6 Data Basics

One of the visible consequences of the rapid proliferation of electronic transaction processing systems is the ubiquity of business data. It has been estimated that, on average, the volume of data doubles about every eighteen months—not surprisingly, data storage and management expenditures can be quite significant. What is, however, surprising is how comparatively little many organizations spend on "transforming" data into actionable insights—typically, only about 5% of data-related expenditures are allocated to analysis of data. It seems there is a tendency to overlook the obvious, namely, that unless it is utilized to guide the organizational decision making, the expenditures necessitated by the ongoing data capture can create a drag on earnings, rather than contributing to the firm's competitiveness. The reason for that is the relative immaturity of the broadly defined marketing analytics, and more specifically—database analytics.

Robust database analytics require a degree of understanding of not only statistical techniques, but also the fundamentals of data, including data structures and the key characteristics of the different types of data. This chapter offers a broad overview of the latter.

Data and Databases

Organizations in virtually all industries allocate considerable economic resources to their database informational infrastructure, all with the goal of trying to become more "fact-driven" in the business decision making process, particularly within the realm of marketing. Capturing, storing and managing of business data is a centerpiece of a thriving industry, which encompasses (data) capture and storage hardware, (data) manipulation and reporting software and (data) exploration-focused consulting. However, the results of billions of dollars of aggregate database infrastructure spending are all too often disappointing—they turn out a barrage of descriptive reports that individually and cumulatively yield disproportionately little predictive, decision-aiding knowledge.

Today, virtually all mid-size and larger business organizations either already have in their possession or have access to a variety of transaction-recording and/or descriptive data. In fact, the vast majority of these organizations own multiple databases, maintained and used by numerous functional areas, such as sales, claims management, human resources, industry analysis, marketing and so on. And, as pointed out in the opening chapter, most organizations subscribe to the flawed belief that "data is an asset," which is to say that they hold on to the belief that the often considerable expense required to capture, store and maintain the ever-growing volumes of diverse data is justified as an investment in "organizational intelligence." Underscoring that unwavering conviction is the fact that, in total, over the past 25 years or so, businesses in the U.S. alone invested in excess of $1 trillion in data-related infrastructure, but with very mixed results. Some, including Walmart, Google, Capital One, Harrah's or Marriott, to name a few, clearly benefited from their data-related investments; many others, however, ended up putting a lot more into the database endeavor then they ever were able to get out of it. In fact, it could be argued that, overall, the database revolution did more for the fortunes of data service suppliers than it for the competitiveness of an average database using organization.

Let me reiterate the point I made in the opening chapter: Data is not an asset—it is a *potential* asset, just as having talent is only an asset if and when it is put to a productive use. In other words, data is a resource, a raw material of sorts, which needs to be made into something useful before it can have value to its holder. Walmart did not overtake K-Mart because it had more data—it did so because it was purposeful and methodical about systematic analysis of its data (which, by the way, was not fundamentally different than K-Mart's). In short, the then-up-and-coming retailer made the exploration of its sales and other data the very heart and soul of their decision making. In other words, Walmart managed to "squeeze" a lot more out of its data, which in turn greatly increased the efficacy of the company's decisions.

Getting more out of data is a function of two, somewhat related considerations. First, it requires what could be called "an intimate knowledge" of data sources. Do not forget—the vast majority of data capture is a by-product of business process digitization, particularly what is broadly termed "electronic transaction processing." It means that business databases tend to be large in terms of size and esoteric in terms of content. Typically, they encompass millions, often billions, of records and hundreds or even thousands of individual metrics, many of which are far from being intuitively obvious. The bottom line: The attainment of robust knowledge of a particular database requires dedicated effort, which is perhaps why an average user will just "scratch the surface"…

The second prerequisite to getting more out of data is the amalgamation of dissimilar data into a singular analytical source. Now, if getting to know a single database seems like a lot of work, getting to know several and finding a way of combining their contents could well be considered a Herculean undertaking. And frankly, it can indeed be a hard and an arduous process. Is it worth it? Any organization not convinced it is, should probably reconsider stockpiling data in expensive databases.

As illustrated throughout this and the remaining chapters, the most significant difference between information-savvy organizations and their data-rich but information-poor counterparts is the data analytical *know-how*. In other words, while virtually the same hardware and software technologies are available to all organizations; it is the power of the subsequent data exploration and utilization that determines the ultimate return on the overall data infrastructure investments. And it all starts with a solid grasp of the available data.

Databases in a Nutshell

A *database* is an organized collection of facts. Although we tend to associate databases with modern computer applications, databases as such existed long before the advent of modern electronic computing. Definition-wise, a telephone book found in nearly every household is as much a database as Walmart's 583 terabyte[1] mega system. In other words, a database can range from a simple listing of your friends' phone numbers written down on a single sheet of paper to a large corporate or governmental computerized system.

A *business database* is defined here as an electronically stored collection of facts that requires its own management system (i.e., commonly known as database management system, or DBMS for short) and specialized query and analysis tools . Furthermore, due to their size and complexity, most business databases also require specialized skills for ongoing reporting and knowledge extraction.

There are multiple ways of describing databases: by data type, purpose, content, organizational structure, size, hardware and software characteristics, etc. From the standpoint of database analytics, the most pertinent aspects of a database are its:

- *scope*, which considers differences between data warehouse and data mart;
- *content*, which specifies the form of encoding, such as text, multimedia or numeric;
- *data model*, which details the basic organizational structure of a database, including entity-relationship, relational and object-oriented.

Each of the key defining qualities of business databases are discussed next.

The Scope: Data Warehouse vs. Data Mart

Even limiting the database definition to business applications, database is still a very general designation. Overall, business databases can be grouped into the following two categories, briefly described in Table 6.1.

