Big data and marketing analytics

Contemporary marketing research and analytics is considerably different from how research was conducted even as recently as a decade ago. A number of trends have driven these changes and will likely lead to even more drastic differences in the future. In order to fully understand data-driven marketing, one needs to understand the processes behind it. New professionals coming into the field should have a basis in both big data and analytical tools for examining that data. In order to do so, it helps to look at the context of these innovations and how they have matured over time.

BIRTH OF LOYALTY PROGRAMS #1

One of the earliest and most effective modern loyalty programs was created in the 1990s when UK retailer Tesco launched the Clubcard. Tesco employed marketing analytics provider dunnhumby to install the card-based system and later acquired the consumer science provider. The Clubcard loyalty system was set up not just to reward consumers but also to gather enough data to allow better understanding of the customer and then create closer relationships. At this time, when Walmart’s low prices and convenient locations threatened to overrun less efficient retailers, Tesco and others saw customer knowledge as an alternative means to compete. With better knowledge, they could serve customer needs better and counter pure price appeals.

At its beginning the Clubcard was supported by information technology (IT) resources though these were fairly limited compared to today’s capabilities. But they did allow for substantial data collection and subsequent analysis. Out of 50,000 or so items (stock-keeping units or SKUs), researchers were able to identify 10–12,000 key ones. These were assigned scores according to a few dozen identified characteristics. In one example,
papayas might be characterized as high on “[a]dventurous, fresh, premium price, short shelf life”.

Classification by characteristics allowed Tesco to analyze entire shopping baskets as it collected transaction data. Associations could be found between items shoppers tended to buy together, initially by hand but eventually by computer as algorithms were developed to spot frequently occurring buying patterns. Based on the similarities in the baskets, Tesco could identify customer segments. Deeper analysis of the shopping baskets and corresponding segments allowed researchers to form hypotheses about who might buy more, what promotional offers might spur purchase, and what non-grocery services might appeal to consumers (e.g. banking, telephone).

Tesco and dunnhumby further enhanced their transactional data and shopping basket descriptions with demographic (age/life stage) and behavioral (promotional responses, shopping time and location) detail. By 1994, according to one source, Tesco could offer 4 million different, distinct offers to consumers based on these variables. Though still based on segments as opposed to individuals (except for a few personal characteristics that could generate an offer), the numbers worked out to an almost one-to-one capability for the grocer’s 10 million Clubcard members.

At this time, Tesco moved advertising dollars away from television and pursued a communication strategy heavy on direct mail and point-of-sale devices such as shelf-talkers. Members would receive quarterly mailings with an update on points, a newsletter, and targeted promotional offers. Additional mailings for special purposes were also made as well as communications aimed at sub-clubs (wine lovers, parents with new babies, etc.). Local stores could also be supported, especially when presented with special circumstances such as a new store opening by a competitor.

Note, again, that all of this was in the mid-1990s. IT was available, including early email capabilities, but not powerful in today’s terms. Communication was chiefly done through mail. And the overall tactics were predicated on identifying and pursuing segments of customers.

Today, Tesco and dunnhumby are not as far ahead of the pack as was the case two decades ago but their capabilities have advanced considerably. dunnhumby serves multiple clients, not just Tesco, analyzing 5–75 terabytes per client; 700 million consumers are being studied, including 200 million weekly shopping baskets and 1.4 billion individual items. In
addition, the agency combines shopper data with additional inputs such as weather, geography, demographics and social media. The data are structured and unstructured. Unstructured, though currently only 10 percent of the database is growing much more quickly, at 10–15 percent per year. dunnhumby notes one client in northern California, Raley’s, using transaction and loyalty card data along with customer comments and social media chatter to target individual customers, improving their experience.

And that is probably the key take-away in terms of how big data and marketing analytics are changing practice in loyalty programs. Previously, loyalty programs were able to sort customers into segments, create offerings for segments, and push offers. Today, not only have organizations moved beyond mail as a principal means of communication, but they are able to identify and target individual customers by collecting all available data on behavior. Internal data are combined with other databases, again often identifiable down to the individual customer. From that understanding of each consumer, marketers can not only design promotional offers but actually enhance the relationship. Ideas for products offered on deal, types of communications, or other individual outreaches may still come from analyzing segments, but the actual marketing proposition can be made to distinct consumers.

dunnhumby sees the future of loyalty programs in those terms. As we’ll discuss, big data and marketing analytics allow the collection, storage and processing of considerably more data than was the case even as recently as five years ago. As such, much more data on loyalty members can be created and stored, from a wide variety of sources. Combining profiles with other databases is also much more easily done. As dunnhumby client director Oliver Harrison points out, the point of loyalty programs is to identify customers (and one could add, identify the best customers, allowing appropriate approaches to every level of customer). But, in the future, rather than push marketing offers to customers based on profiles, loyalty programs will be designed to ease transactions and optimize the entire relationship. Instead of requiring loyalty cards, identification will be made by any number of technologies. Mobile apps, phones themselves and credit cards can all identify individual customers once tied to an account. Once identified, shopping activities can be tracked, including geo-location technologies with iBeacons that can monitor in-store movements. Background data, observed activities and contextual information can all be combined, stored and analyzed to allow real conversations with customers. Hearing each individual voice enables higher levels of customer satisfaction by making it easier for them to express desires and then fulfill them.
One of the key aspects of big data is this granular level of data collection and use. We’ll talk about that in more detail as well as how analytical tools and capabilities can create even further insight. Big data is one of those terms that get thrown around a lot, but not everyone really understands what it means. This chapter should clarify the details as well as those surrounding related concepts such as analytics.

BIG DATA #1

So if big data is a ubiquitous and perhaps overused term, what exactly is it? One thing to note is that big data and marketing analytics (or marketing intelligence) are related concepts but not necessarily the same thing. In this section, we’ll be looking at some aspects of the combination of big data and marketing analytics that apply chiefly to the former. The reasons should become clear but note that the terms are closely related but probably shouldn’t be used interchangeably.

