Scholarly Research in Marketing: Trends and Challenges in the Era of Big Data

Ruby Roy Dholakia and Nikhilesh Dholakia
Scholarly Research in Marketing: Trends and Challenges in the Era of Big Data

A shorter version to appear as a Chapter in *International Encyclopedia of Digital Communication & Society*

Ruby Roy Dholakia, Ph.D.
Professor of Marketing and Electronic Commerce
College of Business Administration
The University of Rhode Island
Kingston, R.I. 02881

And

Nikhilesh Dholakia, Ph.D.
Professor of Marketing and International Business
College of Business Administration
The University of Rhode Island
Kingston, R.I. 02881

2013
Introduction

In 1977, Dholakia and Dholakia (1977) painted a futuristic scenario that would have consumers shopping and paying for groceries on their TV screens and the selected items delivered through a pneumatically-propelled chute into the consumer’s kitchen. In 2013, even in technologically advanced countries like USA, Germany or South Korea, the reality does not quite match the optimistic scenario, although it is being approximated along many dimensions. One consequence of technological advances in electronic commerce is the generation of “massive quantities of information produced by and about people, things and their interactions” (boyd and Crawford 2012). According to EMC Corporation, one of the key players in the Big Data business, the key question is “no longer whether big data can help business, but how can business derive maximum results from big data” (EMC² (nd)). For marketing in particular, practitioners as well as academics, an old debate on prediction vs. explanation has become louder as Big Data has increased pressures for integrating conventional data management methods and governance processes. In this chapter, we examine how marketing is responding to these challenges raised by Big Data and the issues surrounding these responses. First, we provide a brief history of marketing thought and how it has influenced research conducted by marketing scholars and practitioners. Next, we outline the impacts of Big Data on marketing practice as well as on marketing research: in business as well as academic settings. The emergent picture is one of increasing technical sophistication but also of increasing expediency, with firms racing to find ways to discover the influence points and buttons that could be pushed to channel consumer behavior away from competitors and towards the firm’s brands. While such goals – of
influencing and channeling consumer behaviors away from competitors and towards – have always been important for business and marketing strategists, what are new in the era of Big Data are the ontological and epistemic shifts (Zwick and Dholakia 2006a, Zwick and Dholakia 2006b), somewhat away from explanation seeking and market/consumer orientation and towards automated prediction and programming of customer responses.

**History of marketing research: past, present and future**

As a relatively new discipline, marketing has traveled far and fast. The first major change occurred when the ‘marketing concept‘ (McKitteric 1957) propelled marketing practitioners and academics alike to put the consumer in the center. Since then, the role and power of the consumer/customer has become central to the evolving history of marketing thought and practice. Companies that have adopted customer-centric organizational structures have experienced improved financial performance, particularly if they compete in different end markets or have large divisions (Lee et al. 2012).

In the ancient and medieval market activities – when the producer was small, local and in direct contact with the buyers – market knowledge emanated from the buyer-seller interactions and there was no need for organized mass marketing. As the scale of production increased, the distance – temporal, spatial as well as informational – between the buyer and the seller increased. It soon became evident that special efforts were required to connect with and satisfy customers. Intuition and personal connections were no longer sufficient to make successful and profitable business decisions, and marketing as a discipline emerged.

Understanding the consumer in an environment of rapidly expanding and changing consumer needs and wants became key to business success and led to development of various
models of market response (Amstutz 1967; Bass 1969; Kotler 1964). Central to these developments were comprehensive databases constructed from multiple data sources: internal transaction records, market surveys and experiments, and secondary sources such as Census demographics. Practitioner research focused on collecting and analyzing data on customer behaviors and preferences which became a key to sustained competitive advantage. Companies, using a variety of methods, collected data, both soft and hard, which were then combined with internal data such as transactional histories. The challenge of integrating these data types was considerable. As Breur (2011) notes, matching datasets – acquired data such as market research data without individual names and transactional data with names – is a mathematical problem but can provide insights. Organizations that have been able to deploy marketing analytics are able to enjoy its benefits, particularly if they operate in a competitive environment and the needs of their customers change frequently (Germann, Lilien and Rangaswamy, 2013).

Marketing scholars contributed concepts and analytical tools that furthered the theoretical, empirical and analytical underpinnings of the discipline. The concept of customer loyalty, for instance, is not only a major goal driving marketing practice but guides and provides measures of marketing effectiveness: “Without question, loyalty is important. Loyal customers hang on for years, devote a larger share of their wallet to the company, and recommend the company to their friends. Customer loyalty, in short, helps drive profits” (Keiningham et al., 2012).

