

The War on Data

Andrew Gelman and Mark Palko

This column made its debut with an examination of the need for the maximum possible openness in both data and analytic methods. The focus there was on researchers making their data available so results could be replicated and, perhaps more importantly, so new approaches and techniques could be applied, possibly leading to new findings and fresh perspectives.

The value of shared data reaches its logical extreme in high-quality, publicly available databases such as those maintained by the U.S. Census Bureau. These sources do not just support an extraordinary amount of research; they help individuals and institutions make better decisions and give us a set of agreed-upon facts that help keep our discussion honest and productive. For all these reasons, recent threats to publicly available data are cause for concern.

About a month before the recent presidential election, Paul Krugman devoted one of his newspaper columns to the following:

Jack Welch, the former chairman of General Electric, who posted an assertion on Twitter that the [recent unemployment data] had been cooked to help President Obama's re-election campaign. His claim was quickly picked up by right-wing pundits and media personalities.

It was nonsense, of course. Job numbers are prepared by professional civil servants at an agency that currently has no political appointees. But then maybe Welch—under whose leadership GE reported



remarkably smooth earnings growth, with none of the short-term fluctuations you might have expected (fluctuations that reappeared under his successor)—doesn't know how hard it would be to cook the jobs data.

We were curious, so we googled *General Electric historical earnings*. It was surprisingly difficult to find the numbers. Most of the links just went back to 2011,

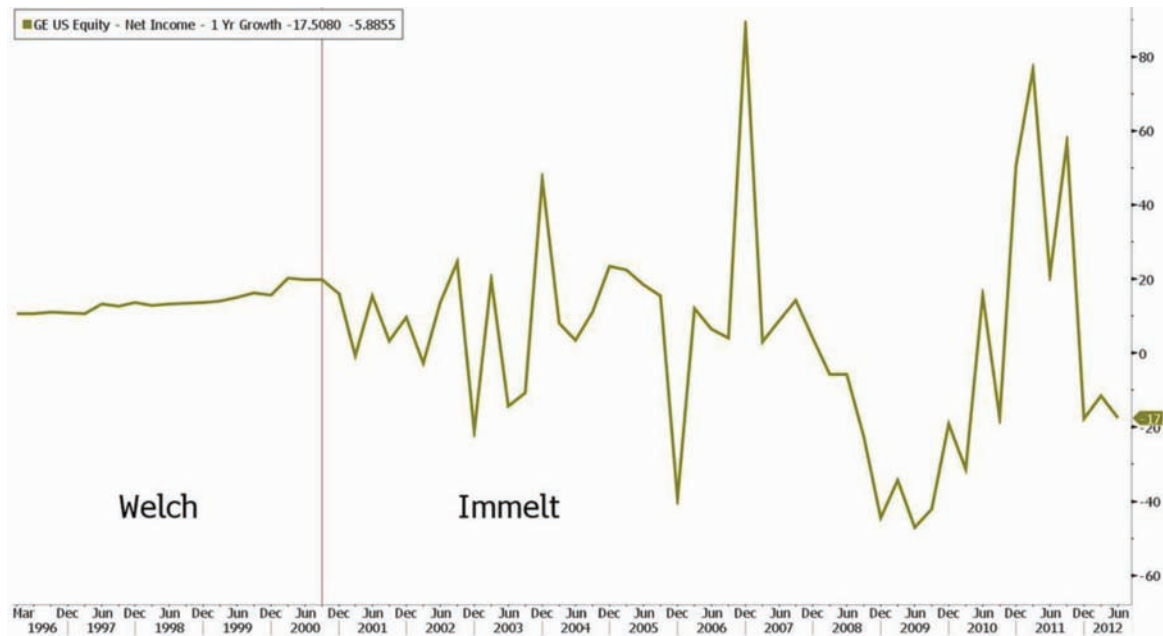


Figure 1. A graph appearing in a blog by Barry Ritholtz showing GE's earnings under two CEOs

or to 2008. Eventually, we came across a blog by Barry Ritholtz that showed the graph in Figure 1.

That looks pretty fishy, indeed. Also a link to a news article from 2002 (shortly after Welch stepped down from running GE) that said:

GE used to feature on university courses as a model of probity. These days, it crops up in the seminars about earnings-manipulation. Everyone agrees that GE practices one form of earnings management: It times one-off asset sales to coincide with one-off write-downs or restructurings. ... Beyond this, the amount of profits-smoothing that GE indulges in is a matter of speculation. GE also manages expectations about its earnings by managing its analysts. ... Managers who are in the habit of smoothing earnings have an especially strong motive to keep the good news coming, whether or not the business warrants it.

