and J. Dozier. 2012. Environmental informatics. *Annual Review of Environment and Resources* 37:449-472E.
- [12] Friedman, T.L. 2005. The world is flat: a brief history of the twenty-first century. Farrar, Straus and Giroux, New York, USA.
- [13] 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.
- [14] 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.
- [15] Golub, T. 2010. Counterpoint: data first. *Nature* 464:679.
- [16] Green, E. D., 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.
- [17] 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.
- [18] Hart, J. K., and K. Martinez. 2006. Environmental sensor networks: a revolution in the earth system science? *Earth Science Reviews* 78:177-191.
- [19] Heidorn, P. B. 2008. Shedding light on the dark data in the long tail of science. *Library Trends* 57:280-299.
- [20] Hey, T., and A. E. Trefethen. 2005. Cyberinfrastructure for e-Science. *Science* 308:817-821.

- [21] 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.
- [22] Huxman, T.E., et al. 2004. Convergence across biomes to a common rain-use efficiency. *Nature* 429:651-654.
- [23] 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.
- [24] Michener, W. K., and M. B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution* 27:85-93.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] Peters, D. P. C., J. Yao, D. Browning, and A. Rango. 2014. Mechanisms of grass response in grasslands and shrublands during dry or wet periods. *Oecologia* 174:1323-1334.
- [30] 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.
- [31] Price, G., and C. Sherman. 2001. The invisible web: uncovering information sources search engines can't see. Information Today, Inc
- [32] Reichman, O. J., M. B. Jones, and M. P. Schildhauer, 2011. Challenges and opportunities of open data in ecology. *Science* 331:703-705.
- [33] 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
- [34] *Science* Staff. 2011. Challenges and opportunities. *Science* 331:692.
- [35] The Economist. 2010. Clicking for gold: how internet companies profit from data on the web. 25 February. The Economist Newspaper Limited, London, England .
- [36] 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
- [37] 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
- [38] Weinberg, R. 2010. Point: hypotheses first. *Nature* 464:678
- [39] 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.