The limitations of the practitioner approaches to customer loyalty quickly became apparent to marketing scholars whose focus on attitudinal as well as the behavioral determinants of loyalty (Jacoby and Chestnut 1978) have led to identification and elaboration of different loyalty concepts such as latent and spurious loyalty (Javalgi and Moberg 1997). This has further
evolved into the concept and measurement of ‘customer lifetime value’ (CLV) which is more in
tune with business goals than short-term transactional benefits. As Kumar and his colleagues
(2013) note, “CLV is a forward-looking metric that does not prioritize loyalty over profit” (p.
340), or as Keiningham and his colleagues note (2012) “When customer value is included in the
measure of loyalty, the goals of improving loyalty and financial performance are synchronized”.
While a valuable concept, CLV fails to incorporate total purchases in a product category since
most managers lack data on their customer purchases from competitors; Chen and Steckel (2012)
proposed a mathematical solution which they believe will allow managers to estimate the
lifetime category value of the customer.

After the rise of customer orientation, a second major change occurred when
technological advances not only altered the competitive context, but transformed the marketer-
customer relationships. Direct response methods such as telemarketing and email marketing as
well as online information and transaction channels expanded consumer choices and contact
points. Direct marketing became possible at low cost. Unsolicited email pitches, for instance,
became cost-effective at very low response rates, as low as 0.5% (Gopal, Walter and Tripathi
2001). Scott (2011) cites the example of Universal Studios Orlando which used seven Harry
Potter ‘super-fan bloggers’ to spread the word about the new Harry Potter attraction, reaching an
estimated 350 million fans at a fraction of the cost of traditional media. One business-to-business
company found that a new customer has 20 “touchpoints” with the company. Pharmacy giant
Walgreen found that a multi-channel user spends six times more than a typical store-only
shopper (Murphy 2013). The growth of direct marketing not only increased contact points
(“touchpoints”) but led to a jump in the variety, velocity, and volume of data.
The explosion of behavioral data, from the multiple touchpoints controlled by the marketer, as well as from external networks such as Facebook and Twitter, is the latest challenge faced by marketing. In this context, marketing has to not only understand consumer motivations but also to demonstrate to the customers that their preferences are remembered (Dholakia and Zwick 2011). It is no longer about a promotion, a script; ‘it’s about a two-way discussion… and every one of us will expect something in return, whether you’re a citizen, or an employee or a customer” (Ginny Rometting, CEO of IBM, quoted in Henschen, 2013). The marketer has to not only gather and analyze data, but has to show consistently and repeatedly that marketing offers are connected to the data generated by individual consumers (Zwick and Dholakia 2012). Personalization and mass customization constitute the new Holy Grail of customer-centric marketing. A review of the approach in the automobile industry notes that the goal is to deliver precisely the car that a customer wants, as and when he wants it: goodbye inventory, hello “mass customisation” (The Economist, 2001).

The explosion of data and the need to develop relevant and competitive insights have created big challenges for data integration and analysis to provide both explanation and prediction. According to McKinsey & Company (2013), the managerial challenge is to determine “which data to use, how to source it and how to get it together into an integrated form”. As Devlin (2012) notes, ‘…marketing analysis has to move from sampling to full dataset analysis, transitioning from demographic segments to “markets of one”, and from longer-term trending of historical data to near real-time reaction to emerging events’ (p. 3). boyd and Crawford (2012) are more critical arguing that all the components of Big Data – technology, analysis and mythology – together “reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information and the nature and categorization of reality”
Similarly, Zwick and Dholakia (2012) argue that “…increase in available consumer data, computational power, and analytical skills have led to a reorganization of the gaze of the marketer… instead of flexibly adjusting production regimes to shifting consumption patterns (what we conventionally refer to as marketing orientation), database marketers collapse the production-consumption dichotomy by manufacturing customers as the actual product (which could be seen as a return to the production concept)” (p. 443).

**Trends and Challenges in Scholarly Research**

Because marketing is considered to be an applied field of study, most of the problems studied by marketing scholars have a managerial focus in that the knowledge generated is ultimately aimed at contributing to managerial practice. Several different routes to address the central concern of understanding customer/consumer behaviors have been adopted by marketing scholars. In general, the approaches can be classified into two categories – those that emphasize explanation and understanding and those that emphasize prediction. These two broad approaches face different challenges in the current context of Big Data.

Those marketing scholars emphasizing the ‘process’ or the ‘black box’ intervening between marketing inputs and consumer responses have focused on ‘explanation’ and ‘understanding’ as the key objectives of scholarly research. Research by these scholars has favored methodologies that use small samples, often conveniently chosen, responding to stimuli carefully constructed and implemented in laboratory settings. For instance, customer satisfaction is a key concept in marketing practice and of central concern to marketing practitioners. Folkes (1984) studied consumer satisfaction in terms of reactions to product failure and her research effort relied on undergraduate students and hypothetical product failure scenarios in order to
develop and test “a theoretical model …. to map out relationships between specific thoughts about product failure and specific complaining behaviors” (p. 398). The conflict between a theoretical model and its practical applications is apparent when the author concludes that the “ability of an attributional approach to predict whether consumer complaining behaviors will actually occur is limited” (p. 409).

