The evolLLution
A DESTINY SOLUTIONS ILLUMINATION

Published on 2017/03/16

Going Wrong with Data

Greg Chmura | Chief Quality Officer, Chmura Economics and Analytics

We often assume that data is infallible, and while that's far from the truth, a better understanding of assumptions required in data analysis—as well as a commitment to observation and critical thinking by users—can help avoid pitfalls.

Going wrong with data is kind of like going wrong with a steering wheel. That is, it is not something you want to do, especially when speeding down the highway.

But aren't data usually reliable? Isn't it difficult to go wrong with data? It turns out, unfortunately, that it is easier than you think.

First a disclaimer: Data can be amazingly valuable, offering a great tool to use in decision making. I work with the team at Chmura Economics & Analytics where data are at the very core of our business, and our mantra is *providing quality data for quality decisions*. Nevertheless, there are uncertainties associated with all data; and the more you appreciate that, the more you'll be able to keep your car squarely on the pavement.

One of the most frequently referenced forms of data uncertainty is the statistical uncertainty typically cited in political polls. The important thing to know about this type of uncertainty is that it only represents one portion of the overall uncertainty.

We don't have to look back far to find an example of an election prediction gone wrong. The 2016 U.S. presidential election polls pegged Hillary Clinton as the favorite to win, even by a wide margin, in battleground states she eventually lost. So, what went wrong?

The answer is hidden in the form of uncertainty that usually isn't even discussed. When poll numbers are presented, they're typically cited with a statistical margin of error, such as "plus-or-minus four percentage points." The problem is that this form of uncertainty only accounts for the uncertainty if all the assumptions used in the analysis are correct.

Furthermore, the more different something is, the more likely assumptions can go wrong due to the lack of precedent. If the 2016 election had featured a rematch between Barack Obama and Mitt Romney, the results would've been much easier to predict since pollsters would have had plenty of data on how the electorate related to those candidates. As for the actual 2016 candidates, there was little precedent for how the electorate would end up voting for them—such as which demographics were more likely to vote.

Another way to go wrong with data is to think you have the correct data set when you actually don't. For example, when companies consider a location for a new facility, they typically look at regional labor supply data to make sure there will be a sufficiently large, skilled labor force from which to draw potential employees. Such companies are clients of ours on a regular basis. In fact, we once had the opportunity of working with a firm that had previously experienced a challenge with this very form of data uncertainty.

This client was looking to locate a new facility in a large metropolitan area. After receiving labor availability data on the metropolitan area they were considering, they saw that it looked sufficient and proceeded with the project. When it came time to hire employees, though, filling the positions was enormously difficult. What went wrong?

It turns out that the data set itself may have been fine, but it simply wasn't the data set that they needed. It was accurate, but it wasn't relevant. The facility location was in a sparsely populated portion of the metropolitan area, with the surrounding terrain possessing transportation barriers such as major bodies of water. Therefore, using the entire metropolitan area as a potential drawing pool wasn't a good starting point. In this case, running an actual drive-time analysis based on the site's location would have been more accurate.

Of course, problems can also arise when there are inaccuracies in the data themselves. We work with very large data sets and have found our share of bloopers that require fixing before offering to the public. For example, while working with government contract data, we once came across what appeared to be an out-of-place $600 million contract. Upon further examination, we noticed the digits showing the value of the contract and the digits describing the company's nine-digit zip code happened to match each other exactly! It could have been a coincidence, but it's far more likely that it was a data entry error—and one that was big enough to completely throw off the resulting analysis.

Many businesses today work with and rely upon what is called "unstructured data," meaning information that cannot be used until it is first extracted from text or other ambiguous sources. The process by which the data are extracted can certainly introduce further uncertainties, as uncertainty is at the heart of these data.

An example of an unstructured data set that Chmura Economics & Analytics works with is online job advertisements. The very nature of the data set requires making necessary assumptions to transform it into something meaningful. For example, these ads are typically gathered from thousands of websites. In this process, duplicates of the same ad are often found. Is it straightforward to determine whether two ads are duplicates of one another? Well, yes, sometimes; but when you're processing millions of ads every day, there are inevitably going to be some that fall into a gray area. This challenge is typical when working with

unstructured data. A black-and-white line needs to be drawn somewhere, and assumptions must be made to figure out where to draw it.

These assumptions can sometimes be innocuous. Other times, they can start to cause problems and then, the next thing you know, you've driven your car into a ditch.

What are decision makers to do, then? Avoid data all together?

No—sometimes data may be all we have to navigate with. But there are two important things that can help. First, it's useful to have a data person in the room when you're using the data—someone able to explain the assumptions made and the limits of the data set. In addition to that, it is smart to also use your own eyes. Take time to look through that windshield for some type of real-world confirmation as to whether the data are steering you in the right direction or that something, somewhere is very wrong.

https://evolllution.com/technology/metrics/going-wrong-with-data/