Table 6.1 Categories of Business Databases

Database Type	*Description*
Data Warehouse	Broadly defined as comprehensive data repositories focusing on enterprise- wide data across many or all subject areas. They tend to be subject-oriented (e.g., purchases, customers), time-variant (i.e., capturing changes across time) and non-updatable in the sense of new replacing the old (i.e., read-only and periodically refreshed with a new batch of data). Data warehouses are usually data- rather than task-oriented, application independent (i.e., can be hierarchical, object, relational, flat file or other), normalized or not (database normalization is a reversible process of successively reducing a given collection of relations to a more desirable form) and held together by a single, complex structure. “Custom” database analytical initiatives typically source their data from a data warehouse, but the analysis itself almost always takes place outside of its confines.
Data Mart	These are specific purpose data repositories limited to a single business process or business group. Data marts tend to be project-rather than data-oriented, decentralized by user are and organized around multiple, semi- complex structures. An example is a direct marketing data mart containing details of the customer base, individual campaigns (e.g., target list, offer specifics, responses), customer contact history, etc. Usually, a data mart contains a sub-set of the contents of data warehouse, which makes it informationally more homogenous and application-ready. Data marts can serve as just data repositories or, in conjunction with business intelligence applications can support ongoing performance dashboarding.

The term “database” is sometimes used to denote a data warehouse, and at other times a data mart . Even worse, it is not uncommon for an organization to expect data mart-like functionality from a data warehouse, simply because in view of some, a database is a database. Yet in the knowledge-creation sense, there is a vast difference between these two general types of databases. A *data warehouse* is merely a repository of facts, usually centering on transactions and related data. A *data mart*, on the other hand, is a specific application of a subset of all data warehouse contents, designed with a particular purpose in mind, such as product/service promotion. The hierarchical and functionality differences between the two are summarized below in Figure 6.1.

Database Content

Content-wise, there are a number of different types of databases, comprising several of distinct categories, detailed in Table 6.2 below.

Bibliographic and full text databases are traditionally associated with library informational services, such as ABI/Inform or LexisNexis, containing summaries or full texts of publicly available published sources, such as newspapers, professional journals, conference proceedings, etc. In a business sense, they offer a referential source of information rather than ongoing decision support. Multimedia and hypertext databases are one of many Internet and the World Wide Web related informational innovations that tend to be used more as businesses communication/(i.e., promotional) vehicles, rather than sources of decision-guiding insights. Although at this point these type of databases offer limited utility to marketing managers, the emergence (albeit, slow) of the Semantic Web[2] may radically reshape that.

The last of the four broad types of databases, numeric, is the primary decision support engine of organizational decision making. The content of numeric databases, such as the earlier described transactions, behavioral propensities or basic descriptors, coupled with the easy-to-analyze coding make these databases both statistically analyzable and informationally rich.

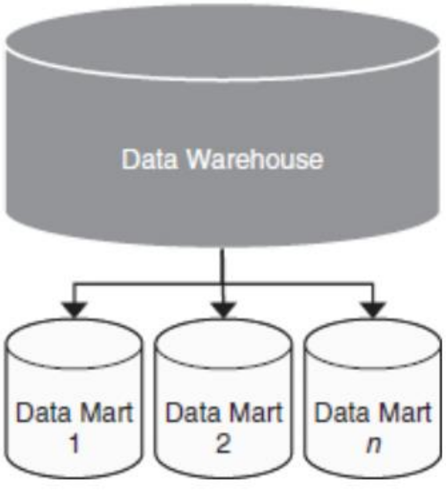

Figure 6.1 Data Warehouse vs. Mart

Table 6.2 Types of Databases

Consideration	*Description*

Bibliographic	Used to organize published sources, such as books, journals or newspaper articles; contains basic descriptive information about those items. Mostly used in library cataloging. Bibliographic databases are reference sources, not analyzable with traditional statistical techniques discussed here.
Full text	Contain complete texts of publications, such as journal or newspaper articles. Examples include Lexis database or *Encyclopedia Britannica*. Great qualitative sources of knowledge, full text databases can be sources of quantitative coded data, but are not directly statistically analyzable.
Multimedia & Hypertext	The most recent database type, largely responsible for the explosive growth of the World Wide Web. It supports creative linking of diverse types of objects, such as text, pictures, music, programs, into a single expression. These types of databases are representative of the modes of electronic, online content delivery, this are indirectly analyzable within the context of campaign management. However, they require a considerable amount of preparation prior to analyses.
Numeric	Used to store digitally coded data, such as purchase transactions, demographic information or survey responses. It is the staple of business database infrastructure and the focus of the analytical processes described here.

Data Models

Business databases that are designed to store transactional and augmenting data are almost always explicitly or implicitly numeric.[3] The information these data reservoirs contain can be organized in accordance with one of several data organizational models, which can be grouped into three general categories of entity-relationship, relational and object-oriented.

1. ENTITY-RELATIONSHIP DATA MODEL

The most basic data model, the entity-relationship model is built around parent–child hierarchical typology; it identifies basic organizational objects, such as a customer, transaction, age, account number and specifies the relationships between these objects. It is the simplest and the oldest of the three models, which means it is relatively easy to set up but also offers limited usability. The model's general logic is illustrated below:

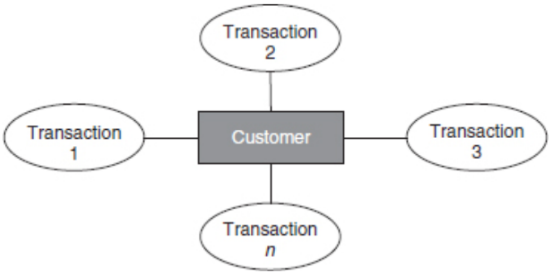

Figure 6.2 Entity-Relationship Data Model

2. RELATIONAL DATA MODEL

The relational model represents facts in a database as a collection of entities, often called tables, connected by relationships in accordance with predicate logic and the set theory. The relational model is more conducive toward automated, templated report generation, but it is also more restrictive in the sense of "on the fly" report generation, as the relationships need to be specified and programmed in advance. The term "relational" conveys that individual tables are linked to each other to enhance the descriptive value of a simple flat file. It is a lot more complex to set up than the entity-relationship model, but offers far greater levels of utility. Its general logic is illustrated in Figure 6.3.

