Does Mining Our Big Data for Terrorists Actually Make Us Any Safer?

By: Shawn Musgrave – June 8, 2013

Whether it's at the NSA, FBI, CIA or some more classified body we mere citizens aren't mature enough to know about, data-mining is the belle of the intelligence ball. The power of statistical prediction to connect the dots, preemptively identify the bad guys and thwart the next terrorist attack has been trumpeted loudly in defense of surveillance programs, including the NSA's latest misadventure.

But many counterterrorism and statistical experts doubt that even the most advanced spot-a-terrorist algorithms can produce anything beyond false positives and mangled civil liberties.

In his address Friday afternoon, President Obama downplayed the recent revelations about NSA surveillance, dismissing much of the ensuing scrutiny as “hype.” He said that the NSA's extensive collection of phone call metadata from Verizon, Sprint and AT&T, as well as its PRISM program to vacuum up server data from Google, Facebook, Microsoft and other Internet service providers (Dropbox coming soon!) were both legal and appropriately supervised. These programs “help us prevent terrorist attacks,” he said, and “on net it was worth us doing.” Senator Diane Feinstein, standing next to Saxby Chambliss, her Republican counterpart on the Senate Intelligence Committee, explained to the citizenry, “It's called protecting America.”

As construction workers put the finishing touches on the NSA's new data facility in Utah—it is said that it will be the largest data center in the world—details continue to emerge that flesh out the exact shape and scope of NSA's various dragnets. As groups like the Electronic Frontier Foundation have been warning for years, it's clear that the agency is pouring considerable resources into collecting and parsing through vast datasets in hopes of neutralizing terrorist threats. But, as has been asked of the TSA and DHS more widely, where's the actual proof these programs offer more benefits than downsides? Where are the thwarted plots to balance against the chill of privacy loss and the threats to, say, activists and the government's political opponents?

Among national security experts and data scientists, there's considerable skepticism that NSA-style data-mining is an appropriate tool for ferreting out security threats. As Ben Smith reported yesterday, finding the Boston bombers relied on old fashioned police work, not troves of data. In a 2008 study, the National Research Council concluded that combing data streams for terrorists is “neither feasible as an objective nor desirable as a goal.” In particular, the report's authors underscore dubious data quality and high risk of false positives as practical obstacles to mining data for signatures of terrorist behavior.

“There's been considerable interest in the intelligence community around using data to identify terrorists,” says Stephen Fienberg, a professor of statistics and social sciences at Carnegie Mellon University, who contributed to the NRC report. “But the specifics have always been elusive, and the claims are rarely backed up by serious empirical study.”

Fienberg insists that the rarity of terrorist events (and terrorists themselves) makes predicting their occurrence a fraught crapshoot. He says that intelligence analysts lack training data – indicative patterns of behavior drawn from observing multiple iterations of a complex event – to verify whether their models have predictive validity.

“These are very, very rare events – terrorist events and terrorists themselves – that you're trying to predict. Clearly there are places where this kind of predictive activity has been very successful – fraud detection in telecommunications, for example – but there we're talking not-so-rare events.”

Jeff Jonas, a data scientist at IBM and senior associate at the Center for Strategic and International Studies, agrees, dismissing such terrorism prediction models as “civil liberty infringement engines." In a 2006 paper co-written by Jim Harper of the Cato Institute, Jonas asserts that sheer probability and a lack of historical data dooms counterterrorism data-mining to a quagmire of false positives.

“Unless investigators can winnow their investigations down to data sets already known to reflect a high incidence of actual terrorist information,” Jonas and Harper write, “the high number of false positives will render any results essentially useless.”

Ethical (not to mention constitutional) issues of wrongly painting people as terrorists aside, Jonas and Harper suggest that chasing down so many bogus leads only detracts from pursuing genuine ones, and thus actually hampers effective counterterrorism.

In a 2006 interview with the *New York Times*, an FBI official confirmed the considerable waste and frustration of running down bogus tip-offs from the NSA's wiretap dragnet, joking that the endless stream of leads meant more "calls to Pizza Hut” or contacting a “school teacher with no indication they've ever been involved in international terrorism - case closed."

Given enough data and fine-tuning of algorithms, of course, other experts emphasize that false positives can be reduced significantly, and insist that data-mining will play a key role in counterterrorism. Kim Taipale of the Center for Advanced Studies in Science and Technology Policy testified to this effect before the Senate Judiciary Committee in 2007, criticizing Jonas and Harper specifically for making “pseudo-technical” arguments that fail to reflect the way actual data-mining algorithms work.

And even critics admit that, with enough data to develop these training sets, analysts might be able to sift out terrorist markers.

“If you can get your arms around a big enough set of data, you'll always find something in there,” says Fred Cate, director of the Center for Applied Cybersecurity Research at Indiana University Law School, another contributor to the NRC report. “It's not unreasonable to think that the more data you can get access to that you might discover something of predictive value.”

The ease of mining personal data may make these systems ripe for abuse, but that ease also lends itself to a “better safe than sorry” mindset. “There's a certain 'because it's there' nature to this,” says Cate. “If you know all these records are there, you worry about explaining why you didn't try to get access to them” to stop a terror plot. As more and more revealing information finds its way online and into commercial databases, the temptation increases for intelligence agencies to gobble up this data just in case.

But the wider the net we cast—and the broader incursion on the privacy of Americans and others—the heavier the burden becomes to produce a terrorist or two. And to Cate's knowledge, despite extensive mining, the NSA has struck no such motherlode. While the government has acknowledged that these latest data surveillance programs are several years old, they have yet to trot out any concrete evidence of their efficacy.

Between the NSA's dismal record, drowsy oversight from the top-secret FISA courts and vague promises from Obama, Feinstein and others that this will all be worth it someday, Washington should buckle up for plenty more “hype” from the civil libertarian set. Absent public exposure, independent oversight, and robust evaluation, it's impossible to determine whether such efforts truly have anything to throw on the scale against citizen privacy.