Duke Data+ 2017 Project Proposal

Data Visualization and Data-Driven Curriculum Development for Long-Term Ecological Records

PI: Emily S. Bernhardt, Professor, Department of Biology

Primary Mentor: Richard E. Marinos, Ph.D. candidate, Nicholas School of the Environment

Other Mentors: Matt R. V. Ross, Ph.D. candidate, University Program in Ecology

Gene E. Likens, president emeritus, Cary Institute of Ecosystem Studies

I. Introduction

Hubbard Brook Experimental Forest, in the White Mountains of New Hampshire, has been the site of some of the most well-known ecological findings of the past century. Prominent examples include the discovery of the environmental impacts of acid rain and the discovery of the role of vegetation in ecosystem nutrient retention. These findings are taught in most every introductory ecology class, and there is a wealth of public data from Hubbard Brook that could be used to teach these canonical experiments in an interactive, exploratory manner. Currently, however, these datasets suffer from existing in diverse formats, as well as lacking a software framework that can be used to examine the data without knowledge of statistical programming languages.

We propose to develop an interactive data visualization web app that will allow students and researchers to explore the richness of sixty years of ecological data collected at Hubbard Brook. Members of the Bernhardt lab have developed a skeleton of this app (http://hubbardbrook.web.duke.edu), but we seek Data+ team members to turn this skeleton into a robust research and teaching tool. This app will consist of two modules, a data stories module and a data explorer module. The data stories module will focus on a data-driven retelling of some of the classic ecological principles learned at Hubbard Brook. The primary audience for this module will be undergraduates and advanced high school students. The data explorer module will allow users to rapidly synthesize the long-term ecological records at Hubbard Brook for exploratory data analysis. The target audience for this module will be advanced undergraduates, graduate students, and researchers.

II. Mentorship

The primary mentor of the Data+ students will be Richard Marinos (rem31@duke.edu). Richard is a fifth-year Ph.D. student who conducts his dissertation research at Hubbard Brook. He will commit to mentoring the sudents for five to seven hours per week. Additionally, Matt Ross (matt.ross@duke.edu), a previous Data+ graduate student mentor, will occasionally provide big-picture guidance on the project. Emily Bernhardt will also periodically meet with students to ensure that the work is focused on meeting curricular goals. Emily is currently a co-PI of the Hubbard Brook Ecosystem Study. Finally, the Data+ team will meet twice with Gene Likens (likensg@caryinstitute.org) to receive feedback on the project. Gene Likens co-founded the Hubbard Brook Ecosystem Study over fifty years ago. He is a recipient of the Presidential Medal of Science and a member of the National Academy of Sciences.

III. Dataset Description

Students will work on the datasets produced at Hubbard Brook under the aegis of the NSF's Long-Term Research in Environmental Biology (LTREB) and Long Term Ecological Research (LTER) programs. The data represent sixty years of ecological research by hundreds of investigators, and have resulted in some of the most significant findings in ecosystem ecology. They also support a robust continuing research program, providing a baseline of data that allow current researchers to continue to formulate and test novel hypotheses of ecosystem functioning.

These datasets include:

- Over fifty years of weekly precipitation and streamwater chemistry measurements for ten gauged watersheds. These 40k+ records comprise the longest continuous streamwater and precipitation chemistry measurements in the world, and Gene Likens will make them publicly available for the first time this spring, allowing Data+ participants a “first crack” at analyzing and visualizing this dataset.
- Sixty-two of years instantaneous streamwater and precipitation volume measurements for ten gauged watersheds, over 2M records.
- Fifty years of bi-decadal vegetation censuses with repeated measures of every tree in over two hundred plots.
- Twenty-five years of bi-decadal soil surveys with six hundred samples taken per survey.
- Twenty-two years of annual leaf litterfall mass and chemistry measurements
- Annual animal census data including fifty-two years of bird census data.
- Twenty-five years annual soil biological process measurements at forty plots.

