The Role of Big Data and Predictive Analytics in Retailing

Forthcoming, *Journal of Retailing*

Eric T. Bradlow
The K.P. Chao Professor
Professor of Marketing, Statistics, Economics and Education
Faculty Director - Wharton Customer Analytics Initiative
University of Pennsylvania, Philadelphia, PA 19104
Email: ebradlow@wharton.upenn.edu

Manish Gangwar
Assistant Professor of Marketing
Indian School of Business (ISB)
Gachibowli, Hyderabad-50032, India.
Email: manish_gangwar@isb.edu

Praveen Kopalle[1]
Associate Dean for the MBA Program
Signal Companies' Professor of Management
Professor of Marketing
Tuck School of Business at Dartmouth
Dartmouth College, Hanover, NH 03104
Email: kopalle@dartmouth.edu

Sudhir Voleti
Assistant Professor of Marketing
Indian School of Business (ISB)
Gachibowli, Hyderabad-50032, India.
Email: sudhir_voleti@isb.edu

[1] The author(s) thank the guest editors, Dhruv Grewal, Anne Roggeveen, and Jens Nordfält, two anonymous JR reviewers, participants at the 2014 Digital Summit, Indian School of Business, 2014 Nordic Wholesale and Retailing Conference, 2015 Marketing Symposium, Tuck School of Business at Dartmouth, and 2016 Wharton Retailing Conference for their thoughtful comments on earlier drafts of this manuscript. The author(s) also thank an anonymous retail chain for providing us access to their confidential data for our analysis.

The Role of Big Data and Predictive Analytics in Retailing

Abstract

The paper examines the opportunities in and possibilities arising from Big Data in retailing, particularly along five major data dimensions - data pertaining to customers, products, time, (geo-spatial) location and channel. Much of the increase in data quality and application possibilities comes from a mix of new data sources, a smart application of statistical tools and domain knowledge combined with theoretical insights. The importance of theory in guiding any systematic search for answers to retailing questions, as well as for streamlining analysis remains undiminished, even as the role of Big Data and predictive analytics in retailing is set to rise in importance, aided by newer sources of data and large-scale correlational techniques. The Statistical issues discussed include a particular focus on the relevance and uses of Bayesian analysis techniques (data borrowing, updating, augmentation and hierarchical modeling), predictive analytics using big data and a field experiment, all in a retailing context. Finally, the ethical and privacy issues that may arise from the use of big data in retailing are also highlighted.

Introduction

According to some estimates, Walmart collects around 2.5 petabytes (1 petabyte = 1,000,000 gigabytes) of information every hour about transactions, customer behavior, location and devices (McAfee et al., 2012). An IT analyst firm Gartner estimates that there will be 20 Billion (13.5 Billion in the consumer sector) devices connected in the "Internet of Things". Imagine the amount of data that will be generated by these devices (Gartner, 2015). Imagine a day where online and offline retailing data provide a complete view of customer buying behavior, and even better if the data is linked at the level of the individual customer to enable "true" customer lifetime value calculations (Gupta et al., 2006; Venkatesan and Kumar, 2004). Imagine a day where data thought only to exist in online retailing, e.g. consumer path data (Hui, Fader, Bradlow 2009a), exists inside the store due to RFID and other GPS tracking-based technologies. Imagine a day where integrated online/offline experiments are being run that provide exogenous variation that enables causal inference about important marketing/retailing topics such as the efficacy of email, coupons, advertising, etc. (Anderson and Simester, 2003). Imagine a day where eye-tracking data isn't just collected in the laboratory from Tobii-enhanced monitors but is collected in the field due to retinal scanning devices embedded within shelves (Lans, Pieters, and Wedel et al, 2008; Chandon et al, 2008).

As futuristic as those data sources sound, all of them exist today (albeit not ubiquitously) and will soon be part of the information that marketing scientists (within and outside of retail) use for customer-level understanding and firm-level optimization. Simply and heuristically put, these data sources will be adding "columns" to our databases (and a lot of columns!) that provide an increased ability to predict customer behavior and the implications of marketing on it. Now, add that to the technology (i.e. IP address tracking, cookie tracking, registered-user log-in, loyalty card

usage, to name just a few) which enables firms to collect this from millions of customers, for each and every moment, linked to each and every transaction, linked to each and every firm-level touchpoint, and linked across distribution platforms, and we have the big data that pervades the popular press today.

While the lure (and lore) of big data is tempting, in this paper we posit that the big data revolution (McAfee et al. 2012) really is a "better data" revolution, and especially so in retailing. Our intent in this paper is to describe the newest forms of data (i.e. "new columns") that exist in retailing, the importance of experimentation and exogenous variation ("better columns"), to describe why data mining and machine learning (despite their obvious value) will never obviate the need for marketing/economic theory (i.e. "where to look in the data"), to describe how managerial knowledge and statistical methods can lead to smart data compression (i.e. "which columns" and summaries of them) that will enable researchers to shrink the data, how better data will feed into predictive models (e.g. CLV, diffusion, choice models), and how firms are likely to use these models for decision making. This framework (both the buckets and the order of them) mirrors the INFORMS (www.informs.org) definition of business analytics which includes descriptive analytics, predictive analytics and prescriptive analytics.

Wedel and Kannan (2016) provide an excellent commentary on marketing analytics past, present and future. Guided by one of Marketing Science Institute's (www.msi.org) top research priorities, they discuss how marketing analytics will shape future decision making by managers in the area of customer relationship management, marketing mix allocation, personalization, customer privacy and security issues. In contrast, our aim in this paper is to highlight the challenges and opportunities facing retailers dealing with big data. The rest of the paper is organized as follows. In the next three sections, we discuss the nature of "big" data in retailing, compare it with

"better" data, and describe new sources of data that also leads to better models. This is followed by a discussion of the importance of theory in the analysis of retailing and various statistical issues involved such as data compression, statistical sufficiency for modeling, and the role of Bayesian inference. Finally, we present results of a case study, i.e., a field experiment that combines predictive analytics and optimization in retailing.

Big Data in Retailing

This section describes "typical" sources of big data in retailing and how there is potential to exploit the vast flows of information in a five-dimensional space: across customers, products, time, geo-spatial location, and channel. We present them in Figure 1 and discuss each in turn.

[Insert Figure 1 here]

Customers

When most people think of big data, they think of data sets with a lot of rows, and they should. Tracking technologies have enabled firms to move from aggregate data analyses which dominated marketing attribution and effectiveness studies when data was limited (Dekimpe and Hanssens, 2000) to individual-level data analyses that allows for much more granular targeting (Rossi, McCulloch, and Allenby 1996). In fact, one could argue that one of the big missions of a firm is to grow the number of rows (via customer acquisition, i.e. more unique IDs) and more transactions per customer with greater monetary value (per row). In retailing, the ability to track new customers and to link transactions over time is key. Loyalty programs (Kopalle et al. 2012; Stourm et al. 2015), widespread today, are the most common way that such tracking exists; however, credit card, IP address, and registered user log-ins are also commonplace. Besides more rows, firms also have much better measures (columns) about each row which typically, in retailing,

might include a link between customer transaction data from a CRM system, demographic data from credit card or loyalty card information, survey data that is linked via email address, and in-store visitation information that can be tracked in a variety of ways. If one includes social media data and more broadly user-generated content (UGC) which can be tracked to individual-level behavior, then customer-level data becomes extremely rich and nuanced.

Products

Product information in marketing, has and likely always will be, defined by a set of attributes and levels for those attributes which define the product. However, in today's data rich environment we see an expansion of product information on two-dimensions. First, this information may be available now for hundreds of thousands of SKUs in the store, making the data set about products have a lot of rows in it. Second, the amount of information about each product need not be limited now to a small set of attributes thus increasing the column-width, if you will, about the product information matrix. Product information along these two dimensions alone (at the store level) can enable a host of downstream analyses - such as that of brand premiums (e.g., Ailawadi, Lehmann and Neslin 2003, Voleti and Ghosh 2013), or of product similarities and thereby grouping structures and subcategory boundaries (e.g., Voleti, Kopalle and Ghosh 2015). Thus, we expect that going forward, retailers will have product information matrices that are both dynamic (see below), and much more descriptive allowing for greater variation of product varieties that are micro-targeted (Shapiro and Varian 2013) towards consumers. Furthermore, since more attributes and levels can be collected about each product, this will allow retailers to gain an understanding of products that were never modeled (in Marketing) before (e.g. experiential goods), because they consisted of too many attributes, or hard to measure attributes, to allow for a parsimonious representation.

Time

While the large data sets described in the above "customer" and "product" pieces may seem large, imagine the third-dimension - "time" which literally multiplies the size of this data. That is, while historical analyses in retailing has looked at data aggregated to monthly or possibly weekly level, data in retailing today comes with a time stamp that allows for *continuous* measurement of customer behavior, product assortment, stock outs, in-store displays and environments such that assuming anything is static is at best an approximation. For example, imagine a retailer trying to understand how providing a discount, or changing the product location changes the flow of customers in the store, how long customers spend at a given store location, what they subsequently put into their shopping basket and in what order? A database that contains consumer in-store movements connected to their purchases (Hui et al 2009a, 2009b, 2009c) could now answer this question because of the time dimension that has been added. In addition, due to the continuous nature with which information now flows to a retailer, the historical daily decision making about inventory levels, re-stocking, orders, etc. aren't granular enough and real-time solutions that are tied directly to the POS systems and the CRM database are now more accessible. In other words, real-time is a compelling option for many firms today.

Location

The famous quote about *"delivering the right message to the right customer at the right time"* has never been truer than in the era of big data. In particular, the first two components (the right message and the right customer) have been a large part of the copy testing, experimental design (e.g. A/B testing) and customized marketing literature for at least the past 40 years. However, the ability to use the spatial location of the customer (Larson, Bradlow and Fader 2005) at any given point in time has opened up a whole new avenue for retailers where customer's geo-

spatial location could impact the effectiveness of marketing (Dhar and Varshney 2011), change what offer to make, determine at what marketing depth to make an offer, to name just a few. When the customer's geo-spatial location is also tied to the CRM database of a firm, retailers can unlock tremendous value where a customer's purchase history (Kumar et al, 2008) is then tied to what products they are physically near to allow for hyper-targeting at the most granular level. However, while this hyper-targeting is certainly appealing, and short-term revenue maximizing, retailers will need to consider both the ethical and potential boomerang effects that many customers feel when products are hyper-localized (e.g., Fong, Fang and Luo 2015).

Channel

This century has seen a definitive increase in the number of channels through which consumers access product, experience, purchase and post-purchase information. Consequently, consumers are displaying a tendency to indulge in 'research shopping', i.e. accessing information from one channel while purchasing from another (Verhoef et al, 2007). This has led to efforts to collect data from the multiple touch points (i.e. from different channels). The collection, integration and analysis of such omni-channel data is likely to help retailers in several ways: (i) understanding, tracking and mapping the customer journey across touch-points, (ii) evaluating profit impact, and (iii) better allocating marketing budgets to channel, among others. Realizing that information gathering and actual purchase may happen at different points of time, and that consumers often require assistance in making purchase decisions, firms now started experimenting on relatively newer ideas like Showrooming - wherein the customer searches in the offline channels and buys online (Rapp et al, 2015), and Webrooming - where the customer's behavior is the opposite. Proper identification and attribution of channel effects thus gains importance and in

this vein, Li and Kannan (2014) propose an empirical model to attribute customer conversions to different channels using data from different customer touch points.

