During the fall semester of 2007, Eppstein (Computer Science) designed and taught the course: CS 195 Data Analysis, Modeling, and Visualization. The intent of the course was to introduce sophomore life science majors to fundamental quantitative and computational concepts to life science majors, using the MATLAB programming language, toolboxes, and computing environment in the context of relevant problems from biological domains, in way that was fun, un-intimidating, and highly integrative. Unfortunately, since this was not a required course for any of these students, we were only able to attract four students into the course, despite advertising and having several biology advisors encouraging their advisees to enroll; however, since this was a pilot offering we elected to run the course anyway to see how well it worked, and to develop the modules for use in future offerings of the course and/or to incorporate into other courses.

Since biology majors were allowed to substitute this for their required statistics course, about one half the semester was spend on statistical concepts. However, unlike a traditional statistics course, all concepts were accompanied by a lot of visual feedback and exploratory data analysis (using Matlab’s excellent visualization capabilities), and students did hands-on exercises that helped them to understand which types of statistical tests to apply, how to apply them in Matlab, what assumptions lie behind them, how to test those assumptions, and what happens when those assumptions are violated. Students learned fundamental concepts of data manipulation and visualization; probability through Bayes’ Rule; various aspects of distributions (normal, exponential, uniform, binomial), including how to display theoretical and actual data distributions, how to assess experimental distributions using $\chi^2$ and Kolmogorov-Smirnov distribution tests (assumptions, trade-offs, sensitivity to sample size), and how to generate random numbers from various distributions for use in simulations. They also learned how to do standard hypothesis testing (using paired and unpaired t-tests with equal/unequal variance, ANOVA), and learned how to use linear and polynomial regression (including understanding and testing for the assumptions of correlation, skedasticity, and independence), and how to use logarithmic transforms to estimate nonlinear equations including exponential, logarithmic, and power-law relationships. In-class computer exercises and homework assignments used real data sets and problems to help motivate the students and make the work feel relevant. Example applications included analyzing global warming data, mouse vs. human DNA segments, investigating the species-area relationship, and basal metabolic rates of species vs. mass. Students were also required to find, read, and discuss papers in the current literature that used some of the statistical techniques they were applying.

During the second half of the semester, students learned fundamental computer programming and modeling skills. Programming concepts (e.g., branches, loops, modularity) were introduced and taught by exploring classical models in mathematical biology, including discrete Markov models such as the Leslie model, continuous models using ordinary differential equations such as logistic growth, predator-prey, and spatially-
explicit discrete models using cellular automata; e.g., Wolfram’s 1D cellular automata and the 2D Game of Life. In some cases, they were able to explore the relative advantages and disadvantages of using different modeling approaches for the same problem. For example, they implemented the SIR (susceptible-infectious-recovered) model of disease spread using both (i) a mean-field differential equation based approach and (ii) a 2D spatially-explicit cellular automata. Students were required to develop some programs from scratch, but for some more difficult problems they were given “skeleton” code and were required to fill in key details of the code, enabling them to focus on the most important concepts without getting bogged down in details. For their final presentations in the course, each student gave a 20-minute oral presentation (using PowerPoint) discussing 3 papers they had individually selected from current biological literature: one using a Leslie model, one using ordinary differential equations, and one using cellular automata; for each paper, they introduced the rest of the class to how computer modeling and statistical analyses were used and what the main take-home message was. In addition to in-class exercises, homeworks, and oral presentations, students were assessed with open-book/open-note mid-term and final exams, each of which required significant thinking, integration, and application of concepts to new situations.

At the start of the semester, students were given a “pretest” to assess what concepts they already know. All of the students reported little or no prior statistics or programming experience. On a set of basic conceptual questions, they scored an average of 10% at the start of the class; at the end of the semester they averaged 87% on a slightly different version of the pretest, covering the same concepts. All students received an A or B in the course. When asked how they perceived the experience of learning Matlab, responses were: “It was very frustrating at first and then it was rewarding when I actually knew what I was doing. It is definitely helpful to use when dealing with biological models. I would use this in my bio lab and in my career.” “It was fairly hard at first, but became easy as time progressed. I feel that is rewarding, because you create things that are actually useful. I am very interested in computational biology [after this course]. I especially liked cellular automata. I am definitely going to use what I learned in future biology classes.” “Very useful, I’ve completely given up using Excel. I would like to take a Matlab/Field Data class. I will very likely use my knowledge in other courses when I can.” “Matlab was very rewarding when applying to my other studies. I can see myself applying this knowledge to something in my future career.” All students reported that they were glad they took this course instead of their required statistics course.

Despite the disappointing enrollment, some materials developed for this course are already being used in some other courses, the software purchased for the course is continuing to benefit students across campus, and there is the potential that this course may become an alternate required course for Biology majors in the future.