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 discipline has failed to embrace these approaches, with repeated recent editorials by members of the National Academy of Sciences criticizing their use. Ecology needs leaders in data science to break down these barriers, make meaningful predictions about ecological systems, and train the next generation of data focused ecologists.

Ecological data is highly heterogeneous and many or the core data types have not experienced the drastic declines in data collection costs of other fields and, where it has, the historical data that is so crucial to understanding global change cannot be recollected. In addition, the field lacks researchers with the necessary training to deal with the volume and velocity of data that is now being collected using sensor networks, satellite collars, and remote sensing. Therefore data-intensive ecology faces three major challenges: 1) mining, assembling, and modeling the large amounts of existing complex and heterogeneous data; 2) integrating these data with the high volume and velocity data from emerging approaches to automated data collection; and 3) training ecologists in the tools and approaches necessary for data-intensive science.

Major Accomplishments

I entered graduate school with a traditional field ecology background, but quickly became frustrated with the limited scope of inference that could be accomplished using typical ecological approaches. It was possible to understand a single system well using observational and experimental approaches, but these results could not be generalize across the globe and across the diversity of life. I joined one of the only labs in the world 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 compilations of ecological data to understand the processes driving ecosystems, test ecological models to determine if they are general or operate differently across different ecosystems and taxonomic groups, and make ecological predictions. We have developed novel approaches to understanding patterns of biodiversity based on dividing species into resident and transient species based on the structure of their population time-series and modeling these groups separately to yield improved predictions for biodiversity (Hurlbert & White 2005, White & Hurlbert 2010, Coyle et al. 2013). We have lead 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). During my first sabbatical I have been learning machine learning and we are beginning to combine this with our work making predictions using general theories to use remotely sensed data in a multi-level modeling framework to forecast the future state of ecological systems and make predictions for unsurveyed locations (McGlinn et al. in prep). ADD GENERAL STATS ON IMPACT OF THIS WORK.

Software Development: Conducting this type of research requires the time-consuming assembly and associations of numerous heterogeneous datasets. Typically every scientist working on these datasets writes custom code to handle the data or worse assembles them manually. This results in a substantial waste of time and effort on data munging instead of science. My group built the EcoData Retriever to address this problem. This extensible software framework automates the tasks of discovering, downloading, cleaning and reformatting ecological data files and stores them in relational databases or as flat text files. It is basically a package manager for ecological data. This simplifies the process of finding, acquiring, and using ecological data making it easier to use the diversity of ecological datasets in combined analyses. Altmetrics by Impact Story show that the Retriever is both highly recommended and highly cited (stars and forks on GitHub in the 93%tile and 89%tile respectively), AltMetric.com shows that the associated paper is in the 97th percentile of impact for scientific papers, and the installers were downloaded over 200 times in the last week.

Training & Mentoring: I have been active in training and mentoring the next generation of data-intensive ecologists and biologists more generally. As part of my current NSF CAREER Award I have developed a suite of university courses that teach core aspects of data science (programming, databases, statistics, visualization) focused on addressing biological problems (http://compb.io). Material for this class is openly available online and has been viewed over 125,000 times. Students and postdoctoral researchers in my lab with backgrounds from field biology to mathematics to statistics receive in-depth training in data science. To help train students beyond my university I am 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 moving biology towards more open, reproducible, and data-intensive approaches by: 1) leading by example through 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 (all written in the open); and 3) through explaining the importance of these approaches on my blog) and through my active presence on Twitter.

Future Research directions

Research: My recent research has focused on developing and testing general theories that make predictions for many patterns simultaneously. I was using this as a route to making predictions about ecological systems at large scales. While this was productive I now plan to broaden my emphasis to even more fundamental questions that are they key to unlocking predictions for most ecological patterns and processes across scales. The two master questions that we will pursue are: 1) What is the number of individuals of every species at every location at all points in time? and 2) How do the traits of individuals vary across species, locations, and times? The answers to these questions can be aggregated to answer most questions about biodiversity (by aggregating the predictions across many species), conservation of individual species or groups of species (by being able to make predictions for how their populations will change), and ecosystem services including biosphere influences on carbon fluxes and global change. To support this broadening focus I am currently spending my first sabbatical learning machine learning and expanding my background in hierarchical modeling approaches.

