Big Data, Big Discoveries, Big Fallacies
Freshman Seminar, 2015-16

A sea change has occurred in science, technology, medicine, politics, and in society as a whole: many of world's biggest discoveries and decisions are now being made on the basis of analyzing massive data sets, referred to as "big data". Early examples were largely in retail: An unexpected correlation was discovered between beer and diaper purchases (or so the lore goes), Victoria's Secret vastly improved efficiency by analyzing buying patterns and shipping the right lingerie to the right location at the right time, and a surprising technology leader in the area was low-end retailer Walmart.

Since those early days, domains ranging from advertising to sports to medical diagnosis have embraced sophisticated techniques for processing and summarizing vast amounts of collected data, and for putting data-driven findings to use. Everyday examples students are familiar with include social-network friend recommendations, and weather predictions far more accurate than a decade ago; both of these applications use vast collections of data to model the past and predict the future. On the other hand, it is surprisingly easy to come to false conclusions from data analysis alone. For example, we might conclude that acne medicine prevents heart attacks and strokes, if we forget to factor in the age of the patients. It turns out the divorce rate in Maine closely tracks the per-capita consumption of margarine, but do we believe there's a connection? Privacy can be a major issue: Target stores analyzed buying patterns to predict with remarkable accuracy which of their shoppers had just become pregnant, but trouble arose when they sent baby ads to the homes of pregnant teens whose parents weren't yet in the know. Recent revelations of data collected by the National Security Agency, and the variety of facts that can be inferred from that data, have undermined many citizens' trust of internet providers and the government.

We will start by surveying the history of data-driven activities, leading up to the recent Big Data explosion. Students will be responsible for finding examples of Big Discoveries and Big Fallacies in areas that interest them, and sharing them with the class. A variety of data analysis techniques will be covered, leading students to appreciate that even simple techniques can go a long way when the data set is large enough. Common stumbling blocks leading to false conclusions will be discussed, and students will be asked to debate the many issues surrounding privacy. Students will complete two significant projects: In the first project, students will compete to see whose analysis techniques can predict user movie ratings most accurately based on past rating behavior. The second project will be individually designed, in an area of the student's choosing.

The seminar will include a mix of assigned readings, small-scale investigations and assignments, classroom discussions, and the two projects. No computer programming or special math skills are required; students will learn the basic techniques and tools they need to complete the data analysis assignments and projects.