Background

Ecology needs data-intensive approaches to predict the responses of biodiversity and ecosystem function to global change, address invasive species, and prioritize areas for conservation. However, the challenges common in many areas of data science are often magnified in ecology. These include highly heterogeneous data across systems and studies, an unprecedented volume and velocity of data that is now being collected using sensor networks and remote sensing, and the lack of ecologists with the necessary training to deal with this magnitude of data. Despite a clear need, the discipline has failed to embrace data-intensive approaches (see e.g., Paine 2010, Lindenmayer & Likens 2011, 2013). Ecology needs leaders in data science to: 1) demonstrate the value of data-intensive approaches in ecology; 2) mine, assemble, and model existing data; 3) integrate existing data with the large amounts of data now being generated by automated data collection and citizen science; and 4) train ecologists in the tools and approaches necessary for data-intensive science.

Major Accomplishments

I entered graduate school with a field ecology background, but became frustrated with the scope of inference that could be accomplished using typical ecological approaches. It was possible to understand a single system well, but these results could not be generalized across the globe and across the diversity of life. I joined one of the only labs doing data-intensive ecology and have spent my career using large data compilations to understand ecological systems at continental to global scales, developing tools to make it faster and easier to conduct this type of research, and training the next generation of data-intensive ecologists.

Research: My research uses large data compilations to understand the processes driving ecosystems, test ecological models, and make predictions regarding ecological systems. We have developed novel approaches to understanding patterns of biodiversity based on dividing species into resident and transient species using the structure of their population time-series (White & Hurlbert 2010, Coyle et al. 2013). We have led the way in strong general tests of ecological theory by creating the largest combined datasets ever used in community ecology and using them to evaluate and compare ecological theories (White et al. 2006, Thibault et al. 2011, White et al. 2012, Locey & White 2013, McGlinn et al. 2013, Xiao et al. 2013). Finally, we have built on these ecological theories by using them to make predictions for biodiversity, rarity, and patterns related to ecosystem level fluxes (White et al. 2012, Xiao et al. 2013). We are beginning to combine this with new work using machine learning and remotely sensed data in multi-level models to forecast the future state of ecological systems and make predictions for unsurveyed locations.

Software Development: Conducting this type of research requires the time-consuming assembly of numerous heterogeneous datasets. Typically every scientist writes custom code to handle the data or assembles them by hand. This wastes time and effort organizing data that could be spent doing science. My group built the EcoData Retriever to address this problem. It automates the discovery, downloading, cleaning, and reformatting of ecological data. This simplifies the process of finding, acquiring, and using data and therefore increases the use of the diversity of ecological datasets. Altmetrics show that the Retriever is both highly recommended and highly cited (stars and forks on GitHub in the 91-99%tile), the associated paper is in the 98th percentile of online impact for scientific papers, and the installers were downloaded over 230 times in the last two weeks.

Training & Mentoring: I actively train and mentor the next generation of data-intensive biologists. Using an NSF CAREER Award I have developed a suite of university courses that teach core aspects of data science to biologists (programming, databases, statistics, visualization). Course material is available online and has been viewed over 125,000 times by users in over 150 countries. Students and postdocs in my lab with backgrounds including field biology, mathematics, statistics, and computer science receive in-depth training in data-intensive approaches to ecology. I am also actively involved in Software Carpentry: developing material, teaching ecology focused workshops (6 in the last 2 years), and serving on the advisory board. Finally, I have taken a leadership role in making biology more open, reproducible, and data-intensive by: 1) openly sharing reproducible code, data, and grant proposals; 2) writing papers on data management and sharing, best practices in computational science, and the need for preprints in biology; and 3) explaining the importance of these approaches on my blog and through my active presence on Twitter.

Future Research Directions

My future research directions involve working in three key areas that will create a foundation for data-intensive ecology: 1) tying data and theory together more tightly by expanding the use of data-driven modeling in ecology (e.g., machine learning, hierarchical modeling); 2) leveraging existing data to test, improve, and make predictions using process based models; and 3) developing tools and personnel to help scientists handle the challenges of ecological data.

My data-driven modeling efforts have so far focused on entropy maximization models, which predict the state of a system based on a set of empirical constraints. We will build on this foundation by expanding more generally into machine learning and hierarchical modeling to make predictions for ecological patterns and processes across scales. We will develop a suite of master models to address core questions about biodiversity, population dynamics, and ecosystem processes, including biosphere influences on carbon fluxes and global change. Specifically, we will develop models for the distribution of individuals of species across the globe, the traits of those individuals, and the dynamics of both counts and traits over time. These models will use large compilations of climate, land use, and ecological data as predictors, including data on the abundances of other species and their potential interactions both within and across taxonomic groups (e.g., birds compete with other birds for food and require plants for food and nesting). These models will be trained and tested using compilations of ecological data from across ecosystems and taxonomic groups, including an expanded version of our current compilation (which currently includes data on the distribution of ~50 million individual organisms), text mining data on interactions among species, and growing compilations of individual level trait data (~2.5 million trait values).

I will leverage the large amounts of data currently being generated in ecology to provide improved testing of process based ecological theories. We will conduct strong tests and comparisons of population, community, and ecosystem level models by evaluating their performance across ecosystems and taxonomic groups, testing all of their predictions simultaneously using independent data, and directly comparing predictions of different models for the same patterns. Specifically, we will focus on models relating population dynamics to environmental drivers, unified community theories making predictions for large numbers of ecological patterns, and ecosystem models that relate properties of the organisms at a site to the system level fluxes. The information from these comparisons will be used to identify the most promising models for making broad scale ecological predictions, to identify aspects of these models that require improvement, and to change or constrain these models to yield improved predictions for ecological systems.

To help ecologists handle the challenges of data-intensive science I will continue to expand my efforts to make data-intensive ecology easier, more robust, and more reproducible. I plan to add provenance tracking and automated reproducibilty features to the EcoData Retriever that will allow data processing steps to be documented and easily reproduced. I will develop new tools to make it easier to combine ecological, environmental, taxonomic, and other datasets in reproducible ways to allow ecologists to quickly make use of the broad array of data that relates to ecological systems. To address the lack of training for scientists in data-intensive approaches I will participate in the development of a new set of Data Carpentry material that builds on Software Carpentry's knowledge and successes to train the next generation of scientists. In combination, this research, tool building, and training will help establish a data-intensive era for ecology.