Keynote speech at the Fall 2012 KMA conference

Big Data 2.0

Greg Allenby*

I am honored to be here at the Korea Marketing Association meeting in Seoul to present some reflections on our discipline over the past 25 years and to offer thoughts of the future. Our field has undergone and will continue to undergo sizable transformations as firms learn to make better use of the data they have at their disposal. The title of my talk, Big Data 2.0, points to what I believe will be the next generation of analysis and insight available from these data.

I have been marketing academic for 25 years, specializing in the development and application of quantitative methods for understanding consumer behavior and marketplace demand. Scanning and bar coding technologies were new to marketing in the 1980's when I was in graduate school, and it was evident that these data would move analysis from the attitudes and intentions of individuals that eventually lead to demand, to the immediate effect of marketplace variables, such as prices, on sales. The availability of retail data, in the form of weekly sales summaries, and panel datasets of household purchases, shifted much of the analysis in marketing from upstream determinates of demand to downstream, or marketplace, analysis of sales. For the first time, researchers were able to accurately represent the competitive environment and began to understand aspects of demand such as cross-price effects, market structure and the effect of promotional variables on sales.

Market segmentation analysis, for example, shifted from its original form of providing an in-depth understanding of the motivations and drivers of behavior useful for guiding product strategy, to analysis focused on how people react to existing offerings. Instead of providing insights into where people are "coming from," market segmentation began to focus on where people are "going to." One limitation of such downstream analysis is that it does not provide insight on unmet demand, and other forms of counter-factual analysis that provide a glimpse of what could be.

I am best known for my work in Bayesian

* Helen C. kurtz Chair Professor, Fisher College of Business, Ohio State University(allenby.1@osu.edu)

statistics. In the early 1990's, a new development in statistical computing known as Gibbs sampling, or Markov chain Monte Carlo methods, made its debut in the academic literature that allowed for the analysis of large datasets using models of demand that more realistically reflected the actual decision environment. Models of demand became more real by being able to incorporate more products, more variables and more complicated decision processes. The opportunity to rationalize retail pricing across many thousands of stock keeping units became a reality as companies started to offer software for price management. Big Data 1.0 was a practical reality.

The Bayesian revolution in marketing also allowed us to change from a macro analysis of weekly demand, to analysis centered on specific individuals. Bayesian statistics revolutionize the application of random-effect models by being able to retain, or "back-out" estimates of individual-level respondent behavior. Having access to coefficient estimates of individual respondents provided insights about the entire distribution of potential demand. For the first time, it was possible to identify respondents who were most satisfied, most likely to switch brands, and who most valued a particular product feature. In statistical parlance, it allowed us to explore the extremes, or tail areas, of distributions and not be confined to their central tendencies (e.g., the mean) when conducting analysis.

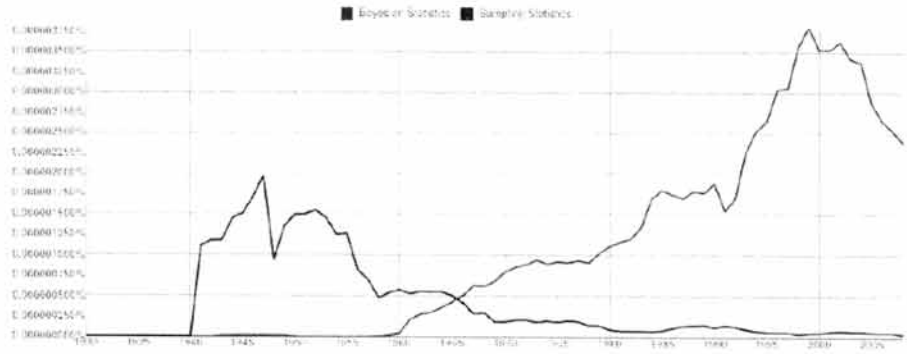
Marketing thrives on the analysis of extremes. People want products that provide them with good value, and value can either be created by increasing a product's benefits or by decreasing its costs. A benefit orientation leads to market niches and local monopolies where consumers are willing to pay extra for products that respond to their specific needs. Identifying the important, unique needs to which a product responds involves understanding extremes, not means. Market segmentation is marketing's response to the distribution of heterogeneous needs, wants and sensitivities in a market by concentrating analysis on a subset of respondents. These respondents constitute the extremes of a distribution, and are often not well reflected by its mean or by the coefficients of a traditional analysis (e.g., a regression coefficient) that describes behaviors in terms of average, or expected outcomes.

The past 25 years has witnessed an incredible growth in Bayesian statistics. Figure 1 demonstrates the magnitude of the Bayesian revolution in comparison to its predecessor, sampling statistics. Plotted on the vertical axis is an index of the count of the word "Bayesian statistics" appearing in books. The horizontal axis is time, and we can readily see the spread of Bayesian methods at the expense of sampling statistics in publications monitored by Google.

