The expansion of high-throughput laboratory techniques and availability of large public databases has made it clear that the ability to generate data has far outpaced most biomedical scientists’ ability to analyze those data. Although many institutions have core facilities that provide statistical and bioinformatic consulting services, these facilities are overrun with clients and are typically a cost sink for their institution. As a solution, researchers attempt to develop their data analysis skills through workshops and online tutorials. The deman and number of learning resources that have become available through organizations like The Carpentries, Data Camp, and Code Academy are a testament to their popularity. Unfortunately, empirical analysis of outcomes from workshops has shown that although learners universally love the format and content, they have minimal long term retention of the material. The lack of repeated reinforcement of the content over time is the most likely explanation for the poor outcomes of these workshops. At the same time, most online tutorials serve their learners poorly because they focus on how to implement individual concepts to answer abstract questions rather than integrating concepts to answer questions that are relevant. Furthermore, both workshops and online tutorials fail to help learners join a broader community that can help them develop their skills. There is a need to create a library of tutorials that present concepts in different contexts that are relevant that allow learners to assess their retention and confidence with employing the concepts. If these tutorials are designed with the intent of also building local and more distributed communities around data analysis (i.e. communities of practice), then it will be possible to significantly improve the retention of material covered in data analysis workshops. As a solution to these problems, my research group has developed the concept of a weekly **Code Club**, which is analogous to a traditional Journal Club, but focused around programming. These interactive sessions have successfully helped bench scientists develop skills in data analysis that are strong enough to go on to careers as data scientists at leading universities and pharmaceutical companies. During the COVID-19 pandemic I have posted weekly Code Club videos on YouTube to help develop the skills of similar bench scientists trying to develop their skills programming in R.

My ***long-term*** goal is to enable bench scientists to analyze biomedical data with robust, rigorous, and reproducible approaches. The ***overall objective*** of this proposal is to develop a video library of virtual Code Clubs. These videos will cover concepts important for performing rigorous and reproducible data science, will be intentionally designed to develop communities of practice, and use robust pedagogical approaches to teaching. This is 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, targeted to trainees and researchers at any career level.” The ***central hypothesis*** is that engaging in the Code Club materials 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. I arrived at 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 videos is that by increasing the diversity and number of videos they will have a broader impact. I am uniquely poised to execute the following ***specific aims*** to achieve my overall objective.

**Specific Aim 1. Produce Code Club videos that highlight concepts important for performing rigorous and reproducible data science.** I will develop an extensive collection of Code Club videos that cover a range of concepts related to rigor and reproducibility in data science. Each video will be composed of a tutorial and the host working through the 3-4 exercises related to the concepts that the viewers should have already worked through on their own. The tutorials and exercises will be motivated by research questions. To help build a community around the videos, we will solicit questions from viewers.

**Specific Aim 2. Assess the level of engagement with video tutorials and whether they improve mastery of computational skills.** I ***hypothesize*** that researchers who take a workshop and engage with Code Clubs will have longer retention of the concepts covered in the workshop than researchers who only take the workshop. I further expect that researchers who participate in the Code Club tutorials will more quickly pursue concepts beyond the scope of the workshop. I will leverage a network of programming workshops to track retention of content with and without the supplemental tutorials.

Successful completion of these aims will significantly enhance the availability and efficacy of materials for improving reproducible data science. In addition to being available through the NIH Clearing House, the tutorials developed through this project will be added to the Riffomonas Project website and YouTube channel. The Riffomonas Project was initiated in 2015 to develop a workshop-style instructional series to help microbiome researchers develop skills that foster reproducibility. The material developed through this proposal will complement the earlier content and help to broaden its audience.