In summary, big data in retailing today is much more than more rows (customers). When one takes the multiplicity of People x Products x Time x Location x Channel data, this is big data. Retailers that have the ability to link all of these data together are ones that will be able to not only enact more targeted strategies, but also measure their effects more precisely. Next, we compare big versus better data (and the potential for better models) and argue that a better data revolution should be the focus of retailers (and others).

Big Data versus Better Data and "Better" Models

The challenges with big data, computationally and housing/hosting/compiling it, are well-established and have spawned entirely new industries around cloud-computing services that allow for easy access and relatively inexpensive solutions. However, although this provides a solution to the big data problem if you will, the problem that the data which is being stored and housed may be of little business intelligence value remains.

For example, imagine a large brick-and-mortar retailer with data that goes back a few decades on each and every customer, which composes an enviable and very rich CRM database. In fact, since the data history at the individual customer level is so rich, the retailer feels extremely confident in making pricing, target marketing, and other decisions towards his/her customer base. "Wow, this retailer really has big data!" However, what we might fail to notice is that much of the data on each individual customer is "old". The data doesn't reflect the needs and wants of the customer anymore, or in the parlance of statistics, a change point (or multiple), has happened. Thus, the retailer with big data actually has a mixture of "good data" (recent data) and "bad data"

(old data) and the mixture of the two makes their business intelligence system perform poorly. More data is not going to solve this problem, and in fact, may exacerbate the problem.

Imagine a second example where the same brick-and-mortar retailer has extensive data on in-store purchases, but is unable to link that data to online expenditures. In such a situation, the retailers loses out on the knowledge that many customers want to experience the goods in-store, but purchase online; hence the retailer underestimates the impact of in-store expenditures as only the direct effects are measured. Thus, the retailer already has big in-store data but without linking (i.e. data fusion, e.g., Gilula, McCulloch and Rossi 2006) of the data to online behavior, the retailer has good, but incomplete data.

Third, and maybe this example is one of the most common in marketing today, imagine that a retailer has a very rich data set on sales, prices, advertising, etc., so that big data is achieved. That is, there is an abundant sample size that will allow the retailer to estimate sales as a function of prices and advertising to any precision that he/she wants. Thus, the retailer estimates this function, utilizes that function within a profit equation, and sets optimal advertising and price and is now "done", right? Well, no. The challenge with this data, while big, is that the past prices and advertising levels that are set are not done so randomly and therefore the retailer is making a decision with what are called endogenously set variables. That is, when thinking about past managers setting the prices, the past manager will set prices high when he/she thinks that it will have little impact and vice-versa (i.e. in periods of low price elasticity). Thus, since this data does not contain exogenous variation, the retail manager erroneously finds that price elasticities are low and he/she raises prices only to find that the customers react more negatively than he/she thought. Thus, the retailer doesn't have a scarcity of data, the retailer has a scarcity of "good data" with exogenous variation (experimental-like randomly assigned data if you will). In fact, as with the

second example, this is a case where big data will exacerbate the problem as the big data makes the retailer very confident (but erroneously) that customers are price inelastic.

In summary, these three examples/vignettes are meant to demonstrate that longer time series are not necessarily better as researchers commonly make an assumption of stationarity. Data that is excellent but incomplete may provide insights under one marketing channel but would fail to inform the retailer of the total effect of a marketing action. Finally, data that is big data but does not contain exogenous sources of variation can be misleading to the retailer and suggests why experimental methods (A/B tests, e.g., Kohavi et al. 2012) and/or instrumental variables methods (Conley et al. 2008) have become popular tools to "learn from data". Next, we describe more relevant data.

New sources of data

New research insight often arises either from new data, from new methods, or from some combination of the two. This section turns the focus on retail trends and insight possibilities that come from newer sources of data. In recent times, there has been a veritable explosion of data flooding into businesses. In the retail sector, in particular, these data are typically large in volume, in variety (from structured metric data on sales, inventory or geo-location to unstructured data types such as text, images and audiovisual files), and in velocity i.e., the speed at which data come in and get updated - for instance sales or inventory data, social media monitoring data, clickstreams, RFIDs etc.), thereby fulfilling all three attribute criteria of being labeled "Big Data" (Diebold 2012).

Prior to the 80s, before UPC scanners rapidly spread to become ubiquitous in grocery stores, researchers (and retailers) relied on grocery diaries where some customers kept track of the what, when and how much of their households' grocery purchases. This data source, despite its

low accuracy, low reliability and large gaps in information (e.g., the prices of competing products and sometimes even the purchased product would be missing), gave researchers some basis in data to look at household purchase patterns and estimate simple descriptive models of brand shares. In contrast, consider the wide variety of data sources, some subset of which retailers rely on, available today. Figure 2 organizes (an admittedly incomplete) set of eight broad retail data sources into three primary groups, namely, (1) traditional enterprise data capture; (2) customer identity, characteristics, social graph and profile data capture; and (3) location-based data capture. At the intersection of these groups, lie insight and possibilities brought about by capturing and modeling diverse, contextual, relevant (and hence, "better") data.

[Insert Figure 2 here]

The first type arises from traditional sales data from UPC scanners combined with inventory data from ERP or SCM software. This data source, marked #1 in Figure 2, enables an overview of the 4Ps (product, price, promotion and place at the level of store, aisle, shelf etc.). One can include syndicated datasets (such as those from IRI or Nielsen) also into this category of data capture. Using this data, retailers (and researchers) could analyze market baskets cross-sectionally - item co-occurrences, complements and substitutes, cross-category dependence etc (see, e.g., Blattberg et al. 2008; Russell and Petersen 2000); analyze aggregate sales and inventory movement patterns by SKU (e.g., Anupindi, Dada and Gupta 1998 in a vending machine scenario); compute elasticities for prices and shelf space at the different levels of aggregation such as category, brand, SKU etc (e.g., Hoch et al. 1995 on store-level price elasticities; Bijmolt et al. 2005 for a review of this literature); assess aggregate effects of prices, promotions and product attributes on sales; etc. These analyses are at the aggregate level because traditional enterprise data

capture systems were not originally set up to capture customer or household level identification data.

The second type of data capture identifies consumers and thereby makes available a slew of consumer- or household-specific information such as demographics, purchase history, preferences and promotional response history, product returns history, basic contacts such as email for email marketing campaigns and personalized flyers and promotions etc. Such data capture adds not just a slew of columns (consumer characteristics) to the most detailed datasets retailers would have from previous data sources, but also rows in that household-purchase occasion becomes the new unit of analysis. A common data source for customer identification is loyalty or bonus card data (marked #2 in Figure 2) that customers sign up for in return for discounts and promotional offers from retailers. The advent of household specific 'panel' data enabled the estimation of household specific parameters in traditional choice models (E.g., Rossi and Allenby 1993; Rossi, McCulloch and Allenby 1996) and their use thereafter to better design household specific promotions, catalogs, email campaigns, flyers etc. The use of household- or customer identity requires that a single customer ID be used as primary key to link together all relevant information about a customer across multiple data sources. Within this data capture type, another data source of interest (marked #3 in Figure 1) is predicated on the retailer's web-presence and is relevant even for purely brick-and-mortar retailers. Any type of customer initiated online contact with the firm - think of an email click-through, online browser behavior and cookies, complaints or feedback via email, inquiries etc. are captured and recorded, and linked to the customer's primary key. Data about customers' online behavior purchased from syndicated sources are also included here. This data source adds new data columns to retailer data on consumers' online search, products viewed

(consideration set) but not necessarily bought, purchase and behavior patterns, which can be used to better infer consumer preferences, purchase contexts, promotional response propensities etc.

Another potential data source, marked #4 in Figure 2, is consumers' social graph information, obtained either from syndicated means or by customers' volunteering their social media identities to use as logins at various websites (such as publishers, even retailers' websites). The increasing importance of the 'social' component in data collection, analysis, modeling and prediction can be seen in all four stages of the conventional AIDA framework - Awareness, Interest, Desire and Action (see, e.g., Dubois et al. 2016 on how social graphs influence awareness and word of mouth). Mapping the consumer's social graph opens the door to increased opportunities in psychographic and behavior-based targeting, preference and latent need identification, selling, word of mouth, social influence, recommendation systems which in turn herald cross- and up-selling opportunities, etc (e.g., Ma, Krishnan and Montgomery 2014; Wang, Aribarg and Atchade 2013). Furthermore, the recent interest to extend classic CLV to include "social CLV" which indicates the lifetime value a customer creates for others is certainly on the forefront of many companies' thoughts.

A third type of data capture leverages customers' locations to infer customer preferences and purchase propensities and design marketing interventions on that basis. The biggest change in recent years in location-based data capture and use has been enabled by customer's smartphones (e.g., Ghose and Han 2011, 2014). Data capture here involves mining location-based services data such as geo-location, navigation and usage data from those consumers who have installed and use the retailer's mobile shopping apps on their smartphones. Figure 2 marks consumers' Mobiles as data source #5. Consumers' real-time locations within or around retail stores potentially provide a lot of context which can be exploited to make marketing messaging on deals, promotions, new

offerings etc more relevant and impactful to consumer attention (see, e.g., Luo, Andrews, Fang and Phang 2014) and hence to behavior (including impulse behavior). Mobile-enabled customer location data adds not just new columns to retailer data (locations visited and time spent, distance from store etc) but also rows (with customer-purchase occasion-location context as a feasible, new unit of analysis), and both together yielding better inference on the feasibility of and response propensity to marketing interventions. Another distinct data source, marked #6 in Figure 2, draws upon habit patterns and subconscious consumer behaviors which consumers are unaware of at a conscious level and are hence unable to explain or articulate. Examples of such phenomena include eye-movement when examining a product or web-page (eye-tracking studies starting from Wedel and Pieters 2000), the varied paths different shoppers take inside physical stores which can be tracked using RFID chips inside shopping carts (see, e.g., Larson, Bradlow and Fader 2005) or inside virtual stores using clickstream data (e.g., Montgomery et al. 2004), the distribution of first-cut emotional responses to varied product and context stimuli which neuro-marketing researchers are trying to understand using fMRI studies (see, e.g., Lee, Broderick and Chamberlain 2007 for a survey of the literature), etc. Direct observation of such phenomena provides insights into consumers' "pure" preferences untainted by social, monetary or other constraints. These data sources enable locating consumer preferences and behavior in psychographic space and are hence included in the rubric of location-based data capture.