The pressures to generalize theory and effects to real-world situations have come from both marketing scholars and particularly from practitioners and students who have to translate marketing theories and market knowledge into specific actions. Much of consumer behavior research is devoted to theory application with high internal validity, notes Winer (1999), but built from data using laboratory experiments and student subjects. He argues that scholarly research “should at least point the way toward more generalization of empirical findings than is usually the case” (p. 352). Devlin (2012) supports this position as he believes that “the ultimate goal is prediction of customer behavior and outcomes of proposed actions such as next best offer” (p. 3). According to O’Driscoll and Murray (2010), ‘synchrony in theory and practice adds value to the management of the enterprise and to the advance of the discipline’ (p.391). Consumer scholars who focus on ‘explanation’ have had to therefore pursue a few additional paths to designing and conducting their research in order to generate scientific as well as managerially useful knowledge.

The first major change in explanatory research was the design of multiple studies in order to expand the domains within which theory applications hold true. Dhar and Nowlis (1999), for instance, examined the general problem of consumer decision making under time pressure. They developed testable hypotheses regarding the factors that affect such decision making, conducted five studies (within one published paper) with multiple product stimuli (binoculars, cordless
phones, cameras, microwave ovens, automobiles, portable computers, TVs, apartments and restaurants) to determine the boundary conditions of their theory. As a result, they were able to draw managerial implications such as “impact of unique features is even greater when decisions are made under time pressure” (p.383) or that “managers should consider not only the attribute values of their brands but also the relationship among the available alternatives” (p. 382).

Lalwani and Shavitt (2013) conducted seven studies to examine price-quality judgments and the influence of self-construal. Despite multiple studies, scholarly articles still emphasized small and undergraduate student samples.

The second major change in explanatory research was to test hypotheses using field generated data. This effort takes two forms. One is to recruit ‘normal’ consumers to participate in the research using a variety of methods. For instance, Banerjee and Dholakia (2014) used a market research agency to recruit a sample of mobile device owners from a national panel with 2 million volunteer participants. Similarly, Chae and Hoegg (2013) recruited members from ‘a national subject pool’ to complete an online study in addition to using student samples to test the theory. To look at the issue of customization, Franke, Keinz, and Steger (2009) conducted two experimental studies using fictitious products with self-administered online questionnaires. Two random samples were drawn from Austria’s leading online panel. Kim (2013) also conducted multiple studies but all the studies tested the hypotheses with consumers who were members of online panels (Study 1 & 2), recruited by a research firm (Study 3), and shoppers interviewed in a convenience store chain (Study 4). Castro, Morales and Nowlis (2013) were innovative in their approach using both undergraduate students and real consumers in their studies but also manipulated the experimental stimuli (shelf display) in an actual store and generated real sales data. The method of recruiting normal consumers has become easier as many online panels –
such as those from Zoomerang and Amazon – have become commercially affordable and are available across multiple countries. Studies using ‘real consumers’ selected or recruited by these methods still rely, however, on relatively small sample sizes.

An alternative approach is to examine marketing and consumer behaviors using data commercially collected by private or research firms. One set of data, proprietary to a specific firm, is often made available to researchers for academic purposes. Kumar and Venkatesan (2005) used data from a large manufacturer of computer hardware and software to examine multichannel buyer behavior. The database was large enough to test the hypotheses among one sample and evaluate the predictive accuracy of the proposed model among a holdout sample with each sample size being over 3,000. Similarly, Dholakia, Zhao and Dholakia (2005) used purchase data from a multi-channel retailer who had recently introduced an internet channel. The database contained information on over 5 million customers and the researchers used specific criteria to select over half a million customer experiences for the analysis.

Other researchers have used customer databases made available widely for academic research. POPAI, an association for the point-of-purchase advertising industry, periodically conducts surveys among shoppers. Inman, Winer and Ferraro (2009) used data collected in 1995 among ‘2300 consumers at 28 grocery stores across 14 geographically dispersed U.S. cities’ to examine in-store consumer decision making. Using data from i-Behavior, a syndicated data aggregator firm, Kushwaha and Shankar (2013) compared profitability of multichannel and single channel customers for two types of product categories. The database had information on 96 million customers of 750 direct marketers and the researchers randomly selected 1 million U.S. customers. The authors feel that “such data are highly representative of the population and allow for empirical generalizability”. In general, these types of studies deal with much larger
sample sizes but are relatively sparse in terms of probing the ‘process’ or ‘black box’ variables that provide richer explanation of the marketing phenomena.

Scholars who emphasize prediction more in their research have frequently built their models and tested them on real data collected by individual firms, research companies, and trade associations. To examine the relationship between marketing mix variables and share of category requirements, Bhattacharya and his colleagues (1996) used two databases compiled by IRI: Marketing Factbook which is a compilation of household-level purchasing data from a panel of 35,000 households in twenty-seven different metropolitan markets, and InfoScan Supermarket Review which provides distribution information on 3,878 brands sold through grocery stores. This research also provides an example of the challenges faced in merging datasets; the authors did not find ‘a perfect correspondence for brands and categories between the Factbook and InfoScan”. To examine the effect of ‘windowing’ in the U.S. motion picture industry, Mukherjee and Kadiyali (2011) built a model of substitution and seasonality and tested it with data from Nielsen VideoScan. Because Walmart, a major retailer of DVDs, was not included in the dataset, the authors note that it is likely that the sample “… understate(s) the importance of larger movies and overstate(s) the importance of smaller movies” (p. 986). The effect of Walmart’s Every Day Low Prices (EDLP) on competitive supermarkets, investigated by Ellickson, Misra and Nair (2012) used data on the supermarket industry from two primary sources: Trade Dimensions TDLinx panel database and Supermarkets Plus Database; the data fields and ranges did not completely overlap.

