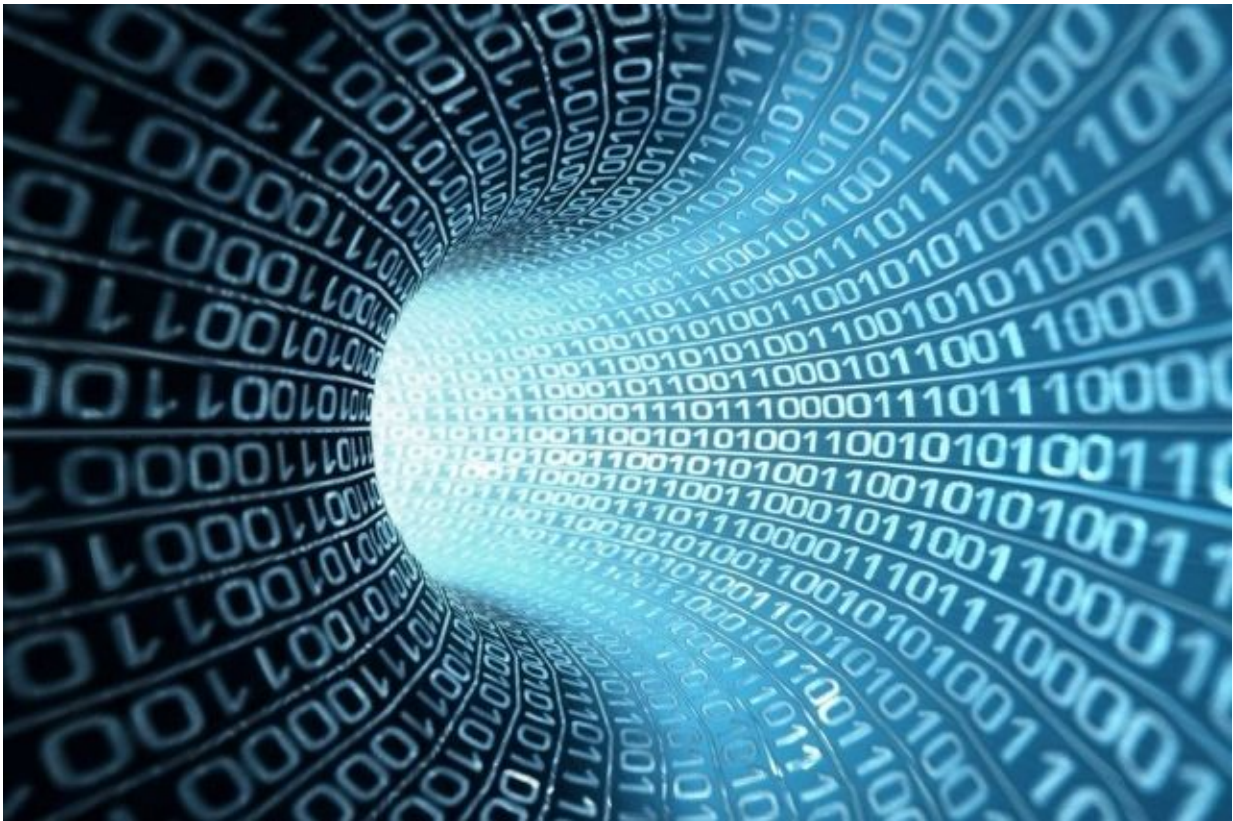# From calorie postings to health ratings, big data does social good

June 17 2015, by Phillip Leslie

Credit: luckey_sun/Flickr

Big data gets a bad rap.

While stories show up practically every day about the novel and

sometimes surprising ways Internet companies can use the massive amounts of data they collect from us (recently, for instance, Belgium criticized the way Facebook shares users' likes with its advertisers), the ability to collect and analyze large amounts of social science data has the capacity to do much social good.

I know this because I'm an economist who has been working with extremely large datasets from a time before we even called this "big data." What first convinced me of how valuable big data could be came in the late 1990s—when we just called it "empirical research."

I was working on a study of what happened after Los Angeles County's Health Department began placing letter-grade signs in the windows of restaurants to rate hygienic food preparation conditions. We had data for some five years on close to 30,000 restaurants, which were all inspected about two to three times per year.

That was hundreds of thousands of data points, which felt like a lot of data to be working with back then. Less than 20 years ago, it was common to see published papers where the number of observations was under 1,000 and even fewer than a couple hundred. That would seem quaint, if not scandalously incomplete, by today's standards.

As economists know, one of the main challenges in working with data has not as much to do with size of the data as the kinds of variation that are in the data—the stock market may seem to jump with news of one war, for example, and may seem to drop with the news of another, if we don't pay attention to what else is going on in the background.

