THE WALL STREET JOURNAL.
WSJ.com

August 26, 2015, 3:03 PM ET

Build Data Quality Into the Internet of Things

By Thomas H. Davenport and Thomas C. Redman

Clocksmiths admitted Wednesday that London's Big Ben has been running fast over the last several weeks. Pictured: Technicians clean on of Big Ben's faces, Aug. 19, 2014.

Ben Stansall/Agence France-Presse/Getty Images

The clock on one of our office file cabinets (Redman's) reads 10:32 a.m. on a Friday. There is every reason to believe it is correct. This clock synchronizes itself at 1:00 a.m. every day using a time signal sent out by the National Institute of Standards and Technology, from Fort Collins, Colo. Redman is already two minutes late for a meeting that always starts promptly. And he must complete some critical preparation before joining. Stressed, he completes the prep and joins the call about 10 minutes late. But no one else is on the call. The clock was off by 20 minutes! He's still 10 minutes early. Other than the unneeded stress, no harm.

The story provides an apt warning of the dangers in connected devices. The Internet of Things (IoT) is bringing billions of new connected devices into our lives. From Fitbit activity trackers to Nest thermostats, to devices embedded in engines and factory machines,

excitement is high and the potential is enormous. But as the clock example suggests, it will be much more difficult to achieve that potential if too much IoT data is bad.

What difference does it make if a connected device like the clock occasionally goes astray? With traditional data, we've become used to the idea that people are the main sources of data quality issues. But how important is it when devices themselves introduce significant errors in data quality?

First of all, simply measuring the error requires significant effort. Back to the clock. After what he thought were isolated instances, Redman began to suspect the clock wasn't always correct and started a log of clock errors. In the past 16 months, he's noted 10 separate instances. As noted above, in one instance the clock was off by 20 minutes, in another by three hours, and in still another, by 19 days. In eight instances, both the time and date were incorrect. The smallest error, 20 minutes, is the only one that caused any real grief. Being early for a meeting is no big deal, but if the clock were driving complex industrial processes it could really lead to problems.

Bear in mind that clocks are remarkably simple in their measurement processes compared to other devices such as accelerometers, locators, and chemical assays. We've seen similar problems with data quality issues with many other connected devices—a health tracking device doesn't count your steps on the treadmill (at least in our experience); a device in the electric grid quits working, then starts again; the gears in a weather vane fill with sand and compromise the measurement.

If one category of problem is that the device doesn't measure as it's supposed to, the second category involves what units of the physical world are being measured. Most such issues individually are pretty mundane—a measurement made in yards (English units) but interpreted in meters (metric units) and a misinterpreted relationship between "steps" and "distance walked" (the faulty health tracking device) are good examples. But there are quite literally thousands of such issues, any one of which can trip you up.

The problem is not just the devices per se. The whole idea of connected devices is that they work with one another to do things they can't do alone. And here problems grow even more acute. For a single device, it is good enough to know whether the units are English or metric. For multiple devices all the units of measure must align to perform even the simplest analytics. It just won't do if your Nest device is measuring your house temperature in

Fahrenheit and transmitting it to a utility company that records temperatures in Celsius. This means that people and organizations have to agree on how they will measure things.

Standards are the obvious answer, but they take a devilishly long time and much effort. For example, the development of the EPCGlobal (electronic product code for radio frequency identification) standard took about 15 years. The development of the ANSI X12 standard for electronic data interchange took about 14 years. We don't want to wait that long for anything these days, and standards development could really slow down the process of the IoT movement.

Beyond understanding the issues and trying to help establish standards, what must a technologist actually do? We take as a general rule of thumb that bad data is like a virus. There is no telling where it will end up or the damage it will cause. With viruses the basic idea is to try to prevent the virus in the first place and do all you can to contain it.

For data quality and the Internet of Things, preventing the virus means excellent design, manufacturing, and installation of the IoT device. Since such devices are typically made by someone (a semiconductor manufacturer, for example) other than the user, buyers must insist the device actually measures what it purports to measure. This implies both specification of the intended measurement and rigorous testing, under both laboratory and real-world conditions, to ensure that is what actually occurs. What is a "step," for example, and does the device actually count them properly?

The specification should spell out operating conditions. Recall we noted that the example of a health tracking device not working so well on the treadmill. The specification should spell out everything you'll need to use the device successfully in practice: what you need to do to install and test it, its expected lifetime, how you'll know when it is time to maintain or replace the device, and so forth.

Insist on two levels of calibration from your device supplier. First, there should be rigorous calibration before the device leaves the factory, and an "on-installation" calibration routine to ensure that the device works as expected. Second, ongoing calibration is required to make sure the device continues to work properly. Ideally, the on-installation and ongoing calibration routines should built-in and automated.

To contain bad data, devices should come equipped with what we call "I'm not working right now" and "I'm broken and must be replaced" features, which do exactly what their names suggest.

Finally, you should not expect perfection, particularly with new devices. But you must insist on rapid improvement. So it is critical that the manufacturer aggregate and analyze the results of all these steps, looking for patterns that suggest improvements. Seek answers to basic questions such as: can the devices really be trusted? Are they lasting as long as expected? What is causing them to fail?

No matter what the domain, measurement is difficult and quality is always an issue. The Internet of Things offers tremendous potential to measure things we've always measured more cheaply, to measure new things, and to connect those measurements and analyze them in powerful, new ways. It's much easier to build quality in from the start.

Thomas H. Davenport is a Distinguished Professor at Babson College, a Research Fellow at the MIT Center for Digital Business, Director of Research at the International Institute for Analytics, and a Senior Advisor to Deloitte Analytics.

Dr. Thomas C. Redman, "the Data Doc," is President of ***Navesink Consulting Group****. He helps leaders craft programs to get in front on data quality, learn to compete with data, and build the organizational capabilities to do so.*