3. OBJECT-ORIENTED DATA MODEL

In many regards, this is the most evolved data organizational model, but it is also least analytically flexible. The data structure is built around encapsulated units—objects—which are characterized by attributes and sets of orientations and rules, and can be grouped into classes and super classes, as illustrated in Figure 6.4. Although the individual "objects" exhibit a considerable amount of usage flexibility, their preparation requires a considerable amount of programming.

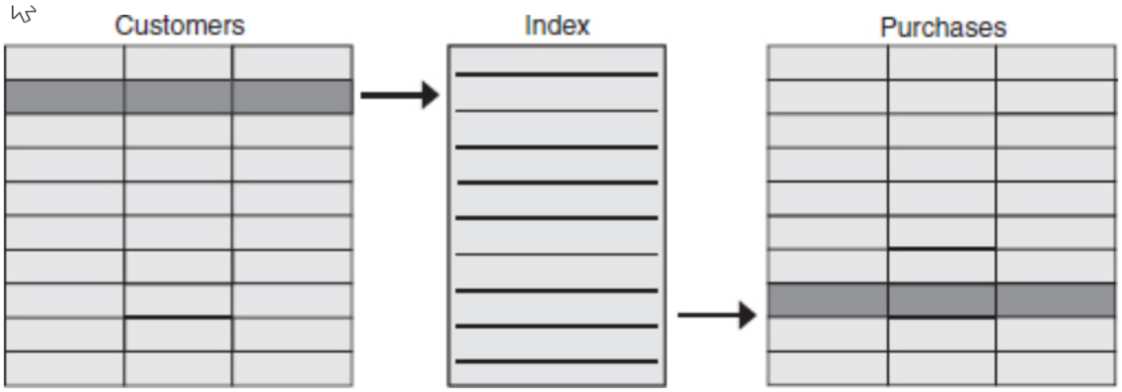

Figure 6.3 Relational Data Model

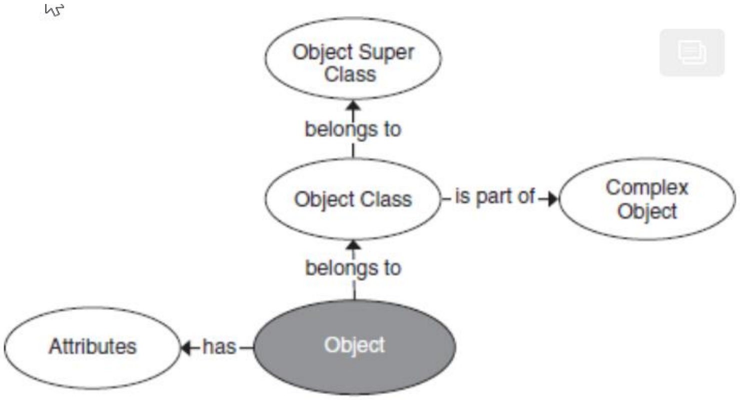

Figure 6.4 Object-Oriented Data Model

Database types and data model differences certainly contribute to the level of difficulty associated with developing a solid level of understanding of the fundamentals of database analytics. However, although these considerations have a significant impact on the database's querying and reporting capabilities, they exert only a relatively marginal impact on analytics. The primary reason for that is that querying and reporting are conducted within the confines of the database itself with the help of the database management system, or DBMS, while data analytics is usually carried out outside of the database. Thus the organizational structure of the database is of interest to database analysis only insofar as the identification and proper classification of the available data.

Data: Types

What is data? Quite simply, *data is facts*. Naturally, there is an enormous, almost infinite variety of data, which can be broadly categorized along two basic dimensions: *qualitative* or *quantitative* (dimension 1) and *root* or *derived* (dimension 2), as shown below in Figure 6.5 (the horizontal axis captures the "quantitative vs. qualitative" distinction, while the vertical axis shows the "root vs. derived" one).

Qualitative vs. Quantitative

Qualitative data represents facts recorded in a non-numeric fashion. In the business context, qualitative data often represents opinions, beliefs or feelings captured through structured or unstructured interviews, ethnographic research or simple observation—nowadays best exemplified by social networking data (e.g., Facebook, Twitter, etc.). In a scientific sense, it is typically used as the basis for deriving initial hypotheses, which are then tested on a more representative and larger sample.[4] While qualitative data captures a greater depth of a single individual's state, such as individual beliefs, it tends to be non-projectable (i.e., extrapolated onto a larger population) due to either small, usually non-representative sample size or cross-record informational dissimilarity. Also, due to the fact that it often needs to be extracted or inferred by a researcher, it is additionally prone to the researcher's interpretative bias. The most common business application of the qualitative data format takes the form of focus groups, frequently used in new product/idea testing. Due to its non-numeric format, qualitative data is not statistically analyzable[5] at the level required by robust knowledge creation.