Data source #7 in Figure 2 draws on how retailers optimize their physical store spaces for meeting sales, share or profit objectives. Different product arrangements on store shelves lead to differential visibility, salience, hence awareness, recall and inter-product comparison and therefore differential purchase propensity, sales and share for any focal product. Slotting allowances (e.g., Lariviere and Padmanabhan 1997) and display racks testify to the differential sales effectiveness

of shelf space allocation, as do the use of planogram planning software, computation of shelf space elasticities (Curhan 1972, 1973) and field experiments to determine the causal effect of shelf space arrangements on sales (e.g, Dreze et al. 1995). More generally, an optimization of store layouts and other situational factors both offline (e.g., Park, Iyer and Smith 1989) as well as online (e.g., Vrechopoulos et al. 2004) can be considered given the physical store data sources that are now available. Data source #8 pertains to environmental data that retailers routinely draw upon to make assortment, promotion and/or inventory stocking decisions. For example, that weather data affects consumer spending propensities (E.g., Murray et al. 2010) and store sales has been known and studied for a long time (see, e.g., Steele 1951). Today, retailers can access a well-oiled data collection, collation and analysis ecosystem that regularly takes in weather data feeds from weather monitoring system APIs, collates it into a format wherein a rules engine can apply, and thereafter outputs either recommendations or automatically triggers actions or interventions on the retailer's behalf. One recent example wherein weather data enabled precise promotion targeting by brands is the Budweiser Ireland's Ice cold beer index (promotions would be proportional to the amount of sunshine received in the Irish summer) and the fight-back by local rival Murphy's using a rain-based index for promotions (Knowledge@Wharton, 2015). Another example is Starbucks which uses weather-condition based triggering of digital advertisement copy[1].

Finally, data source #9 in Figure 2 is pertinent largely to emerging markets and lets small, unorganized sector retailers (mom-and-pop stores, for instance) to leverage their physical location and act as fulfillment center franchisees for large e-tailers (Forbes 2015). The implications of this data source for retailers are very different from those in the other data sources in that it is partly B2B in scope. It opens the door to the possibility (certainly in emerging markets) for inter-retailer

[1] http://www.weatherunlocked.com/resources/the-complete-guide-to-weather-based-marketing/weather-based-marketing-strategies

alliances, co-ordination, transactions and franchise-based relationships (provided the retailers in question are not direct competitors, of course). For example, Amazon currently partners with local retailers as distribution centers which allows for same day delivery of many SKUs via predictive analytics and therefore advanced shipping locally.

Better Models

The intersection between the customer specific and location based data capture types enables a host of predictive analytics and test-and-learn possibilities, including more sophisticated predictive models that were previously unavailable because they were "data under-supplied". For instance, a customer's past purchase history, promotional-response history and click-stream or browsing history can together inform micro-segmentation, dynamic pricing and personalized promotions for that customer that could be inferred from tree-based methods that allow for the identification of complex interaction effects. The advent of geo-coding, whereby consumers can be identified as belonging to well-defined geo-codes and hence can be targeted with locally content sensitive messages (e.g., Hui et al. 2013), geo-fencing whereby consumer locations are tracked real time within a confined space (usually the store environment) and targeted promotions are used (e.g., Luo et al. 2014; Molitor et al. 2014), geo-conquesting (Fong, Fang and Luo 2015) whereby retailers will know consumers are moving towards rival retailers and can entice them with offers at precisely that point to lure them away, etc. points to the use of technology to mesh with consumers' locational context to build relevance and push purchase and other outcomes. However, the use of location-based targeting and geo-fencing etc. is predicated on the availability of real time statistical models such as Variational Bayes approaches (see, e.g. Beal 2003; Dzyabura and Hauser 2011) that allow for computation fast enough to make location-context exploitable. However, given the large numbers of variables, the rapid speed of analysis and processing and the

automatic detection of effects now attainable, the question remains about how relevant theory would be in such a world. Despite the advances in machine learning and predictive algorithms, a retail manager's decision making will be far from being fully automated. As we describe next, the role of theory in retailing is to help "navigate big data" and this today may be more crucial than ever.

Theory driven Retailing

Big data provides the opportunity for business intelligence, but theory is needed to guide "where to look" in the data and also to develop sharp hypotheses that can be tested against the data. Predictive algorithms essentially rely on past observations to connect inputs with outputs, (some) without worrying about the underlying mechanism. But rarely does one have all the inputs that affect outcomes managers are interested in influencing, consequently when machine learning is used as a black box approach, without a sound understanding of the underlying forces that drive outcomes, one typically finds it to be inadequate for predicting outcomes related to significant policy changes. This is where theory can guide managers. Theory helps put structure on the problem so that unobserved and latent information can be properly accounted for while making inferences. The data mining approach of - analyzing granular data, generating many correlations at different aggregation levels, visualizing and leveraging the power of random sampling - may not be adequate in our interconnected, omni-channel world. Managers should strive to understand the underlying *cause* of emerging trends. They not only need to understand what works under the current scenario but also why it works, so that they know when something may not work in other contexts. Theory (and as we discuss below empirical exogenous variation) enables managers to uncover cause-and-effect, to identify the real drivers of outcomes, the underlying factors producing new trends, and to parse out spurious patterns. In the rest of this section, we emphasize three points

regarding Marketing theory's role in big-data retailing, and illustrate each with one or more examples.

First, there are risks inherent in ignoring theory and relying entirely on a data driven approach. Often managers sitting on a lot of data not knowing what to do, and may fall into the trap of apophenia, i.e., human tendency to perceive meaningful patterns and correlations in random data. The predictions purely generated from patterns have limitations of going beyond the learnings from a training set. A good case in the point is the Google flu detector, an algorithm intended to predict new or as yet unrecorded flu cases in the US faster than the Center for Disease control (CDC) that potentially can help retailers manage their inventory better. This algorithm, using vast troves of search data (Ginsberg et al. 2009) was essentially a theory-free pattern matching exercise. The popular press bubbled with claims that big data had made traditional statistical techniques to establish causality obsolete and some even declared the end of theory (Anderson, 2008). Later a deeper analysis showed that these 'theory-free' projections were over-predicting actual cases by 100%, among other problems that arose (see Lazer et al. 2014 for a review). It is therefore crucial that in most cases, when possible, out-of-sample validation been used as a first measuring stick for the business intelligence and possible causal interpretation given to predictive models. This case (among others) also serves to caution that theory-free predictions based merely on correlation patterns are fragile. Alternative approaches, such as the use of structural dynamic factor-analytic models on (for instance) Google trends data, may provide better results. Du and Kamakura (2012) analyze data of 38 automobile makers over a period of 81 months and find the aforementioned approach useful for understanding the relationship between the searches and the sales of the automobiles.

Second, theory could be useful in identifying and evaluating the data required to mitigate inconsistencies and biases in model estimates that mislead decision-making efforts. Consider the issue of endogenous variables. This is one of the areas where theoretical considerations have helped identify (and avert) bias in what might otherwise look like perfectly fine data for analysis, prediction, and optimization. In a nutshell, the endogeneity problem arises when nonrandom variables (which were either set strategically or are the outcome of processes that are not random) are treated as random variables in a statistical model. Consider a manager looking to optimize advertising promotions. We know that the effectiveness of an ad depends on whether the ad is viewed or not. However, the data pertaining to whether an ad is actually viewed is seldom collected owing to logistical and practical reasons. Modeling ad effectiveness without this crucial piece of data (on actual viewership) biases model estimates and severely curtails the usefulness of decisions based on the data. But now, armed with the knowledge of what data are required, the manager can choose to explore options such as using an eye-tracking technology to record responses from a random sample of customers. Subsequently, with data on actual ad viewership coupled with statistical inference techniques that project responses from the random sample to the population at large, the manager can obtain better estimates of advertising effectiveness, and accordingly make better decisions. Note that in this instance, more data (greater volume) would not have helped mitigate the problem. But greater variety of data (the eye-tracking information) does help mitigate the bias. Thus, the answer to the endogeneity challenge critically depends on our understanding of the sources of endogeneity, which in turn requires the theoretical underpinnings of the phenomenon of interest.

Third, theory itself is not set in stone but is routinely updated as underlying trends and foundations change. For example, after observing pricing patterns in IRI marketing dataset, that

significantly differ from theoretical predictions, Gangwar, Kumar and Rao (2014) updated classical price promotion theory. The new theory not only matches the empirical observations better but also provides useful guidelines to mangers on how to accommodate important considerations (consumer stockpiling) while developing promotional strategies. Highlighting frameworks which bridge theory and practice, remain relevant and useful. For example, consider the AIDA framework which has influenced theory and practice for decades that has helped managers understand the customer purchase process. Below, we show how a technology-savvy retailer (Starbucks, in this example) uses some of the data sources in Figure 2 to better understand the consumer purchase process and boosts its sales and profits. In the process, the old AIDA theory is updated to accommodate changes in consumer behavior and motivations.

The central idea in AIDA is that the consumer's awareness of (or attention to) a need, often by way of exposure to advertising, is the first step of the purchase process. This 'Awareness stage' precedes the construction of a 'consideration set' of plausible solutions (brands, products, services) that fulfill the need (the 'Interest' stage). Next follows an evaluation of the consideration set of products, the formation (or uncovering) of preferences over these products (this constitutes the 'Desire' stage) and finally, the purchase itself (the 'Action' stage, which may sometimes also include actions such as word of mouth propagation and recommending the focal product to others). Traditionally, the AIDA was envisioned as a linear process, progressing from one stage to the next with marketing interventions exerting influence at each stage. However, recent work on customer "journeys" and the availability of new data (Van Bommel, Edelman and Ungerman 2014) is helping firms discover the non-linearity of the AIDA process in today’s interconnected word.

Today's consumer in a retail setting could be exposed to latent needs, product ideas or recommendations etc. from a wide variety of sources, not just advertising. One example relates to

consumers (almost instinctively) searching the web for solutions to needs or problems. Web-search results potentially expose the consumer to a large variety of need-solutions (perhaps cutting across product categories) which might otherwise not have occurred via traditional advertising or serendipitous discovery on store shelves. Other examples include product recommendations made by recommendation systems, exposure to particular brands within one's social circle etc. Previously, the 'Action' stage resulted in hard data - purchase - that could be recorded and analyzed as secondary data. Anything short of purchase was, conventionally, not recorded. Thus, data pertaining to the other stages (A, I and D) had to be obtained through primary means for limited samples in target segments. However, the advent of big data and the possibilities that arise from concatenating data from a variety of sources enable the creation of a better, and a more complete picture of the typical customer's journey. This in turn, enables retailers to better deploy marketing efforts via targeted marketing interventions and thereby realize better returns on marketing investments because the path to purchase, and where it is “stopped” is now better understood.

