1. Fundamental Question and Five-Year Impact

1.1 Impact

Despite the critical need to forecast how nature will respond to global change, ecology is still primarily a descriptive science focused on understanding, rather than predicting, nature. Little research generates meaningful predictions and, when predictions are attempted, they tend to focus on a single ecosystem. This lack of prediction is slowing progress in ecology (McGill 2012a), because the field lacks meaningful benchmarks for comparing and improving models and fails to develop models that can be applied to understudied ecosystems and species. Despite having recognized the importance of forecasting for over a decade (Clark et al. 2001), ecology has failed to embrace a predictive approach to science.

Now is the perfect time to lead ecology from a descriptive to a predictive science. We now possess the abundance of data necessary to tackle prediction, the methods to develop predictive models from these data (e.g. machine learning, hierarchical modeling), and the mathematical and computational technology to develop and assess models from across the data-driven to theory-driven spectrum (Luo et al. 2011). I will develop and test predictive models for ecology across different levels of organization, ecosystems, and the diversity of life. I will build tools to make data-intensive science easier and train the next generation of scientists in data-intensive approaches. All of this work will be conducted in an open, collaborative, and reproducible manner. This will broaden its impact by increasing knowledge of the approaches and building a community to further the development of the tools. In combination, **the impact of my work will be to make ecology a more predictive, data-intensive, and open science that is capable of addressing the major ecological and environmental challenges of our time.**

1.2 Fundamental Question

I will focus on one of the fundamental questions in ecology: **what will nature look like in the future?** This question is fundamental to knowing how well we understand ecological systems because, if we cannot predict how they will change, we do not understand how they operate. As such, answering this fundamental question requires understanding what governs the structure, dynamics, and fluxes of ecological systems. This question is also fundamental to the application of ecology in management and policy decisions. To protect at-risk species, conserve biodiversity, and maintain ecosystem services, we need to be able to predict how nature responds to stresses such as climate and land-use change, and potential interventions such as the creation of reserves and the removal of invasive species (Clark et al. 2001, Evans et al. 2013).

I will make forecasts for three major areas of ecology: 1) the abundance of individual species; 2) the structure of communities (groups of species in the same region); and 3) the function of entire ecosystems. Populations, communities, and ecosystems are three of the central levels of organization in ecology. They are all amenable to data-intensive approaches due to large amounts of existing data (from individual studies, citizen science projects, and coordinated government sampling) and to a massive influx of new data from ecological observatory networks. Focusing on these dimensions of ecology will allow us to predict how species distributions, biodiversity, rarity, ecosystem fluxes, and ecosystem services will respond to anthropogenic pressures including shifts in climate, changes in land-use, and invasive species.

I will seek to capture how these different aspects of ecology change through time and vary across ecosystems and the diversity of life. I will develop predictive models using both data-driven and theory-driven approaches, using large compilations of ecological data to evaluate and improve the models. I will use far more comprehensive and sophisticated data compilations than have been brought to bear on these questions before, including data on hundreds of millions of species' occurrences, high temporal resolution climate and land-use data, and newly available text-mining and compilation-based data on species traits and interactions.

1.3 Measuring Progress

The core measure of progress is how well my research group can predict independent data and how effectively we can forecast the future state of ecological systems. We will measure the progress of our research at three levels based on our ability to: 1) predict the state of ecological systems in different locations; 2) forecast and hindcast within existing time-series when training models only on data from the beginning or end of the time-series; and 3) forecast the future state of ecological systems. Each year we will publish predictions for the state of ecological systems one to ten years into the future and evaluate those predictions every year as new data is collected.

Beyond my own research progress, a fuller measure of my impact will be whether ecology as a field focuses more on prediction and data-intensive approaches. I hope to create an environment that fosters forecasting by explicitly publishing forecasts and evaluating their accuracy. This will help create a culture focused on producing better forecasts, like that in disciplines with successful forecasting, such as weather and climate (McGill 2012b; see e.g. Kalnay 2003). To accelerate this transition I will conduct this research in a fully open manner, using open notebooks, public code repositories, and social media as outreach for the ideas and to encourage collaboration. All code will be open source, all training material will be open access, and all papers will be open access and posted as preprints prior to submission.

I will measure the influence of these efforts, and further the outreach, by running a series of "forecasting challenges" (similar to Kaggle competitions but with a focus on ecological forecasting). Impact on the field will be measured using the level of participation as an indication of how interesting and popular these approaches are among ecologists, and using the collective performance of the competitors to measure our success as a field in using data-intensive approaches to predict the future state of ecological systems.

