WALL STREET JOURNAL

Making the Case for the 'Long Tail' of Big Data

By Randy Bean | Contributor | August 27, 2015

Around the time Chris Anderson introduced the "long tail" to the masses in a 2004 Wired cover story, MIT's Erik Brynjolfsson and others were studying how products in low demand could produce a larger market share than higher demand items — if the distribution channel was large enough. The "long tail" came to reference these harder to find items that, taken collectively, could create a big market. Now we can apply the term to Big Data.

I recently had the opportunity to spend some time with Michael Stonebraker, a pioneer in the field of data management and the 2014 recipient of the ACM Turing Award, which is often called the "Nobel Prize of Computing." Prof. Stonebraker, a member of the faculty at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), has drawn upon Prof. Brynjolfsson's statistical research to apply the "long tail' concept to Big Data.

For most large companies, Big Data is less about managing the "volume" of data they have, and much more about integrating the wide "variety" of data sources that are available to them – which can include data from legacy transaction systems, behavioral data sources, structured and unstructured data, and all sizes of data sets. Prof. Stonebraker estimates that corporations manage to capture a small fraction of this data within their enterprise data warehouse systems. He's calling for companies to shift their focus to "the long tail" of data that may actually be the key to their most critical business insights.

Prof. Stonebraker believes that corporations minimize and misunderstand the difficulty of data integration, which he calls the foundation of data management. "Data integration is damn hard," he said at the MIT Chief Data Officer Symposium in Cambridge, Mass. last month. "Data warehouses work for less than 25 data sources – they don't scale." He continued, "The notion of a global data model and data standards alone being sufficient for data integration is fantasy. It doesn't work!"

Instead, the future of data management lies in "data curation," which he describes as being "aimed directly at the 'long tail' – the hundreds or thousands of data silos not captured within the traditional data warehouse, and which can only be captured and integrated at scale by applying automation and machine-learning based on statistical patterns.

While many firms are embracing the notion of the "data lake" as a staging area for data management, Prof. Stonebraker views the data lake as "just a bunch of un-curated data, a junk drawer that, on its own, is not solving any significant problem." Data curation relies upon machine learning systems that use statistical techniques to learn and build knowledge over time, he says. As business analysts continue to demonstrate an insatiable appetite for more data, data curation holds the potential to release firms from the "bondage of traditional schema."

Prof. Stonebraker acknowledges that he is challenging conventional wisdom about data warehousing, a discipline that has grown in stature over the past two decades with many thousands of practitioners. He foresees a changing data and information landscape, part of the Big Data revolution, where complex data analysis supplants the simple data analysis that he sees as the current state of analytics today. He concludes, "We are in the midst of an explosion of new ideas that will change the data landscape. We are going to be at this for a while." Given the continuing proliferation of data and new data sources, he may be exactly right.

Randy Bean is CEO and managing partner of consultancy NewVantage Partners. You can follow him at @RandyBeanNVP.