<Name>

Discrete Mathematics

Dr. Lund

Thoughts on the “Rise of Big Data”

The 1890 census was a very important census in American history because it was the first census that was aided by a mechanical computing machine. A tabulator and key punch machine that was invented by Herman Hollerith “reduced a ten-year job to three months.” This machine ushered in a new age of computing large data sets. This is not to say that things changed overnight. It would be over a hundred years before the phrase “big data” would be coined. However, it’s worth noting the importance of the census and its use as a statistical tool when discussing the concept of big data as a phenomenon.

Culkier and Mayer-Schoenberger discuss big data in terms of the size of the data sets involved. Until the rise of big data (and the ubiquitous digital technology that allows us to harness it), there were really only two widespread datasets that could be said to approach “N=ALL.” Those two things were the ten year census cycle and political elections. Sure exit polls are sampled, but elections are not. Every vote is counted. The result of an election is the set of all voters who cast a ballot. The census is a collection of data on every household that fills out the census card and this is nearly all households in America.

Through the years, the census has been one of our only truly large data sets, but no matter what, there was really only so much we could do with those data. In its earliest days, it was really little more than a headcount. It certainly wasn’t until the advent of the digital age and massive parallel computing that those data could ever be analyzed in the ways that they are now. One might imagine Hollerith shaking his head in wonder at the things that we are able to do with the large datasets that his company (which would become IBM) paved the way for.

There’s an old quote popularized by Mark Twain, concerning “There are three kinds of lies: lies, damned lies, and statistics.” The misconceptions surrounding statistics lay solely at the feet of those who misuse statistics for dishonest ends, knowingly or not. Statistics are sometimes the only thing that matters. Consider quantum mechanics, arguably the fundamental stuff of reality, where everything is probabilistic. In a world where the certainty of a causal relationship is sometimes not possible to extract from reality, correlation is sometimes our only way to get things done. But more importantly, the simple truth of the matter is that sometimes, the statistical correlation is really all you need. It wouldn’t do you any more good to find the true causal relationship. You would gain no further insight. Culkier and Mayer mention the UPS method of maintaining their vehicles which involves a purely statistical correlation between certain temperatures and vibrations of certain parts in the engine that would lead to engine failures. They don’t know or even care what actually causes the failure. All that matters is that when certain conditions in the engine exist, parts fail. A simple correlation is all that is needed to save significant amounts of money and time.

There are valid concerns about Big Data. There are many reasons to mistrust very large organizations, be they corporations or governmental bodies having access to massive massive data sets, particularly when it has been shown time and again that these correlations can and do prove things about specific instances and individuals. Orwell was one of the first to warn about big data—and its sinister cousin, big surveillance—and he had no idea just how massive the collection of data would grow in just a few decades. And not only has the size of that collection grown, but our ability to make use of it has developed at an incredible pace. Bulk collection of data by government agencies is very much in the public eye right now. Or at least it should be. And it will require all of our efforts to ensure that such a powerful tool does not become the plaything of tyrants who don’t understand it. Or worse, tyrants who do.

The closing words of Culkier and Schoenberger ring true. They discuss the need for always being aware of the need for human innovation and inspiration. Data is dumb. It’s enticing and useful and powerful, but while a hammer can drive a nail, if one doesn’t understand its function, it’s nothing more than a heavy stick, and in the wrong hands, it’s a weapon.