Harvard
Business
Review

Analytics And Data Science

Data's Credibility Problem

by Thomas C. Redman

From the Magazine (December 2013)

Summary. Reprint: R1312E Fifty years after the expression "garbage in, garbage out" was coined, we still struggle with data quality. Studies show that knowledge workers waste a significant amount of time looking for data, identifying and correcting errors, and... **more**

Artwork: Chad Hagen, Nonsensical Infographic No. 7, 2009, digital

As a rising product-management executive prepares for an important presentation to her firm's senior team, she notices that something looks off in the market share numbers. She immediately asks her assistant to verify the figures. He digs in and finds an error in the data supplied by the market research department, and the executive makes the necessary corrections. Disaster averted! The presentation goes very well, and the executive is so delighted that she makes an on-the-spot award to her assistant. She concludes, "You know, we should make it a policy to double-check these numbers every time." No one thinks to inform the people in Market Research of the error, much less work with the group to make sure that the proper data is supplied the next time.

I've seen such vignettes play out in dozens of companies in my career as a data doctor. In telecommunications, the maintenance department might have to correct bad addresses inputted by Customer Service; in financial services, Risk Management might have to accommodate incorrect loan-origination details; in health care, physicians must work to improve patient outcomes in the face of incomplete clinical data. Indeed, data quality problems plague every department, in every industry, at every level, and for every type of information.

Much like our rising executive, employees routinely work around or correct the vast majority of these errors as they go about their daily work. But the costs are enormous. Studies show that knowledge workers waste up to 50% of time hunting for data, identifying and correcting errors, and seeking confirmatory sources for data they do not trust.

And consider the impact of the many errors that *do* leak through: An incorrect laboratory measurement in a hospital can kill a patient. An unclear product spec can add millions of dollars in manufacturing costs. An inaccurate financial report can turn even the best investment sour. The reputational consequences of such errors can be severe—witness the firestorm that erupted over problems with Apple Maps in the fall of 2012.

How Data Get Dirty

When crude oil is thick, one of the major costs of working an oil field is steam-heating the crude in the ground to make the oil easier to pump. To figure out how much steam is needed, field technicians point an infrared gun at the flow line, take a reading, and send the data to the reservoir engineer. On the basis of those data, the engineer determines the right amount of steam and instructs field technicians to make any adjustments.

But the flow line can get dirty, which insulates the line and causes readings to be as much as 20°C lower than the true level. A dirty flow line means dirty data. This was a big problem at one oil company, whose field technicians had no idea how inaccurate their readings were—or that bad readings routinely caused reservoir engineers to use more steam than necessary, jacking up operational expenses by tens of millions of dollars.

This story is all too typical of data quality problems that plague every industry. Yet the solution is usually quite simple: Make sure that the employees involved in creating the data understand the problem. Once managers at the oil company specified that employees had to clean the flow lines, the errors stopped.

When data are unreliable, managers quickly lose faith in them and fall back on their intuition to make decisions, steer their companies, and implement strategy. They are, for example, much more apt to reject important, counterintuitive implications that emerge from big data analyses.

Fifty years after the expression “garbage in, garbage out” was coined, we still struggle with data quality. But I believe that fixing the problem is not as hard as many might think. The solution is not better technology: It’s better communication between the creators of data and the data users; a focus on looking forward; and, above all, a shift in responsibility for data quality away from IT folks, who don’t own the business processes that create the data, and into the hands of managers, who are highly invested in getting the data right.

Connect Data Creators with Data Customers

From a quality perspective, only two moments matter in a piece of data's lifetime: the moment it is created and the moment it is used. The quality of data is fixed at the moment of creation. But we don't actually judge that quality until the moment of use. If the quality is deemed to be poor, people typically react by working around the data or correcting errors themselves.

But improving data quality isn't about heroically fixing someone else's bad data. It is about getting the creators of data to partner with the users—their "customers"—so that they can identify the root causes of errors and come up with ways to improve quality going forward. Recall our rising executive. By not informing Market Research of the error and correcting it herself, she left others to be victimized by the same bad data coming from the department. She also took it upon herself to adjust the numbers even though she was far less qualified to do so than the creators of the data.

The good news is that a little communication goes a very long way. Time and time again, in meetings with data creators and data users, I've heard "We didn't know that anyone used that data set, so we didn't spend much time on it. Now that we know it's important, we'll work hard to get you exactly what you need." Making sure that creators know how data will be used is one of the easiest and most effective ways of improving quality.

Even better news is that addressing the vast majority of data quality issues does not require big investments in new technologies or process reengineering, as the sidebar "How Data Get Dirty" illustrates. To be sure, disciplined measurement, automated controls, and methodologies like Six Sigma are helpful, particularly on more sophisticated problems, but the decisive first step is simply getting users and creators of data to talk to each other.

Focus on Getting New Data Right

Once a company realizes that its data quality is below par, its first reaction is typically to launch a massive effort to clean up the existing bad data. A better approach is to focus on improving the way new data are created, by identifying and eliminating the root causes of error. Once that work has been accomplished, limited cleanups may be required, but ongoing cleanup will not.

Take the drilling department at Chevron, a $230 billion energy giant. Although the system it used for collecting data to evaluate drilling, plan new wells, and develop safety programs was the industry standard, the data often came up short. For example, managers couldn't determine from the data whether the drilling of a well had been completed on budget. The company launched a program to clean up the most critical data associated with the wells, but leaders very quickly realized that a comprehensive cleanup would take as long as five years—and that unless they made changes, everything created over those five years would be no better than today's data.