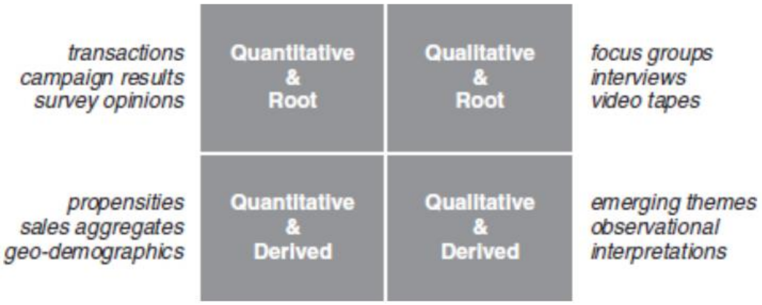

Figure 6.5 High-Level Data Categorization

Quantitative data are facts recorded in a numerical fashion. In business, the single most valuable source of quantitative facts are ongoing operations, which generate transactional records often referred to as the "secondary" data (so-named because its capture was secondary to business operations, such as sales, that generated it). Other, often used sources of quantitative business data include *surveys* (usually referred to as the "primary data," which reflects its purpose-specific origination) and the so-called *individual differences*, such as demographics and lifestyles or propensities to engage in specific behaviors. In general, quantitative data is statistically analyzable[6] and projectable, which is to say that it can support broader (i.e., stretching beyond the confines of a particular sample) generalizations.

The distinction between the quantitative and qualitative data types points to fundamentally different methods of collecting and codifying facts. The former offers a high degree of respondent representativeness and a numeric, objectively analyzable format, while the latter delivers non-numerically coded depth of an individual respondent's insight. As previously mentioned, qualitative data is often used as a basis for forming hypotheses, which are then tested with the help of quantitative data. For example, focus groups (a commonly used qualitative data capture mechanism) can be used to pinpoint a desired set of characteristics for a new product (i.e., to generate hypotheses about bundling individual product attributes), and a subsequent in-practice pilot (a good source of quantitative data) can offer an objective, generalizable estimate of the new product's market potential.

Although in principle both the qualitative and quantitative data types are potentially important sources of insights, applied marketing analytics tends to be primarily concerned with the analyses of the latter, for a couple of obvious reasons. First and foremost, corporate databases are essentially repositories of quantitatively coded or quantitatively codable data, which is a consequence of the systemic properties of most of the electronic business systems. Secondly, digitally coded data is easier to analyze objectively, as data and data analytical methods are independent of one another, which is to say that two different analysts applying the same method to the same dataset should arrive at identical or nearly identical results (assuming invariance in terms of any data imputation or re-coding). Analyses of qualitative data, on the other hand, are essentially inseparable from the analyst, as they represent subjective evaluation and/or interpretations, both of which can vary considerably across analysts. The inability to replicate and thus cross-validate a particular set of findings is a significant hindrance to marketing analytics; it effectively makes it impossible to compile sufficient amount of corroborating evidence for findings that might be hard to accept on face value.

Root vs. Derived

In contrast to the *qualitative–quantitative* distinction, the *root* vs. *derived* continuum represents the difference of "degree," rather than "type." Specifically, this distinction points to the degree of data aggregation, which is the process of combining multiple, typically narrowly operationalized elements into a new, single and more aggregate data point. In view of the fact that the aggregation process usually involves mathematically manipulating data, it is typically associated with quantitative data. The purpose for summarizing disaggregate data may vary across situations, but usually it is at least in some way related to the desire to increase the data's informational value, aid handling of large data files or to comply with legal requirements.[7]

Root data represents the most disaggregate form of storing facts in a database and it usually corresponds to the form in which data was originally captured. Depending on the source, such as individual purchase transactions, survey responses or product inquiries, this general data form can exist at varying levels of detail and abstraction.

Let's consider the purchase transaction data. The vast majority of larger retail outlets are equipped with electronic barcode scanners, which record individual transactions at the Stock Keeping Unit (SKU) level, which is the actual physical product being purchased.[8] The root data collected by those systems are the individual product–time– place–cost conjoints, which represent unique product transactions captured at a particular place and point in time. For instance, a 6-pack of 12 oz. cans of Diet Coke sold at 5:45 pm on March 21, 2012 at Kroger store #21034 for $2.99 is an example of root transactional data.

Clearly, root data can be quite voluminous. An average supermarket, for instance, stocks around 50,000 individual SKUs on a single store basis, while an average supermarket chain carries in excess of 150,000 SKUs overall (the difference reflects regional as well as store size variability). In terms of data, a mid-size chain of 500 or so stores captures several million transactions[9] daily, which adds up to several billion records in just a single year.

Although not quite that voluminous, accident-related claim data is another example of disaggregate root data. To stay with a retailer example used above, a large volume of "foot traffic" will also generate a considerable amount of (mostly minor) accidents, such as "slip and fall" and others. A large retailer with hundreds or thousands of location will quickly accumulate thousands of individual liability records.

Thus although powerful in terms of its informational content, data stored in root format can be extremely cumbersome to work with, in large part because of its size and detailed character. It has high potential value, but the insights it contains are far from self-evident. On a practical level, the management is rarely interested in tracking performance at the individual item level (e.g., individual accident claim, SKU) and in many cases it is simply not feasible due to the very high numbers of individual items. As a result, it is common to aggregate the root data to a more meaningful (and manageable) levels, such as a brand or a category. Doing so effectively converts root into derived data, because the sales levels attributed to the more aggregate entities of a brand and a category are in effect *derived* from the sales of their disaggregate components.

As previously remarked, the distinction between root and derived data types is best looked at in terms of difference of degree, or specifically, their respective levels of aggregation which ultimately translates into measurement precision. In general, the lower the level of aggregation of data, the higher its informational precision, simply because the process of data aggregation invariably inflates the amount of noise in the data, becuase exact disaggregate quantities (e.g., individual-level accidents) are replaced with less exact aggregate-level quantities. Hence, the root vs. derived level of precision differential tends to be inversely proportional to the amount of aggregation that might be deemed necessary. For example, rolling up individual item data (e.g., 2-liter bottle of diet Coke) to progressively more aggregate levels as represented by, for instance, the brand (Coca-Cola) and then the manufacturer (The Coca-Cola Company), will render the resultant data successively less accurate. Hence while on one hand data aggregation is necessary from the standpoint of informational parsimony, it carries with it an (often significant) amount of precision decay.