Big data is sometimes referred to as a dataset too big to be handled by traditional software tools (the phrase “too big for Excel” is sometimes used). That is somewhat imprecise, but even that has a purpose as processing and storage tools are expected to change, so the software programs will get more powerful but the databases are also likely to continue to grow. Further, big data differs by context. What might be considered big data in one industry would be considered a relatively small dataset in another. So, in some sense, we know big data when we see it but there is also a sense of not wanting to be pinned down in defining it.

But there are more precise descriptions that are helpful. In particular, big data is sometimes referred to according to the “three V’s”: volume, velocity and variety. Volume refers to the amount of data generated by various sources, fed into the system, and stored. McAfee and Brynjolfsson report that in 2012, 2.5 exabytes of data were created every day and that amount was doubling every three years or so. An exabyte is one billion gigabytes. We’ll talk about sources shortly, but the bottom line is that the amount of data being generated and captured is exploding.

Velocity is the speed of the data collection and transfer. Quarterly reports on sales have been replaced with transactional data updated by the nanosecond. Retailers can track what is happening in stores in real time. Website operators of all kinds can do the same. Data are much more available than was previously possible. Variety has to do with the nature of the
### Table 1.1 Useful terminology

<table>
<thead>
<tr>
<th>Big Data</th>
<th>Volume</th>
<th>Gigabyte</th>
<th>1000 Megabytes (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Terabyte</td>
<td>1000 Gigabytes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Petabyte</td>
<td>1000 Terabytes</td>
</tr>
<tr>
<td>Velocity</td>
<td></td>
<td>Speed at which data enters, is processed, and exits system (e.g. annual, weekly, hourly, seconds, micro-seconds)</td>
<td></td>
</tr>
<tr>
<td>Variety</td>
<td>Structured</td>
<td>Ordered data, usually quantitative</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unstructured</td>
<td>Unordered data, at least in raw form (e.g. text, images, video)</td>
<td></td>
</tr>
<tr>
<td>Enterprise Systems</td>
<td>Enterprise Resource Planning (ERP)</td>
<td>Information technology system for managing all resources of the firm’s operations (inputs, processing equipment, human resources, etc.)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Supply Chain Management (SCM)</td>
<td>Information technology system for managing the firm’s supply chain, ensuring right inputs arrive in right place at right time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Customer Relationship Management (CRM)</td>
<td>Information technology system for managing all interactions with individually identified customers (transactions, communications, etc.)</td>
<td></td>
</tr>
<tr>
<td>Big Data Systems</td>
<td>Data Lakes</td>
<td>Combination of all databases of firm in one place, often requiring considerable work to rationalize different data systems and formats</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Key Performance Indicators (KPIs)</td>
<td>Metrics identified as key to achieving performance objectives, monitored on a regular basis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dashboards</td>
<td>Screens designed to report important metrics, often KPIs and often in a visual format appealing to the decision-maker</td>
<td></td>
</tr>
<tr>
<td>Analytics</td>
<td>Deeper analysis using advanced tools capable of handling the volume and complexity of big data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
New methods of market research and analysis

data inputs. Practitioners in the field often refer to structured (traditional, quantitative) vs. unstructured (not quantitative) data. The latter category includes all sorts of inputs that previously weren’t stored in databases or, if stored, were very hard to organize and analyze. Today, unstructured data such as recorded voice, text, log files (such as machine performance or activity tracking), images and video are all regular parts of databases. Monitoring things like customer attitudes with video, call center comments, or social media chatter is both do-able and subject to analysis. Some key observers believe the ability to enter unstructured data into databases is the key piece of the big data concept, and it may be. But unstructured data usually takes up a huge amount of storage and processing capacity (volume) and is valuable, in part, because it can be monitored in real time (velocity), so all three V’s do have a part in our understanding of the concept. Some observers add more V’s to the definition, specifically veracity (truthfulness or accuracy) and value, which can have their place as well.

From where do all the data come? Firms installed enterprise systems in the late 1990s and into the new century. These included enterprise resource planning (ERP), supply chain management (SCM), and customer relationship management (CRM) systems. ERP has to do with tracking and bringing together everything needed for operations (raw materials and components, labor, machine availability, etc.) in the most efficient manner. SCM does much the same with inputs from suppliers, while CRM records all interactions with customers. These systems generate tremendous amounts of data, everything to do with supply chains, operations, distribution channels, vendors and consumers. Early generation installations created the data but it wasn’t necessarily all tracked or stored after the fact. But one key reason for the advent of the big data phenomenon is the drop in costs of computing storage and processing. Firms can afford to collect and keep everything from all systems; they can also afford to distribute it more widely and/or subject it to deeper analysis.

This is part of why the cloud is an interesting part of the package. Cloud computing is nothing magical, it’s just a matter of moving data storage and processing to third party systems. So organizations rent space on systems provided by Amazon, Microsoft, Google, or another provider, and that solution is cheaper than installing their own servers and such (as well as personnel needed to service them). Cloud computing is more an indicator of the tremendous drop in computing costs mentioned earlier.

With an affordable structure to take in and process ever more data, the door was opened to new sources of inputs as well. As noted, social media
(text, images, video), customer comments (audio, text), and other data-heavy inputs that we’ve referred to as unstructured data also joined the party. The digital world that has developed over the past decade has substantially increased the amount of potentially relevant data for organizations, particularly for marketers.