Taken together, these data provide one of the most complete pictures of ecosystem growth and function of nearly any forest in the world. All of the data to be used in the project are available (or will be available by spring 2017) through the Hubbard Brook LTER data portal (http://www.hubbardbrook.org) and are free for public use with proper attribution. Richard Marinos will also provide a Github repository where participants can download all of the data to be used in the project before the start of the Data+ program.

IV. Project Goals

Primary Goal: *Data+ participants will create an interactive education and research web application that synthesizes the sixty years of ecological data collected at Hubbard Brook.* This tool will be composed of two modules. In the first module, the “Data Stories” module, participants will create data visualizations that will allow students to learn the classic ecological lessons from Hubbard Brook in an interactive an exploratory manner. The module will consist of these visualizations combined with explanatory text, videos, and curricular resources for teachers that Richard Marinos, Emily Bernhardt and Gene Likens will prepare. The second module, the “Data Explorer” module, will allow researchers to combine and synthesize the long term records at Hubbard Brook in novel ways, in order to develop novel hypotheses of ecosystem function. In particular, this module will focus on streamlining the incorporation of new data into the existing datasets for rapid analysis of incoming data. The data visualizations will be created using the *shiny* package in R, which creates web-ready data visualization mini-servers.

"Stretch" Goal: *Data+ participants will develop a toolkit that will allow users to load all LTER datasets into the Data Explorer Module and analyze them in arbitrary, user-specified manners.* All other LTER and LTREB sites also make their data publicly available, and we will seek to extend the data explorer module to be able to download these data sources, reconcile formatting differences in a computer-assisted manner, and compare datasets across LTER sites. In particular, one tool that we would like to develop is a formula parser that can generate arbitraty graphs. (E.g. typing "formula: Annual Tree Growth ~ Annual Precipitation at: FernowLTER from: 2000 to: 2010" would tell the software to generate a plot of annual tree growth as a function of precipitation for the specified site and date ranges.) This would allow more advanced users to rapidly and flexibly combine datasets in novel ways without having to know the intricacies of a statistical programming language.

Deliverables: The primary deliverable of the project will consist of the data visualization and teaching app, which will be hosted on Duke OIT web servers. In addition, a public Github repository of the data and source code for the project will be created, allowing others to adapt the software to their own uses. Furthermore, the Github repository will allow Data+ participants to show their individual contributions to the source code as well as to document their collaborative process.

Timeline:

By the end of...	*Data+ participants will...*
Week 1	Understand the scope of the project, learn the ecological methods and major findings of Hubbard Brook, and begin learning the software tools required for the project (base R, ggplot2, and shiny software packages, git)
Week 2	Continue learning the requisite tools for the project and begin curating the datasets for interoperability and synthesis.
Week 3	Finish curating the datasets and storyboard the data visualization tools we will develop.
Week 5	Create data visualization tools that show the twelve "canonical" lessons from Hubbard Brook that researchers have identified.
Week 6	Add ancillary curricular content developed in conjunction with the mentors to create a well-rounded teaching tool.
Week 7	Continue to develop curricular content, post a live beta version of the website, and solicit feedback from other researchers and educators.
Week 9	Develop the data explorer module, including tools that allow researchers to update the long-term records with new data.
Week 10	Improve the Data Stories module based on the feedback solicited in week 7, make sure all code is thoroughly documented, and publish the first stable version of the project.

V. Software Needs

Participants in the project will only require the latest version of R statistical software and a Git client for collaborating via Github. Personal computers should be sufficient for processing data. The data visualization website will be hosted on a Stevedore-based Linux VM from Duke OIT.

VI. Skills

The essential skills that all team members should possess include:

- Proficiency in a high-level programming language, preferably one used for statistical analysis and visualization (e.g. R, MATLAB, Python, SAS)
- An ability to understand and summarize scientific literature.

Desirable skills that team members may possess include:

- Fluency in R.
- Experience using Github for collaborative software development.
- A working knowledge of general chemistry concepts (Chem 101DL or equivalent level.)
- Previous coursework in either ecology or environmental chemistry.
- Basic Linux and virtual machine administration skills.
- Basic statistical literacy.
- Javascript coding experience.