Yup. Also this from Marie Leone and Tim Reason:

[In 2009,] after a four-year investigation, GE settled accounting fraud charges with the SEC for allegedly misleading investors with improper hedge accounting and revenue recognition schemes. Specifically, GE was charged with violating accounting rules when it changed its original hedge documentation to avoid recording fluctuations in the fair value of interest rates

swaps, which would have dragged down the company's reported earnings-per-share estimates.

In addition, the SEC charged GE with concocting schemes to accelerate the recognition of revenue from its locomotive and aircraft spare parts business, to make the company's financial results appear healthier than they actually were.

Without admitting or denying guilt, GE paid a fine of \$50 million and agreed to remedial action related to internal control enhancements. "GE bent the accounting rules beyond the breaking point," noted Robert Khuzami, director of the SEC's Division of Enforcement, in a statement.

As statisticians, what interests us about this story are the different attitudes on data manipulation. The Bureau of Labor Statistics (BLS), Census Bureau, etc., take their data pretty seriously, and we agree with Krugman that it's hard to imagine them manipulating the numbers in any way. For one thing, they don't have much direct personal incentive to do so. As Krugman notes, they are civil service workers, not political appointees. And it's not as if better numbers would increase their budget line. (In contrast, we can understand the motivation for those military guys who faked the data on missile tests: Success can lead to more funding.) Beyond this, it just doesn't seem that this sort of fraud is part of the culture of government

statistics in the United States. In contrast, Leone and Reason report:

The SEC complaint relates several instances of round-robin email discussions among GE accountants, internal auditors, executives, and the company's external auditor, KPMG, debating whether aggressive accounting would pass muster with regulators.

So, it's not about Welch being some sort of data sociopath; rather, data manipulation is part of his corporate culture. We presume the vast majority of CEOs are more ethical than Welch (otherwise there would be nothing ethically noteworthy about Welch's tenure), but data is a secondary concern for even the most honest executive. For statisticians at the Census and other data-gathering institutions, the mission is to produce timely, trustworthy information. For heads of corporations, the mission is to maintain stock prices.

And, indeed, these guys have lots of motivation to fake the numbers (i.e., "aggressive accounting"). The executives and accountants personally make millions of dollars from it. Millions of dollars in win, very little personal risk if they are caught. (It was GE that paid the fine, right? Jack Welch is still loose on Twitter.) That's what we call an incentive.

The ethics of data are inextricably entangled with the decisions people make based on that information, and the impact of manipulating, discrediting, or suppressing data has to include the choices that resulted (which is one reason it makes sense that there is a taboo on perjury).

During Welch's tenure as CEO, investors chose to pay a large premium on GE stock because they believed the company had a history of strong, steady growth that was almost unprecedented for a major corporation. (According to Ritholtz, "GE's revenues grew 385% under his watch, but the company's market cap grew 4,000%.") Directly or indirectly, Welch was responsible for releasing incorrect information to the public, and that information caused investors to make decisions they would not have made with symmetric information.

On the subject of BLS, Welch also presented bad data (or more accurately, bad metadata, specifically that information used to decide how to weight a data source) to the public by saying there was a high likelihood that certain statistics were manipulated despite there being no evidence to support the claim and extensive reason to assume otherwise. There are, however, a couple of fundamental differences between misrepresenting GE's record and impugning BLS.

First, the bureau was and is a major ongoing source of data. Second, the group affected by its data is much broader. Manipulating earnings reports primarily affects the decisions of investors; propagating misinformation about the state of the economy affects the decisions of investors, business leaders, governments, and, most relevantly in this case, voters.

If you start out with certain broadly pro-democratic assumptions (that democracy is the best form of government and works best with well-informed voters and a high level of participation), reliable, publicly available data is vital for a healthy country, just as symmetric information (which includes publicly available data) is essential for efficient markets.

If, on the other hand, you start from the opposite view—that democracy is a fundamentally unworkable system and that the masses aren't smart enough or mature enough to make their own decisions (a position that goes back at least to Plato)—you also will have a different view of institutions such as the BLS.

One of the logical consequences of assuming typical voters can't evaluate information on their own is that data sources that are recognized as reliable are a threat to society. They can't be spun. They encourage people to make their own decisions.

If the masses can't handle the truth and need to be fed a version crafted by the elite to keep the people happy and doing what's best for them, the public's access to accurate, objective information has to be tightly controlled.

