

Data Intensive Science: A New Paradigm for Biodiversity Studies.

In Press *BioScience*

Steve Kelling¹, Wesley M. Hochachka¹, Daniel Fink¹, Mirek Riedewald², Rich Caruana³, Grant Ballard⁴, Giles Hooker⁵

- 1) Cornell Lab of Ornithology, Cornell University
- 2) College of Computer and Information Science, Northeastern University
- 3) Department of Computer Science, Cornell University
- 4) PRBO Conservation Science
- 5) Department of Biological Statistics and Computational Biology, Cornell University

ABSTRACT: The increasing availability of massive volumes of scientific data requires new synthetic analysis techniques to explore and identify interesting patterns that were otherwise not apparent. For biodiversity studies a “data driven” approach is necessary due to the complexity of ecological systems, particularly when viewed at large spatial and temporal scales. Data intensive science organizes large volumes of data from multiple sources and fields that are then analyzed using techniques tailored to the discovery of complex patterns in high-dimensional data through visualizations, simulations, and various types of model building. By interpreting and analyzing these models, truly novel and surprising patterns that are “born from the data” can be discovered. These patterns in turn provide valuable insight for concrete hypotheses about the underlying ecological processes that created the observed data. Data intensive science allows scientists to analyze bigger and more complex systems efficiently, and complements more traditional scientific processes of hypothesis generation and experimental testing to refine our understanding of the natural world.

Biodiversity research is a branch of ecology that identifies and predicts patterns of organism distribution and abundance, and explains the causes of these patterns. Ecological systems are extremely complex, and a multitude of processes may impact organisms (McMichael et al. 2003). These processes can vary over time (Delcourt and Delcourt 2005) and through space (Tuomisto et al. 2003). Consequently, to understand the determinants of biodiversity, data need to be collected over long periods of time (Gaston and McArdle 1994) and at appropriate, potentially large, spatial scales (Doak et al. 1992). Further, because we must often guess at the environmental features that can affect distributions, tens if not hundreds of potentially important predictors must be screened. Given these challenges, we believe that processes different from those typically used by ecologists are necessary to best understand patterns in biodiversity.

Traditional ecological research has relied on *expert-centered parametric analysis*. While many variants of this approach exist, they all fundamentally rely on extensive domain knowledge to allow a scientist to identify a problem, and formulate and test hypotheses. This is accomplished by developing an experimental design to gather the data needed to test the validity of the hypothesis. However, for many biodiversity studies this expert-centered parametric analysis alone is inherently limiting because collecting data for hypothesis-testing analyses at the spatial and temporal scope needed is logistically, financially, or ethically challenging and likely not feasible for a single individual expert.

Gaining insights into the patterns of species occurrence in complex ecological systems will require new synthetic analysis of massive amounts of disparate data (Brown 1995). Recently there has been much discussion about the need for the organization of large volumes of data and their use in scientific analysis in both the scientific (Lynch 2008), and popular (Anderson 2008) press. This need has led to the creation in the US of the \$100-million DataNet program (NSF 2007), and in Europe to the creation of the Alliance for Permanent Access (Angevaere 2008). The goal of these initiatives is to develop cross-domain data standardization and curation strategies to make scientific data available—from particle colliders to counting birds at a feeder—and preserve these data for long-term and unanticipated use over time and across

disciplines. While these initiatives focus on the cyberinfrastructure to organize and provide access to massive volumes of data, there has been less discussion on how this will impact scientific processes.

In this paper we introduce a new analysis paradigm for biodiversity studies that takes advantage of access to massive quantities of data. *Data intensive science* (Newman et al. 2003) takes a “data driven” approach, where information emerges from the data as opposed to a more traditional “knowledge-driven” approach that examines hypothesized patterns expected from the data. Data intensive science is emerging in the face of similar challenges across multiple scientific domains due to the accumulation of large quantities of data, and the need for new analysis techniques to study them. For example, recent advances in access to species occurrence data and environmental features data (e.g. climate, weather, land cover, human demographic data) are providing the impetus for this new paradigm in biodiversity studies. Here we describe our experiences in developing a data intensive science workflow for identifying the factors that influence the distribution and abundance of bird populations across North America.

Data intensive science uses digital data that are well documented, suitably protected, dependably preserved, and which can be integrated, explored, and analyzed through visualizations, simulations, and various types of model building. A well defined work flow harnesses the availability of these data and facilitates pattern discovery, hypothesis generation, and confirmation for complex systems. In this section we explain the data intensive science workflow and outline the processes that will be more completely described in later sections of this paper.

