Article from

Predictive Analytics and Futurism

December 2015
Issue 12

Big Data or Infinite Data?

By Dave Snell

There was a young fellow from Trinity
Who took $\sqrt{\infty}$
But the number of digits
Gave him the fidgets;
He dropped Math and took up Divinity.
(from One Two Three … Infinity: Facts and Speculations of Science, by Georg Gamow)

One of my favorite books of all time is one I read in high school a half century ago: One Two Three … Infinity by George Gamow. Prior to that time, I had a limited understanding of the concept of infinity. Gamow, an expert in theoretical physics, introduced the idea of infinity by describing a tribe of Hottentots, who had words for one, two, and three; but none for higher numbers. Anything larger than three was considered "many"—our rough equivalent of infinity. Through the tribe analogy, he addressed the issue of how to compare one infinity with another infinity. If you have many beads and many coins, how do you determine which is your larger collection? Gamow related how Georg Cantor, the founder of set theory, compared two "infinite" sets. Cantor proposed pairing the objects of the two collections and see which, if any, ran out first. If each object in the beads collection can be paired with an object in the coins collection, then the two collections are the same size. However, if you arrange them in pairs, and some unpaired objects are left over in one collection, then it is said to be larger, or stronger, than the other collection. Thus, he introduced the "arithmetics of infinity," where the infinite set of all even numbers is the same size, or cardinality, as the infinite set of all odd plus all even numbers. And while you are still wrapping your mind around that non-intuitive result, they both are smaller (less strong) than the cardinality of the set of real numbers, which, in turn, is less strong than the cardinality of the number of geometric curves.

The many years since *One Two Three … Infinity* (I read his 1961 edition; the first edition was published in 1947) have seen a dramatic increase in the number of collections we count and analyze and compare to other collections. Indeed, according to former Google CEO, Eric Schmidt, "Every two days now we create as much information as we did from the dawn of civilization up until 2003." He said that on Aug. 4, 2010 at a Techonomy conference in Lake Tahoe, California and I have to believe the figure today would be even more astounding.

We are clearly into an age of "Big Data"; and it is a term so overused that my Google search for it today yielded 795 million results. Yet, in some respects, we understand this no better than the tribe of Hottentots that George Gamow described in 1947—perhaps no better than how Georg Cantor explained it in 1874. In fact, according to Dan Ariely, the author of *Predictably Irrational*, and other excellent behavioral science books,

> "*Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it*"—Dan Ariely

What is Big Data? Some companies brag about being able to handle big data of millions of rows of information. Others claim they process over a billion data items and boast about their big data capability. WalMart was supposedly the first commercial enterprise to store a terabyte of data, in 1992; and then we thought that was truly big data. Now, you can buy a USB thumb drive on Amazon, for your keychain, which stores a terabyte of data. The Human Genome Lab stores petabytes of DNA information. Many database manufacturers claim the ability to store exabytes of data. The NSA stores … oops! That is a classified size, but obviously a lot! Cisco, the multinational technology company that makes and sells networking equipment, predicts that by 2016, about the time you receive this issue in the mail, global internet traffic for the estimated 3.4 billion Internet users will reach a staggering 1.3 zettabytes annually.

What distinguishes Big Data from just large, or very large, or very, very large data?

I'd like to propose a new term: Infinite Data. Infinite data is data that is so large that the mere acquisition of it overwhelms our ability to process it with classical statistical methods.

Take, for example, weather indicators. Our ability to forecast the weather today or tomorrow seems quite good; but our best estimates of weather next month seem unimpressive because the amount of data coming in is so voluminous that the so-called butterfly effect cannot be analyzed in real time.

Another example is the streaming data regarding insurability of a cohort of applicants from the Internet: Tweets, wearables, etc., and other information preclude calculating a classic mean or standard deviation because the data is changing before you even have an opportunity to count it. Like Cantor, we may need to eventually differentiate between Infinite Data of cardinality $\aleph_0$ (read aleph-naught or aleph-zero), the

smallest type of Infinite Data, and $\aleph_1$ (aleph-one, a stronger set of infinite data), or $\aleph_2$ (still stronger).

We can also describe big data with more modern terms such as the three Vs: volume, velocity, and variety. Sometimes we add a fourth V, variability, or even a fifth, veracity, to the mix. When these characteristics combine—especially when they are expanding at an increasing rate, we feel that we have Big Data. Yet, actuaries should not feel intimidated by the newer terms. In most cases, we can relate them back to basic techniques we studied years ago under different names.

Take the case of velocity. The data may be coming in so fast that by the time we count it, the count has increased. In these situations, we could throw up our hands and say that a mean, a standard deviation, and a random sample are impossible to calculate. Alternatively, we can use stream algorithms, Reservoir sampling, and other algorithms to compute stats on the fly based on the data received to date, and then project the trends. This is conceptually similar to the rolling average that actuaries have used for decades in their experience studies.

Volume has always been a concern for actuaries. Before computers were fast enough to process a block of business on a seriatim basis, we had to employ grouping and sampling techniques. Likewise, the variety of various types of policy benefits (consider disability income policies with their differing benefit periods, definition of disability, elimination periods, waiting periods, occupational classes, etc.) required classification techniques, and some of what data scientists now call feature engineering. Veracity has always been a challenge. Insurance applicants understate the amount of alcohol they consume, how often they smoke, how heavy they are; and different sources of data (from physicians, motor vehicle records, credit reports, paramedical exams, lab results, policy applications, etc.) often show inconsistent or even conflicting information. We have had to apply credibility factors and techniques for years.

Similarly, our experience with graduation techniques, mortality table construction, demography, complex variables, stochastic (and stochastic on stochastic) projections, multivariate contingency analysis, and many other 'standard' components of the actuarial education can be applied to work with big, or infinite, data. Yes, we may have to learn some new names for techniques we already know. Yes, we may have to supplement those techniques with more current research. Yes, we may have to gain a comfort level with some data science tools such as R and Python and others beyond our basic Excel models (although, Excel is a lot more impressive in this arena than most data scientists assume; and actuaries are often experts using it). Yes, yes, yes. We cannot just rest on previous accomplishments and expect to compete on future opportunities. Please read the following article, by Dihui Lai and Richard Xu, about tools such as Spark, to help with the processing speed and volume issues.

The bottom line is that actuaries are entering a new era where they can be pioneers and leaders and highly valued; or they can be followers and Luddites and marginalized. The choice is ours; but only if we are willing to learn to count beyond "many." ■

Dave Snell, ASA, MAAA, is technology evangelist at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *dave@ActuariesAndTechnology.com.*