The reason for the precision decay is the error additivity associated with the data aggregation process. It increases the amount of imprecision embedded in the data because it is based on compounding of disaggregate data values (many of which may contain some level of imprecision), which typically involves averaging. Hence, the necessary evil of data aggregation should be considered in the context of the following:

1. Averaging creates data error, as the more precise disaggregate value are replaced with less precise aggregate ones. For example, computing average income for a particular consumer segment will lead to less accuracy for every individual household in that segment.
2. Compounding can magnify the already existing data errors, particularly systematic omissions, such as not capturing sales of certain SKUs, non-traditional channels, or counting promotional giveaways as sales (e.g., buy one, get one free). Overall, summary-based derived data may be inappropriate for certain transactional databased statistical applications, such as the estimation of action-attributable impact. In that sense, correct data categorization can have a strong impact on the accuracy of analytic results.

The reasoning presented above may seem somewhat counterintuitive in view of the conventional statistical "error cancellation" provision, where randomly distributed error terms sum up to zero as "over" and "under" estimates cancel each other out. This notion is one of the central provisions of the least squares estimation methodology employed by some of the most commonly used statistical procedures, such as regression. However, when applied without a proper due diligence to the analysis of large, transactional databases, the otherwise sound error cancellation rationale can lead to erroneous conclusions as it gets inappropriately stretched beyond its applicability limits. It all stems from a handful of fundamental differences between the nature and the behavior of statistical estimates and the actual data often found in large corporate systems.

First of all, as the name implies, *statistical estimates* are mathematically derived values, while data are facts. The former is arrived at by means of (usually) unbiased computational processes calibrated to ensure a random distribution of errors around the expected values, while the latter has no built-in error randomization provisions. Secondly, the term "error" carries vastly different meanings in the context of statistical analyses and transactional data.

Webster's Dictionary defines *error* as a "*departure from what is true, right or proper*," in other words, a mistake. In statistics, *error* is a difference between two values, such as actual and expected or a single estimate and a group mean. It is interpreted as (a typically expected) variability rather than a (typically unexpected) mistake, which means that quite often the former is a desired and a necessary prerequisite to statistical analyses. In contrast to that, an *error* in the context of raw, transactional data is simply a mistake in the everyday sense of the word. It is neither desired nor necessary, rather it is an unintended consequence of data capture and much effort is put into eliminating and correcting it.

What is of more importance to robust data analysis is that a purposeful statistical error exhibits predictable properties and behaves in an equally predictable fashion, which effectively eliminates the likelihood of aggregation-related skewing[10] (assuming, of course, a sufficiently large sample size). However, the "behavior" of fact-based data aggregation is quite different from that, insofar as data mistakes do not necessarily behave in a predictable fashion (i.e., are not randomly distributed). Thus it is not reasonable to expect that under-reporting of the purchase frequency of Brand A will be cancelled out by over-reporting of the purchase frequency of Brand B, when both are aggregated up to the level of a store or a region. Frankly, it would make more sense to expect that a process which under-reports the purchase frequency of one brand will also under-report the purchase frequency of another brand.

In the end, there will always be compelling reasons to aggregate detailed data up to a more meaningful level. It is important, however, to keep in mind the proportional increase in the likelihood of data errors associated with rolling up of disaggregate data. And since much of the database analytical effort ultimately translates into numerical conclusions, such as sales lift quantification or a propensity score, appropriately categorizing the data inputs on the "root vs. derived" continuum will go a long way toward, well, reducing the amount of error in those conclusions.

Data: Contents

There are multiple ways of categorizing data: by source, type, usage situation, etc. From the standpoint of (potential) informational content, data can be broadly divided into the *behavioral core* and the *causal layer augmenting the core*. The former are essentially transactions (such as purchases) or other behavior-denoting interactions; the latter encompasses a somewhat broader, type-wise, variety of metric classes, ranging from individual or geo-level demographics (or firmographics for businesses), attitudes and lifestyles (or industry classification for businesses), satisfaction surveys, field information (e.g., outgoing telemarketing for consumer markets or direct sales for businesses) and other types of action- or actor-describing details. Hence while the behavioral core metrics convey observed or observable outcomes, the (outcome) augmenting causal layer variables hide the potential causes, or at least correlates of the observed behaviors.

These rudimentary differences between the behavioral core and the augmenting causal information are indicative of the informational value of the two data types. Behaviors, particularly purchases or promotional responses, are typically the central component of any database analysis since they are most closely tied to revenue (i.e., they reflect actions tied directly to earnings) and as such can attest to the ultimate success or failure of business strategies. More specifically, from the standpoint of marketing strategy, the attainment of competitive advantage is analogous to realizing the greatest revenue or profitability gains at the lowest possible cost; hence it follows that explaining and predicting behavioral outcomes is at the nexus of marketing analytics.

This is not to say that the causal data is of trivial value—far from it. The root causes of behavioral outcomes and patterns can rarely, if ever, be understood without the explanatory power of causal descriptors, such as demographics, lifestyles, attitudes or NAIC (North American Industry Classification) codes for businesses. In that sense, the behavior-augmenting causal data contributes the "why" behind the "what." It is intuitively obvious that without the ability to single out specific precipitators of observed behavioral patterns, those patterns could not be understood beyond a series of hard-to-predict outcomes spuriously associated with business actions. Needless to say, it would be difficult, if not altogether impossible to derive competitively unique knowledge.