[Insert Figure 3 here]

In all but high-involvement product categories, situational (or immediate "context") factors surrounding a consumer gain importance in determining response propensities to marketing interventions aimed at different AIDA stages. For example, in mobile advertisements and couponing, Andrews et al. (2015) use a series of field experiments to study mobile ad effectiveness under "hyper-contextual" situational factors such as 'crowdedness' on subway trains; Luo et al. (2014) study the effects of contexts such as current distance-from-store of consumers and the time-lag between receiving a promotion and its activation; Bart and Sarvary (2014) study the effect of product category type on consumer attitudes and intentions etc. To illustrate context-based targeting, consider the mobile coupon promotion shown in Figure 3. This mobile coupon appeared

on mobile-screens in a few relevant and neighboring geocodes around a Starbucks store in central London. The attempt to build relevance and thereby enhance response propensities through the use of three distinct contextual factors can be seen in Figure 3. First, time is used as a contextual factor ("It's lunchtime!"), followed by a (mobile) printable promotional code for a 10% discount to spur purchase, and finally directions to the store (for locational context and relevance). If customers are already in the Awareness stage of AIDA, then the receipt of such a time-sensitive and contextual ad could potentially move them through the I, D, and A stages fairly rapidly. Note that we may add other contextual factors to Figure 3 based on the availability of more information about the target customer. For instance, if the target customer is a sociable office-goer and has come to Starbucks in the past as part of a larger group, then it could offer a group discount ("Groups of 4 or more get 15% off!"). Or if the customer is known to be a soccer fan, the ad could further say, “Catch the Manchester United vs Barcelona match at Starbucks", etc. The data sources, technology, business processes and ecosystems required to send such contextually targeted ads to customers are currently available and rapidly gaining currency (e.g., Nesamoney 2015). On the data front, piecing together data sources for hyper-contextual targeting would likely involve big data analytics given the volume, variety, and dynamics (velocity) of data involved.

Statistical Issues in Big Data and Retailing

When dealing with big data in retailing, a number of statistical issues arise including two key ones: data compression and Bayesian inference for sparse data. As data volumes rises, so do the time and cost (effort, resources) required to process, analyze and utilize it. But armed with the knowledge of the business domain, theory and statistical tools, practitioners and researchers can shrink the volume of the data - 'smart' data compression - mitigating the cost and time needed from

analysis to decision making without much loss of information. That is, when cleverly done, big data can be made smaller.

With regards to Bayesian inference, as we gain more data on certain dimensions (more rows), we still may have sparse data at the level of the individual retail customer (not enough columns if you will) which Bayesian methods handle naturally by borrowing information from the population of other users for which we may have abundant information. That is, the promise of big data science typically lies in individual-level targeting; but, in some cases (e.g. especially for new customers), the data supplied doesn't match the data needed and methods which optimally combine a given customer's information with that from the population is needed. We now describe both data compression and Bayesian estimation in retailing in detail.

Data Compression

Data compression, shrinking the volume of data available for analysis can broadly be classified into two types: (i) technical, which deals with compressing file formats and originates in computer science, and (ii) functional, which shrinks data volume using econometric tools and originates in Statistics. Functional data compression can further be subdivided into (a) methods that transform the original data, and (b) methods that provide a subset of the original data without any transformations (e.g. sampling-based approaches).

In signal processing, data compression (also referred to as source coding, or bit-rate reduction) involves encoding information using fewer bits than the original representation. Compressing data enables easier and faster processing of signals above a minimum quality threshold. This definition extends very well to the business context, particularly in the era of big data. From a logistical point of view, data compression reduces storage requirements while

simultaneously enhancing other capabilities such as "in-memory" processing, function evaluations, and the efficiency of backup utilities.

More importantly, from a modeling perspective, data compression of a functional nature enables output and insights of the same quality and sharpness as could be had from the original, uncompressed big data. This is because models are at some level probabilistic representations of the data generation process which include the behavior and responses of units of analysis (model primitives) to various (external) stimuli. A small dataset having the same information content as a larger one would allow simpler, more elegant models to be estimated and yield estimates that are at least as good as those from the larger dataset. Also, simpler models tend to be more robust, more intuitive and often more extensible (or generalizable) across data and problem contexts, a well-known issue in econometrics. Hence it is no surprise that a slew of methods to reduce the dimensionality of the data have been proposed and applied in the marketing literature.

Accordingly, functional data compression involves either a transformation of the original data in some form, or more commonly, the use of sampling procedures to obtain and use subsets of the untransformed original dataset for analysis, or some combination of the two. A simple categorization of data transformation for functional data compression could be in terms of (i) the extent of information loss entailed, and (ii) the particular dimension being compressed. Regarding the extent of information loss entailed, compression can be either what is called lossy or lossless. Lossy compression involves a cost-benefit evaluation of keeping versus discarding certain data values, whereas lossless compression reduces data-bits by identifying and eliminating statistical redundancy. In either case, identifying statistically redundant or partially redundant bits of data would be key, and this often requires managerial domain knowledge in addition to statistical tools. Regarding the particular dimension being compressed, data compression could act either on the

variables (data columns) or the rows (records, or observations of the units of analysis). Figure 4 shows the different data compression categorizations.

[Insert Figure 4 here]

Consider the following example. An online retailer has precise clickstream data - on page views and navigation, access devices, preferences, reviews, response to recommendations & promotions, price elasticities, as well as actual purchases - on a large number of identified customers, as well as point of sale data for the entire product assortment. We know that a data column's information content is both proportional to the variance of the data in the column, as well as its correlation with other variables in the data set. For example, if the price of an SKU in a category shows zero variation in the data (i.e., is constant throughout), then the information content is minimal - the entire column could be replaced by a single number. If the SKU sales show variation despite no variation in price, then the sales figure is explained by factors other than price. Thus, price would be redundant information that does not explain sales. If on the other hand, both price and sales show variation, then a simple correlation between these two columns could signal the extent to which these quantities are 'coupled' together. If they were perfectly coupled, then both data columns would move perfectly in tandem (correlation would be either +1 or -1) and deletion of either would result in no loss of information. We next consider data compression first among the columns (variables) and then the rows (units/customers).

Consider data reduction along the variables (or column) dimension. Traditional techniques such as principal components and other factor analytic methods attempt to uncover a latent structure by constructing new variables ('factors') as functions of constituent, (tightly) inter-correlated variables. These are lossy in that not all the information in the original uncompressed dataset is retained and some is invariably lost (unless the variables are perfectly linearly dependent

in which case the compression would be lossless). Thus, one can consider Principal components methods as a way to navigate between lossy and lossless data compression methods. However, the researcher must evaluate and decide if the compressed representation should be adopted, the original uncompressed one retained or consider some mix between the two which in most cases turns out being more of a business domain problem and less of a statistical one. High uniqueness scores for particular variables would imply that they be considered independent factors in their own right whereas the factor solution would serve to dimension-reduce the other variables into a handful of factors. In big data contexts, especially with the advent of unstructured data such as text, images or video, with thousands of variables and (typically) no ready theory to reliably guide variable selection, understanding and visualizing latent structure in the data columns is a good first step to further analysis. For instance, in text analytic applications, large text corpora typically run into thousands of words in the data dictionary. Many if not most of these words may not be relevant or interesting to the problem at hand. Methods for text dimension reduction such as latent semantic indexing and probabilistic latent semantic indexing, text matrix factorization such as the latent Dirichlet allocation (Blei et al. 2003) and its variants have since emerged and become popular in the Marketing and Retailing literatures (e.g. Tirunellai and Tellis 2015).

Many variable selection techniques have emerged that attempt a data-driven (and theory-agnostic) way to identify redundant variables - or variables that contribute nothing or almost nothing to explaining dependent variables of interest. Examples include stepwise regression, ridge regressions, the LASSO (e.g., Tibshirani 1996), the elastic net (e.g., Rutz, Trusov and Bucklin 2011) as well as stochastic variable search methods. All of these methods attempt to automatically identify variables that correlate with an outcome of interest, but also retain parsimony.

Now, consider data reduction along the row dimension. This can be achieved in broadly one of two ways - (i) grouping together "similar" rows and thereby reducing their numbers, or (ii) using sampling to build usable subsets of data for analysis. Traditionally, some form of cluster analysis (e.g., the k-means) has been used to group together units of analysis that share similarities along important variables (called basis variables), based on some inter-unit distance function in basis variable space (see Jain 2010 for a review of clustering methods). Thus for instance, the online retailer's customers who share similar preferences, behavior and/or purchase patterns could be grouped together into a 'segment' for the purposes of making marketing interventions.

Sampling reduces data dimensionality in a different way. A sample is a subset of population data whose properties reveal information about population characteristics. Because different samples from the same population are not likely to produce identical summary characteristics, assessing sampling error helps quantify the uncertainty regarding the population parameters of interest. Thus for instance, analyzing a random sample of a few hundred customers' purchase patterns (say) in a set of categories from the online retailers' population of tens of thousands of customers could throw light on the purchase patterns of the population as a whole. The key to sampling, however, is that the probability that a unit is included in the sample is known and preferably (and strongly so) *not* related to the parameter of interest. In those cases where being in the sample is related to the underlying quantity of interest (e.g. shoppers who use a website more often are more likely to be in the sample), then the sampling mechanism is said to be non-ignorable (Little and Rubin, 2014) and then more sophisticated analysis methods are needed. It is for these reasons that analysts take probability samples, even if not simple ones with equal probability, as it allows us to generalize easily from the sample to the population of interest.

Bayesian Analysis and Retailing

Although Bayes theorem is over 250 years old, Bayesian analysis became popular in the last two decades, at least partly due to the increased availability of computing power. The Bayesian paradigm has deeply influenced statistical inference in (and thereby, the understanding of) natural and social phenomena across varied fields. In particular, its promise for retailing, where the ever-increasing desire to provide optimal marketing decisions at the level of the individual customer (Rossi and Allenby 1993), has never been higher. That is, individual-level customization, en-mass, is no longer a dream of retailers, and Bayesian methods sit at the heart of that dream. We now lay down three properties and advantages of Bayesian analysis and inference that are relevant from a big data retailing perspective. Where applicable, we relate these properties to our discussion on data compression as Bayesian methods may allow less information loss with smaller datasets.

Bayesian Updating

The big advantage of Bayesian analysis is that it has the inherent ability to efficiently incorporate prior knowledge and the sharing of information across customers. This is particularly useful when analysts have to deal with a large amount of data that updates frequently. While re-running a model every time on full dataset (updated with new data) is time consuming and resource intensive, Bayesian analysis allows researchers to update parameters at any point of time without re-running the model again on the full dataset. Any new information that comes in the form of new data can be easily accommodated by running a model on just the new dataset, and the old knowledge can be simply incorporated as a prior (albeit the representation of the old analyses as a prior can be non-trivial), thus reducing potential time and required resources. This can be seen as a form of efficient data use and compression where the old data/analyses are represented as a prior.

Hierarchical modeling and household or individual level parameter estimation

Unlike in other disciplines where researchers are typically interested in estimating averages (and for whom individual level heterogeneity is a nuisance to be controlled for), Marketers place a premium on the analysis, inference and understanding of individual level heterogeneity in any sample (Allenby and Rossi 1998). This allows marketers to then efficiently group individuals or households into a manageable number of segments and thereafter design effective segment-specific marketing campaigns and interventions; or individual-level ones if cost effective. One of the challenges in inferring individual level parameters is that marketers typically do not have enough data to estimate individual level parameters in isolation; albeit more columns as discussed in this research is changing that to some degree. However, this challenge plays into a big advantage of Bayesian analysis - estimation of and inference over individual level parameters by 'borrowing' data from other units of analysis. In this manner, as data is being used more efficiently, smaller data sets can yield the same level of precision as larger ones; hence, a form of data compression. Today researchers in marketing (and more widely, in the social sciences) have widely accepted and adopted the Bayesian as a preferred tool of estimation.

