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 models 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). We are now combining these approaches with remotely sensed data and machine learning 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).

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. 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, AltMetric.com shows that the associated paper is in the 97th percentile of impact for scientific papers, and data from the Python Package Index suggests ~100 downloads per month.

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). 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.