Increasingly, data collected on the Web is becoming available for scholarly analysis. One approach is to directly collect data available on various sites. Bizrate (www.bizrate.com), for instance, compiles customer ratings on purchases from over one thousand stores. Instead of
using fictitious stores rated by student respondents, Dholakia and Zhao (2010) and Posselt and Gerstner (2005) used Bizrate data to examine the determinants of customer satisfaction and isolate those attributes that influenced the product ordering interaction and those attributes that influenced post-delivery experiences. To examine the effect of online reviews on the sale of books, Chevalier and Mayzlin (2006) collected data – reviews and book characteristics - from the public Web sites of Amazon.com and bn.com. Unable to access sales data from the companies directly, the authors adopted several additional steps to procure data from other sources in order to ‘approximate’ sales from ‘a random sample of 3587 books from Global Books in Print (see www.GlobalBooksInPrint.com)’ and from ‘data on all 2818 titles that appeared in Publishers Weekly best-seller lists (see www.publishersweekly.com)’ for a relevant period of time. Katona, Zubcsek and Sarvary (2011) examined the diffusion process in an online social network in an unnamed country; computational limitations restricted the number of observations (the study had 250,000 users that generated over 13 million friendship links) and also influenced the researchers’ assumptions about relationships among variables. In 2012, Zhu and her colleagues (2012) published their research on online community participation and risky financial behaviors. They investigated their research hypotheses among members of a peer-to-peer online lending community (Prosper.com); then conducted an experiment on a German site of eBay (www.eBay.de) that included analyzing behaviors of eBay customers for a period of over two years in order to test whether results from Prosper.com would generalize to a different online community.

In addition to collecting data from public websites as described above, scholars have been also taking advantage of digital data made available by various firms such as Twitter that collect data. Researchers can access data made available by Twitter (through its API) and access it from
other sources as well. Tweet Scan, for instance, allows users to search public Twitter posts in real time using either a customized search engine or Firefox’s search box. Data aggregators such as Gnip (http://pages.gnip.com/twitter-data/) claim it “offers everything the Twitter API offers, only better” and one can get started with $500 to access tweets including historical data back to 2006 when Twitter was first established.

Bollen, Mao and Zeng (2011) analyzed tweets to predict the stock market. Specifically, they correlate moods with daily Dow Jones Industrial Average (DJIA) prices by focusing on ‘tweets that contain explicit statements of their author’s mood states’ and used a publicly available software – OpinionFinder (OF) – with another mood analysis tool, Google Profile of Mood States (GPOMS), created by the authors themselves. They find specific dimensions of mood included in GPOMS to be more predictive than OF’s bipolar measure of moods. Even though this analysis improves predictive accuracy, it does not provide ‘information on the causative mechanisms that may connect public mood states with DJIA values’ (p.7).

**Trends in and Challenges for Managerial Research**

Marketing practitioners, faced with an entirely different set of problems, practice data mining unevenly. The issue is rarely lack of data; more often the challenge is of finding and analyzing relevant and timely data to generate insights that lead to superior decisions as well as lack of expertise. In 2009, the industry consulting firm Forrester Research reported a deep chasm that limited the use of analytics and a key constraint was ‘human resources required for analyzing data and converting information into necessary action’ (Lovett 2009, p.2). The same constraint was noted in 2013 by McGuire of McKinsey & Company (2013) who notes in an interview, analytics ‘is a highly math-intensive, analytic-modeling exercise. Getting that right,
getting the right skills and capabilities, getting people who really know how to use the latest mathematical techniques and the latest statistical methodology to get inside that data and find the real nuggets of gold’ (p.2) is the second of three challenges faced by management.

In the current context of data explosion in terms of volume, variety, velocity and veracity, the management of data has become an even bigger challenge. Since most of the data mining and management practices by private companies are proprietary in nature, our knowledge is limited to reports and white papers from consulting companies, reports in the business press, and a few academic articles.

The first challenge is mining internally generated data. From the very early years of organized marketing, transactional (behavioral) data is generated as part of the business process. The ability to take advantage of such data, however, has been quite limited. In retail banking, for instance, customers provide a lot of data in the process of creating, managing or closing an account. It has taken some effort to pull data on a customer (across retail banking products) in order to determine profitable product bundles that satisfy customers and build positive relationships. O’Rourke (2012) recounts the success at Wells Fargo Bank which has beaten the industry average and suggests that the practice can be duplicated elsewhere by ‘pulling 4-5 database reports related to product penetration, onboarding cross-sell, customer retention, and profitability’.