Data augmentation and latent parameter estimation

A third advantage of Bayesian estimation techniques comes from its ability to simplify the estimation procedure for latent parameter models. By efficiently using data augmentation techniques, Bayesian estimation avoids resource intensive numerical integration of latent variables (Tanner and Wong 1987). For example, the Probit model doesn't have a closed form solution. Consequently, the most popular "frequentist" method for its estimation – the GHK simulator - relies on costly numerical integration but the Bayesian model vastly simplifies the estimation procedure (McCulloch and Rossi 1994). The data augmentation approach has been very useful in estimating Tobit models with censored and truncated variables (Chib 1992) and has shown

tremendous promising in detecting outliers and dealing with missing data. From a retailer's perspective, having an augmented data set can help answer business problems that were hard to do otherwise such as understanding relationships among preferences of related brands.

In summary, the future of big data and retailing will necessitate the creation of data compression and statistical methods to deal with ever increasing data set sizes. When done in a "smart way", this can yield significant benefits to retailers who want actionable insights but also those that can be run in real time. In the next section, we describe a case of predictive analytics in retailing that exemplify the use of statistics in big data to estimate model parameters and the corresponding results of price optimization that show a managerially significant enhancement in retailer profitability.

Predictive Analytics and Field Experimentation in Retailing

"What gives me an edge is that others go to where the ball is whereas I go to where the ball is going to go."—Pele, world soccer player

In this section, we provide a real-life application at a retail chain that ties together our earlier discussion on big data in retailing, the sources of data, role of theory and the corresponding statistical issues—including Bayesian inference, with the objective of enhancing retailer profitability. In this regard, we report the results of a field experiment that evaluates whether the use of predictive analytics in terms of price optimization increased the profitability of retail stores owned by one particular chain, which chooses to remain anonymous.

Predictive analytics in retailing is part of business intelligence – it is about sensing what's ahead – but alone does not provide firms the insights that they need. To support our claim, we

present a case study of a pricing field experiment conducted at a large national retail chain involving forty-two stores, randomly allocated equally to test and control.

Note that typical predictive analytics in practice are typically (i) Simply extrapolative, for example, moving average methods, i.e., next month sales = average of last three months' sales, (ii) Judgmental, e.g., Delphi method, and/or (iii) Explanatory: lack the incorporation of causal variables. Causal research purports to uncover the causes, the "why" of the effects of interest. One major challenge in empirical research is ensuring that the independent or input variables actually be exogenous. This necessitates (or at least is the most straightforward way) that a controlled test (i.e., a test of the probable cause or 'treatment') be carried out to measure the effect of a deliberately (and hence, exogenously) varying treatment on outcomes of interest and compared against those for a 'control' group that was not exposed to the treatment. Although a lab environment helps exercise better control of environmental factors, in a marketing context especially, it may influence the outcomes themselves. Hence, a field experiment would be preferable in retailing contexts (Sudhir 2016). We discuss a field experiment next that describes an application of randomization and model-based inference in retailing.

Figure 5, reproduced from Levy et al. (2004), depicts a flowchart that the case follows to enable a customer-based predictive analytics and optimization system for pricing decisions. This figure lays out the steps involved in applying big data to achieve retail objectives, identifies the sources of data—in this case, store level panel data combined with pricing data from competing stores, the importance of a theory driven approach in terms of determining the demand function and incorporating the psychological aspects of pricing, strategic role of competition, and the objective to be maximized—in this case, the overall store profitability.

[Insert Figure 5 here]

The approach outlined in Figure 5 highlights the key role field experimentation plays when firms use available data to make managerial decisions. Consider, for example, the recent case of J. C. Penney, which attempted to change its pricing structure based on its analysis of available data. We believe that J. C. Penney's new pricing strategy backfired because of the following two reasons: (1) Lack of a broader experimental testing with its customer base at its retail stores and (2) Not taking into consideration psychological aspects of pricing from a customer's point of view. Below, we describe the field experiment along with the big data that were used, the theory-driven demand model, incorporation of psychological aspects of pricing, and the corresponding price optimization.

Field Experiment

In this field experiment, we partnered with a large national chain in the United States that prefers to remain anonymous. Utilizing an A/B testing approach, we chose forty-two stores that were divided randomly into test versus control. For our randomization to provide greater balance, we ensured that all the selected stores were similar in terms of store size, demographics of clientele, annual sales etc. Fourteen product categories, covering 788 SKUs were selected. The main criterion for selecting the categories and the corresponding SKUs within each category was the existence of significant variation in their price history. One hundred and two weeks of data were used to estimate the model (described below) parameters and prices then were optimized for a period of thirteen weeks. Twenty-one stores were used as test stores where our recommended prices were implemented in those stores whereas the rest of the stores were used as control where the store manager set prices on a "business-as-usual" basis. Table 1 lists the categories and the number of SKUs optimized in each category.

[Insert Table 1 here]

Table 2 lists the various categories studied, along with the number of SKUs in each category, the number of SKUs optimized, and the total number of observations.

[Insert Table 2 here]

Basic Econometric Model. Our predictive econometric model allows a retailer to estimate demand for a given SKU in a given store for a given time period using a number of input variables including price, feature, and display. The basic model is a standard logit-type, aggregate-based SKU-level, attraction model (Cooper and Nakanishi, 1989; Sudhir 2001). The objective in the modeling framework is to understand price and promotion elasticity after accounting for all other factors including seasonality. The model formulation for market share (MS) is given by:

$$MS_{i,c} = \frac{e^{u_{i,c}}}{\sum_{i,c \in s} e^{u_{i,c}}}$$

where the attraction of SKU i in category c given by $u_{i,c} = \sum_{k} \beta_{i,c,k} X_{i,c,k}$, where u_{ic} is the deterministic component of consumer utility, and **X** is the vector of independent variables, which consists of price (additional details below), promotion, and product features that are all under the control of the retailer. In addition, we model market share (include in u) using time trends, seasonality, demand shifters and special events (such as snowstorms in the Northeast, a new store opening, or a new highway being built) which allows for time-dependent demand in the estimation process. Details of the variables operationalization are available upon request.

Incorporating Psychological Aspects of Pricing. The demand model we use is also modified to include basic psychological aspects of pricing. The key psychological aspect of pricing that we use in our estimation is that of reference price effects (Greenleaf 1995; Kopalle, Rao, and Assunção 1996; Kopalle et al. 2012). Reference prices are modeled as an exponential smoothed average of

past prices. If the reference price in time t is greater (less) than the observed price in time t, it is considered a loss (gain). Accordingly, we estimate the corresponding gain and loss parameters. In other words, our model captures the reference prices as an exponentially smoothed average of past price (Kopalle, Rao, and Assunção 1996) and takes into consideration what the unit sales (reference sales) would be when the price is at parity with the reference price. The model parameters are estimated using maximum likelihood estimation. Based on the estimated parameters, we can easily compare the forecasted sales versus actual sales over time. Figure 6 provides a summary performance measure of our demand-estimation model in three categories. In the first stage, 75% of the data are used to estimate the model. Second, the parameters obtained in the first stage are used to forecast unit sales in the remaining 25% of the data. Third, the forecasted sales are compared with the actual data during the forecast time in order to check the validity of the forecasts. The results indicate an excellent fit with out-of-sample R^2 ranging from 76.3% to 88.8%, as well as superior fit to a multiple regression benchmark.

[Insert Figure 6 here]

Price Optimization. In the next step, we optimized prices over the aforementioned thirteen week out-of-sample period by maximizing total profitability across all categories and SKUs. We incorporated various constraints including limits on the margins and price changes, price families (for example, pricing all six-packs of Coca-Cola products similarly), appropriate gaps between store brands and national brands, same pricing within price zones, etc. The levels of other independent variables (feature and display) were kept at their observed levels. Thus, the optimization problem for each category may be summarized as follows:

$$\underset{price_{it},}{Max}[\sum_{i}\sum_{t}\{(p_{it}-c_{it})S_{it}\}]$$

where subscripts *i* and *t* denote product and time (week or month) respectively, *P* is price, S is unit demand, and c is unit cost.

The optimized prices for the various SKUs were implemented in the twenty-one test stores. At the end of each week, per the algorithm in Figure 5, the econometric model was re-estimated and prices re-optimized for the following week. The test ran for 13 weeks. The results were analyzed using a difference model where the gross margin dollar of SKU i in week t was the dependent variable and the key independent variable was whether the observation was from a test store relative to control. The analysis includes many control variables including category dummies, average gross margin in the three-month period before the test, whether the SKU's price was optimized, category purchase frequency, unit share, and dollar share.

The results (Table 3) show that there is a significant improvement in the gross margin dollar per SKU per week of 40.7 cents ($p < .01$). Table 4 extrapolates the corresponding profit enhancement to the enterprise level with 10,000 SKUs per store and one hundred stores.

[Insert Tables 3 and 4 here]

The key message from our field experiment is that model-based elasticity price optimization improved gross margin dollars both managerially and statistically significantly at the test stores over the control stores and unit sales at approximately the same levels as the control. This combination of experimentally generated exogenous variation, statistical modeling, and optimization (in this order), we hope is a significant part of the future of business intelligence.

Conclusion

One goal of this paper is to examine the role of and possibilities for big data in retailing and show that it is improved data quality ('better' data) rather than merely a rise in data volumes that drives

improved outcomes. Much of the increase in data quality comes from a mix of new data sources, a smart application of statistical tools and domain knowledge combined with theoretical insights. These serve also to effect better data compression, transformations and processing prior to analysis. Another goal is to examine the advent of predictive analytics in a retailing context. Traditionally theory-agnostic predictive analytics tools are likely to have larger impact and lesser bias if they are able to smartly combine theoretical insights (akin to using subjective prior information in Bayesian analysis) with large troves of data. Hence overall, whereas the role of big data and predictive analytics in a retailing context is set to rise in importance aided by newer sources of data and large-scale correlational techniques that of theory, domain knowledge and smart application of extant statistical tools is likely to continue undiminished.

Ethical and Privacy Issues

There is a need for self-regulation on part of profit-maximizing firms which use Big Data lest litigation, a PR backlash and other such value-diminishing consequences result. A few recent examples bring out this point very well. Retailer Target's analysts, in an effort to pre-emptively target families expecting a newborn, used Market Basket Analysis to identify a number of products bought typically by pregnant women, developed a pregnancy status prediction score for its female customers, and thereafter used targeted promotions during different stages of pregnancy (Duhigg 2012). In a controversial turn of events, Target knew about and sent coupons related to a teenager's pregnancy even before her family members were aware of the same. The resulting furor created a lot of negative publicity for Target. Orbitz, a travel website, showed higher priced hotels for customer searches originating from Apple computers than those from PCs. Products like Amazon Echo keep listening for the keywords from the customer once it gets activated when customer say 'Wake'. The product then starts recording the audio and streams it to a cloud server. A related concern is that if such products start recording the private conversations and determine the presence of the people in the house. A key concern is how long the primary collectors hold the

data and resell the same to other aggregators/providers. A clear opt-out policy should be displayed prominently on the websites and apps, especially for the companies with default opt-in policy. In October 2016, the Federal Communications Commission (FCC) voted in favor of the privacy rules which require an explicit permission of the user before the websites/apps start collecting the data on web browsing, app usage, location and email content. Anonymization or masking of the Personally Identifiable Information (PII) should be a top priority for the companies. Companies might have to adhere to the ethical standards as mentioned in the Menlo Report (Dittrich, 2012).