Ultimately, it is the synergistic effect of the *behavioral core– augmenting causal data* combination that is the most fertile source of business insights. However, to realize its full potential, a strong understanding of each of two rudimentary data sources is necessary.

The Behavioral Core

Most databases have highly operational origins and it is particularly true to the behavioral outcome-focused ones. Consider the rapid proliferation of bar-code-based point-of-sales (POS) scanner technology (first introduced in mid-70s), the electronic order processing systems and the virtual explosion of Web-related capabilities—all of those developments and resultant trends output massive amounts of data, largely as a byproduct of their operations. As discussed in the opening chapter, organizations tend to gravitate toward the belief that data, in and of itself, is an asset that warrants investing considerable monetary and human resources. As a result, quite a few organizations amassed tremendous amounts of behavioral information, often warehoused in multiple systems or a single, distributed albeit networked data reservoir. For instance, the Walmart Corporation has what is by many believed to be the largest non-governmental transactional database in the world, with its size estimated to have surpassed the 500 terabytes[11] mark. And although Walmart's database obviously represents an extreme in terms of its sheer size, many transaction-intensive organizations, such as those in retail or banking industries have amassed transactional databases ranging in size from 10 to around 50 or so terabytes of data. This is an incredible amount of analyzable data, which as shown later, can benefit the organization's knowledge pursuit on one hand, while limiting the applicability of some “staple” statistical techniques on the other.

Entities as Attributes of Events

Behavioral metrics are essentially actions, and depending on the source of data they can take on a different form, reflecting the two key dimensions of “who” (e.g., an employee, a customer, an organization) and “what” (e.g., a sale of a 48oz. box of Cheerios). From the standpoint of a database, behaviors represent “transactions” and as such, an individual record is comprised of a specific “event,” such as a purchase, as well as a varying number of “attributes” associated with that event, such as the date or the amount. However, because the capture of these individual data elements is a by-product of electronic transaction processing, many of the individual data tidbits may reside in separate transaction files, or even multiple data-recording systems, rather than a single, central data warehouse. For example, customer purchase, demographic and promotional response details routinely reside in separate files. Although largely an infrastructure-related consideration, it also has a profound impact on the subsequent analyses of the behavioral data, insofar as it leads to *data organizational schema inversion*, where *customers are coded as attributes of events* (e.g., transactions, promotional responses, etc.), rather than *events being attributes of customers*. Consider Figure 6.6.

In the hypothetical diagram shown above, the “event,” such as an in-store purchase (a product) is the central unit of analysis of the resultant record , while the multiple “entities” represent individual customers. In other words, the database transaction is structured in such a way that the “who” is an attribute of “what.” This is a tremendously important consideration, both from a philosophical as well as practical data manipulation and analyses standpoints. In regard to the former, it is simply illogical for customers to be attributed to products for the obvious reason that it leads to lumping of the otherwise dissimilar purchasers. Doing so diminishes the analyst's ability to reliably differentiate between different customers, which is ultimately the goal of behavioral predictions. In other words, the customers-as-attributes-of-events type of data organizational schema forces analysts to look at data from the standpoint of “what transpired” rather than “who engaged in a particular behavior,” which severely limits the efficacy of the resultant insights.

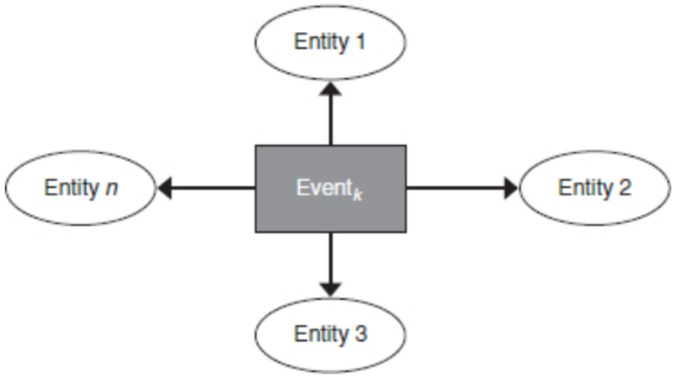

Figure 6.6 Entities as Attributes of Events

Attributing customers to events also carries with it important practical data manipulation considerations. On the one hand, it is conducive to event-level reporting (hence the widespread use of reporting-oriented business intelligence tools), while at the same time, it makes customer-level analyses far more difficult, because of the extensive data re-coding requirements (detailed later in this chapter). In a sense, it "pushes" data users into outcome-based database reporting (i.e., generic information) and impedes the more competitively advantageous explanatory analytics, and ultimately, the edge-producing knowledge creation.

However, because so much of the behavioral data is passively collected by the various electronic transaction processing systems, the customers-as-attributes-of-events data organizational schema is very common, as are reports summarizing product purchases without tying any of these behavioral outcomes to specific causes.

Events as Attributes of Entities

As pointed out above, viewing customers as properties of events is clearly neither rational nor conducive to insightful analyses; hence the relationship needs to be reversed so that actions associated with different events (i.e., behaviors) can be attributed to those engaging in the behaviors of interest. However, before that can be done, a considerable amount of data engineering is needed, in part because of the re-coding task itself, but also because there is quite a bit of invariance among the individual behavioral core data types. The *cross data type invariance* is particularly evident when attempting to amalgamate attitudinal "causes," best exemplified by survey-captured opinions, with transactional "effects," such as in-store purchases.