Real life has so many varied factors at play that it can be hard to distinguish whether two events just happened at the same time by coincidence, or whether there's a true cause-effect relationship between them. Economists wish we could run randomized trials like drug

companies, testing only one variable at a time, but instead we try to uncover natural experiments in the available data.

The reason I was drawn to studying the restaurant cards was that there was a clear separation between the time before restaurant grade cards and the time afterward.

It took the health department only a few months between a hidden-camera newscast exposé of disgusting conditions in some restaurants' kitchens and the county launch of the grade-card program in January 1998. This meant the grade cards were unanticipated, allowing us to say that changes in restaurant and customer behaviors were caused by the grade cards.

We knew we were trying to measure subtle, small effects. And anyone who works with data will tell you that data tends to be noisy, which means you try to follow a lot of leads that go nowhere. We first started to look at how revenues changed according to whether a restaurant got an A, B, or C, and then we kept pushing the data more and more.

The biggest surprise for us in this research was the ability to connect how grade cards changed health outcomes. We realized we could look at the number of people hospitalized with food-borne illnesses in Los Angeles County hospitals before and after the grade cards; and the signal turned out to be clear and strong: a 20 percent decrease in hospitalizations for illnesses associated with poor food safety, such as staphylococcal food poisoning, after the grade cards appeared. It was astonishing that we could make the connection.

Of course, studies like this raise many questions. Life isn't always easily packed into a standardized grid. What if a restaurant got a particularly strict instructor rather than a more lenient one for their inspection? What if the restaurant was having a particularly bad day when the inspector

showed up? Should they have to display that grade card for the next few months even if it's not representative?

These questions went beyond the scope of our study, but they're worth asking. We have to recognize that studies like this show the trade-offs we are making. And to me, they're OK because we can clearly see that there was a reduction in illness. In the years after this study, many other jurisdictions have adopted similar grade cards.

Big data can also help with important health problems whose influences are hard to track. Obesity is one of those. In 2008 and 2009, I worked with colleagues on a project studying the effect of a new New York City law that mandated the display of [calorie counts](#) on menus.

Starbucks, home of decadent caramel lattes and "skinny" lattes, agreed to share transaction-level data on purchase behavior with us. We didn't have any names attached to the data—and I'm not sure if Starbucks itself knew. We compared the data we got from New York City to Philadelphia and Boston, where there was no calorie posting.

We found that consumers at Starbucks reduced calories by 6 percent when the calories were posted. While this doesn't sound like a huge amount—typically choosing a drink that was 232 calories rather than one that was 247 calories—we could tell the reduction was real and not some fluke variation.

But the study also revealed to us that New Yorkers weren't all identical automatons. When we indexed our results by zip code, we found that the zip codes with wealthier, more educated, and less obese people tended to cut their calories to a larger degree than those from the poorer, less educated, more overweight zip codes. We were disappointed, of course, that those who could benefit most from losing weight weren't very responsive to the campaign.

Of course, just because there are millions of data points doesn't mean the interpretation of these data points isn't slippery. Take the different analyses of the 10-year-long "Moving to Opportunity" study, which observed what happened when 4,600 families from low-income parts of Baltimore, Boston, Chicago, Los Angeles and New York City were given the chance to move into more well-to-do neighborhoods.

Early analysis of the data from the late 1990s suggested that moving to a new location didn't matter all that much in terms of how likely a kid was to graduate from college or how much he or she earned as an adult. But recently, economist Raj Chetty at Harvard came up with a new way to slice the data: If you looked at how young the children were when they moved to better neighborhoods, you could see that every extra year of childhood spent in a better neighborhood mattered.

Still, no matter a study's merits, there's always the privacy issue to consider. I recognize that people are very sensitive to companies using data they consider personal. It was a little disconcerting when Target started sending ads to young women that their data-crunching algorithms found were likely to be pregnant.

Do companies and governments screw that up? Of course. But it's a mistake to presume that the intent inside these organizations is to misuse, or abuse, that information. In my experience working with businesses, they are very sensitive about how they can keep their customers' trust. And there are valuable uses of the data companies collect—especially when tracking customer behavioral shifts in response to outside factors like new laws or regulations, such as in the case of New Yorkers suddenly confronting posted calorie counts.

I, for one, am quite happy that an online retailer may know I'm a cycling enthusiast and lets me know when helmets or other gear are on sale. I am also willingly sharing data through a smartphone app with a website for

cyclists called Strava. The website tells you what rides you have been on and compares your rides to others'.

And the benefits to me are clear: I'm connecting with friends through the site because it enables me to discover new places to ride and compares my performance to others. And it helps me stay in shape: I get on the bike a lot more than I would have otherwise.

Provided by University of California, Los Angeles