2. Advancing Data Science Methodologies and Human Capital

We have too much data and too many important problems to be addressed by the small number of individuals with the requisite skills to work with large amounts of heterogeneous data. Realizing the potential of data-intensive approaches requires us to both bring the data to the researchers by developing improved tools for the acquisition, assembly, and analysis of data, and bring the researchers to the data by providing training in computational, statistical and other data science methodologies. Over the last five years I have been actively building these bridges between researchers and data as part of an NSF CAREER award and I plan to significantly expand these efforts by: 1) developing methodologies for working with the variety dimension of big data by building software that automates the acquisition and assembly of heterogeneous data sources; 2) developing approaches for modeling complex data and making them available in easy-to-use software; and 3) training scientists in data science skills. To maximize the impact of these efforts, all tools and training material will be developed in public GitHub repositories using open source and open access licenses. I will use this openness to actively encourage collaboration from both scientists and members of the technology community. In combination, these efforts will allow more scientists to engage in data-intensive approaches, and will let them spend more time focusing on doing science and less time wrestling with data.

2.1 Methodologies for Automatically Combining Heterogeneous Datasets

Combining heterogeneous data from disparate sources and formats is a core challenge in many areas of data science, and one that is particularly prevalent in my research. Typically this involves individual researchers developing custom scripts to download, cleanup, and restructure individual datasets, followed by even more custom scripts for combining datasets. This is error prone, time consuming, and does not allow scientists to benefit from each other's knowledge and effort. We can do better. By building tools to automatically handle

the data side of data science we can remove impediments to data-intensive approaches and allow scientists to focus on doing science.

My lab developed a platform for acquiring, cleaning, and restructuring heterogeneous data sources in reproducible ways, and is building a community who are adding and improving datasets (Morris & White 2013). We are expanding the platform to non-ecological data and plan to expand its provenance and reproducibility functionality. The next step is to tackle the challenge of combining datasets. I will lead the development of a general tool for automatically combining multiple heterogeneous datasets in reproducible ways. This tool will build on our successes in solving the individual dataset problem, using generalized routines to automate the handling of standard tasks involved in assembling datasets while leveraging human collaboration to develop the metadata describing how to combine datasets. This tool will interface with efforts for acquiring and streaming data, such as dat, rOpenSci, rOpenGov, NEON and our Data Retriever, to allow data from all three dimensions of big data to be easily combined to answer fundamental scientific questions.

2.2 Methods for Complex Data

Most data science methodologies assume that while data may be large and heterogeneous the data themselves are relatively simple: responses are linear, there is a single response variable, and data points are independent and identically distributed. However, many data-intensive questions involve data that violate all of these assumptions. For example, my research requires simultaneously predicting the interrelated abundances of hundreds of species that respond to climate in non-linear ways (Harris 2014), with context-dependent interactions among species (Poisot et al. 2014), where standard cross-validation fails due to strong spatial correlations in both features and outcomes (Bahn & McGill 2012). These challenges apply to many areas of data science. They require complex approaches that are capable of simultaneously handling non-linear responses and predicting high-dimensional joint distributions as outcomes (e.g., stochastic neural networks, Markov random fields), and methods for handling complexities such as spatial autocorrelation, irregularly sampled time-series, and missing data. We will build on existing methods (Le Rest et al. 2014) to provide general solutions to cross-validation in spatially and temporally autocorrelated contexts, build general implementations of our approaches for forecasting the distributions of species and ecosystem services, and extend methods for dealing with missing and irregularly sampled data. The solutions we develop will be broadly useful to any field that deals with complex data. Our core focus will be developing both tools and training to allow scientists across disciplines to take advantage of these approaches.

2.3 Building Human Capital

Tools can help bring data to scientists, but they cannot overcome the lack of individuals with the skills to conduct data-intensive research. To build human capital, we need to train scientists at all levels in the tools and approaches for tackling data-intensive problems. Just like open source software projects, training initiatives benefit from collaboration and community. This is why I focus my training efforts as a core member of the Software Carpentry team. While data science skills overlap with software development skills, major aspects of data science approaches are not covered in the current Software Carpentry curriculum. I am part of a core group that is in the early stages of developing a Data Carpentry curriculum that focuses on the tools and approaches of data science. I would use support from this award to help build both beginner and advanced curricula, to teach this material in workshops and university courses, and to develop approaches to engaging scientists in collaborative open-source communities. This will help produce a new generation of data-intensive scientists with the ability to work collaboratively to address fundamental questions using the variety, volume, and velocity of data that are now available.