The basic steps in a data intensive science workflow for biodiversity research are shown in Figure 1. Observational data are collected from various sources. These data have to be organized, validated, and made available via a distributed network of resource providers. Next these data sources must be synthesized into a single federated data structure. This is a non-trivial process, because datasets come in various formats and are often collected with different

protocols, which requires sophisticated techniques for joining them in a meaningful way. After data have been prepared and made available, exploratory analysis techniques are used to discover interesting patterns. If non-parametric or semi-parametric approaches are used, flexible and powerful models can be trained that require little or no tuning by domain experts. By interpreting and analyzing these models, truly novel and surprising patterns that are “born from the data” can be discovered. These patterns in turn provide valuable insight for concrete hypotheses about the underlying ecological processes that created the observed data. Finally, the support for these hypotheses is assessed to identify the most interesting biodiversity patterns. Once the scientist has narrowed the search to a few promising patterns and hypotheses, they can employ traditional data collection and hypothesis testing techniques to verify their findings.

The creation of data intensive science workflows provides the opportunity for novel descriptions of the world, more detailed and extensive than previously possible. The opportunity now to discover patterns in massive quantities of data promises empirical scientific study of more complex and comprehensive systems, with immense potential for new and innovative scientific discovery. Potentially, a data intensive science process can even facilitate the opening of new paradigms in science (Kuhn 1962).

In the next sections, we will discuss the main steps of data intensive science workflows in more detail.

Acquiring data for data intensive science in biodiversity allows the study of complex ecological systems at ever-finer resolution and increasingly larger extents. However, the ability to access and combine large volumes of data from multiple sources presents both opportunities and challenges. In this section we define observational data, and identify the inherent challenges of using these data.

Data-collection methodologies affect the robustness with which causal effects can be assigned. While data from appropriately designed experiments provide the strongest evidence of causation, such data are often not well suited for use in data intensive research. This is

because experiments often manipulate the data collection environment, making data synthesis difficult if not impossible. Additionally, hypothesis-driven experiments are typically conducted on small spatial and temporal scales by researchers working autonomously. This creates a network of heterogeneous data repositories with little opportunity for data integration or reuse (Michener 2006). Alternatively, non-experimental data—that we will hereafter refer to as “observational data”—can be used to identify correlations and possibly suggest, but usually not demonstrate causation (Winship and Morgan 1999). These data are useful for generating or refining hypotheses that can then be tested experimentally to provide strong inference of causation.

Observational data needed to understand biological processes come from multiple and diverse sources distributed across a network of data providers (Fig 1). The mechanisms used to collect these data vary across data types: much remote sensing data are available to describe land cover, vegetation, and forest canopy characteristics; geochemical and soil characteristics are described by continental sensor networks; and anthropogenic data are available through a variety of government census products. Despite inroads in the use of sensors in ecology (Hamilton et al. 2007), the vast majority of existing species occurrence data are gathered by human observers (Kelling 2008). Thus, observational data may offer the only practical source of data for many biodiversity studies due to the logistical or ethical challenges of implementing large-scale hypothesis-driven experiments (Hargrove and Pickering 1992); even those large-scale experiments that do exist (Gardner et al. 2001) are conducted at a relatively small scale in relation to the ranges of most species.

Analyses of observational data must account for biases that cannot be or may not have been controlled during the data-collection and synthesis process (e.g. Parmesan and Yohe 2003, Root et al. 2003). Understanding these biases begins by understanding the provenance (i.e. information regarding the origins, identification, ownership, and structure) of the data sets of interest. Without this information it would be extremely difficult to understand the factors that may have affected the data collection process, know how the data were measured and reported, and identify biologically-relevant factors that affect organisms’ distribution and

abundance. With sufficient provenance metadata, potential sources of bias may be investigated and accounted for as part of the exploratory analysis (see later section).

While tremendous volumes of data are being made available, much care must be made when they are combined, because of the potential of joining data that have major incompatibilities. For example, the same spatial extent (e.g. the lower 48 United States) could be represented by large numbers of observations from very few sites or by a large number of sites with few observations per site. Similarly, temporal replication can be at widely varying scales, ranging from seconds to much longer intervals between records. How the samples were gathered, stored, and synthesized will constrain the types of questions that can be answered and the robustness of conclusions. Nevertheless, huge amounts of data can be combined because a common spatio-temporal context underlies all biodiversity studies.

The use of observational data for biodiversity research requires access to a broad and diverse array of relevant resources. In some scientific domains ubiquitous public access to data has become essential to research. For example, the US National Virtual Observatory (<http://www.us-vo.org>) allows researchers, students, and interested citizens to find, retrieve, and analyze data from a worldwide network of ground- and space-based telescopes. Developing similar resources for use in ecology presents significant challenges, because of the very heterogeneous nature of the data. What is required is the adoption of formal standards within an informatics framework of descriptive metadata and semantic organization, for archiving, access and synthesis (Jones et al. 2006). While much of this organizational effort in ecology is still being developed (Madin et al. 2007), we now describe the processes we took to organize data necessary to understand the patterns and abundance of bird populations.