All of that has to do with data internal to the organization or at least internal to it and its network of collaborators. But the trends propelling interest in big data have also created opportunities for firms to collect, store and process other sources of data, both publicly sourced (e.g. government, visible social media) and commercial. Not only does this add to the amount of data available to the system but enables combinations of data. The term “data lake” has come into the conversation recently, referring to the ability of firms to combine all kinds of databases into one huge, accessible collection. Harmonizing the databases into a single format in a data lake is no small task but, again, contemporary technology and reduced costs have enabled firms to do so.

What these capabilities allow from a marketing standpoint, in particular, is an ability to create even more detailed profiles of individuals within a database. So if an organization has records on an identifiable consumer, it can supplement those records with items from other sources. From government records, for example, voter registration, professional credentialing, and real estate transactions can be harvested. Commercial sources may have all sorts of data, in general or identifiable to an individual consumer, from online and offline activities (product and service registrations, social media profiles, or from a multitude of other communications and observations). All it takes is one connector between databases, such as an address, to join together even anonymized profiles. One Federal Trade Commission report asserts that a profile exists for virtually every consumer in the United States.

Once a profile exists, firms can add to it. In particular, transactions and communications can supplement what the organization knows about any individual customer, provided they can be identified by means of loyalty check-ins, mobile apps, registered credit cards, and other records already on file. Essentially, the technologies are there to allow firms like dunnhumby and its clients like Tesco to collect massive amounts of data. Affordable tools are also available to store, organize and analyze all of this data, and the end result is personalization on a level we couldn’t even conceive just a decade ago. Where else that leads, we’ll discuss shortly.
A second well-known pioneer of loyalty programs is Caesar’s (formerly Harrah’s), the casino/resort operator. The firm introduced its Total Rewards loyalty program in 1997 (then called Total Gold). At that time, just about all casinos had some sort of rewards program, especially for the high rollers referred to as “whales”. But Harrah’s went further, seeking to do something about the legendary fickleness of customers (even loyalty members reportedly spent only 36 percent of gambling dollars at Harrah’s at that time) by means of the greater geographical spread of its properties. The firm also found that the vast majority of its revenues came from gambling (over 87 percent in 2001) rather than hotel rooms, shows, stores or restaurants. Total Gold tracked demographics and transactional data from members, including gambling spending and preferences (cards were inserted in slot machines, for example, to earn points, and machines of choice could be tracked).

Total Rewards was installed to increase visits to Harrah’s properties, thereby increasing loyalty. Looking more deeply into the existing database, managers found that 26 percent of customers generated 82 percent of revenue. Essentially, different customers had different value to the organization. Moreover, the most valuable ones weren’t the whales but middle-aged and senior slot players, often current or former professionals. These customers didn’t necessarily stay overnight, playing on trips home from work or on weekend outings and typically responded better to promotional offers centered on the gambling experience (free chips) than on other perks (room discounts, free meals).

The loyalty program evolved to include both more detailed descriptions of member preferences and an enhanced service capability to improve delivery on those preferences. Different information systems were linked together into a winner’s information network (WINet) so that a single database included all available data, including transactions, gameplay, hotel management, reservations and background demographics. This database could be updated hourly at the time. Records were analyzed to determine long-term potential, the customer lifetime value. These data showed different customer groups with different values, and loyalty program levels were designed to match those groups. Rewards were established not just as rewards but as a means of increasing lifetime value. These were personalized to the individual, specific promotional offers for each member.

But the marketing proposition wouldn’t work without service excellence, so the customer database was also linked back to the hospitality and
management staff. Names were used by everyone from the valet to casino floor staff. Rooms were set up as guests desired. Members were steered to their preferred activities. And, again, the service level could vary by loyalty program level, with some guests experiencing virtually no line at check-in or at restaurants and others having a normal check-in experience. According to at least one manager, customers didn’t have a problem with different treatment; they understood that in the gaming industry the highest value players warranted the best service.

Today the system has evolved to even higher capabilities. Even though Caesar’s has had some rough years following the 2008 economic meltdown (a bad time to be highly leveraged as the entire gambling industry took a hit), its big data program is still highly respected. Customer information records were estimated to be worth $1 billion, the highest valued asset in Caesar’s bankruptcy portfolio. The firm continues to collect all available data about member behavior. Loyalty cards, for example, are now loaded with funds, and when at a gambling station can be inserted and tracked. So not just networked slot machines, but all forms of gambling can be monitored, including time spent, wins and losses. Video images can track lines, especially for high-value individuals. If a particularly valuable member has a bad night, staff can intercept them before they leave the casino, providing a personal interaction and incentive to return.

So the emphasis on data and excellent service continues but the data have become even more granular. Membership incentives represent an even wider range of options, not just promotional offerings like free rooms or shows but pricing differences (reservation agents can price based on supply, demand, and the individual customer’s data profile) and product differences (what’s included in the reservation package). Mobiles have entered the equation, both as a data-gathering tool (location data, social media chatter) and as a service device (alerting members to show times, game availability and time-sensitive offers). Algorithms have also been installed so that the system automatically takes action to improve the customer experience when it senses something has gone off track. When a member with an established pattern (visit to a specific property once a month with $X spent) halts that pattern (no visit in three months or reduced spending per visit), customer service reps can be alerted to make an outreach with a specific offer.

The Caesar’s example again illustrates the nature of big data, that deep and detailed data are available on individual customers. Those data can be monitored for established behaviors, and changes in trend can be
spotted while time is still available to do something about them. Further, such actions can be “automated” with algorithms and other decision-making rules, providing prompts to customer service reps or even independently fixing a problem, making responses even more rapid. Finally, as we’ll discuss, such rich databases can also be analyzed for even further insights, those that might not be apparent without a deeper dive into the data lake.

