**Parables of the Big Data Age (500 Word Abstract)**

We live in a world in which individuals produce and record more data about their behavior than in any time before. Every day, humans now created 2.5 quintillion bytes of data, and it is estimated that 90% of the world’s data created throughout history has been generated in the past two years. This has led some scholars to predict that big data will overturn our entire way of thinking about statistics. While big data does hold enormous promise, it has also raised new perils in its analysis. Issues of overfitting, sample bias, and even sabotage loom large in the big data arena.

This research looks at the promise of big data and its potential problems to produce more educated consumers of big data. Three main problems can arise in the analysis of big data. The first is “big data hubris,” where the size of data is taken to be an adequate substitute for smaller, carefully collected data. Examples from disease detection, automated news reading programs, and city planning demonstrate that big data can be misleading when it is not paired with carefully collected “small” data. In addition, these examples show that assertions by scholars that causality is no longer important – that “correlation is enough” or “the data can speak for itself” – are misguided and potentially dangerous for analysts.

The second is what we call “blue team dynamics.” Most big data sources are from what scholars have labeled “data exhaust” – the byproduct of systems whose essential operation is not about producing data. [Note for translation: “exhaust” in this case means the waste gasses or air expelled from an engine when it is used.] For example, the Google search engine is not designed to collect information about people’s behavior, it is meant to drive advertising so Google can make money. This means that the people who set up these systems are constantly changing them to meet goals that are, at best, only loosely related to the goals of the researcher. We have little understanding of what these changes do to the data generating process for big data. This can lead to large changes in the results from data tracking systems and misleading analysis.

The final problem we call “red team dynamics.” As we become better at tracking big data systems, incentives arise for people to manipulate the signals from these systems. For example, news reports on how many Twitter followers a political candidate has, as well as Twitter’s own recommendation algorithm, has led to candidates either creating fake accounts to follow them or purchasing so-called “bot” accounts to re-tweet their posts. Similar incentives arise whenever there is a material interest in the results of big data analysis. While a lot of work is going into detecting “bots” and adjusting results accordingly, being unaware of potential incentives for manipulation has led to problems for analysis of a range of systems, from crowd sourcing to urban planning.

In sum, while we think that big data is an invaluable source of information, researchers and consumers must understand the potential problems and take steps to remedy them. Big data is not a substitute for careful and well-reasoned research.