We have organized more than 60 bird occurrence data sets, which are part of the Avian Knowledge Network (AKN) (<http://www.avianknowledge.net>). These data are provided to the AKN via a distributed network of governmental and non-governmental organizations. Each dataset has a complete descriptive metadata record that fully describes the provenance, access

level, geographic scope, and information on how the data were gathered and are stored. The descriptive metadata are available on the AKN website.

The AKN uses a data warehouse where all the dissimilarly structured source data are stored in a single format (Phokion et al. 2006). Creating a data warehouse requires considerable effort up front, but it eases subsequent use of these data by a wide range of analysts. Effective mechanisms for data dissemination should make data available without requiring advanced knowledge of data manipulation. Therefore the warehouse structure must be carefully developed to ensure efficient access to current data, provide sufficient descriptive metadata to document both the structure, contents, and use constraints, and allow the addition of new data when it becomes available. Note that data intensive workflows are not tied to a centralized warehouse architecture, and can be extended to federated data repositories (e.g., virtual data warehouses).

We have built a data warehouse for avian distributional data to support decision and analysis processes. The model is similar to other efforts to integrate diverse data from multiple fields of ecological research (McGuire et al. 2008). Our warehouse is multidimensional and consists of an event table (i.e. the information detailing the observation of a bird), and multiple predictor tables (i.e. variables that impact the observation of the bird) (Fig. 2).

The first step in creating our warehouse was bringing together bird observational data into the observation event table (Fig. 2). This was very challenging because data collection techniques are diverse, the original data resources were widely dispersed and owned by a variety of organizations, and the data formats varied dramatically. To overcome these issues we developed a common data model, the Bird Monitoring Data Exchange (BMDE) (Lepage et al. 2005). The BMDE captures as many data elements as possible to describe the bird observation event (i.e. who, what, how, and when an observation was made). The BMDE schema is based on Darwin Core (Wieczorek 2007), with extensions to describe characteristics of bird occurrence. Each contributing organization mapped their data to the BMDE. We then used DiGIR (the Distributed Generic Information Retrieval protocol;

<http://sourceforge.net/projects/digir>) to automatically transfer data from contributing organizations to the data warehouse.

While bird observation data come from multiple sources, each environmental attribute (i.e. predictor variable) associated with a bird observation came from a single, uniformly collected data set. Collection mechanisms for these variables varied from remote sensing (e.g., land cover) to anthropogenic surveys gathering human demographic information (Fig. 1). Because all bird observations and attributes have latitude, longitude, and date as shared context, we could join observations of species with observations of environmental features at individual locations.

The data warehouse is available through at the AKN website, and contains more than 55 million bird observations gathered at over 425 thousand locations in the US. In addition more than 500 predictor variables that describe the birds' environments are linked to each location where an observation event occurred. Finally, each data set contributed to the AKN has an associated level of access control, and all contributing members have complete authority over the use of the data they have provided, and can withhold data at any time from any party or application. It is recognized that the investment and time committed to the collection of a data set entitles the owner to the fundamental benefits of its use.

Within a data intensive science workflow such as the one developed for the AKN, creating the data warehouse was not an end in itself. The potential of our data organization processes is realized only when the data are used, and we have made inroads in the ease of access to analyses of the warehoused data.

Exploratory analyses for data intensive science have a fundamental goal of discovering patterns in the data that account for potentially complicated relationships between very large sets of predictors and the responses of interest. Conventional analyses of ecological data using well-established methods from the statistical sciences are geared toward hypothesis confirmations that require specific, detailed input in the form of parametric models that are subsequently confronted with data. When candidate sets of important environmental features

number in the tens, if not hundreds, these confirmatory tools are inefficient methods for discovering patterns (Hochachka et al. 2007). Instead, powerful, flexible, and efficient exploratory (model creation) analyses are needed to provide insights leading to hypothesis generation.

New exploratory data analysis tools emerging from the fields of machine learning, data mining, and statistics can automatically identify patterns in large and complex biodiversity data sources. For example, bagged decision trees have been used to identify the patterns of winter bird distributions across North America (Caruana et al. 2006). Several other sophisticated nonparametric tools for data exploration have been introduced to ecologists (Elith et al. 2006, Hargrove and Hoffman 2004, Phillips et al. 2006). All of these techniques share an ability to automatically identify patterns in data, making these techniques especially well suited for exploratory analysis.

Exploratory analysis of biodiversity data requires two main steps. First, the analyst must train a model to predict organism distribution or abundance with good generalization performance. Then she uses tools to explore the model and to find important patterns. We now discuss these two steps in more detail.

