Abstract: The R programming environment is used by researchers and students alike to perform all manner of statistical analysis. This talk will describe a project at Amherst College that is using R to simulate and analyze protein structure perturbation, an area of study that has wide ranging implications for many branches of medical research. While writing the code for these simulations in R is relatively straight forward, running the code on a single machine can be prohibitively time consuming. In this situation, the typical approach is to rewrite the software so that it can be deployed on a Hadoop- or MPI-based cluster.

With this project, we took a different approach and used Spark, one of the newer distributed computational platforms. The primary advantage of Spark in this case is its tight integration with the R execution environment, meaning the R code did not need to be substantially changed in order to run across a cluster of worker machines. Using the protein simulation project as a case study, this talk will explore how Spark allows data scientists to make use of existing R-based code both to analyze very large datasets and run highly parallelized computations.