EVERYBODY LIES

BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE

SETH STEPHENS-DAVIDOWITZ

FOREWORD BY STEVEN PINKER

3

DATA REIMAGINED

At 6 A.M. on a particular Friday of every month, the streets of most of Manhattan will be largely desolate. The stores lining these streets will be closed, their façades covered by steel security gates, the apartments above dark and silent.

The floors of Goldman Sachs, the global investment banking institution in lower Manhattan, on the other hand, will be brightly lit, its elevators taking thousands of workers to their desks. By 7 A.M. most of these desks will be occupied.

It would not be unfair on any other day to describe this hour in this part of town as sleepy. On this Friday morning, however, there will be a buzz of energy and excitement. On this day, information that will massively impact the stock market is set to arrive.

Minutes after its release, this information will be reported by news sites. Seconds after its release, this information will be discussed, debated, and dissected, loudly, at Goldman and hundreds of other financial firms. But much of the real action in finance these days happens in milliseconds. Goldman and other financial firms paid tens of millions of dollars to get access to fiber-optic cables that reduced the time information travels from Chicago to New Jersey by just four milliseconds (from 17 to 13). Financial firms have algorithms in place to read the information and trade based on it—all in a matter of milliseconds. After this crucial information is released, the market will move in less time than it takes you to blink your eye.

So what is this crucial data that is so valuable to Goldman and numerous other financial institutions?

The monthly unemployment rate.

The rate, however—which has such a profound impact on the stock market that financial institutions have done whatever it takes to maximize the speed with which they receive, analyze, and act upon it—is from a phone survey that the Bureau of Labor Statistics conducts and the information is some three weeks—or 2 billion milliseconds—old by the time it is released.

When firms are spending millions of dollars to chip a millisecond off the flow of information, it might strike you as more than a bit strange that the government takes so long to calculate the unemployment rate.

Indeed, getting these critical numbers out sooner was one of Alan Krueger's primary agendas when he took over as President Obama's chairman of the Council of Economic Advisors in 2011. He was unsuccessful. "Either the BLS doesn't have the resources," he concluded. "Or they are stuck in twentieth-century thinking."

With the government clearly not picking up the pace anytime soon, is there a way to get at least a rough measure of the unemployment statistics at a faster rate? In this high-tech era—when nearly every click any human makes on the internet is recorded somewhere—do we really have to wait weeks to find out how many people are out of work?

One potential solution was inspired by the work of a former Google engineer, Jeremy Ginsberg. Ginsberg noticed that health data, like unemployment data, was released with a delay by the government. The Centers for Disease Control and Prevention takes one week to release influenza data, even though doctors and hospitals would benefit from having the data much sooner.

Ginsberg suspected that people sick with the flu are likely to make flu-related searches. In essence, they would report their symptoms to Google. These searches, he thought, could give a reasonably accurate measure of the current influenza rate. Indeed, searches such as "flu symptoms" and "muscle aches" have proven important indicators of how fast the flu is spreading.*

Meanwhile, Google engineers created a service, Google

Correlate, that gives outside researchers the means to experiment with the same type of analyses across a wide range of fields, not just health. Researchers can take any data series that they are tracking over time and see what Google searches correlate most with that dataset.

For example, using Google Correlate, Hal Varian, chief economist at Google, and I were able to show which searches most closely track housing prices. When housing prices are rising, Americans tend to search for such phrases as "80/20 mortgage," "new home builder," and "appreciation rate." When housing prices are falling, Americans tend to search for such phrases as "short sale process," "underwater mortgage," and "mortgage forgiveness debt relief."

So can Google searches be used as a litmus test for unemployment in the same way they can for housing prices or influenza? Can we tell, simply by what people are Googling, how many people are unemployed, and can we do so well before the government collates its survey results?

One day, I put the United States unemployment rate from 2004 through 2011 into Google Correlate.

Of the trillions of Google searches during that time, what do you think turned out to be most tightly connected to unemployment? You might imagine "unemployment office"—or something similar. That was high but not at the very top. "New jobs"? Also high but also not at the very top.

The highest during the period I searched—and these terms do shift—was "Slutload." That's right, the most frequent search was for a pornographic site. This may seem strange at first blush, but unemployed people presumably have a lot of time on their hands. Many are stuck at home, alone and bored. Another of the highly correlated searches —this one in the PG realm—is "Spider Solitaire." Again, not surprising for a group of people who presumably have a lot of time on their hands.

Now, I am not arguing, based on this one analysis, that tracking "Slutload" or "Spider Solitaire" is the best way to predict the unemployment rate. The specific diversions that unemployed people use can change over time (at one point, "Rawtube," a different porn site, was among the strongest correlations) and none of these particular terms by itself attracts anything approaching a plurality of the unemployed. But I have generally found that a mix of diversion-related searches can track the unemployment rate—and would be a part of the best model predicting it.

This example illustrates the first power of Big Data, the reimagining of what qualifies as data. Frequently, the value of Big Data is not its size; it's that it can offer you new kinds of information to study—information that had never previously been collected.

Before Google there was information available on certain leisure activities—movie ticket sales, for example—that could yield some clues as to how much time people have on their hands. But the opportunity to know how much solitaire is being played or porn is being watched is new—and powerful. In this instance this data might help us more quickly measure how the economy is doing—at least until the government learns to conduct and collate a survey more quickly.

Life on Google's campus in Mountain View, California, is very different from that in Goldman Sachs's Manhattan headquarters. At 9 A.M. Google's offices are nearly empty. If any workers are around, it is probably to eat breakfast for free—banana-blueberry pancakes, scrambled egg whites, filtered cucumber water. Some employees might be out of town: at an off-site meeting in Boulder or Las Vegas or perhaps on a free ski trip to Lake Tahoe. Around lunchtime, the sand volleyball courts and grass soccer fields will be filled. The best burrito I've ever eaten was at Google's Mexican restaurant.

How can one of the biggest and most competitive tech companies in the world seemingly be so relaxed and generous? Google harnessed Big Data in a way that no other company ever has to build an automated money stream. The company plays a crucial role in this book since Google searches are by far the dominant source of Big Data. But it is important to remember that Google's success is itself built on the collection of a new kind of data.

If you are old enough to have used the internet in the twentieth century, you might remember the various search engines that existed back then—MetaCrawler, Lycos, AltaVista, to name a few. And you might remember that these