Training of accurate nonparametric models is highly automated. The main user specification is the set of predictors that the modeling process will consider. For exploratory analysis, the best strategy is to include as many uncorrelated predictors as possible. Doing this broadens the scope of ecological exploration, opening the door for unanticipated discoveries. It also provides a mechanism for assessing and accounting for biases caused by the observation process. Thus, for data intensive biodiversity studies, it is important to include predictors that describe the data collection, measurement, and organization processes *in addition to* predictors that describe important ecological and environmental processes.

After the model is developed, cross-validation is used to estimate the model's ability to generalize to unseen, new data. A model that generalizes well avoids over-fitting to spurious

patterns and noise in a given sample of observational records and hence produces better inferences on the underlying ecological processes.

Opening up the data mining black boxes. Although nonparametric predictive models can make highly accurate predictions, many are essentially “black box” methods, not designed to easily reveal how predictors are related to an accurate result. In the second step of exploratory analysis, the model itself is analyzed to elucidate what it has “learned.” Variable importance measures can be used to identify and rank the most important predictors (Caruana et al. 2006, Hastie et al. 2001). Once important predictors have been identified, their effects on the response can be studied by summarizing the model, e.g., with partial dependence functions (Friedman 2001, Hastie et al. 2001, Hooker 2007).

Partial dependence functions capture the relationship between selected predictors and the response, while accounting for the average effect of all other predictors. For biodiversity research, some of these functions might hold the key for the discovery of a new relationship between a predictor and a species’ occurrence. However, due to the volume of partial dependence functions generated when a multitude of predictors are analyzed for different “slices” and “dices” of the data space, it is infeasible to manually search through all possible predictor groups. This has led us to the development of a novel system for searching massive collections of model summaries, including partial dependence functions. These discovered patterns then provide the inspiration for the next steps: hypothesis generation and confirmatory analysis.

Confirmatory analyses for data intensive science seek to determine the support that observational data offer for specific models and hypotheses either generated from an exploratory analysis or by elucidating expert opinion. To undertake confirmatory analysis, statistical tools must be developed to test the correspondence between data and user-specified hypotheses, often given in the form of restrictions on nonparametric models. These must be translated into inferences and confidence sets based on formalized probabilistic estimates of uncertainty.

In a data intensive setting, specification of all but the simplest models is often complicated by the large numbers of predictors. To alleviate this problem, we are developing new semiparametric models that combine complementary aspects of parametric and nonparametric methods by incorporating flexible nonparametric model components within a parametric framework. For example, Additive Groves (Sorokina et al. 2007, Sorokina et al. 2008) is a decision-tree based method that allows an analyst to enforce additivity between arbitrary groups of predictors. Using this method, the analyst may allow large groups of predictors to influence the response without restriction while isolating the effects of other predictors for more focused inquiry. As another example, Hierarchical Predictive Models (HPM) (Fink and Hochachka 2009) use a hierarchical modeling framework to separate nonparametric estimation of fixed effects from parametric estimation of random effects. Viewed as an extension of more traditional semiparametric (mixed) models, HPM allows an analyst to specify a wide variety of semiparametric models. Together these methods substantially increase the scope of confirmatory analysis possible in a data intensive setting, by allowing the analyst to specify models for confirmatory testing of only those components that are of specific interest.

Considered from a statistical sampling perspective, uncertainty may be quantified simply by using resampling techniques like bootstrap or Monte Carlo techniques (Efron and Tibshirani 1994). We have used these techniques successfully to compute estimates of confidence for trends (Fink and Hochachka 2009) as well as for partial effects (Hochachka et al. 2007). However, resampling entire exploratory analyses is computationally intensive, and additional research is needed to improve the efficiency of estimating uncertainty so that confirmatory analyses can be carried out as a routine part of data intensive analysis. Confirmatory analysis along these lines must also account for the large number of summaries that are implicitly being tested, either by adjusting the uncertainty estimates or by the use of held-out validation data.

Visualizations are an essential part of exploration and analysis in a data intensive science workflow. Creating visualizations can be highly interactive, because much of the analysis functionality can be made available through desktop applications (e.g. Kepler (Ludäscher et al.

2006) that can link web-enabled data resources with computational and visualization resources to process the data.

To illustrate the role of visualization in data intensive biodiversity studies we present examples based on data on the presence or absence of wintering birds in North America (Fig 3). The bird occurrence data were collected in Project FeederWatch, a winter-long monitoring project in which members of the general public record the maximum number of birds seen at one time between mid-November and early April (Hochachka et al. 2007). These bird occurrence data were linked with 197 predictor variables, and we explored the relationship between the occurrences of 89 species of birds across 27 Bird Conservation Regions (i.e. regions that encompass landscapes that have similar bird communities) between 1994 and 2004. For our exploratory analysis we built 89 bagged decision tree models that allowed us to examine more than 125,000 partial dependence plots describing the additive effects of all the predictors for all species by Bird Conservation Region combinations. Next, ranking measures, such as slopes of regression relationships between each predictor and the predicted probabilities of bird species' occurrences, were used to search for the strongest relationships. To foster exploration and visualization of these results, we have made this library of partial dependence relationships and the ranking tool available at the Avian Knowledge Network (<http://www.avianknowledge.net/content/toolbox/>).