The introduction of credit cards and loyalty cards has generated additional internal data that offers marketers opportunities to glean customer insights and tailor offers to specific segments, even specific individuals. Casinos, for instance, have very successfully offered incentives to lure their patrons to spend more time and money (see Giles (2012) for Caesars Entertainment example, p. 17) based on data gathered from their loyalty cards. Supermarkets and
drugstores routinely offer coupons and other incentives based on present and past purchase behaviors. Some companies have been successful in using Business Intelligence (BI) that is built on mostly internal, structured data and business processes. As Mr. Loveman (CEO of Caesars Entertainment and a former academic) observes, the casino using loyalty cards encouraged its customers “to share as much transactional information as possible; then we built inferential models that predict how we could move them best from their current purchase behavior to a more appealing pattern of purchase behavior” (Giles 2012, p. 18).

Generally, it appears that the ability to extract insights from data through meaningful analytics is limited. In a survey of 228 database companies in the Netherlands, Verhoef and his colleagues (2002) report limited use of data modeling. In terms of data availability, internal data—such as purchase history, source of the customer, and the channel of purchase—is more commonly stored than external data such as purchase behaviors at competitors or socio-demographics or lifestyle data, which need to be purchased. The use of statistical techniques is limited to cross-tabulations, RFM (recency, frequency, monetary) counts, and to some extent linear regression and cluster analysis; more sophisticated modeling is positively related to company size and size of the database. In a review of quantitative model research for direct marketing, Bose and Chen (2009) report that while many advanced statistical and data mining techniques are available, researchers mostly use regression analysis and ANN (artificial neural network) modeling techniques. Similarly, Germann, Lilien and Rangaswamy (2013) note that their survey respondents did not indicate extensive use of marketing analytics (average was 3.4 on a 7-point scale (SD=1.6), p. 122).

The addition of company websites provides browsing and clickstream data that differ in terms of volume and velocity of internal data and pose additional challenges. Use of other
digital strategies such as Search Engine Optimization (SEO) requires the ability to exploit digital behaviors. Of course, many of these sites provide the tools (such as Enterprise-wide Web Analytics offered by Google) but additional analysis is required to gain a competitive edge.

The early users of data mining techniques have been Internet-based companies – such as Google, Amazon, Netflix, and Facebook – that collect massive amounts of data as part of their business operations. These companies have been at the forefront of data mining practices and have been able to not only improve their relationships with existing customers but identify new opportunities and products. Facebook runs ‘thousands of different versions of the site at any given moment’ (Henschen 2013) because real-time measurement has been key to coming up with new iterations of products very quickly. Google, by analyzing flu-related search terms, has been able to identify possible flu breaks ahead of official health statistics and Google is attempting to promote the use of tools such as Google Flu Trends by government agencies (Bollier 2010). But the task of mining these enormous volumes of data remains challenging enough that even data mining leaders such as Netflix have sought external talent to develop analytical tools in order to improve their recommendation algorithms (Shmueli 2012).

The second set of data mining challenge occurs with data from external sources such as research commissioned by companies or research purchased from specialized research companies. Survey, panel and scanner data, for instance, is used by consumer goods companies such as Procter and Gamble, Johnson and Johnson, and Unilever. McKinsey & Company (2013), the consulting giant, sees ‘real value in understanding what other data sources are available and bringing external data into play—whether that is weather and climate data, whether that is traffic-pattern data, whether that is competitive data—to understand what other prices are being offered in the market’ (p. 2). Realizing such inherent value is another story. Rossi, DeLurgio and
Kantor (2000) note the challenges faced by marketers in exploiting the benefits of scanner data: “Most manufacturers spend a lot of money to get detailed data on product sales, but they lack the statistical tools necessary to relate the sales numbers to promotional activities at the account level”. No wonder that the Gartner Group estimates that ‘through 2015, 85% of Fortune 500 organizations will be unable to exploit big data for competitive advantage’ (Laney 2012, p. 2).

Merging multiple datasets collected by many different methods poses another challenge. As Breur (2011) notes, matching datasets – acquired data such as market research data without individual names and transactional data with names – is a complex mathematical problem but can provide insights. When datasets generate qualitative data such as from Facebook, Twitter or various blogs, the analytical challenges are magnified. As the use of social networking sites in marketing strategies become common, analyzing the data generated and merging the data with existing data is another hurdle that needs to be overcome. Qualitative data generated by company activities such as conversations on a brand community site pose similar problems. Because of these multiple challenges, opportunities for a variety of data mining services, both large and small, have opened up and have also created opportunities and hurdles for scholars to address these issues. Yahoo, for instance, just signed an agreement with DataSift to provide insight to advertisers on what consumers are saying on the Tumblr site based on both real-time and historical data. DataSift has similar agreements with Facebook and Twitter (Reisinger 2013).