The use of big data and predictive analytics in retailing will raise underlying ethical and privacy issues. Government intervention in the form of new regulations to protect consumer privacy is a distinct possibility. The New York Times story about Target's use of data analytic techniques to study the reproductive status of its female customers has certainly raised the level of debate about big data in retail versus consumer privacy. Retailers may proactively address consumer privacy and the corresponding ethical issue via three ways: (1) Allow a clear opt-in policy for their customers with respect to collecting and using their data. For example, almost all loyalty programs are on an opt-in basis, (2) Show the benefits of predictive analytics to their customer base. For example, customers of Amazon.com find it harder to switch from Amazon because they clearly see the benefits of the personalized recommendation system at Amazon, and (3) Reward loyalty, i.e., it should be obvious to customers that a retail store rewards customer loyalty. For example, when Amazon ran a pricing experiment many years ago, there was consumer backlash to that experiment because Amazon was not rewarding loyalty, i.e., it offered lower prices to its switching segment and higher prices to its more loyal segments. Thus, while work remains to be done in the realm of ethics and consumer privacy, the concept of strategic usage of big data for the benefit of both retailers and its customers seems viable and worthy of the effort required to more fully understand it.

References

Ailawadi, K. L., Lehmann, D. R., & Neslin, S. A. (2003). Revenue premium as an outcome measure of brand equity. *Journal of Marketing*, 67(4), 1-17.

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1), 57-78.

Anderson, Chris. "The end of theory." *Wired magazine* 16, no. 7 (2008): 16-07.

Anderson, Eric T. and Duncan Simester. 2003. Effects of $9 Price Endings on Retail Sales: Evidence from Field Experiments. *Quantitative Marketing and Economics*. 1(1): 93-110.

Andrews, M., Luo, X., Fang, Z., & Ghose, A. (2015). Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. Marketing Science.

Anupindi, Ravi, Maqbool Dada, and Sachin Gupta. "Estimation of consumer demand with stock-out based substitution: An application to vending machine products." Marketing Science 17, no. 4 (1998): 406-423.

Bart Y, Stephen A, Sarvary M (2014) Which products are best suited to mobile advertising? A field study of mobile display advertising effects on consumer attitudes and intentions. Journal of Marketing Research. 51(3):270–285.

Bijmolt, Tammo HA, Harald J. van Heerde, and Rik GM Pieters. "New empirical generalizations on the determinants of price elasticity." Journal of marketing research 42, no. 2 (2005): 141-156.

Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. "Market basket analysis." Database Marketing: Analyzing and Managing Customers (2008): 339-351.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning Research*, 3(Jan), 993-1022.

Chandon, Pierre, J. Wesley Hutchinson, Eric T. Bradlow, and Scott H. Young, ―Measuring the Value of Point-of-Purchase Marketing with Commercial Eye-Tracking Data, in *Visual Marketing: From attention to action*, edited by Michel Wedel and RikPieters (New York: Lawrence Erlbaum, Taylor & Francis Group, 2008), 223-258.

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51(1-2), 79-99.

Conley, T. G., Hansen, C. B., McCulloch, R. E., & Rossi, P. E. (2008). A semi-parametric Bayesian approach to the instrumental variable problem. Journal of Econometrics, 144(1), 276-305

Cooper, L. G., & Nakanishi, M. (1989). Market-share analysis: Evaluating competitive marketing effectiveness (Vol. 1). Springer Science & Business Media.

Curhan, R. C. (1972). The relationship between shelf space and unit sales in supermarkets. Journal of Marketing Research, 406-412.

Curhan, R. C. (1973). Shelf space allocation and profit maximization in mass retailing. The Journal of Marketing, 54-60.

Dekimpe, Marnik G., and Dominique M. Hanssens. "Time-series models in marketing:: Past, present and future." *International Journal of Research in Marketing* 17.2 (2000): 183-193.

Dhar, Subhankar, and Upkar Varshney. "Challenges and business models for mobile location-based services and advertising." *Communications of the ACM* 54.5 (2011): 121-128.

Dhar, S. K., Hoch, S. J., & Kumar, N. (2001). Effective category management depends on the role of the category. *Journal of Retailing*, 77(2), 165-184.

Diebold, Francis X., On the Origin(s) and Development of the Term 'Big Data' (September 21, 2012). PIER Working Paper No. 12-037. Available at SSRN: http://ssrn.com/abstract=2152421 or http://dx.doi.org/10.2139/ssrn.2152421

Dittrich, David, and Erin Kenneally. "The Menlo report: Ethical principles guiding information and communication technology research." *US Department of Homeland Security* (2012).

Dreze, Xavier, Stephen J. Hoch, and Mary E. Purk. "Shelf management and space elasticity." Journal of Retailing 70, no. 4 (1995): 301-326.

Dubois, David, Andrea Bonezzi, and Matteo De Angelis. (2016) "Sharing with Friends versus Strangers: How Interpersonal Closeness Influences Word-of-Mouth Valence." Journal of Marketing Research (Forthcoming).

Fong, Nathan M., Zheng Fang, and Xueming Luo (*2015*) Geo-Conquesting: Competitive Locational Targeting of Mobile Promotions. *Journal of Marketing Research*: October 2015, Vol. 52, No. 5, pp. 726-735.

Forbes (2015), From Dabbawallas To Kirana Stores, Five Unique E-Commerce Delivery Innovations In India, Accessed 15 April 2015, http://tinyurl.com/j3eqb5f

Gangwar M, Kumar N and Rao RC (2014), Consumer Stockpiling and Competitive Promotional Strategies. *Marketing Science*, 33(1):94-113

Gartner "Gartner Says 6.4 Billion Connected Things Will Be in Use in 2016, Up 30 Percent From 2015", Gartner Press Release, November 10, 2015, http://www.gartner.com/newsroom/id/3165317, accessed 31-10-2016

Ghose A, Han SP (2011) An empirical analysis of user content generation and usage behavior on the mobile Internet. Management Science, 57(9):1671–1691.

Ghose A, Han SP (2014) Estimating demand for mobile applications in the new economy. Management Science, 60(6):1470–1488.

Gilula, Zvi, Robert E. McCulloch, and Peter E. Rossi. "A direct approach to data fusion." *Journal of Marketing Research*, 43, no. 1 (2006): 73-83.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data." *Nature* 457, no. 7232 (2009): 1012-1014.

Greenleaf, Eric A. "The impact of reference price effects on the profitability of price promotions." *Marketing science* 14, no. 1 (1995): 82-104.

Gupta, Sunil, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V. Kumar, Nathaniel Lin, Nalini Ravishanker, and S. Sriram. "Modeling customer lifetime value." Journal of service research 9, no. 2 (2006): 139-155

Hall, J. M., Kopalle, P. K., & Krishna, A. (2010), "Retailer dynamic pricing and ordering decisions: category management versus brand-by-brand approaches", *Journal of Retailing*, 86(2), 172-183.

Hoch, Stephen J., Byung-Do Kim, Alan L. Montgomery, and Peter E. Rossi. "Determinants of store-level price elasticity." Journal of marketing Research (1995): 17-29.

Hui, Sam K., Peter S. Fader, and Eric T. Bradlow. "Path data in marketing: An integrative framework and prospectus for model building." *Marketing Science* 28.2 (2009a): 320-335.

Hui, Sam K., Eric T. Bradlow, and Peter S. Fader. "Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior." *Journal of consumer research* 36.3 (2009b): 478-493.

Hui, Sam K., Peter S. Fader, and Eric T. Bradlow. "Research note-The traveling salesman goes shopping: The systematic deviations of grocery paths from TSP optimality." *Marketing Science* 28.3 (2009c): 566-572.

Hui SK, Inman JJ, Huang Y, Suher J (2013) The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies. J. Marketing 77(2):1–16.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, 31(8), 651-666.

Knowledge@Wharton (2015), "The Destiny of a Brand Is in Your Hand", http://knowledge.wharton.upenn.edu/article/thedestinyofabrandisinyourhand/

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012, August). Trustworthy online controlled experiments: Five puzzling outcomes explained. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 786-794). ACM

Kopalle, Praveen K., Ambar G. Rao, and João L. Assunção (1996), "Asymmetric reference price effects and dynamic pricing policies." *Marketing Science* 15, no. 1 60-85.

Kopalle, P.K., Kannan, P.K., Boldt, L.B. and Arora, N., 2012. The impact of household level heterogeneity in reference price effects on optimal retailer pricing policies. *Journal of Retailing*, *88*(1), pp.102-114.

Kopalle, Praveen K., Yacheng Sun, Scott A. Neslin, Baohong Sun, and Vanitha Swaminathan (2012), "The joint sales impact of frequency reward and customer tier components of loyalty programs." *Marketing Science* 31 (2) : 216-235.

Kumar, V., Rajkumar Venkatesan, Tim Bohling, and Denise Beckmann. "Practice prize report-The power of CLV: Managing customer lifetime value at IBM." *Marketing Science* 27, no. 4 (2008): 585-599.

Lariviere, Martin A., and V. Padmanabhan. "Slotting allowances and new product introductions." Marketing Science 16, no. 2 (1997): 112-128.

Larson, Jeffrey S., Eric T. Bradlow, and Peter S. Fader. "An exploratory look at supermarket shopping paths." International Journal of research in Marketing 22, no. 4 (2005): 395-414.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. "The parable of Google flu: traps in big data analysis." *Science* 343, no. 6176 (2014): 1203-1205.

Lee, Nick, Amanda J. Broderick, and Laura Chamberlain. "What is 'neuromarketing'? A discussion and agenda for future research." International Journal of Psychophysiology 63, no. 2 (2007): 199-204.

Levy, Michael, Dhruv Grewal, Praveen K. Kopalle, and James D. Hess. "Emerging trends in retail pricing practice: implications for research." *Journal of Retailing* 80, no. 3 (2004): xiii-xxi.

Li, Hongshuang, and P. K. Kannan. "Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment." *Journal of Marketing Research* 51, no. 1 (2014): 40-56.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Luo X, Andrews M, Fang Z, Phang CW (2014) Mobile targeting. Management Science, 60(7):1738–1756.

Ma, Liye, Ramayya Krishnan, and Alan L. Montgomery. "Latent homophily or social influence? An empirical analysis of purchase within a social network." Management Science 61, no. 2 (2014): 454-473.

Malhotra, N. K. (2008). *Marketing research: An applied orientation*, 5/e. Pearson Education India.

McAfee, Andrew, Erik Brynjolfsson, Thomas H. Davenport, D. J. Patil, and Dominic Barton. "Big data: The management revolution," *Harvard Bus Rev*90, no. 10 (2012): 61-67.

McAlister, Leigh (2005), "Cross-brand pass-through? Not in Grocery retailing", Marketing Science Institute Report #05-0113.