**BIG DATA #2**

So firms are taking in high volumes of data, of all types of variety, at increasing velocity. That’s the input part of the big data equation. What do the organizations do with the data and what are the outputs? As noted repeatedly, one key feature of big data is the unstructured data coming into the firms. These types of data are converted into an analyzable form by various programs such as one known as Hadoop. Once in appropriate digital form and organized in a manner that can be accessed and manipulated in a number of ways, the data can be fed out and reported, subjected to deeper analysis, or used for modeling. This section chiefly focuses on the first: data monitored and reported on a regular basis or data harvested for a specific purpose.

If monitoring data is the primary purpose of a system, it is set up to take in the data, perhaps summarize it or reduce it to descriptive statistics, and then report it out to appropriate parties. This could take many forms. It might be transaction data, sales of specific products, for example, or sales by location at retail stores. It might be image or video data, as would be the case with Caesar’s’ attention to lines in its casinos. It might be log data, perhaps from the internet of things. GE, for example, is able to monitor the performance of its airplane engines in real time. When any of the key indicators moves out of tolerance, action can be taken to correct the problem.

When that is the case, the data system is set up only to process, store and distribute the data. As noted, summary statistics might be reported but no major transformations of the data take place. When the system is set up, decision-makers do need to settle on what data are important to track. Based on historical data, benchmarks or target levels of the indicators can be determined. These are typically referred to as key performance indicators (KPIs). Monitoring systems are established to feed the KPIs to decision-makers, who then make necessary adjustments if those KPIs are out of line. Typically, these will be at the operational (adjust the machinery,
adjust how web pages load) or marketing (adjust customer communications, adjust the price) level.

To illustrate, Tesco, discussed earlier in this chapter, uses KPIs to monitor daily outcomes at individual stores. The “corporate steering wheel” tracks data related to customers, community, operations, people and finance. For the customer category, for example, the grocer’s high-level objectives include:

- earn lifetime loyalty;
- aisles are clear;
- I can get what I want;
- the prices are good;
- I don’t queue (stand in line);
- the staff are great.

At the store level, the related KPIs reported by managers, again for customers are:

- gaps on shelves (product outages);
- I don’t queue;
- mystery shoppers.

These may be measured at different times. Managers might do an hourly sweep, for example, to look for empty shelves. Lines at cash registers could be watched by managers, again perhaps on an hourly basis. Or they could be monitored in real time by video cameras. Mystery shoppers might appear on a daily, weekly or monthly basis. And these KPIs are undoubtedly supplemented with the massive data Tesco generates on transactions. Cash registers can be monitored constantly concerning category or individual product sales, use of promotional offers, responses to price experiments, and other matters. But the main point is that systems are set up to constantly feed raw or summarized data to decision-makers. These data are often presented on “dashboards”. Dashboards are set up to monitor and communicate pertinent data, typically KPIs, to decision-makers. As indicated by the term, they provide a quick look at current performance, just as a car dashboard lets you know immediately important indicators such as speed, RPM and temperature. In some cases, the dashboard may be set up to provide alerts as performance goes outside set limits (e.g. temperature gets too high). Embedded analytics may even contain an algorithm that prompts a remedial action without a direct action from the decision-maker (e.g. engine automatically corrects to bring temperature down).
In Tesco’s case, this could involve some of the KPIs noted above, customer data from loyalty cards and transactions, or other real-time performance indicators. Dashboards might be set up, for example, to monitor and communicate data on hourly (or even minute-to-minute) sales, use of Clubcard offers, stock levels of key items and so on. If an item on offer gets below a certain safety stock level, the dashboard might alert a store manager. Or, if embedded analytics are programmed into the system, an order to a supplier or distribution center might be sent out automatically.

But, again, the key point is that some big data systems are set up in this way to collect, organize and distribute information to decision-makers in real time. These types of systems might do some basic processing of the data, such as converting it to summary statistics or tables/charts. Or they may just transmit the raw data to be viewed on dashboards. And all of this can be quite useful, improving operations and marketing processes, increasing efficiency, and reacting to events quickly, perhaps even before a problem occurs. But deeper analysis of the data is not necessary and not necessarily completed. Not everyone needs the more complex systems in order to benefit from big data, though some do.

**SPOTIFY**

Spotify is a firm using big data in some obvious ways as well as some not so obvious. The obvious part includes the capabilities we just discussed: gathering and distributing big data. The less obvious part has to do with deeper analysis and new insights, something that can’t be accomplished with the data sharing systems in isolation.

By late 2015, Spotify had 75 million subscribers to its digital music service, some paid and some taking advantage of the free service. Their activity amounted to 1.7 billion hours of listening per month or around 20 billion hours per year. If you have an account, the service is able to track everything you listen to and has also accumulated data from your registration, from other web-tracking information, and, perhaps, Facebook if accounts are linked. From tracking behavior, Spotify is able to gather even more data, aiding its profile of you and on the music played, such as location, time and the full range and depth of musical interests. All told, this amounts to 4 terabytes of storage just for music files, 600 gigabytes of listening data added daily, and 28 petabytes of total storage. Spotify possesses one of the largest data warehouses in the world. In fact, given our earlier discussion about the economics of the cloud, it’s not surprising that
the service recently announced it would move its data storage and processing to Google Cloud.

Spotify’s initial success came from copying pirating services, down to the level of replicating some of their software. From such peer-to-peer connections, in addition to a centralized streaming ability, Spotify sought to offer the quickest “click to sound” experience in music streaming (utilizing the network of users and stored music rather than just a centralized storage location). Operational data helped to further optimize the listening experience. Another point of differentiation was the playlist. While listeners could easily move across platforms to competitors, whether iTunes, other pay services, or pirate sites, the playlist was something that was a chore to replicate, once built. A listener taking the time to construct numerous playlists on Spotify would be less likely to move to a different provider.

Big data stores are clearly of help in providing an excellent listening experience to subscribers. Categorizing music in numerous ways allowing easy search and identification, the operations are seamless, and listeners today are provided with their own personalized statistics on listening patterns. What is often not seen is the value provided to artists and labels.