**Critical Issues**

According to Calder and Tybout (1987), consumer research “seeks to produce knowledge about consumer behavior” using whatever methods. The use of laboratory settings is supported
because a theory can be rigorously tested by controlling internal, construct, and statistical-conclusion validity (Calder, Phillips and Tybout 1981). As the preceding discussion on scholarly research has shown, the trend over the years has been to conduct multiple studies, implement field experiments, and collect data from real consumers in an effort by consumer researchers to increase generalizability of their results and develop theories that explain real-world situations. Scholars more interested in prediction have built models and tested them with ‘real’ data; by emphasizing greater statistical generalizability, employing stimuli and settings that provide more ecological validity, these researchers have more directly addressed the issue of managerial relevance. But as Bose and Chen (2009) note, few researchers have incorporated managerial issues such as costs into their direct marketing models or have been able to validate their models with current data. They also note that if real data was made available, because it tends to be large, it creates its own problems: “new concerns emerge such as computational time, software and hardware support, appropriate explanation of results, etc.” (p. 14). As we saw in the case of the study of diffusion process in an online social network, Katona and his colleagues (2011) restrict the number of observations due to computational limitations.

The proliferation of data sources and the massive volumes of data will lead to even more emphasis on quantitative models and use of behavioral data (Bose and Chen 2009; boyd and Crawford 2012). Despite the emphasis on quantitative models, significant differences exist in data collection, analysis and interpretation methods prevalent among marketing scholars and those that are more popular among marketing practitioners. These differences highlight the critical issues that are likely to govern future research by marketing scholars.
Data for Analysis

As ecological validity is increasingly rewarded in scholarly research, it is tempting to use data when external sources promise data from ‘real’ consumers, especially since collecting data from ‘real’ consumers is time consuming and expensive. Such data, usually dealing with large numbers, is not necessarily free of limitations. boyd and Crawford (2011) are critical of large datasets because of the emphasis on quantitative modeling: “data are often reduced to what can fit into a mathematical model” (p. 670), or according to Zwick and Dholakia (2012), consumers are ‘constructed’ (‘manufactured’, from datasets) to match what is on offer. Bose and Chen (2009) find that most researchers that emphasize model building do not spend much effort on data preparation even though research has shown that data preparation methods influence data mining techniques (Crone, Lessman and Stahlbock 2006).

Even when real data is available from companies owning or aggregating large sets of data, access is neither universal and ‘free’ nor ‘representative’ of the relevant population. In describing Facebook’s approach to its data, Varian of Google (in Bollier 2010) comments: “Our basic operating procedure is to come up with an idea, build a simulation and then go out and do the experimentation” (p. 20) and then test hundreds of variations. Only insiders have access to this full set of data. While Twitter’s API makes tweets available for research, it is not “clear what tweets are included in these different data streams or sampling them represents” (boyd and Crawford 2011, p.669). Chen, Chen and Xiao (2013) have tried to show through a simulation approach how estimating social intercorrelations is affected by the sampling method used and the topology of the social network.

Even when marketing scholars have access to company’s internal data, they are restricted to past data after its proprietary value has been diminished. As a result, model validation is
limited to historical data and has lower managerial relevance: “since no data was collected after the implementation of their dynamic multi-mailing models it was not clear how their devised models actually influenced customers’ behaviors” (Bose and Chen, 2009, p.15). When data is incomplete or do not completely overlap, researchers have to adopt various methods to overcome the data limitations. In the case of missing data to calculate share-of-wallet (SOW), Chen and Steckel (2012) tried to combine externally available data (geodemographics) at the zipcode level, even though the credit card data is at the individual level. While their mathematical approach proposes one solution to the missing information problem, it has many other limitations. The authors note the “absence of time-varying covariates such as outstanding balance and credit availability from our dataset. Having these variables would likely have improved both model fit and predictive performance” (p.667).

Furthermore, easily accessible, seemingly ‘public’ data does not free the researchers from considering the ethical issues in using that data: “data may be public (or semi-public) but this does not simplistically equate with full permission being given for all uses” (boyd and Crawford, 2011, p. 673). These issues become aggravated when managerial interests in personalization, customization and measuring the effectiveness of marketing efforts require data at the individual level. “In real applications, it is often required to predict who the next adopters will be and not only provide probability predictions” (Katona, Zucsek and Sarvary 2011, p. 435). Many of these ethical issues are ignored in scholarly research as researchers are eager to work with ‘real’ data.
Analytical Approaches

While academic research in marketing is broader than modeling, the ultimate goal is to answer whether a specific study or stream of research contributes to our understanding of the consumer and the consumer-marketer interactions. This has often been referred to as the explanation vs. prediction debate. Shmueli (2010) values predictive modeling as a scientific endeavor because it can assess ‘the distance between theory and practice’ and notes that a larger sample size is needed for predictive modeling than explanatory modeling. In the current context of Big Data, it would seem this kind of argument would favor predictive and mathematical modeling.

One of the issues raised is regarding the size of the data needed to test any hypotheses. Varian (of Google) comments on company engineers’ temptation to use the entire dataset because a day’s worth of data can be run in half an hour; instead, he recommends a random sample. But as he notes, “the trick is making sure that it’s really a random sample that is representative” (quoted in Bollier 2010, p.15). Very few academic researchers have access to the full dataset to draw a really random sample or the ability to insist that a sample be drawn randomly.