McAlister, L. (2007), "Comment-Cross-Brand Pass-Through: Fact or Artifact?", *Marketing Science*, 26(6), 876-898.

McCulloch, R., & Rossi, P. E. (1994). An exact likelihood analysis of the multinomial probit model. Journal of Econometrics, 64(1), 207-240.

Molitor D, Reichhart P, Spann M, Ghose A (2014) Measuring the effectiveness of location-based advertising: A randomized field experiment. Working paper, New York University, New York

Montgomery, Alan L., Shibo Li, Kannan Srinivasan, and John C. Liechty. "Modeling online browsing and path analysis using clickstream data." Marketing Science 23, no. 4 (2004): 579-595.

Murray, K. B., Di Muro, F., Finn, A., & Leszczyc, P. P. (2010). The effect of weather on consumer spending. *Journal of Retailing and Consumer Services*, 17(6), 512-520.

Nesamoney, D. (2015). *Personalized Digital Advertising: How Data and Technology Are Transforming How We Market*. FT Press.

Park, C. W., Iyer, E. S., & Smith, D. C. (1989). The effects of situational factors on in-store grocery shopping behavior: The role of store environment and time available for shopping. Journal of Consumer Research, 15(4), 422-433.

Rapp, Adam, Thomas L. Baker, Daniel G. Bachrach, Jessica Ogilvie, and Lauren Skinner Beitelspacher. "Perceived customer showrooming behavior and the effect on retail salesperson self-efficacy and performance." *Journal of Retailing* 91, no. 2 (2015): 358-369.

Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby. "The value of purchase history data in target marketing." Marketing Science 15, no. 4 (1996): 321-340.

Rossi, Peter E., and Greg M. Allenby. "A Bayesian approach to estimating household parameters." Journal of Marketing Research (1993): 171-182.

Russell, Gary J., and Ann Petersen. "Analysis of cross category dependence in market basket selection." Journal of Retailing 76, no. 3 (2000): 367-392.

Rutz, O. J., Trusov, M., & Bucklin, R. E. (2011). Modeling indirect effects of paid search advertising: which keywords lead to more future visits?. *Marketing Science*, 30(4), 646-665.

Shapiro, Carl, and Hal R. Varian. *Information rules: a strategic guide to the network economy*. Harvard Business Press, 2013.Stourm, Valeria, Eric T. Bradlow, and Peter S. Fader. "Stockpiling Points in Linear Loyalty Programs." *Journal of Marketing Research* 52.2 (2015): 253-267.

Steele, A. T. (1951). Weather's Effect on the Sales of a Department Store, *Journal of Marketing*, 15(4), 436-443.
Sudhir, K. (2001), "Competitive Pricing Behavior in the Auto Market: A Structural Analysis", *Marketing Science*, 20(1), pp. 42-60.
Sudhir, K. "Editorial—The Exploration-Exploitation Tradeoff and Efficiency in Knowledge Production." *Marketing Science* 35, no. 1 (2016): 1-9.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398), 528-540.Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51(4), 463-479.

Van Bommel, E., Edelman, D., & Ungerman, K. (2014). Digitizing the consumer decision journey. *McKinsey Quarterly*.

Van der Lans, Ralf, Rik Pieters, and Michel Wedel. "Research Note-Competitive Brand Salience." *Marketing Science* 27.5 (2008): 922-931.

Venkatesan, Rajkumar and V. Kumar (*2004*) A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. *Journal of Marketing*: October 2004, Vol. 68, No. 4, pp. 106-125.

Verhoef, Peter C., Scott A. Neslin, and Björn Vroomen. "Multichannel customer management: Understanding the research-shopper phenomenon." *International Journal of Research in Marketing* 24, no. 2 (2007): 129-148.

Voleti, S., & Ghosh, P. (2013). A robust approach to measure latent, time-varying equity in hierarchical branding structures. *Quantitative Marketing and Economics*, 11(3), 289-319.

Voleti, S., Kopalle, P. K., & Ghosh, P. (2015). An Interproduct Competition Model Incorporating Branding Hierarchy and Product Similarities Using Store-Level Data. *Management Science*, 61(11), 2720-2738.

Vrechopoulos, A. P., O'Keefe, R. M., Doukidis, G. I., & Siomkos, G. J. (2004). Virtual store layout: an experimental comparison in the context of grocery retail. Journal of Retailing, 80(1), 13-22.

Wang Jing, Aribarg Anocha, Atchadé YF (2013) Modeling Choice Interdependence in a Social Network. *Marketing Science*, 32(6):977-997.

Wedel, Michel, and Rik Pieters. "Eye fixations on advertisements and memory for brands: A model and findings." *Marketing science* 19.4 (2000): 297-312.

Wedel, M. and Jie Zhang (2004), "Analyzing Brand Competition Across Subcategories", *Journal of Marketing Research*, 41(4), 448-456.