Conclusion: Anticipating and mitigating large scale threats to biodiversity such as climate change or land use change associated with human population expansion will require a thorough understanding of how environmental features structure ecological systems. One challenge to such understanding is the potential variation in importance and effects of different environmental features on species occurring in different geographic regions and through time. The trans-disciplinary nature of this work (Wake 2008) is also challenging, requiring knowledge of physical and biological processes as well as human impacts on these systems.

Faced with complex systems, a “data driven” approach to research allows information to emerge from the data as opposed to a more traditional “knowledge-driven” approach that is

most efficient when a small number of clear hypotheses can be stated and examined. We argue that a “data driven” approach is necessary in biodiversity studies due to the complexity of ecological systems, particularly when viewed at large spatial and temporal scales, and data-driven research requires a well-designed data intensive science workflow (Fig. 1). For a data intensive workflow solution to be successful access to large volumes of data from multiple sources and research domains must be acquired and coordinated. Efforts such as DataNet or National Ecological Observatory Network (NEON) in the US, the European Alliance for Permanent Access, and other specific scientific domains or biodiversity data clearinghouse initiatives are timely, and through a data intensive science process have the potential to change the paradigm for how science is carried out.

While there are inherent challenges in organizing and analyzing massive and heterogeneous data sets, leveraging them to unravel the complexity of ecological systems is essential if we are to understand the profound effects that humans have on earth’s natural systems, and develop science-based environmental policies. The significance of a data intensive science approach is that it allows scientists to put any and all available data on biodiversity and potential explanatory variables into model sets to analyze bigger and more complex systems than ever before. Data intensive research is an efficient method for generating valuable new hypotheses about complex systems, thus complementing and facilitating more traditional scientific processes of hypothesis generation and experimental testing to refine our understanding of the natural world.

Many aspects of a data intensive scientific workflow create opportunities for engaging a larger number and wider scope of people in the process of biodiversity research. The need to organize and document large and complex data sets creates an opportunity for making data resources widely available in forms that do not require advanced knowledge of database management. Similarly, methods that do not involve presupposition of specific models that relate environmental features to species occurrence can be highly automated. Automation permits on-demand web-based data analysis and visualization portals allowing researchers with strong biological intuitions but limited technical background in data management and analysis

to explore large quantities of data. Further, web-based tools provide opportunities for students at various levels and the general public to explore and gain a better understanding of ecological systems and processes.

We believe that online environments, like we are building into the Avian Knowledge Network, will foster and facilitate inquiry-based exploration by a broad spectrum of users that includes students, educators, citizens, resource managers, scientists, and policy makers. Ultimately our goal is to develop decision support tools that enhance interactive data exploration and ultimately empower users to develop their own new and valuable insights. Further, by providing “Web 2.0” functionality the results of analyses and visualizations can be dynamically incorporated into web sites or project reports for even broader dissemination and consumption.

In conclusion, we believe that overcoming the challenges in organizing and analyzing massive and heterogeneous data, and leveraging the resources to unravel the complexity of ecological systems is essential for understanding the profound effects that humans have on earth’s natural systems, and for developing science-based environmental policies.

Acknowledgements: This work was funded by the Leon Levy Foundation, and the National Science Foundation (Grant Numbers ITR-0427914, DBI-0542868, DUE-0734857, IIS-0748626, and IIS-0612031). In addition the authors would like to thank Rick Bonney, Art Munson, John Wiens, and 3 anonymous reviewers for comments on the manuscript, Walt Koenig for comments on the biology of Acorn Woodpeckers, Will Morris for developing the figures, and all of the individuals and organization who have contributed their observations to the AKN.

References:

- Anderson C. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired* 16.
- Angevaere I. 2008. Keeping the Records of Science Accessible: Can We Afford It? Annual Conference of the Alliance of Permanent Access. Budapest.
http://www.alliancepermanentaccess.eu/documenten%5CAlliance2008conference_report.pdf.
- Brown JH. 1995. *Macroecology*. Chicago: University of Chicago Press.
- Caruana R, Elhawary M, Munson A, Riedewald M, Sorokina D, Fink D, Hochachka W, Kelling S. 2006. Mining Citizen Science Data to Predict Prevalence of Wild Bird Species. In Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining 909-915.
- Delcourt H, Delcourt P. 2005. The legacy of landscape history: the role of paleoecological analysis. Pages 159-166 in J. Weins, Moss M, eds. *Issues and Perspectives in Landscape Ecology*. Cambridge: Cambridge University Press.
- Doak D, Marino P, Karieva P. 1992. Spatial scale mediates the influence of habitat fragmentation on dispersal success: Implications for conservation. *Theoretical Population Biology* 41: 21.
- Efron B, Tibshirani R. 1994. *An Introduction to the Bootstrap* Boca Raton FL: Chapman and Hall
- Elith J, et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.
- Fink D, Hochachka W. 2009. Gaussian semiparametric analysis using hierarchical predictive models. Pages 1011 in Thomson D, Cooch E, Conroy M, eds. *Environmental and Ecological Statistics special monograph on "Modeling Demographic Processes in Marked Populations*, vol. 3. New York: Springer.
- Friedman J. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29: 1189-1232.
- Gardner RH, Kemp WM, Kennedy VS, Petersen JE. 2001. *Scaling Relations in Experimental Ecology*. New York: Columbia University Press.
- Gaston K, McArdle B. 1994. The Temporal Variability of Animal Abundances: Measures, Methods and Patterns. *Philosophical Transactions: Biological Sciences* 345: 335-358.
- Hamilton M, Graham E, Rundel P, Allen M, Kaiser W, Hansen M, Estrin D. 2007. New Approaches in Embedded Networked Sensing for Terrestrial Ecological Observatories. *Environmental Engineering Science* 24: 192-204.
- Hargrove WH, Pickering J. 1992. Pseudoreplication: a *sine qua non* for regional ecology. *Landscape Ecology* 6: 251-258.
- Hargrove WH, Hoffman FM. 2004. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environmental Management* 34: S39-S60.
- Hastie T, Tibshirani R, Friedman J. 2001. *The elements of statistical learning: data mining, inference, and prediction*. New York, New York, USA: Springer Verlag.
- Hochachka W, Caruana R, Fink D, Munson A, Riedewald M, Sorokina D, Kelling S. 2007. Data mining for discovery of pattern and process in ecological systems. *Journal of Wildlife Management* 71: 2427-2437.

Hooker G. 2007. Generalized Functional ANOVA Diagnostics for High Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics* 16: 709-732.

Jones M, Schildhauer M, Reichman O, Bowers S. 2006. The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology Evolution and Systematics* 37: 519-544.

Kelling S. 2008. Significance of organism observations: Data discovery and access in biodiversity research. Report for the Global Biodiversity Information Facility, Copenhagen. Report no.

Kuhn TS. 1962. *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Lepage D, Kelling S, Ballard G. 2005. The Bird Monitoring Data Exchange Schema <http://www.avianknowledge.net/content/about/bird-monitoring-data-exchange>.

Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, Lee E, Tao J, Zhao Y. 2006. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience* 18: 1039-1065.

Lynch C. 2008. Big data: How do your data grow? *Nature* 455: 28-29.

Madin J, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F. 2007. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics* 2: 279-291.

McGuire M, Gangopadhyay A, Komlodi A, Swan C. 2008. A user-centered design for a spatial data warehouse for data exploration in environmental research. *Ecological Informatics* 3: 273-285.

McMichael AJ, Butler CD, Folke C. 2003. New Visions for Addressing Sustainability. *Science* 302: 1919-1920.

Michener W. 2006. Meta-information Concepts for Ecological Data Management. *Ecological Informatics* 1: 3-7.

Newman HB, Ellisman MH, Orcutt JA. 2003. Data-intensive e-science frontier research. *Communications of the ACM* 46: 68-77.

NSF. 2007. Sustainable Digital Data Preservation and Access Network Partners (DataNet) <http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.htm>.

Parmesan C, Yohe G. 2003. A globally coherent fingerprint of climate change impacts across natural systems. *Nature (London)* 421: 37-42.

Phillips S, Anderson R, Schapire R. 2006. Maximum Entropy Modeling of Species Geographic Distributions. *Ecological Modelling* 190: 231-259.

Phokion GK, Jonathan P, Wang-Chiew T. 2006. The complexity of data exchange. *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Chicago, IL, USA: ACM.

Root T-L, Price J-T, Hall K-R, Schneider S-H, Rosenzweig C, Pounds JA. 2003. Fingerprints of global warming on wild animals and plants. *Nature (London)* 421: 57-60.

Sorokina D, Caruana R, Riedewald M. 2007. Additive Groves of Regression Trees. Pages 323-334 in Kok J, ed. *18th European Conference on Machine Learning*. Warsaw, Poland: SpringerLink.

Sorokina D, Caruana R, Riedewald M, Fink D. 2008. Detecting Statistical Interactions with Additive Groves of Trees. *Proceedings of the 25th International Conference on Machine Learning*.