Viewed from this perspective, debates over government data sources take on an additional dimension. Take this news from October of 2012: "The Congressional Research Service [a nonpartisan arm of the Library of Congress] has withdrawn an economic report that found no correlation between top tax rates and economic growth, a central tenet of conservative economic theory, after Senate Republicans raised concerns about the paper's findings and wording."

Meanwhile, budget analyst David Kendall, a former Democratic congressional staffer, estimates that under budget plans of the Republican ticket, the National Oceanic and Atmospheric Administration would be cut so that, "Our weather forecasts would be only half as accurate for four to eight years, until another polar satellite is launched. For many people planning a weekend outdoors, they may have to wait until Thursday for a forecast as accurate as one they now get on Monday. ... Perhaps most affected would be hurricane response. Governors and mayors would have to order evacuations for areas twice as large or wait twice as long for an accurate forecast."

And what are weather forecasts but statistical data and analysis? In the political context of an election campaign, it is not clear how seriously to take these particular partisan claims, but, in the aftermath of hurricanes Katrina and Sandy, it is disturbing that these cuts would be considered at all.

Several months earlier came this report: "The House Committee on Appropriations recently proposed cutting the Census budget [including] a \$20 million cut in funding for this year's Economic Census, considered the foundation of U.S. economic statistics."

Why would it make sense to propose such a cut in the face of opposition from economists and the business community? One reason, we fear, is that public data represent a common factual ground on which people can argue policy.

This is not simply a partisan issue. There have been lapses on both sides when it comes to data, such as JFK running on a mythical missile gap or left, right, and center selectively ignoring information coming out of the Soviet Union for decades. (e.g., Samuelson's classic economics textbook notoriously predicted in successive editions throughout the 1960s that the gross national product of the Soviet economy would catch up to America's in 20 years.)

More recently, political figures on both sides have stepped up to defend the government's collecting and publishing of reliable data, with defenders on the right including Sen. Tom Coburn, *The Wall Street Journal*, the Chamber of Commerce, and the American Enterprise Institute's Norman Ornstein, who compared the cuts to the census to "eating our seed corn."

To diminish this shared space is to reduce the ability for policy to be debated based on reasoned arguments. ■

Further Reading

Chinn, M. 2012. Romney/Ryan on FEMA and NOAA. Econbrowser, Oct. 29, www.econbrowser.com/archives/2012/10/romneyryan_on_f.html.

Economist. 2002. The Jack and Jeff show loses its lustre. *The Economist*, May 2, www.economist.com/node/1111969.

Gelman, A. 2012. The Russians are coming! The Russians are coming! Statistical Modeling, Causal Inference, and Social Science, Jan. 13, http://andrewgelman.com/2010/01/the_russians_ar.

Krugman, P. 2012. Truth about jobs. *The New York Times*, Oct. 7, www.nytimes.com/2012/10/08/opinion/krugman-truth-about-jobs.html.

Leone, M., and T. Reason. 2009. GE settles accounting fraud charges. CFO.com, Aug. 4, www.cfo.com/article.cfm/14162632.

Ornstein, N. 2012. Research cuts are akin to eating seed corn. Roll Call, May 16, www.rollcall.com/issues/57_136/Research_Cuts_Are_Akin_to_Eating_Seed_Corn-214527-1.html.

Philips, M. 2012. The Ryan budget may cut economic data. *Bloomberg Businessweek*, May 3, www.businessweek.com/articles/2012-05-03/the-ryan-budget-may-cut-economic-data.

Ritholtz, B. 2012. GE's Jack Welch knows about cooking the books. The Big Picture, Oct. 5, www.ritholtz.com/blog/2012/10/ges-jack-welch-on-bls-book-cooking.

Smith, M. S. 2012. Ryan budget would cut more from NASA, NOAA budget functions. Space Policy Online, Mar. 21, www.spacepolicyonline.com/news/ryan-budget-would-cut-more-from-nasa-noaa-budget-functions.

Weisman, J. 2012. Nonpartisan tax report withdrawn after G.O.P. protest. *The New York Times*, Nov. 1, www.nytimes.com/2012/11/02/business/questions-raised-on-withdrawal-of-congressional-research-services-report-on-tax-rates.html.

About the Authors

Andrew Gelman is a professor of statistics and political science and director of the Applied Statistics Center at Columbia University. He has received many awards, including the Outstanding Statistical Application Award from the American Statistical Association and the award for best article published in the *American Political Science Review*. He has coauthored many books; his most recent is *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*.

Mark Palko is a statistician in the private sector who has worked with companies such as Capital One, Earthlink, and Kaiser Permanente. He blogs on analytic topics at "West Coast Stat Views" and on math education at "You Do the Math—K thru Calculus."