Figure 1: Dimensions of Big data in retailing

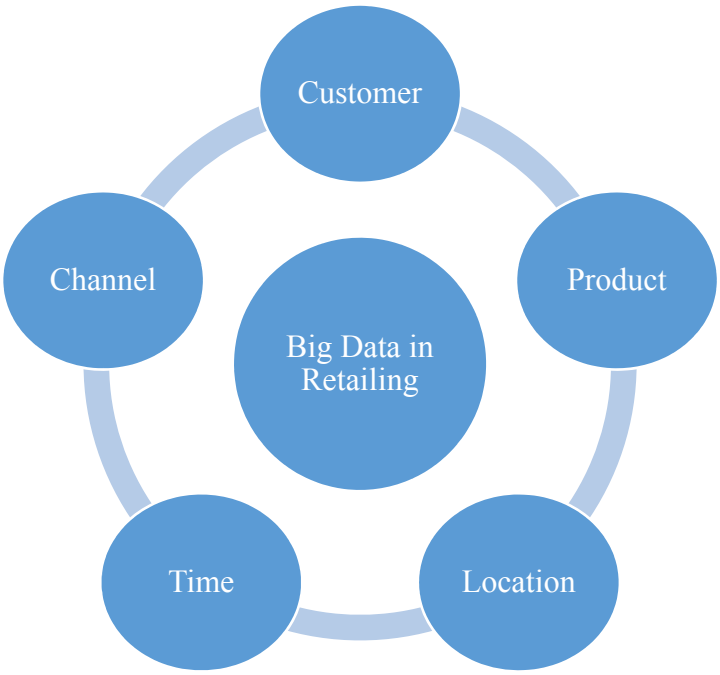

Figure 2: New Sources of Retailer interest Data

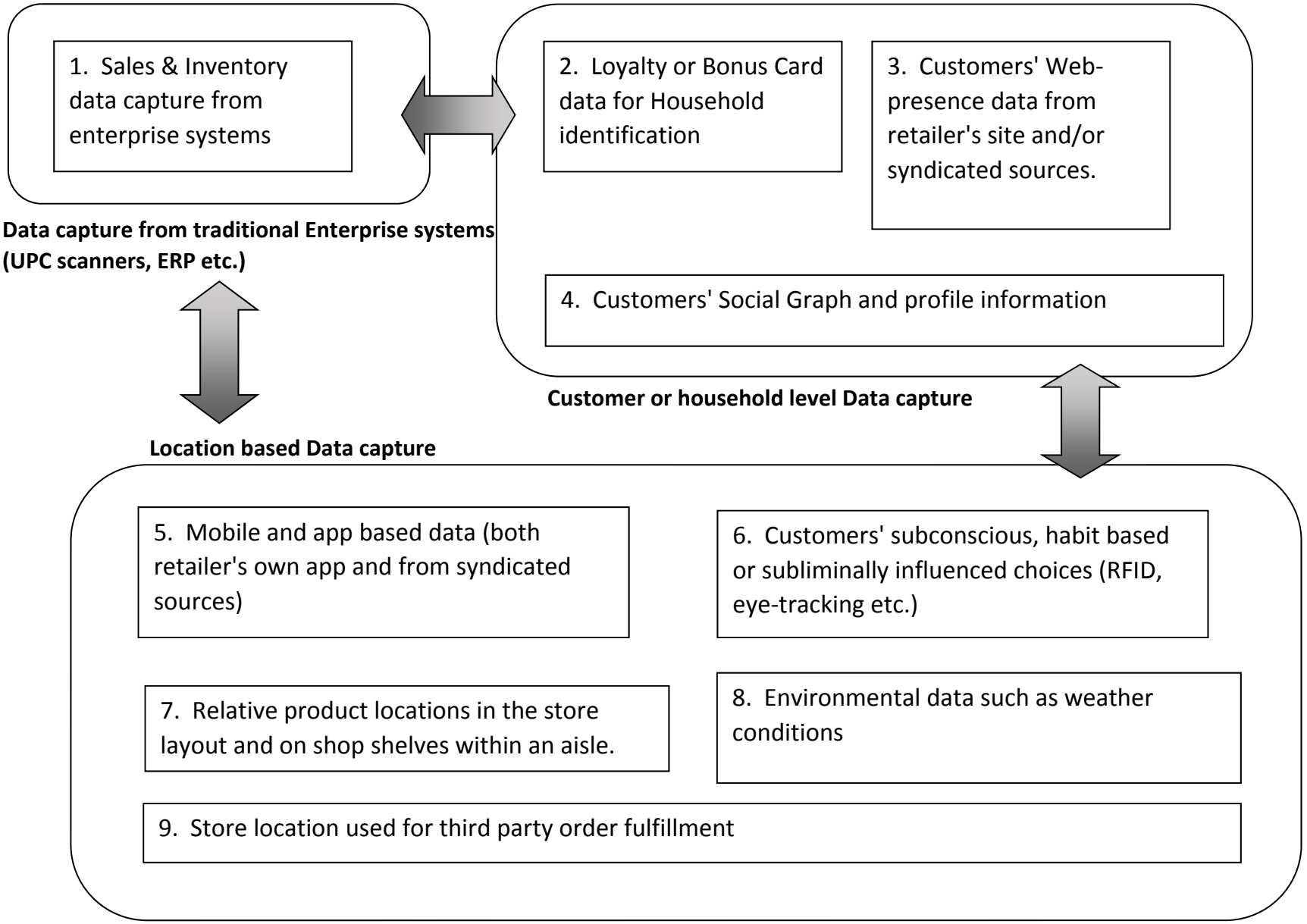

Figure 3: Illustrating Contextual Targeting on Mobile Devices

Figure 4: A Typology for Data Compression Approaches

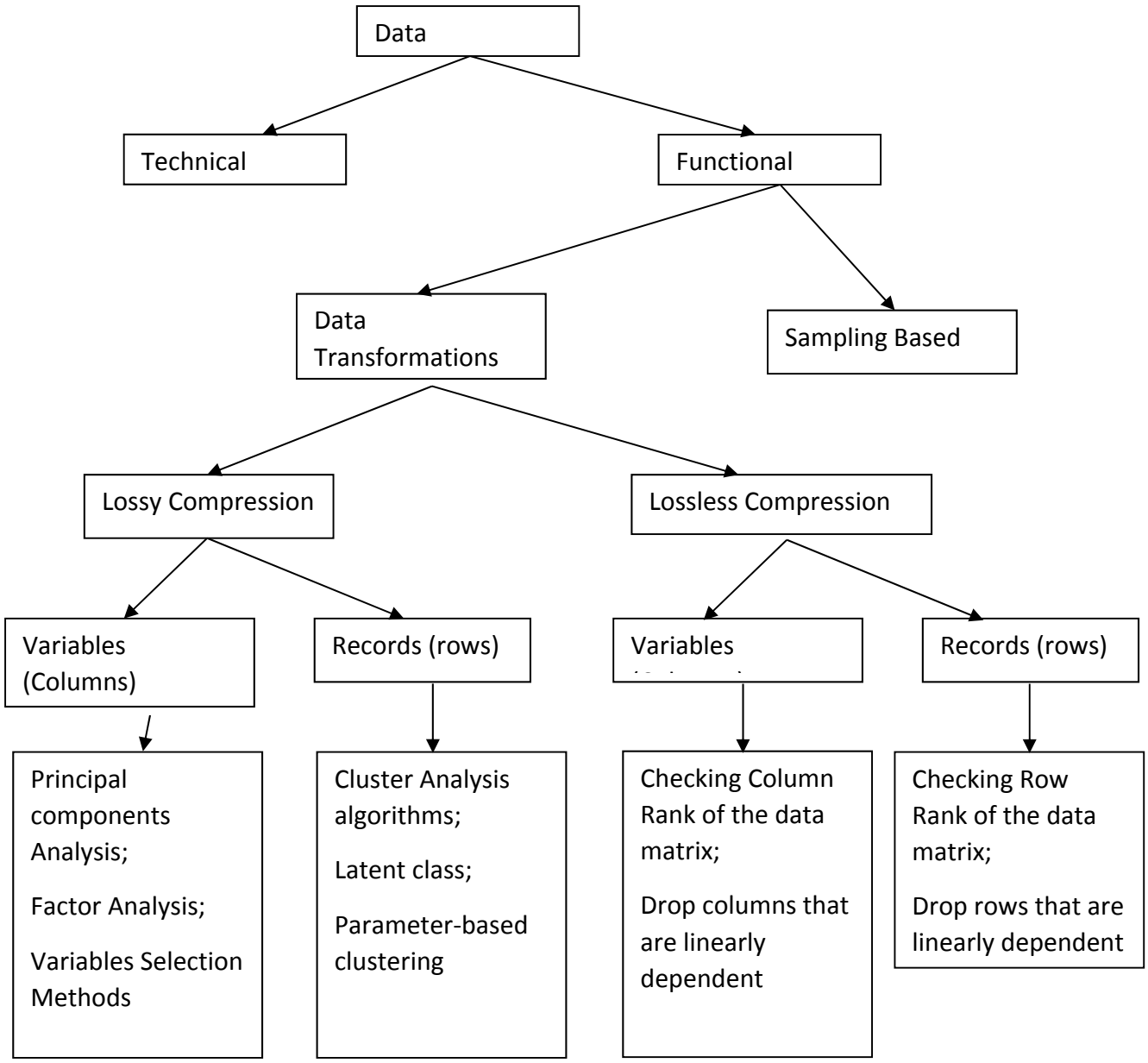

Figure 5: Flowchart for a Customer-Centric Predictive Analytics and Optimization System for Pricing Decisions

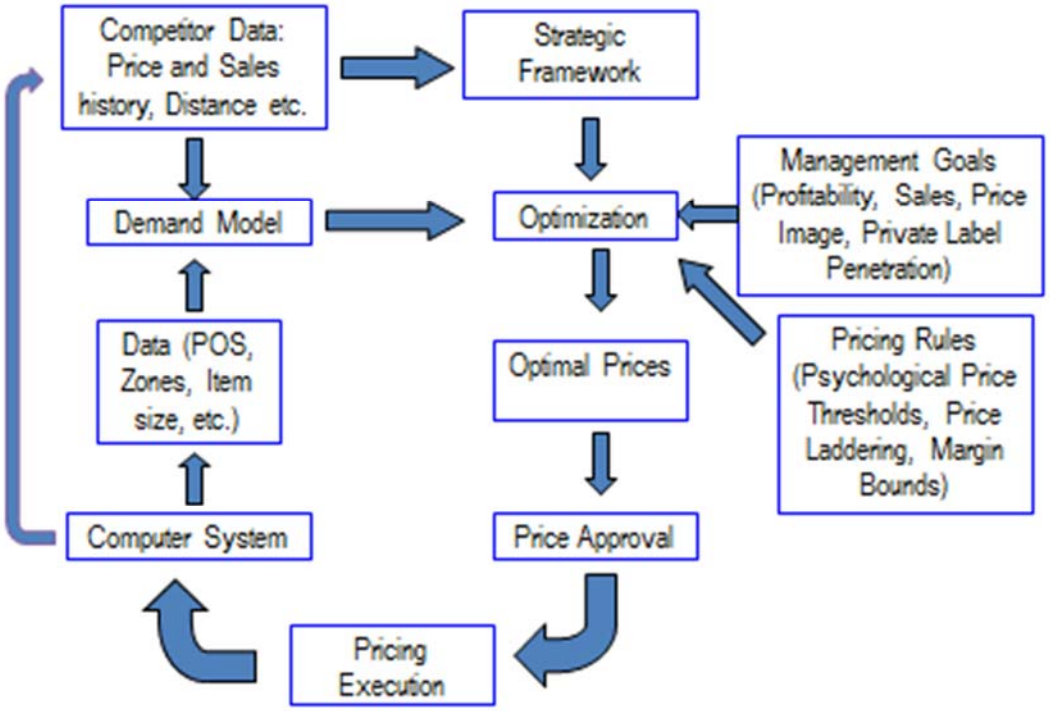

Figure 6

Hold Out Sample Results

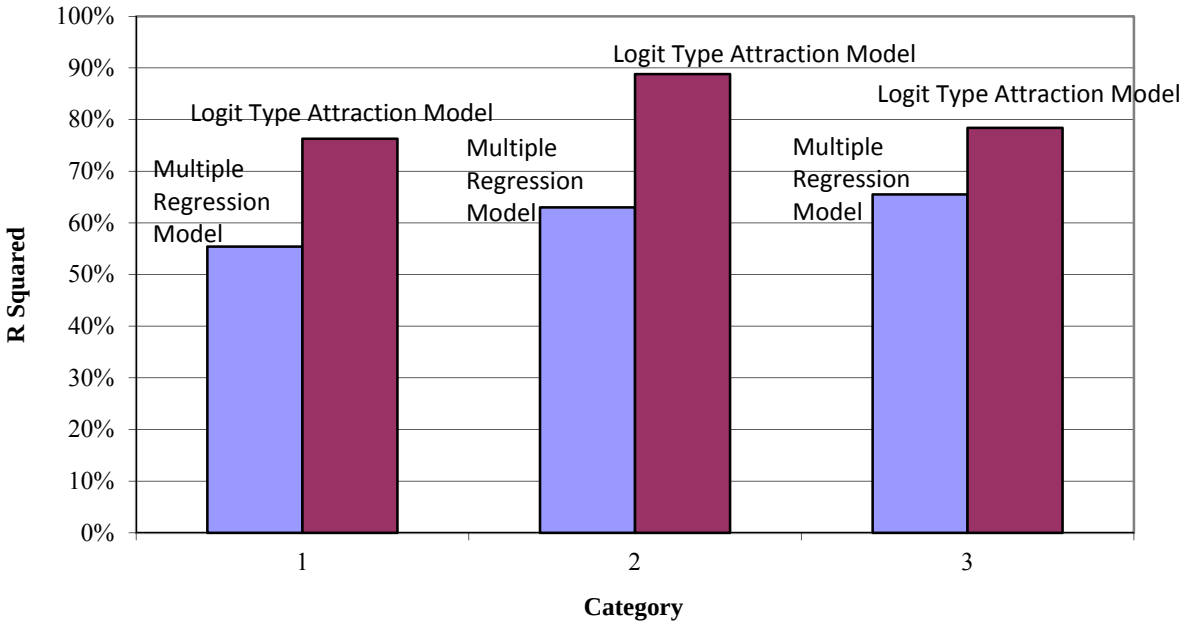

Table 1: Overview of Field Study

	Test	**Control**	**Total**
Number of Stores	21	21	42
Number of categories	14	14	14
Number of Weeks	12	12	12
Total number of SKUs	761	764	788
Number of SKUs optimized	512 (67.3%)	512 (67.0%)	512 (65.0%)
Total number of Store, SKU, Week combinations (observations, *n*)	154,020	154,440	308,460
Number of observations that contain optimized SKUs	118,320 (76.8%)	118,788 (76.9%)	237,108 (76.9%)

Table 2: Categories Studied

Category	**# of SKUs**	**# of SKUs Optimized (%)**	***n***	**Optimized Data (%)**
1. Vitamins	139	109 (78.4%)	59,616	50,952 (85.5%)
2. Office Supplies	48	17 (35.4%)	17,880	8,472 (47.4%)
3. Spices	58	50 (86.2%)	24,984	23,568 (94.3%)
4. Canned Soup	32	19 (59.4%)	11,136	7,752 (69.6%)
5. Sauces and Oil	34	22 (64.7%)	9,096	6,912 (76.0%)
6. Rice	14	11 (78.6%)	3,852	3,180 (82.6%)
7. Nutrition	137	68 (49.6%)	52,872	32,520 (61.5%)
8. Light Bulbs	20	11 (55.0%)	6,840	5,292 (77.4%)
9. Feminine Hygiene	22	21 (95.5%)	10,704	10,260 (95.9%)
10. Foam Cups and Bags	42	25 (59.5%)	15,816	12,084 (76.4%)
11. Deodorant and Skin Care	90	54 (60.0%)	34,224	25,344 (74.1%)
12. Cookies & Snacks	51	34 (66.7%)	20,256	15,540 (76.7%)
13. Cereal	18	18 (100.0%)	8,436	8,436 (100.0%)
14.Body Soap and Laundry	83	57 (68.7%)	32,748	26,796 (81.8%)
Total	788	512 (65.0%)	308,460	237,108 (76.9%)

Table 3: Regression Results

	Dependent Variable
Independent Variables	**Gross Margin $**
Test (Relative to Control)	.407***
Controls (Category dummies included but not shown below)	
GROSS MARGIN IN PRE-PERIOD	.545***
OPTIMIZE (0 or 1)	3.341***
PURCHASE FREQUENCY	.388***
UNIT SHARE	-3.627**
DOLLAR SHARE	14.445***
Intercept	-5.649***
R-square	0.759
Sample Size	235,564

Table 4: Extrapolation to the Enterprise Level

Avg margin increase/week per SKU per store ($)	0.407
Number of SKUs	10000
% SKUs where margin improvement is realized	0.74
Average Margin Improvement per week per store ($)	3011.8
Number of weeks	52
% weeks where margin improvement is realized	0.5
Average Margin improvement per store per year ($)	78306.8
Number of stores	100
Total margin improvement per year	$7,830,680
Lower Limit for 99% Confidence Interval	$4,713,396
Upper Limit for 99% Confidence Interval	$10,947,964