**Summary**

In ecology and watershed sciences, large datasets often come from a variety of sources like continuous automated sensors, water grab samples, and community-collected scientific data. Overcoming these challenges is critical to explore the prevalence, persistence, and impact of degraded water quality on human society and wildlife. This project exposes students to approaches for merging and cleaning two disparate data sources, basic tools for statistical analyses, and data visualization.

The course consisted of two meetings – the first an informal walk-through of the [Data+ web portal product](https://bigdata.duke.edu/projects/a-city-and-its-river-contaminant-risk-in-durham/), to give context into the dataset we would be diving into, some examples of how to visualize data, and finally some time for students to explore their own hypotheses using the website (e.g., correlations between two water quality indicators of their choosing).

We subsequently hosted a formal session where we walked through an R tutorial covering:

1. Workflow
2. Why R?
3. Cleaning Data
4. Visualization and Analysis
5. Case Study: Exploring a Federal Dataset

**Guiding Questions**

* What are some ways to understand and begin to merge and manipulate disparate environmental datasets? How can we identify the limitations of a merged dataset, particularly when collected by different researchers?
* How do different measures of water quality vary across time in Ellerbe Creek and how does it vary across space? What are some ways we can visualize these changes? What are the limitations to these findings?

**The Datasets**

Two datasets are used for this analysis. The first (“**Duke Synoptic Sampling Data**”) was collected during 3 synoptic sampling events by the Duke Bass Connections team (2021-22) focused on Ellerbe Creek. 34 sites were sampled for ~20 different analytes ranging from major ions, heavy metals, nutrients, and physical characteristics.

The second dataset (“**Durham Ambient Sampling Data**”) is sourced from the City of Durham’s ambient water sampling program. The dataset has data for approximately 10 analytes collected on a monthly to biweekly basis over approximately 5 years from 3-6 sites.

**Course Materials**

*Includes instructions to load the curated dataset as well as visualizations/graphics/etc*

* Rpubs site: <https://rpubs.com/margaret-swift/ellerbe-creek>
* GitHub site: <https://github.com/margaret-swift/ellerbe-creek>