Tuomisto H, Ruokolainen K, Yli-Halla M. 2003. Dispersal, Environment, and Floristic Variation of Western Amazonian Forests. *Science* 299: 241-244.

Wake M. 2008. Integrative Biology: Science for the 21st Century. *BioScience* 58: 349-353.

Wieczorek J. 2007. Darwin Core TDWG Task Group

<http://www.tdwg.org/activities/darwincore/>. <http://www.tdwg.org/activities/darwincore/>)

Winship C, Morgan SL. 1999. The Estimation of Causal Effects from Observational Data. *Annual Review of Sociology* 25: 659-706.

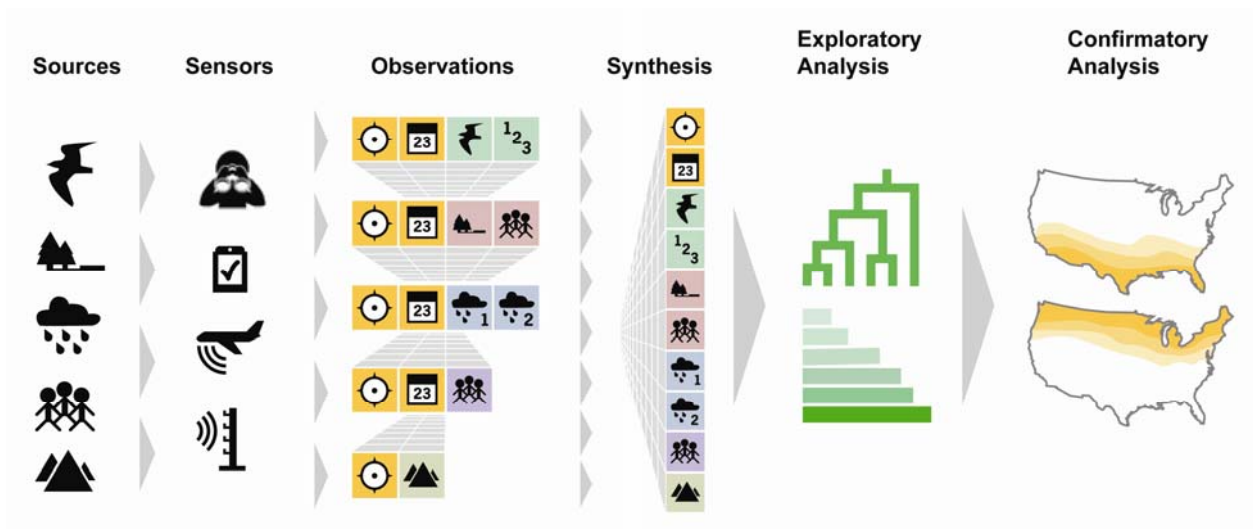


Figure 1. High level processing workflow for integrative data intensive biodiversity research.

Physical events and objects are recorded through sensor, observer, and survey networks. These data are stored in heterogeneous repositories. Informatics processes allow heterogeneous data to be synthesized for processing. Exploratory analyses (analyses useful for generating hypotheses) can drive confirmatory summative analyses. A variety of visualization tools allow these data to be viewed by a broad public.