With purchased music, whether CDs or digital rights such as iTunes, all the provider really knows is when and where the product was bought. Whether the customer listens once or hundreds of times is opaque unless some other tracking ability is available (e.g. a trackable app). With Spotify, when, where, how much, and context can all be discerned. This sort of data can be invaluable to musicians, publishers and live event planners. Greeley (2011) provides the example of Jay Z sales being very high in London, UK, but listening rates actually showed him to be more popular in Manchester. Similarly, by city, Spotify can identify who is being played on Friday and Saturday nights (at parties), what music spikes after being added to radio playlists or being featured on a television show, and what tracks or playlists are being shared with friends. This sort of data can drive more effective concert planning, promotional efforts and media appearances. Artists have a much better handle on where and with whom they are popular.

In fact, more recently, Spotify has introduced an artist service to identify “high passion” fans. Its own data indicate that such fans are five times more likely to attend shows, so such data is of even more use to artists than just the basic data. Further, such users, though only 10–20 percent of the fan population, contribute to the majority of the artist’s sales (streaming
revenues, in this case). Accumulating data on different types of usage helps to identify the high passion fans. Spotify can identify fans who have listened to an artist every day in the last week. It can identify fans who listen to an artist more than any other artist. And it can identify fans who have listened to an artist for the majority of days in the last month. Add it all up, and you have a pretty good idea of who your best customers might be if you are a musician. You also have a way to reach them, and a stream of constantly updated data allowing you to track these numbers and spot trends. How this fits into what we discussed earlier concerning dashboards and KPIs should be readily apparent.

But Spotify is also indicative of higher abilities of big data, the applications beyond just monitoring, sharing and informing that we’ve been discussing. With all that data, firms are capable of deeper analysis. There are various approaches to examining the data for unexpected insights, and we’ll discuss them in this book, but collectively they make up what is referred to as marketing analytics or marketing intelligence. More broadly, outside the marketing context, these would be business analytics or business intelligence. In some cases, the approach may simply be “cutting” the data differently by introducing a different variable providing a different perspective. When data are sorted by an additional variable or two, they can sometimes provide a totally different insight. Or the approach may be advanced statistical analysis allowing researchers to spot patterns in the data. We’ll discuss several of these in more detail.

In Spotify’s case, the data analysis capabilities are substantial. In terms of cutting the data, context was mentioned earlier. Spotify Labs employs a program named Cassandra to analyze listening preferences. One example provided concerns a heavy metal fan according to general listening patterns. But the individual is a young married professional with young children, so heavy metal recommendations from Spotify in the evenings would not be appropriate. Context matters, and the deeper analysis provided by adding that variable alerts the firm to the need for better recommendations depending on time of day, listening device or other variables.

Discover Weekly is one of Spotify’s newest products, a personalized weekly playlist of recommended songs. The songs tend to be new to the listener, as are most of the artists. The recommendation is based on a mix of the listening history of the individual, the listening history of other individuals with similar profiles, and linkages between songs and artists noticed across the Spotify database (again, especially for those with comparable profiles). All of this is aggregated into a brand new playlist every week for each customer.
Further, even if the customer doesn’t like everything on the playlist, Spotify collects additional data every time they click on a song on the playlist, listen to more material from one of the included artists, or add a song or artist somewhere else in their collection of playlists. All of this constantly provides additional data for the listener’s profile, allowing Spotify to learn and even better customize the experience. As noted earlier, this is very different from a big data system simply compiling and sharing out results. This is deeper analysis, looking for insights in the data, creating knowledge or intelligence rather than just the raw data or information. In the manner in which the system itself learns from new data, processes and acts, it’s a form of artificial intelligence or machine learning. That’s big data at a different level, more along the lines of marketing analytics.

MARKETING ANALYTICS

Deeper analysis of databases can be done in a number of ways. We'll cover some of the basic techniques when we look at some statistical software later in this text. For now, it's just a matter of having some sense of how marketing researchers can manipulate data to get answers to questions or to uncover new insights. So Tesco trying to describe the best customer groups for a new product offering such as banking services, or Spotify trying to predict what new music you might like, based on your listening habits and those of similar customers, are both examples of instances where the data needs to be processed in order to come up with answers.

At the most basic level, such processing might just be a matter of “cutting” the data differently, adding a variable to a tabulation and observing the results. One of the most famous examples of this approach has to do with fatal automobile accidents. If insurers could identify the highest risk groups and charge accordingly (they can’t always as discrimination by some of these variables is illegal), their marketing mix could be much more effective. Consider the data in Table 1.2.

<table>
<thead>
<tr>
<th>Table 1.2</th>
<th>Fatal passenger vehicle crash involvements, by gender, April 2001–March 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>33 733</td>
<td>14 633</td>
</tr>
</tbody>
</table>
With no other inputs, one might reasonably conclude that females are safer drivers than males. But add some additional variables to the mix, and you get the results in Table 1.3.

By calculating in miles driven, one can immediately see that the differences between men and women aren't as dramatic as they seem. As opposed to twice as many fatal crashes (33,733/14,633 or 2.3:1), the ratio is only 2.5/1.7 or 1.47:1. Much of the explanation behind higher crash rates for males has to do with them driving more miles.

