# Project Summary

The development of high throughput data generation tools used across the biomedical sciences has led to a situation where researchers with excellent bench skills struggle to appropriately and reproducibly analyze their data. With the increased size of the datasets, the complexity of the analyses has also grown. Although many institutions provide bioinformatic and statistical consulting services, the reality is that these services are overburdened and ultimately require the researcher who generated the data to also analyze the data. Researchers that once used paper notebooks to record data and spreadsheet-based tools to analyze their data now struggle to use command line tools. The ***long-term goal*** of this work is to enable bench scientists to analyze their biomedical data with robust, rigorous, and reproducible approaches. Traditional training programs have not been able to meet the needs of these researchers. Although very popular and well rated, workshops and bootcamps have proven ineffective at establishing lasting competency. *The lack of repeated reinforcement of the content over time is the most likely explanation for the poor outcomes of these workshops and resources.* Code Clubs have proven critical for providing this repeated reinforcement. Code Clubs are weekly activities that are analogous to a traditional Journal Club, but that focus on developing data analysis skills. The ***overall objective*** of this proposal is to develop a collection of virtual Code Club sessions that researchers can use on their own or with colleagues at their own institutions to improve reproducible data analysis skills. These sessions will cover concepts important for performing rigorous and reproducible data science, will be intentionally designed to develop local communities of practice, and will implement robust pedagogical approaches to teaching. These efforts are aligned with the overall goal of this RFA to create “exportable training modules designed to enhance the rigor, reproducibility, and responsible conduct of biomedical and behavioral data science research.” The ***central hypothesis*** is that completing Code Club sessions will improve the retention of concepts covered in prior workshops and allow learners to more quickly develop their skills expand beyond those covered in a workshop. This hypothesis based on 20 years of experience helping bench scientists learn to do their own data analysis and the excitement of colleagues who have run their own Code Clubs. The ***rationale*** for developing additional Code Club sessions is that by increasing the diversity and number of videos available, researchers will make quicker and deeper gains in their knowledge of reproducible research practices. This project will yield a ***significant vertical step*** in the field because it will put tools into the hands of researchers performing the analyses, empowering them to perform sophisticated and reproducible analyses. The approach taken in the proposed research is ***innovative*** because it represents the first concentrated effort to develop materials that use on repeated engagement of the same content in different contexts to help researchers develop data analysis skills.