The most important, and at times most difficult step is that of attributing multi-sourced data elements to one another to establish a "cause–effect" relationship. In many instances it requires extensive additional data coding, particularly when analyzing high repurchase frequency products. Consider the product purchase cycle (e.g., awareness, consideration, trial, etc.) and the related, though separate in the database sense, promotion response behaviors (e.g., exposure, processing, etc.). At the time, a purchase will represent a response to promotional exposure, yet that "cause––effect linkage" is not expressly set up in the data. Unless individual behaviors or actions are appropriately coded and the corresponding behaviors are attributed appropriately, it may not be known which, if any, action caused the specific outcome.

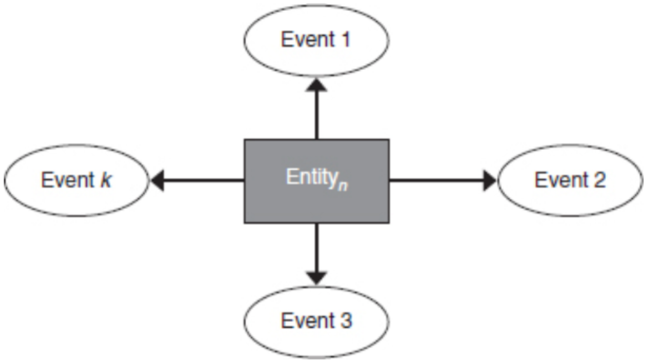

Figure 6.7 Events as Attributes of Entities

Once the source-specific data has been properly set up, the resultant data structure needs to be redefined from event-centric to entity-centric, as shown in Figure 6.7.

Expressing events as properties of entities, rather than the other way around, is more reflective of the underlying processes. Marketing management is about paying attention to specific events that impact the outcomes of interest to the organization. Some transactions represent events directly impacting those outcomes, while others may represent attempts at trying to manage their effects. To maximize the productivity of marketing programs—i.e., to bring about the greatest possible revenue or profit gains at the lowest possible cost—the organization has to have a way of quickly and efficiently "drilling" into the cost-precipitating events, which in turn requires an appropriately structured behavioral data. Ultimately, it means that database analytics initiatives need to be built around outcome-causing entities.

The desirability of focusing source-specific data on the "who" (i.e., customers), as opposed to the "what" (i.e., products) is also evident from the standpoint of informational efficacy. In particular, it is critical to making a transition from a mostly single-topic reporting (such as sales or promotional response rates) to multi-source data analytics (e.g., delineating and quantifying the drivers of sales variability or response rates). The latter requires the merging together of multiple, typically dissimilar data files, which in turn calls for the common-to-all organizational unit needs to be identified. In the vast majority of cases, the only consistent threat connecting the otherwise dissimilar data files is the "who," which means that expressing events as properties of entities is a necessary precondition of multi-source analytics.

Reconciling Dissimilar Data Attributional Schema

In most data types, there is a need to establish attributional relationships, akin to the cause–effect connection. For instance, the rudimentary unit of analysis in a transactional database is a transaction, which necessitates the attributing of an event, such as a store purchase, to another entity, such as a household or an individual consumer. The reason this seemingly redundant step is necessary has to do with the disparity between how transactional data is collected and how it should be analyzed. Regarding the former, virtually all transactions are captured one-at-a-time, as individual records. In other words, two distinct purchases made at different times by the same buyer will be treated as two distinct database records—in fact, in some databases each product within a "shopping basket" (all products purchased at the same store, at the same time and by the same consumer) may comprise a separate record. In a database sense, two purchases made by the same customer are treated the same way as two purchases made by two different customers. Obviously, the two scenarios yield different conclusions in the context of frequency and severity management. For that reason, the initial data capture steps have to include express rules for attributing (i.e., connecting) outcomes to sources. This is the *data source attributional schema* due diligence consideration.

In the realm of the emerging enterprise data infrastructure, a need typically exists to combine multiple sources of data in a single database environment, which gives rise to somewhat more involved *enterprise data attributional schema* considerations. The bulk of marketing systems' transactional data is a by-product of the digitization of business processes discussed earlier, where the "natural unit of analysis"—or the basic organizational structure of data—is an individual purchase, response or other action, rather than entities (e.g., buyers, promotional responders) behind those actions. As noted earlier, in those systems two or more purchases from the same source (e.g., a person) are treated as separate entities. Of course, each such database record typically contains the appropriate outcome-to-source linking identifiers, yet the outcome-to-source attribution is usually not expressly established. But again, because this data was not captured for analytic purposes per se (hence the often used name, "secondary data" given to transactional metrics), explicit steps need to be taken to make sure that the differences between independent, source-wise, and dependent events is expressly recognized.

This is generally not the case with behavioral, lifestyle or attitudinal surveys, all of which are examples of data that has been captured expressly for its own value (often referred to as "primary data"). In the database sense, it means that the aforementioned outcome-to-source attribution is usually already in place.

Aside from the collection method, these two data sources—i.e., primary vs. secondary—are separated by an important distinction: The former's organizational schema implies that outcomes are attributes of sources (which is the way we would typically think about the outcome-to-source attribution), while the latter's schema implies the opposite—namely, that sources are attributes of outcomes. This attributional schema difference—or the dissimilarity in the way that differently sourced data elements relate to one another is an important consideration in making the enterprise data model more "analysis-friendly."

Figure 6.8 visually contrasts the two schemas.

