NAXION thinking

Viewpoint

April, 2018

Smart-Size Big Data to Get Better Answers Faster

By Mike Kelly, Ph.D.

The Challenge of Tying Down Big Data

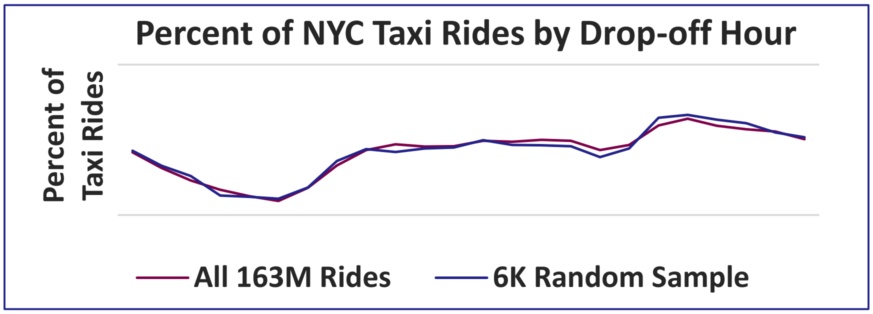
Although the promise of Big Data is exhilarating, the practical burdens − logistical and analytical − are equally daunting. The problem is that Big Data can be, well, big – very big. Customer transaction databases often contain tens of millions of records. Moreover, Big Data often accumulates rapidly in real time, and is populated with diverse types of information, such as timestamps, spatial coordinates, and text from social listening. Even when powered by the formidable processing heft of today's corporate-owned or cloud-based IT infrastructure, Big Data computing requires an enormous amount of time. It can take many hours, potentially even days, to run a single model. And since modeling is typically an iterative process, the overall endeavor can be prohibitively time-consuming. In struggling to tie down Big Data, we can feel like Lilliputians next to Gulliver.

Various strategies have been employed to address the "time sink" of Big Data, but algorithms known to be effective for small data sets (e.g., Markov chain Monte Carlo) don't scale well. One line of attack has been to optimize analytic operations for Big Data contexts (e.g., design algorithms that process or compress data more efficiently); another boosts the computational power of IT infrastructure through parallel computing (e.g., Hadoop). But such approaches involve substantial investment in both IT infrastructure and the talent needed to architect and manage it.

An Elegant Solution to the Heavy Lifting Problem …

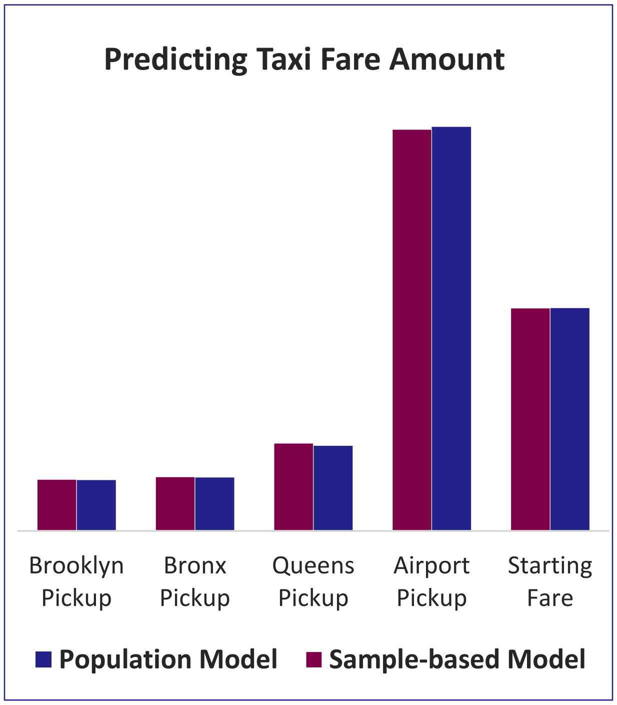
While Big Data bottlenecks seem to be an occupational hazard, they are not necessarily intractable if we revisit a basic assumption − that *every bit of it* must be analyzed to wring full value. Just as we can efficiently and accurately measure the characteristics of a survey population with systematic sampling techniques, so too can we apply principles of statistical sampling to Big Data. To illustrate, we analyzed a publicly available dataset of 163 million taxi rides in New York City. The dataset contains a variety of information:

- Temporal *(passenger pickup/ drop-off times)*
- Spatial *(latitude/longitude of pickup/drop-off coordinates)*
- Numerics on different scales *(trip distance, fare amount)*



The chart plots number of rides by drop-off hour for the full population and a random sample of 6,000 rides. It's clear that the small random sample mirrors the pattern in the full universe. The same story holds with other metrics like trip distance.

… with the Heft to Build Sturdy Models



Modeling with Big Data can be particularly time-consuming, much more so than calculating summary information as in our first example. A sampling approach to Big Data could be especially helpful in modeling situations. To demonstrate, we developed a *population model* and a *sample-based alternative* that predicted taxi fare amount from various characteristics of the ride such as pickup location. Predictions from the two models are essentially identical but were much faster to produce with a sample-based approach.

Sampling revolutionized – in some sense, created – the field of market research. We are ripe now for a similar transformation in our approach to Big Data.

Be Careful How You Sift and Weigh the Data

Although Big Data sampling will deliver significant cost and timing advantages over a Big Data census, practitioners need to consider their sampling procedures carefully to avoid drawing the wrong conclusions. Theory and proven best practices from the science of survey sampling can help.

Depending on the nature of the business questions to be answered and the type of information available in the Big Data universe, a particular stratification may be required (e.g., by customer demographics such as census region, or spending history) along with random sampling of records within each stratification cell. Weighting adjustments may be needed if certain types of records are over or under sampled compared with their incidence in the Big Data universe. These activities are essential to ensure the accuracy – as well as efficiency − of sample-based approaches to Big Data.

The Ultimate Big Data Pay-off: Agility and Accessibility

Computing efficiency solutions need to be less about bandwidth than about agility. Once we cut Big Data down to size, it becomes easier to make effective use of it, directing efforts where they are most needed: toward the extraction of insight from lighter loads rather than processing heavy loads faster. It will democratize the use of Big Data by making it more broadly accessible, putting it in the hands of people who know their markets well enough to apply the fruits of “smart-sized” Big Data analysis and modeling to real business problems.

About NAXION

NAXION is a broadly resourced, nimble boutique that relies on advanced research methods, data integration, and sector-focused experience to guide strategic business decisions that shape the destiny of brands. The firm is distinguished by a truly effective synthesis of authoritative market research and consultative marketing application, and a dedication to solving problems in ways that are both inventive and pragmatic. NAXION's hybrid "enterprise DNA" is rooted in our origins as the world's first business intelligence firm and subsequent decades as the National Analysts division of Booz•Allen & Hamilton. And our exceptional commitment to partnership reflects a unique, employee-owned organizational culture scaled to provide highly effective solutions to clients' most challenging marketing problems.

About the Author

Mike is a Senior Group Director at NAXION who designs and manages major engagements for clients seeking to develop B2B and B2C business strategies based on customer insight and advanced market analytics. Leveraging his skill in devising new modeling techniques and integrating multiple data streams, Mike has helped clients in Information Technology, Energy, Consumer Electronics, and Manufacturing optimize pricing strategies, guide product bundling, sharpen targeting, and prioritize service improvements that drive customer loyalty. Mike has notable subject matter expertise in the fields of cognition and computational linguistics, where his highly regarded academic work has been a platform for innovation on behalf of clients, while building the firm's intellectual capital in advanced methodologies.

mkelly@naxionthinking.com
215.496.6842