Week 1 Readings [Rich Gruss]

The readings for this week attempt to provide a concrete definition for a term that has burst from obscurity in the past two years: data science. Zumel defines data science as "managing the process that can transform hypotheses and data into actionable predictions." Although this does little to differentiate it from the tasks that business analysts have been performing for decades, it is a decent start. Five important points about data science emerge from the readings: 1) We're in the midst of an exuberance about "data scientists" similar to the exuberance about web developers in the late 1990s. Demand is extremely high, and the pay reflects that. 2) Modern Ph.D. research, because much of it relies on data analysis, is preparing people nicely for the role, and since academic jobs are increasingly scarce, more and more astrophysicists, biostatisticians, and even cognitive psychologists are becoming data scientists. 3) Data science derives it theoretical basis from statistics, computer science, and operations research. Computer science, my personal interest, contributes database theory, distributed programming, machine learning, and artificial intelligence. 4) Data science has commercial goals, because making better predictions than your competitor is the key to staying in business. Examples of successful data science project include Amazon's product recommendation system, Twitter's trending topics, and Wal-Mart's consumer demand projections. 5) The prevalence of data science is due to the availability of enormous quantities of data. As storage media got cheaper, it became cost-effective to store everything, even when its strategic value was not yet clear. My employer—the Virginia Tech Math Emporium—has 14 years worth of student online testing clicks, but nobody has a clue how to extract any insight from it.

I'm intrigued by several questions, but to select three: 1) Some academic disciplines have already put their stamp on the practice of data science—biostatistics, astrophysics, and psychology were mentioned in the readings—but what other fields have specialized algorithms that can contribute? I know that physical anthropology has special similarity metrics for 3D images that could be useful, and I can imagine that sports betting firms have some impressive ranking algorithms. 2) As companies grow increasingly capable of anticipating our every desire, will our lives get better, or will the web become unusable? 3) How does IBM's Watson work? Does this proprietary knowledge extend way beyond what academia knows (like Google vs. Information Retrieval)?