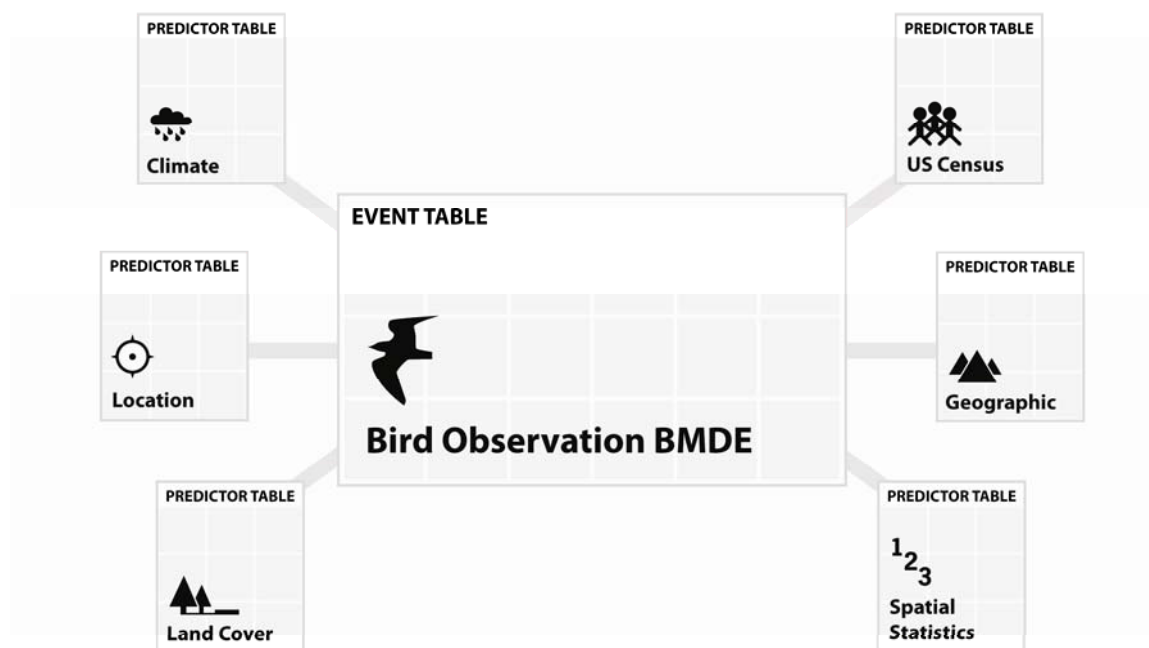


Fig 2. Multidimensional data warehouse for bird occurrence. Data were stored in a star schema consisting of an event table, which consists of observations of birds, with multiple predictor tables, which consists of more than 500 variables that could influence the bird observation. The data warehouse is accessible via the Avian Knowledge Network website (<http://www.avianknowledge.net>).

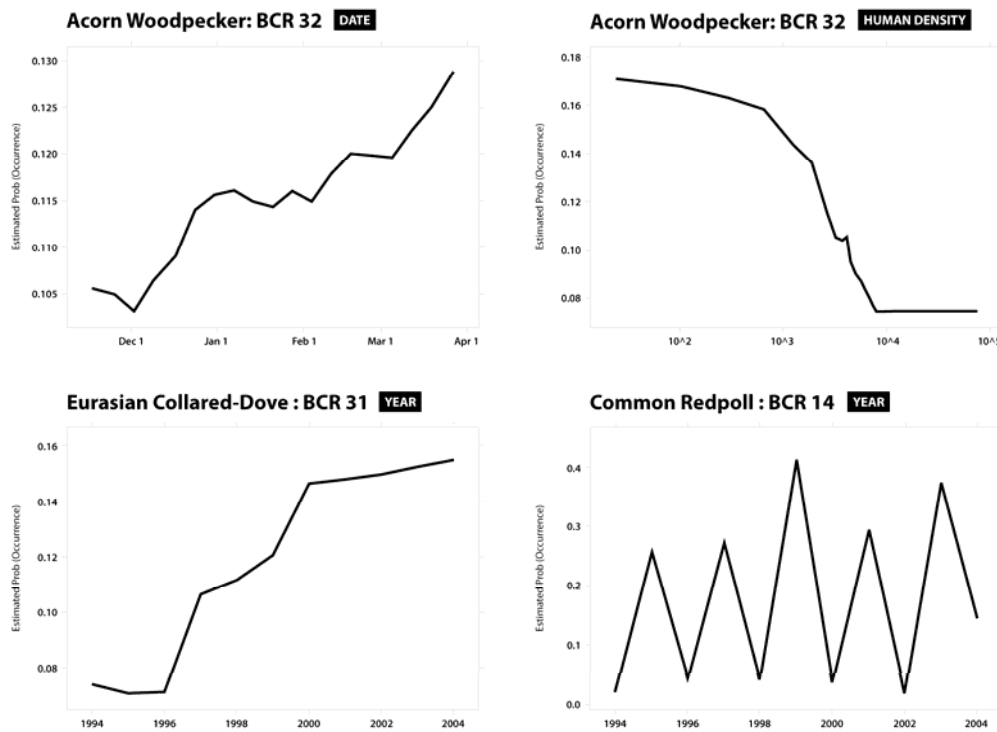


Figure 3. Examples of interesting patterns that emerged from exploratory analysis of winter bird populations. The top 2 partial dependence plots represent patterns of occurrence of Acorn Woodpecker (*Melanerpes formicivorus*) in California. The graph on the upper left shows that Acorn Woodpeckers visit bird feeders more frequently during the later part of winter. This may be due to a premium for being in good physical condition, and thus the need for more food, in preparation for the breeding season. The plot on the upper right shows a drop in the probability of Acorn Woodpecker occurrence as human population density increases above 1,000 people per square mile. It is hypothesized that habitat competition between the woodpeckers, which need dead or dying branches to store acorns, and humans who remove these branches could be the cause for this decline. The plot on the bottom left shows the rapid

expansion of the invasive Eurasian Collared-Dove (*Streptopelia decaocto*) in Florida. The Eurasian Collared-Dove was originally introduced to the Bahamas in the late 1980s, but has now expanded into to British Columbia. See text for explanation of methodology. The graph on the bottom right shows the biennial winter irruptive migration of Common Redpoll (*Carduelis flamma*) into New England, likely caused by biennial cycles of production of tree seeds in Northern Canada.