Pre-read: privacy and ethics (part 2)

The ethics of predictive decision making comprise many issues including fairness, justice, power, and accountability. Today we'll primarily focus on fairness to keep the scope manageable. For a deeper look at fairness in machine learning, see COS 534: Fairness in Machine Learning (offered next semester) and/or the resources at fairmlbook.org.

The first reading is a quick primer that describes a few reasons why we should expect machine learning to be unfair, unless explicit corrective measures are taken, when it is used to make decisions about people. The fact that potential unfairness is ever present, and that there is no general method for avoiding it, suggests a normative limit to predictive decision making which we will revisit in some of the following readings.

Next, we'll turn to a recent high-profile investigation of bias in a predictive algorithm used in healthcare. The problem, in short, is that the system used healthcare costs as a proxy for needs. This relates to a recurring theme, the difficulty of measuring the target variable. Think about which, if any, of the factors in Hardt's blog post are relevant here. Along with this article, we'll read a perspective by Ruha Benjamin (Associate Professor of African American studies at Princeton) who argues that the seeming neutrality of automated systems both hides and entrenches systemic racism.

The next pair of readings are about the so called fairness impossibility theorems. It turns out that there are many fairness criteria that all intuitively seem desirable, but aren't simultaneously satisfiable (under mild assumptions that almost always hold true in practice). How do you interpret this impossibility? Do you believe that are incompatible fairness criteria or are some of these criteria not actually important? What are the implications for predictive decision making?

In the final reading, we will return to a theme we have touched on repeatedly: prediction versus intervention. The conflict between the two perspectives is nowhere more sharp than in criminal justice. As you read this paper, make note of the ways in which optimal predictions may not translate to optimal decisions. A striking point the authors make is that the rise of predictive tools warped the values and goals of the criminal justice system itself. How and why did this happen?