But when age is added in, the picture grows even more complex. The difference between males and females is much more pronounced at younger ages but totally disappears by the time both genders move toward retirement age. And if we just look at the age differences exclusive of genders, the dif-

Table 1.3  Fatal passenger vehicle crash involvements, by gender, mileage and age, April 2001–March 2002

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crashes</td>
<td>Miles</td>
<td>Rate (crash/</td>
</tr>
<tr>
<td>16–19</td>
<td>4257</td>
<td>45 billion</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>1852</td>
<td>35 billion</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>6109</td>
<td>81 billion</td>
<td>7.5</td>
</tr>
<tr>
<td>20–29</td>
<td>8949</td>
<td>225 billion</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>3172</td>
<td>156 billion</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>12,122</td>
<td>382 billion</td>
<td>3.2</td>
</tr>
<tr>
<td>30–59</td>
<td>15,027</td>
<td>845 billion</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>6946</td>
<td>551 billion</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>21,973</td>
<td>1396 billion</td>
<td>1.6</td>
</tr>
<tr>
<td>60–69</td>
<td>2,097</td>
<td>128 billion</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>1,008</td>
<td>64 billion</td>
<td>1.6</td>
</tr>
<tr>
<td></td>
<td>31,05</td>
<td>193 billion</td>
<td></td>
</tr>
<tr>
<td>70 and above</td>
<td>3,148</td>
<td>76 billion</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>1,571</td>
<td>39 billion</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>4,719</td>
<td>116 billion</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>33,733</td>
<td>14,633</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>13,24 billion</td>
<td>847 billion</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>48,638</td>
<td>21,72 billion</td>
<td></td>
</tr>
</tbody>
</table>

ferences between the youngest group of drivers (7.5 crashes/million miles) and those in middle age (1.6 crashes/million miles) are substantial. And if one looks at multiple variables, the widest apparent difference is between males 16–19 (4257 crashes, 9.2 crashes/million miles) and females 30–59, (6946 crashes but only 1.3 crashes/million miles). It should be clear that an insurance company looking to set auto rates based on risk profiles would want to charge those young males considerably more than the more mature females, if possible.

The problem illustrated is referred to as “omitted variable bias”, where missing a key piece of data leads you to an incorrect conclusion. It’s also the reason why one aspect of marketing analytics, particularly predictive analytics, is repeatedly adding new variables to the process. So rather than just cutting the data by gender, in this case you would also want to add miles driven and age. Further detail from the statistics would let you know that alcohol consumption, speed and seat belt use are also critical variables that you would want to introduce. Once added to the analysis, all provide critical new insights into how you divide the market into smaller and smaller segments in order to better understand it and provide it with insurance products. If the true high-risk individuals could be identified and charged appropriately, prices would go down for everyone else. Different product packages could also be created for different segments.

Another illustration of the omitted variable problem, also referred to as Simpson’s Paradox, comes from a famous court case in the 1970s. The Visualizing Urban Data Idealab at the University of California, Berkeley has an interesting, interactive representation of the data (http://vudlab.com/simpsons/), which come from the university itself. The data show that the school had accepted 44 percent of male applicants into its graduate programs but only 35 percent of female applicants. The surface data prompted a sex discrimination suit.

A deeper look at the data, however, uncovered a “lurking variable”, one that provided a very different perspective when introduced. The data were cut not only by gender but also by department. There were departments with relatively high acceptance rates (typically math, science, engineering, etc.) and those with relatively low acceptance rates (more humanities and related disciplines). Women applied in much higher percentages to the latter. So, as illustrated in the online example, even though their per-department acceptance rates were higher than or virtually identical to those of men, the overall acceptance rate was significantly lower. As investigators pointed out, sexism may be apparent but it came from
factors pushing women into the disciplines with higher rejection rates and occurred long before anyone ever applied to Berkeley. The university was not at fault as the deeper analysis of the data showed no evidence of institutional discrimination.

Cutting the data to better understand circumstances and, for marketing in particular, customer segments that may be revealed can be a fairly simple process, as in these examples. But remember that when one is doing it with millions of items and dozens or even hundreds of variables, it’s not necessarily as easy as creating cross-tabulation tables. When the process is intended for what we call predictive analytics, specific techniques appropriate to large data sets are employed. We’ll talk about some of these in more detail later, but basic regression, decision trees, or even neural networks can be used to correlate and group variables that might predict a specific outcome (e.g. purchase).

Conceptually, these processes can be similar to cutting the data but also different in some ways. Again, the main point is identifying and aggregating the variables predicting an outcome. With such identification, marketers can target customer profiles that are most likely to react with a specific action (again, purchase, visit to a website, enrollment in a loyalty program, etc.). Further, if certain marketing initiatives can be correlated to that specific action, those can also be employed most effectively. Essentially, exploring the data in depth with predictive analytics can engender much more effective marketing, down to the individual customer.

Perhaps the most notorious example of predictive analytics involved retailer Target. Marketing decision-makers there approached one of the firm’s statisticians, Andrew Pole, to help them identify expectant mothers. Research suggests that many things bought at a store like Target are habit purchases. If those habits include buying certain products elsewhere (e.g. groceries), they are hard to break, even if the consumer is a Target shopper otherwise and the store carries the same products. But research also shows that major life events can trigger new habits. Nothing says major life event like a first baby. Marketers were interested in identifying expecting mothers at the beginning of their second trimester, then giving them incentives to develop a habit of shopping at Target for everything they would need before and after the newborn’s arrival.

Target has a sophisticated big data program, tracking all points of contact with customers indicated by a Guest ID number. Credit cards, website visits, and other identifiable interactions with customers are fed into
the system and combined with other demographic and behavioral data, perhaps obtained externally from commercial suppliers. Pole started by looking at customers in the baby shower registry, identifying variables that might indicate a pregnancy. A common purchase like a natural body lotion or calcium supplement indicates little by itself but starts to point to pregnancy when combined with other variables. Eventually, 25 variables, when viewed together, were found to provide a fairly accurate “pregnancy prediction score”. A hypothetical example provided in a *New York Times* article suggested that cocoa butter lotion, zinc and magnesium supplements, a purse large enough to also carry diapers, and a bright blue rug would generate a high probability that the buyer was about three months pregnant. In terms of what we were discussing earlier, Target was able to bring together these variables and use them to predict purchase behavior in a specifically defined segment, making a strong case for the power of predictive analytics. The results were accurate enough to result in at least one complaint from a father about his teenage daughter receiving promotional offers related to pregnancy. It turned out she was pregnant, and Target knew it before the father did.