So the drilling group selected a veteran manager, Nikki Chang, to head up a new data management group. It was clear to Chang that the organization had to focus on the future. "Cleaning up data is non-value-added work," she says. "We're a rich company, but not...[that] rich." Her first step was to make changes in the way that errors in new data were measured. "In our [existing] metrics, if one value in a data record was wrong and nine were correct, that record scored 90%," she says. "But we can't use the record when it has even one error. It should score zero. When I adjusted the metrics to reflect this, we saw a truer picture. The metrics confirmed we had a real problem."

Rather than launch a massive effort to clean up existing bad data, companies should focus on improving the way new data are created.

Chang and her team soon crafted new targets for reducing the incidence of unusable records, keeping two main objectives in mind: "First, we wanted something simple," she says. "Second, we wanted business units to improve—and fast." At the same time, she wanted to focus on identifying root causes of big issues. "I didn't want to penalize someone for a random error or two, at least initially."

She set a first-year goal for each unit: All basic data for 95% of new wells had to be created correctly the first time. The second-year target was 100%. "For most, that was a demanding but achievable target," Chang observes. Now her team updates a scorecard regularly. "Everyone can see how they're doing at all times. This is important—when they try to improve something, they get to see whether or not they were effective. And they can see how they're doing relative to their peers."

Chang was careful not to specify how business units should pursue the new targets. Many approaches emerged: One group simply set up daily reviews to go over the previous day's data; another used Lean Sigma (a variant of Six Sigma); a third set up an internal competition among the various rig groups. "Chevron people are creative and competitive," Chang says. "Give them a target they buy into and they'll figure out how to meet it."

Not surprisingly, most units have done just that. Eight months into the initiative, 13 of the 15 business units had met the year-one target, and the other two were on track to do so. Those results are impressive—but they're by no means unusual. Indeed, I find that most companies that address data quality in this way show similar results.

Put Responsibility for Data in the Hands of Line Managers

Very often, data creators are not linked organizationally to data users. Finance creates data about performance against quarterly goals, for example, without considering how Sales will want to use

them or Customer Service analyzes complaints but fails to look for patterns that would be important to product managers.

When quality problems become pervasive or severe, the organizational response is often to task the IT department with fixing them, usually by creating a special unit in the group to spearhead the initiative. This may seem logical, since IT is a function that spans all silos. But IT departments typically have little success leading data quality programs. That's because, as I've noted, data quality is fixed at the moment of creation. With rare exceptions, that moment does not occur in IT. To address problems, IT people can talk to creators and users, but they can't change the offending business processes. All they can do is find and correct errors, which, as we've seen, is not a long-term solution.

Fixing the Metadata

Many data quality problems are rooted in metadata, something that has been much in the public eye thanks to the recent NSA scandals. A good working definition of "metadata" is "data about data"—for instance, units of measure. High-quality metadata makes it easier for people to find the data they need, combine information, and draw the appropriate conclusions—and errors in metadata can have a big impact. For example, NASA notoriously lost the $125 million Mars Climate Orbiter because one group of engineers used English units (such as feet and pounds) while another used metric units for a key operation.

One firm that has done an exceptional job with its metadata is Aera Energy. It identified 53 common business terms, such as "contract" and "customer," and then brought people from across the organization together to hammer out definitions of those terms, which serve as the core of its metadata.

The work was time-consuming and demanding, but it has paid enormous dividends. The productivity of Aera's most critical resource—its engineers—has more than doubled. As CEO Gaurdie Banister observed, "High quality metadata makes everything we do easier, from internal communications to planning new applications to making better decisions."

The incentives for IT are weak as well. Business departments benefit tremendously from having access to good data to improve products, services, and decision making. IT reaps little reward, and it doesn't feel the pain when the data are wrong. It is the business units and managers who must face angry customers, make good on promises, or explain poor results to shareholders.

Smart companies place responsibility for data quality not with IT but with data creators and their internal data customers. IT folks readily admit that "the business owns the data," I find; once companies understand that, alignment comes quickly. Liz Kirscher, formerly the president of the data business at Morningstar, the Chicago-based provider of mutual fund and other financial markets data, explains it this way: "We would no more have Tech run data than we would have Research run Tech. They're different kinds of assets."For most companies, the real barriers to improving data quality are that some managers refuse to admit their data aren't good enough, and others simply don't know how to fix poor-quality data. The first bit of progress occurs when a manager somewhere in the organization (possibly a senior executive, but more often someone in the middle) gets fed up and decides that "there has to be a better way." The manager launches a data program and, if the prescriptions noted here are followed, usually gets good results.

But that manager often has little motivation or is unable to push beyond his or her department. The company is left with superior quality data in a few areas and poor data everywhere else.

Getting past that plateau takes the commitment of senior leadership. Some 20 years ago, Joseph Juran made the case in his seminal HBR article, "Made in U.S.A.: A Renaissance in Quality," that leadership for quality could not be delegated. Juran was, of course, talking about quality in manufacturing, but his words ring equally true for data. If anything, the data quality challenge is both tougher and more urgent. It's time to amplify the call for leadership.

A version of this article appeared in the December 2013 issue of *Harvard Business Review*.

Thomas C. Redman, "the Data Doc," is President of Data Quality Solutions. He helps companies and people, including start-ups, multinationals, executives, and leaders at all levels, chart their courses to data-driven futures. He places special emphasis on quality, analytics, and organizational capabilities.