These questions are widely recognized as important but rarely pursued with any depth. Most efforts focus only on a single group of species in single region, only model whether a given species is present or absent, ignore interactions between species that are known to be important, and weights the trait values at the species level ignoring the large differences in the numbers of individuals between species and the variance in traits among species. This is because answering these questions in the most general and useful ways is an inherently difficult problem requiring novel combinations of disparate data sources and the development of novel models for approximating missing information. For example, individual level trait data is often not available at large scales, so ways of modeling individual level variance in traits and combining this with large scale data are necessary. We are leading the way in developing these approaches (Thibault et al. 2011).

Question 1 will be addressed using machine learning to model the number of individuals of each species across space based on a combination of climate, land use, the number of individuals of other species in both the same and disparate taxonomic groups (e.g., birds compete with other birds for food and require certain kinds of plants for food and nesting). We will conduct this research on compilations of data from large numbers of datasets the global and the diversity of life by expanding our current compilation which includes data on ~50 million individual organisms. Meaningfully comparing counts of individuals from different datasets is a challenging problem since different methods and sampling intensities do not produce equivalent counts. We will address this using by using more limited high quality datasets and methods for estimating true densities from samples with models to link these corrections to large scale data.

Question 2 will use machine learning and hierarchical modeling to understand variation in individual traits. Traits can be used either directly or in combination with models to determine most of the aggregated properties of individual organisms. For example, the traits of individual trees in a forest can be used to estimate the carbon flux of that forest and the amount of carbon that would be released if the forest were cut down or burned. To accomplish this we will develop methods for scaling small scale information on how traits vary within species to global scales, by building on our work modeling individual size variation in birds at these scales (Thibault et al. 2011).

Software Development: Development of the EcoData Retriever was funded by an NSF CAREER award that is reaching its end. Future development the Retriever includes four major additions: 1) Provenance tracking. Science is experiencing a reproducibility crisis as it is becoming clear that published results often cannot be reproduced. Part of the challenge of reproducibility in data science research is that most of the steps related to data: downloading it, cleaning it up, restructuring it, are either done manually or using one-off scripts and therefore this phase of the scientific process is not reproducible. The EcoData Retriever already solves many of these problems, but it doesn't currently keep track of exactly what has been done and therefore fails to support full reproducible workflows. We will add both tracking of manipulations and the ability to easily reuse the exact versions of the code and raw data to reproduce all data munging steps exactly. 2) While simplifying the acquisition and use of individual datasets is a major step forward the real challenge for data intensive approaches in ecology is the ability to combine numerous medium and small datasets to make inferences and predictions. General solutions to this extremely difficult problem remain a long way off, so we will develop an intermediate term solution that will work today for combining datasets in a shareable and reproducible manner. The Data Blender will use simple recipes to describe which tables from each dataset to combine and how to combine them. While, the human involvement in writing recipes limits the ability to access the full long tail of ecological data, it will make it easy to integrate dozens of datasets of into single combined analyses in a fully reproducible way. This is a general problem from across data science and the Data Blender will be developed as a separate project targeted at data scientists in general.

Training & Mentoring: I and my lab will continue build on our training and mentoring efforts to bring ecology into the data science era.

Notes:

- Propose rebranding EcoData Retriever as Data Retriever

Bone yard

2) Taxonomic name resolution. One of the challenges of ecological (and evolutionary) data is that the names of species are constantly being redefined. This makes it difficult to combine datasets to do interesting science. We will add automated reconciliation of different species names as part of the process of accessing the data by using existing web APIs for resolving species names (e.g., iPlant's Taxonomic Name Resolution Service).

3) Creation of simplified datasets. Many of the major databases in ecology are very complicated and contain far more information than most ecologists want to work with. We will automate the creation of

simplified datasets that provide the core data for exploring the common questions in ecology.

Research in ecology often fails to tackled questions in general ways The central focus of my research over the next 5 years is making robust predictions for population, community, and ecosystem level phenomena at continental to global scales. Conceptually this requires the identification of “master questions” for ecology; i.e., questions that if answered provide answers to large amounts of both basic and applied research.

The cutting edge of this area has just started to focus on interactions between species, but this work focuses on interactions within a taxon (e.g., interactions with other birds). Most important interactions are across taxa (food, predators, etc.),