One last major approach to marketing analytics is clustering. Again, this is conceptually similar to cutting data and to predictive analytics. In fact, there can be overlap, as we’ll discuss. But the point is to use multiple variables to group subjects together that have similarities. In a particular dataset of shoppers, for example, you might have a substantial segment all in the range of 24–30 years of age, female, single, employed in service industries, college degree, urban, and income $35000 – $50000. Another substantial segment might all be 55–60, male, married, self-employed, some graduate study, suburban, and income >$80000. Other descriptive data, behaviors and attitudes might be added to these descriptions, including how good a customer they might be. The point is that with big data you might find more and richer variables concerning who they are, enabling you to better understand differences between the segments and how to approach them.

Clustering is apparent in what Spotify does with its listener pool. The idea of identifying listeners with similar tastes in music is based on grouping them together when their playlists overlap. So if two customers both like Jason Isbell, Chris Stapleton and Rhiannon Giddens, they can be grouped together and studied for other similarities, as well as how they might be dissimilar compared to another group of customers. It’s much more complex than that, of course, and might get into hundreds of artist preferences and contextual details in defining the clusters. But the concept is very simple.
As noted, clustering could be used for predictive analytics as well. If one of the variables is a choice of interest (if most of a cluster who like artists A, B and C also like artists X, Y and Z, then others who are only identified with A, B and C might also like the latter choices), the extension can easily be made. But it has other purposes as well, mainly the deeper understanding that comes with ever more precisely defining clusters by descriptive variables, attitudes, behaviors and other items.

One final point to make in this section is the ability of these large databases to take in new information, analyze it, and adjust descriptions, predictions or other outputs accordingly. In traditional statistics, this is referred to as Bayesian statistics, specifically the belief that knowledge about a true condition is based on a probability but that the probability could and should change as new data about the true condition become available. In more contemporary terms, big data and marketing analytics are able to take advantage of artificial intelligence and machine learning. The analytical processes are able to learn from successes and failures. If predictions are right, that is incorporated into future predictions. The same thing is done if predictions are wrong. The waves of data provided by these big data systems enable continuous learning and real-time adjustments in analysis to take place.

**BLOOMBERG**

Bloomberg, owner of *Businessweek* magazine, a cable business news network and, perhaps most importantly, the ubiquitous Bloomberg terminals used throughout financial services industries, is a pre-eminent digital/big data player. The firm generates 1.3 billion data points per day from its various digital properties. Some of this is available to advertising clients, clients generating tons of data themselves (though not always used effectively, generating opportunities for Bloomberg’s operation).

In 2015 Bloomberg launched an offering called B:Match. The aim is to work with advertisers using its media outlets to identify appropriate target customers, add other available data to profiles, and then use the profiles to drive the media plan, tailoring messages and media to specific target profiles. As Bloomberg explains, its strategy is the opposite of Facebook. Instead of a single point of contact for virtually everyone, Bloomberg provides multiple points of contact for a very select group of consumers: affluent businesspeople. This group is in great demand for advertisers, not only for business-oriented products they might buy in their job-related
Big data and marketing analytics

roles but also for luxury or high-end products they might buy for their personal lives.

The purpose of Bloomberg, whatever the medium, is to inform its consumer base. Even before digital video became commonplace, the provider was comfortable with it, given its various distribution options (television, online as a supplement to the magazine, and online independently). It had the experience and talent to supplement and combine all the various outlets, adding to the experience of customers. Essentially, it can provide the information in whatever format customers desire. Consequently, it has a substantial and loyal customer base of affluent business professionals. And most of them are registered or otherwise identifiable by Bloomberg. Their business media consumption habits are trackable, by person, and can be supplemented with other data, often from external sources. Bloomberg typically has deep and detailed records of its customers.

Bloomberg has developed a number of initiatives to try to take advantage of its customer base, its knowledge of its customer base (big data), and what it can do with that knowledge, especially regarding advertisers. Advertising, of course, is its principal revenue source, so development of the database is essentially a means to make advertisers happy. B:Match is designed to find appropriate individual targets, flesh out their descriptions, and then help clients tailor their media plans to reach them. Context could also be supplied, including variables such as time of day and favored medium (for that time).

Examples provided include an unnamed airplane manufacturer that could use B:Match to identify Bloomberg users who might be customers, allowing them to track visits and activity at various Bloomberg sites. The advertiser could monitor topics in which they might have an interest (e.g. oil prices), observe the target individuals’ activities on sites related to these topics, and then quantify interest in various advertising outreaches. Similarly, an advertiser seeking individual investors was able to identify them via Bloomberg’s data on use of home computers to visit properties and tracking data showing previous stops at Yahoo Finance. Again, further data could be added to the descriptions once they were identified.

What can be done with the data and profiles once acquired? One client was targeting young, individual investors with display advertising (e.g. banner ads). With B:Match, they were able to identify targets, fill out descriptions, and then start to collect even more data on reactions to different ad choices. Choices such as optimal times of day, what types of content best
complemented the ads, and even creative matters like color palette could be determined from the resulting database. KPIs for the advertising rose 200–400 percent after optimizing these and other variables.

B:Match is also easily matched with other Bloomberg offerings for advertisers. Social Connect 2.0 allows segmentation according to the social media platform from which the user was drawn to the Bloomberg site. Trendr provides a way for advertisers to “buy in” to a trending story widget, tracking user behavior from when they first noticed the topic through their access of the story and subsequent behavior. Finally, Bloomberg Denizen offers advertisers a platform from which to use the media company’s data to create infographics or other branded content to be run as “native advertising”, sponsored content that looks similar to regular editorial pieces at the outlet (though explicitly identified as sponsored advertising). Insurance provider Zurich, for example, created 40 pieces of content with Bloomberg’s help, including articles, infographics and video, that could be run across media platforms (online, print, etc.). Data were also provided to Zurich that helped with targeted specific pieces according to user profiles and behavior, social media trends and other variables.

