Your on the Street Reporter

Uyless Black

Big Data

Big Data

February 9, 2014

Hello from Your on the Street Reporter, continuing the series on Internet Privacy and security, and NSA/private industry surveillance. The subject for this report is *Big Data* (BD). The term is not accurate. If data were big, the word *big* would be bigger as shown here:

Big

The more accurate term is a *Lot of Data* (LD). But I will stay with *Big Data* because it is the popular term. The term and what it conveys is the product of converting many of the written images in the world (from hard copy) to digital (electronic) images. The product is impressive. One estimate claims "less than two percent of all stored information is nondigital."[1]

Through the use of massive data bases containing digital information, powerful computers, and sophisticated software, Big Data systems can *infer* much from this data, an idea introduced in in the previous report in this series. This capability is where the NSA's collection effort of a Lot of Data comes into play. In the past, a small sample of data was analyzed (hammered, if you will) by applying statistical analysis to the samples in order to gain inferences. Big Data does not bother with the tedious task of sampling small amounts of data. Big Data hammers massive amounts of data to obtain its inferences.

Big Data uses big hammers: hundreds of high-speed computers, massive amounts of data containing millions of emails and other files, billions upon billions of words and numbers. The speed of electronic computation can do wonders.

In the past, the people who designed the sampling systems for these methods had to know beforehand what data to collect. For example, while employed at the Federal Reserve, I wrote software that emulated America's money supply. The Fed economists selected specific pieces of data (checking accounts, CDs, passbook accounts, savings, etc.) for the software to hammer. The results were used by the Federal Open Market Committee (FOMC) to set goals for regulating money levels and interest rates.

If we had been using Big Data techniques, the economists would not have spent (arduous) hours selecting the sample. They would have continued to use statistical analysis software and determine the data to be used. But they would not have worried so much about the correctness of all the data. Big Data techniques allow messier data than do sampling techniques.

With Big Data, it is not necessary to know ahead of time how the data will be used. Just collect <u>all</u> of it, and decide how it is to be used later.[2]

Big Data is NSA's approach to some of its snooping. Is this practice troublesome to you? That is a question each citizen must answer. These essays have made this point several times: We

[1] Kenneth Cukier and Victor Mayer-Schoenberger, "The Rise of Big Data," *Foreign Affairs*, May/June 2013, 28-29.
[2] Or as much data that is legally and/or technically possible to collect.

citizens should not give up our constitutional right to privacy just because our letters are digitized into 0s and 1s instead of ink-script. And the power of Big Data systems and the ease of using BD software on massive databases to (increasingly) reveal our private lives should give us pause.

During the Bush II administration, former Vice-President Dick Cheney and his chief of staff David Addington played major roles in the implementation of warrantless surveillance. The post 9/11 time in American history was one of fear, especially among the political leaders in Washington, D.C. With admirable intentions, Bush---with Cheney pulling substantial strings---approved measures that:

> Allowed the government to obtain "roving wiretaps" targeting individuals regardless of which telephone they used. It empowered federal agents to obtain orders allowing them to seize "any tangible things" related to a security investigation, including business customer records…the language [of the measure] was expansive enough that the government would have vast new abilities to peer into the lives of citizens it suspected of extremist ties.[3]

These measures also fostered the collection efforts at NSA. It appears they contributed to NSA's Big Data program. They served to take advantage of the digital world to exploit Big Data capabilities in the following way:

Many of the Big Data concepts make a startling change in how humans have viewed "data," such as information stored in books and letters. Humans focus on trying to understand *why and how* something has occurred. For example, many famous and influential people read about wars. They digest varied opinions of why and how events of the past led to a war or sets of wars. In so-knowing, they (as do their less-powerful fellow citizens) hope to come up with these whys and hows in order to avoid wars in the future.

Big Data does not look for causation. Big Data looks for relationships of the data pieces to establish correlations of events that can be gleaned from the data. For this example, *Big Data's focus is not on the subjective reasons of why and how wars come about. Its focus is on understanding the correlation of certain events that make a war more likely to occur*.

Democracies such as the United States have built-in checks on invasion of the government into citizens' private lives. But in nondemocratic countries, Big Data makes the state more powerful and intrusive. Even in America, the well-intentioned zeal of powerful people, such as Cheney and Addington, can lead to semblances of Big Brother's intrusion into our rights. (I do not doubt the intentions of such people. I question their ways of going about implementing their intentions.)

The questions we citizens must put to ourselves and to our leaders: Do we allow Big Data to foster Big Brother? More modestly, do we draw a line in the sand, and say, "No more"? More realistically, do we ask, what are the trade-offs between civil liberty and physical security?

[3] Peter Baker, *Days of Fire* (New York: Double Day, 2013), 171.

Let's not be passive about this issue. Internet privacy is a subject that is vital to our republic and to each of us personally. It reaches into an aspect of what we humans have long treasured: the age-old assumption that we can be ourselves---but unto ourselves, warts and all, privately and unobtrusively.

PS

Recent press releases state the NSA program is collecting only between twenty to thirty percent of the calling data from Americans.[4] It is reasonable to ask if this semi-Big Data database is sufficient to establish calling patterns pointing to potential threats. If not, what is the point of capturing the data in the first place? I am now removed from knowing about either Big Data or Little Data technologies. But I do remember that semi-random samples could yield questionable results. I think you and I can agree that we do not wish our files, computers, hard disks, letters, and Valentine Day cards seized because the surveillance systems had a *hunch* we were making unpatriotic phone calls.

[4] Stephen Braun, "Report: NSA Gets under 30 percent of Phone Data," (CDA Press, February 9, 2014), 12A.