Another issue is regarding use of statistical vs. data mining techniques based on machine learning. In using statistical modeling, researchers have to specify the relationships between variables a priori. Data mining models based on machine learning, on the other hand, extract rules from the data based on algorithms. According to Anderson of Wired magazine, the “real challenge is not to come up with new taxonomies or models, but to sift through the data in new ways to find meaningful correlations…. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world
has ever seen and let statistical algorithms find patterns where science cannot” (quoted in Bollier 2010, p.4-5).

In this new world of ‘throw the data into the hopper and see what comes out’, practitioners have been quite innovative in attracting global talent to detect patterns in the data. Netflix for instance, organized a contest in 2006 to improve its recommendation system (http://netflixprize.com) and offered a prize of one million dollars. A huge dataset of user movie ratings was made public. An Australian company (Kaggle) has been formed to run public competitions to analyze large datasets and such competitions have “attracted more than 3000 statisticians, computer scientists, econometrists, mathematicians and physicists from approximately 200 universities in 100 countries” (Carpenter 2011). Development of analytical solutions using Big Data is now open to a wide variety of disciplines dispersed across the world and without any domain knowledge about marketing. One can even say that in an intensely algorithmic world, machine learning and strategizing wins and conventional creative marketing dies. Meyer (2013) notes that advice for real-word marketing problems is “as likely to be sought from scholars in computer science, economics and psychology as marketing academics, for whom such topics were historically their bread and butter” (p. 1). This is another challenge faced by marketing scholars.

**Theory Building and Knowledge Generation**

The systematic changes in consumer research incorporating multiple data sources that include multiple methods such as surveys, laboratory and field experiments have increased our understanding of market behaviors as well as the underlying psychological processes. Research by Thomas, Simon and Kadilyali (2010) provide an illustrative example. They look at how
buyers perceive and act on real estate prices in terms of precision with which the prices are presented. They conduct five studies that include laboratory experiments using students (Study 1-3), an online study using ‘real’ consumers who are part of Zoomerang’s panel, and analysis of residential real estate transactions data. Liu-Thompkins and Tam (2013) were able to distinguish attitudinal loyalty from habit and used transactional data from a convenience store chain’s loyalty program. They conducted four studies; they were able to operationalize habit strength at the individual level and demonstrate the impact of the two drivers (attitudinal loyalty and habit) on repeat purchase as well as responses to promotional offers. As a result, the authors feel they are able to contribute to both scholarly research and marketing practice.

Quantitative modeling approaches generally spend less effort on causal mechanisms for a variety of reasons. Efforts to include demographic data in direct marketing models have provided poorer predictions than using behavioral data alone (Bose and Chen 2009) which further promotes the development and use of better and more predictive analytical techniques. In advocating the use of Behavioral Analytics, Montibeller and Durbach (2013) emphasize its “prescriptive, action-orientated analysis, aimed at improving individual and organizational decision making” (p. 16). Academic research that improves predictions – even if it fails to show the causal mechanisms – would not necessarily be a disadvantage to practitioners, with their emphasis on prediction.

Scholarly research in marketing has made great strides from its early focus on developing tools and techniques to generating knowledge and understanding but the availability of Big Data and its emphasis on predictive modeling is putting pressures on these developments. This highlights the differences within the scholarly community as well as the distance between
academia and practice. There are additional considerations that affect these differences in goals and approaches.

First, within the scholarly community itself, there are some fundamental differences in approaches to generating market insights. Cayla and Arnould (2013) advocate ethnographic stories as an explanatory (not exploratory) approach to ‘unpack and comprehend more abstract forms of market knowledge’ (p.9) and to complement other research approaches. Their interviews with commercial ethnographers indicate that this approach can ‘flesh out the multidimensionality of consumer lives’ and create understanding of market complexity. This approach is, of course, based on small rather than large sample sizes and goes against Big Data and data mining techniques. It also goes against the conventional consumer research approach based on laboratory and field experiments.

Second, despite efforts to improve predictability of market behaviors and the availability of large sets of data, marketing practitioners face problems calculating the return on their investment (ROI); the availability of large datasets has not made it any easier because they lack ‘an integrated system to manage the necessary data’ (Marketingcharts.com 2013). As Devlin (2012) notes, “integration of data from a variety of sources, both traditional and new, with multiple tools, is the first prerequisite. A well-integrated process around all data, both big and small, is a further necessity to extract business value from information” (p.4). Davenport and Patil (2012) recognize the complexity of the challenge and see the value of the data scientists ‘who make discoveries while swimming in data’. It is clear, however, no single discipline is sufficient to address the challenges. Devlin (2012) suggests a team approach that includes “one from the business, one who understands analytics and likes to play detective, and an IT expert
from the BI team, who can access data from the EDW and integrate it into the new big data technologies” (p. 9).