Bloomberg Media is an example of a number of the practices described in this chapter. On one level, it constantly takes in data on use of its sites, apps and other content outlets. This is wide-ranging data, including background on users, their behavior while engaged with Bloomberg, and additional data from external sources. As noted, 1.3 billion additional data points every day meets just about anyone’s definition of big data. These data can be collected and shared, as with the big data systems we discussed, evaluated against key performance indicators set by Bloomberg itself (media circulation and ratings, website traffic, app usage, Google Analytics) or by its advertisers (views, click-throughs, or other actions by targets). The data can also be analyzed for further insights and/or adjusted according to algorithms. If a particular piece of advertising is getting better results than others given certain situational variables, the system might automatically adjust to run it more in those circumstances and less in other circumstances. And, as noted, Bloomberg and its advertisers can analyze data and results to predict responses, such as users fitting a specific profile will click on a specific piece of native advertising or will respond more positively to a banner ad in a certain shade of red. Application of just about any big data or business analytics technique covered in this ad should be seen as potentially applicable to what Bloomberg and its advertisers are doing.
SUMMARY

This chapter has provided background on the concepts we refer to as big data and marketing analytics. As suggested by the structure of the chapter, these can be two different phenomena though they are also complementary and effective when used together. With an understanding of how big data works, you are more prepared to see how it affects marketing research and decision-making. Similarly, some exposure to analytics techniques prepares you for some of the advanced data processing you’ll see in practice in upcoming chapters.

The key points to note are the wide range of inputs now available for data coming into organizations; how such data are organized, stored and shared; and then how the data might be analyzed. On a regular basis, data are coming into organizations related to all customer touchpoints. In the real world, registered or self-reported demographics, psychographics and other lifestyle data, shopper transactions (cash register data), geo-location data (where they are and have been), customer comments (live or by phone or web) and multiple other trackable interactions are readily available and freely surrendered by consumers. In the virtual world, web browsing patterns, app use, social media, customer comments or contacts (email or social media) and, again, any number of other trackable interactions are also available. These data can be traditional quantitative inputs, collected and stored in digital format. Or they can be unstructured data such as images, videos, text or other such inputs that can now also be reduced to a digital format and catalogued for sharing or analysis.

Modern technology allows all of these forms of data to be collected and stored in real time. Again, big data is about volume (lots of data), variety (in different forms) and velocity (coming in on an increasingly frequent basis). Retailers and service providers such as Tesco and Caesar’s monitor an increasing number of data inputs from their physical facilities, tracking activity in multiple ways (admission to locations and events; digital connections to machines, cameras, checkout) as it happens. Virtual service providers like Spotify can constantly monitor not only customer music choices but also matters like context (time of day, type of device, location). As can Bloomberg, accumulating all kinds of data from its variety of web properties regarding customer activities (news, up-to-the-minute financial information, video, click-throughs) and preferences (site, device, format). All of this can be gathered, transmitted, categorized and kept because of the plummeting costs of storing and processing data, especially in the cloud.
From a big data standpoint, all of these inputs could be monitored and probably are. But the most important data are designated as key performance indicators and delivered to decision-makers, probably on specially designed dashboards. These dashboards present the data in whatever manner is desired, perhaps raw data, perhaps tabulated or cross-tabulated, or even in visual formats such as charts, graphs or gauges. KPIs are chosen by the organization to reflect its priorities. We’ve already discussed a number of Tesco’s KPIs as well as some key indicators tracked by Caesar’s, such as the length of lines or heavy losses taken by designated priority customers. Similarly, Spotify can track highest played songs, activity in specific locations, device trends and other matters. Bloomberg can track content popularity, content format, click-throughs, device choices and similar activities of interest. And the data requests/dashboard visualizations can be changed as priorities change. Again, it is up to strategists and decision-makers to specify the data they want to see on an ongoing basis.

Researchers will also be asked to look more deeply into the data. Sometimes this will be for a specific project, looking for a specific answer to some question. Sometimes it will just be investigation with no particular objective, just mining the data to see what can be found that could be interesting. Analysts can do this by cutting the data in different ways, by different variables to see what distinctions might be apparent in the results. Spotify can do this by slicing data sets into ever smaller groups, perhaps by location or musical genre. At a very specific level, it could even do so by particular artists or songs. Once you’ve separated Drake from Meek Mill listeners or Apple from Samsung users in Helsinki, you can delve into all the other variables to see what differences there might be (demographics, location, lifestyle, app usage and so forth).

Alternatively, analysts can conduct predictive studies, studying groupings of variables to see which are associated with some outcome of interest. Or clusters of variables might be identified that merit further examination even without a clear current objective. Bloomberg can find variables identifying decision-makers in energy or the airline industries. Their news preferences, including topics viewed, or choice of metrics to be regularly streamed across their devices could indicate such segments. These variables might also indicate a receptiveness to targeted advertising, including some of the new communication offerings discussed earlier. Longer views of online ads, click-through rates or other metrics might be connected to such designated segments. An ad targeted to the airline industry may generate better outcomes when shown to micro segments identified by behavior...
on Bloomberg properties and related demographics (from registration or combined with other data from other sources).

We’ll talk about all of these things as we move through the book, providing details and context. Again, the main objective is for you, as a marketer, to have a good understanding of what big data and analytics have to offer. You can see the outcomes, from successful applications, combined with enough background on the underlying data management and analytical processes to understand how practitioners got there. We’ll start with the different research designs employed in traditional marketing research.

**BIBLIOGRAPHY**