Even more striking is the comparison of

〈Figure 1〉

Bayesian Statistics versus Sampling Statistics

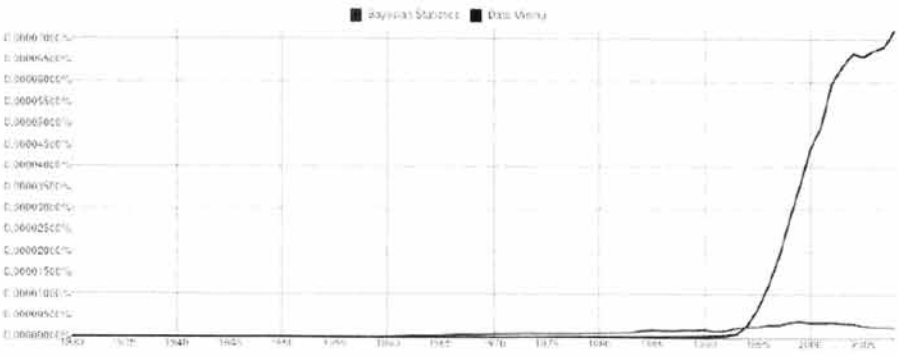


Source: Google Books Ngram

"Bayesian statistics" to "Data mining" presented in Figure 2. The data mining revolution has already dwarfed the Bayesian revolution, and promises to have an even greater impact on field of marketing. The data mining revolution currently involves the collection, storage and exploration of many forms of data of interest to firms beyond sales data recorded by bar-

〈Figure 2〉

Bayesian Statistics versus Data Mining



Source: Google Books Ngram

coding technologies of the 1980s. It includes Internet and web-browsing data, direct marketing data and even data from on-going tracking studies.

While firms currently have access to a greater variety of data reflecting various aspects of demand and its determinants, the analysis of these data is still at a primitive stage of development, e.g., aggregate summaries of the records presented in graphical form and simple measures of association. Firms can now readily retrieve sales figures for a specific offering from a specific period of time and compare it to sales of other items and other time periods. Firms also examine text data to understand words used to describe consumer's brand experiences and their opinions regarding service encounters. These data are used to form associations that point to patterns, trends and cross-selling opportunities. However, they have yet to be used to their full potential because these data lack rich insight into the determinants of demand.

Data stored in archival retrieval systems record what was purchased at a particular point in time, but these data typically lack information about the context in which the purchase was made, who made the purchase and for what purpose. It lacks information on the attitudes and opinions of the purchaser, and lacks information on the competing alternatives that were also considered. This data does not provide researchers with the ability to formulate product strategy, such as whether a product-line extension is a good idea. It also does not inform a firm as to an appropriate response to a competitor setting a new price, and whether an advertising message would work well in a new medium. Many channel problems are also uniformed, such as the desirability of a new point of distribution or retail format. Big data is good at answering the "what" and "when" questions, but often falls short on "who," "where" and "why." The answers to these questions are needed to draw causal conclusions and suggest optimal courses of action. What and when provides some guidance on tactical decisions, but lacks important information needed for strategic decisions.

Big Data 2.0 is a next-generation data source that combines what and when information available in existing company databases with who, where and why information from customized surveys and analysis using primary data.

One example of a combination of data will occur in the context of text analysis. Text analysis can be thought of as an analysis of open-ended responses. In contrast, the analysis of data from questionnaires can be thought of as closed-ended responses where answers to specific questions are desired. An interesting issue is how best to treat respondent answers in a survey using traditional binary and fixed-point rating scales (e.g., a 7-point scale). In a

typical questionnaire, a proposition is provided to a respondent and they indicate their agreement with it. Key words in the prompting sentence can be treated as words similar to the unaided, open-ended responses. Or, the entire proposition sentence could be considered as a "word." With later this orientation, the specific items measured with the scale would form the vocabulary, or dictionary of terms from which respondents are allowed express their views. I believe that combining open-ended text with closed-ended questionnaire responses will be a fruitful area of research and application.

A second example involves conjoint analysis studies in which respondents are presented with an array of choice options and asked to verbalize what is on their mind as they make a decision. Verbalized text can be used to inform why a respondent prefers specific attributes and benefits, providing deeper insights into the concerns and interests associated with product usage.

We are on the threshold of a new generation of analysis in marketing that moves beyond tactical, aggregate analysis of purchase records to individual-level analysis that mines the depth of market extremes and provides strategic guidance to firms. I look forward to the future of research in marketing, and believe there is much work to be done. Thank you for listening.