Where does the role of marketing academic lie in this partnership? We noted some of the issues concerned with access to internal data for scholarly purposes. It is not common practice for organizations to share insights about proprietary research and these insights are closely guarded. When academics operate as consultants, there are restrictions on public sharing of knowledge gained from their experiences. The barriers grow when the pressure is to generate ‘now’ knowledge such as real time forecasting of consumption and these pressures conflict with traditional academic focus on data quality and other methodological issues. This has been emphasized by boyd and Crawford (2012): “Just because Big Data presents us with large quantities of data does not mean that methodological issues are no longer relevant. Understanding sample, for example, is more important now than ever” (p.668).

**Concluding Observations**

Information technologies – Internet, mobile communications, electronic commerce, social media, geographical positioning systems, and more – are reshaping markets and marketing in fundamental ways. Just by leading their day-to-day lives, consumers leave a massive trail of data points, every day and often every minute. Soon, the ‘Internet of things’ will generate data in terms of volume and velocity that will tax the ability of companies and scholars to process such data. The challenges of dealing with Big Data - to understand, to explain, and to predict the behavior of consumers from millions of data points – will be magnified.

These challenges exist at various levels. Business practitioners want ways to deal with the data deluge so that they can devise competitively superior ways to approach and influence
consumers. Research firms seek methods – based on statistics, artificial intelligence, machine learning, neural networks, algorithms, indeed based on ‘whatever works’ – to uncover patterns lurking in the tsunami of Big Data. Academic researchers, especially in marketing, struggle to balance the needs of in-depth understanding, causal explanation, and predictive accuracy; and these competing epistemic demands are fragmenting the academic research fields and often creating end-runs around well established academic research traditions, replacing decades long research programs with clever algorithmic devices or machine learning methods that “work” (in predictive terms) without achieving understanding or explanation.

In the near future, the challenges of Big Data will be met by a variety of expedient methods, but over a longer term, the need for in-depth understanding and explanation will reassert itself. This is because consumers are humans and they are not quite programmable – they change, adapt, stray, experiment, shift gears, get bored, contradict themselves, and more. For the academic research enterprise in marketing, the challenge thus is not to keep sticking to traditional ways of researching but to seek new ways to collaborate with multiple science-based fields such as computer science, biotechnology, and neuroscience; as well as with other knowledge fields such as cultural anthropology, philosophy, and the arts. It is only through such collaborations that we would achieve fuller understandings of how people, products and processes interact with one another.
References


Banerjee, Syagnik (Sy) and Dholakia, Ruby Roy (2014) Situated or Ubiquitous? A Segmentation of Mobile Internet Shoppers *Int. J. Mobile Communications*, forthcoming.


Bollen, Johan, Mao, Huina and Zeng, Xiao-Jun ( ) Twitter mood predicts the stock market. Available at: http://arxiv.org/pdf/1010.3003&


Liu-Thompkins, Yuping and Tam, Leona (2013) Not all Repeat Customers are the same: Designing Effective Cross-Selling Promotion on the Basis of Attitudinal Loyalty and Habit. Journal of Marketing, 77 (September), 21-36.


Zhu, Rui (Juliet), Dholakia, Utpal M., Chen, Xinlei (Jack) and Algesheimer, René (2012), Does Online Community Participation Foster Risky Financial Behavior? Journal of Marketing Research, XLIX (June), 394–407.


Founded in 1892, the University of Rhode Island is one of eight land, urban, and sea grant universities in the United States. The 1,200-acre rural campus is less than ten miles from Narragansett Bay and highlights its traditions of natural resource, marine and urban related research. There are over 14,000 undergraduate and graduate students enrolled in seven degree-granting colleges representing 48 states and the District of Columbia. More than 500 international students represent 59 different countries. Eighteen percent of the freshman class graduated in the top ten percent of their high school classes. The teaching and research faculty numbers over 600 and the University offers 101 undergraduate programs and 86 advanced degree programs. URI students have received Rhodes, Fulbright, Truman, Goldwater, and Udall scholarships. There are over 80,000 active alumnae.

The University of Rhode Island started to offer undergraduate business administration courses in 1923. In 1962, the MBA program was introduced and the PhD program began in the mid 1980s. The College of Business Administration is accredited by The AACSB International - The Association to Advance Collegiate Schools of Business in 1969. The College of Business enrolls over 1400 undergraduate students and more than 300 graduate students.

**Mission**

Our responsibility is to provide strong academic programs that instill excellence, confidence and strong leadership skills in our graduates. Our aim is to (1) promote critical and independent thinking, (2) foster personal responsibility and (3) develop students whose performance and commitment mark them as leaders contributing to the business community and society. The College will serve as a center for business scholarship, creative research and outreach activities to the citizens and institutions of the State of Rhode Island as well as the regional, national and international communities.

The creation of this working paper series has been funded by an endowment established by William A. Orme, URI College of Business Administration, Class of 1949 and former head of the General Electric Foundation. This working paper series is intended to permit faculty members to obtain feedback on research activities before the research is submitted to academic and professional journals and professional associations for presentations.

An award is presented annually for the most outstanding paper submitted.