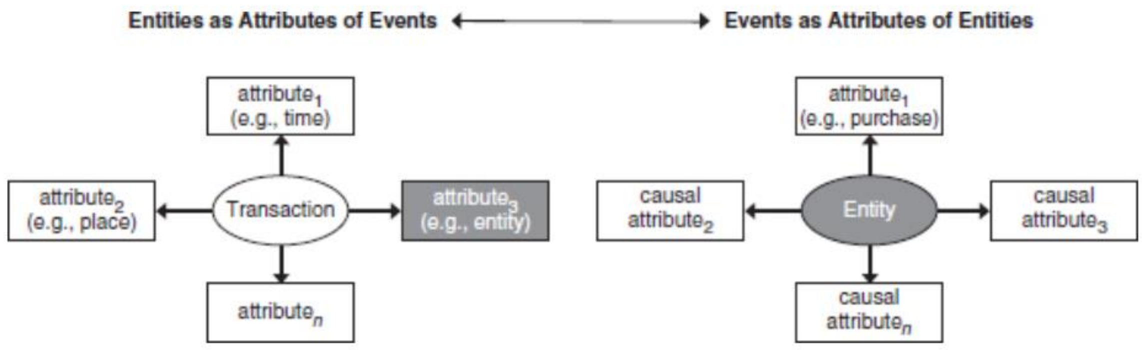
Entities as Attributes of Events
Events as Attributes of Entities
attribute$_1$ (e.g., time)
attribute$_2$ (e.g., place)
Transaction
attribute$_3$ (e.g., entity)
attribute$_n$
attribute$_1$ (e.g., purchase)
causal attribute$_2$
Entity
causal attribute$_3$
causal attribute$_n$

Figure 6.8 Entities as Attributes of Events vs. Events as Attributes of Entities

The above distinction is important because the knowledge creation process typically involves the culling of dissimilar data types, which requires resolving attributional schema disparities, which is one of the reasons the otherwise powerful databases often yield little unique knowledge. Frankly, it is far simpler to analyze individual data sources separately, in effect limiting the informational insights to a single, or a handful of schematically similar data sources. Taking information to the next level—which is turning it into competitively advantageous knowledge—usually calls for a satisfactory resolution to the seemingly inescapable data schematic dissimilarities. Overall, the inability to systematically synthesize multi-sourced data is among the most significant impediments to making that transition to the higher level of information utility (in practice, it is one of the key impediments getting in the way of enterprise databases models or the so-called "360° view" of customers becoming operational realities). As a topic of a considerable importance to the overall database analytics endeavor, transforming schematically dissimilar data types into a single format is fully discussed in the next chapter.

Let's look at a couple of specific and commonly encountered data types. The first is the transactional data which exemplifies the *entities as attributes of events* organizational schema. Here, purchases captured at an item level constitute the primary unit of measurement, with other aspects of behaviors acting as descriptors contextualizing the transaction. So, for instance, the purchase of item #12345 is associated with store $_{p}$, time $_{t}$, location $_{l}$, price $_{c}$, etc. In terms of the data matrix, rows constitute individual purchases while columns (variables) are the individual purchase characteristics. This type of organizational schema makes the basic outcome reporting relatively straightforward (hence, the widespread use of the automated business intelligence tools in this realm), but explanatory or predictive analyses (discussed in latter chapters) cannot be attempted without a considerable amount of data rearranging.

The second main organizational schema is one where *events are expressed as attributes of entities*, and is exemplified by behavioral and attitudinal surveys. In contrast to the above, sources (e.g., consumers) are the central unit of measurement to which everything else is attributed. In the practical sense, this means that in a basic data matrix, database records are individual survey responders, while columns are the record-describing variables. In other words, unlike the *entities as attributes of events* organizational schema describe above, this data model makes predictive analyses relatively easy to carry out. The challenge, of course, lies in establishing attributional unity between the two data types—doing so, entails two distinct data engineering steps:

1. MASTER-TO-EVENT THREADING

The process of structuring multi-sourced data into a coherent organizational schema requires *master-to-event threading*. In a sense, this process can be thought of as the establishment of a *hierarchical attributional schema*, where temporal or causal connections are established linking individual *events* and the *master* record. For instance, individual purchases or promotional responses (i.e., events) are all "threaded" to a single *master* (typically, a customer or a household identifier) linking together all of those data pieces. The purpose behind this data preparatory step is to establish cross-record standardization to support future *cause–effect* analyses. This is particularly pertinent for organizations wishing to establish an enterprise-wide view of their marketing activities.

2. VARIABLE CODING

Yet another data capture-related preparatory consideration relates to *variable coding*, which is the assignment or an explicit recognition of specific measurement properties to individual data elements. As a general rule, if measurement properties need to be assigned, the most analytically flexible scale should be used, which might be intuitively obvious, but unless the measurement property specification rules are expressly stated, less- than-ideal choices can be made during the initial data capture efforts.

From the statistical data analysis standpoint, numeric data can be either *continuous* or *categorical* (also called *metric* or *non-metric*, respectively). The former is comprised of variables measured either with interval or ratio scales, often referred to as *metric* data, while the latter can be measured with *nominal* or *ordinal* scales, and it is also called *categorical* data. In the knowledge creation sense, metric data has far greater informational value than the non-metric data, primarily because it permits a wider range of mathematical operations, making it analyzable by a wider range of analytical techniques. Table 6.3 summarizes the types of variables along with their most important characteristics.

In general, interval and ratio scales-measured variables are both assumed to be continuously distributed for the purposes of data analysis, while the nominal and ordinal ones are treated as categorical. It is important to note that while continuously distributed variables (i.e., metric) can be easily converted into categories[12] (i.e., non-metric), discrete metrics cannot be made into continuous ones. As a result, once data has been coded as categorical, certain commonly used statistical techniques cannot be deployed to analyze it.[13] Given its obvious importance, this topic is discussed in more depth later.

Table 6.3 Types of Scales

Trait	*Categorical*		*Continuous*	
	Nominal	*Ordinal*	*Interval*	*Ratio*
Definition	Labels of states that can be expressed numerically, but have no “numeric” meaning	Rank-ordering-based categorization; intervals between adjacent values are indeterminate	Fixed distance (with respect to the attribute being measured) rank-ordered categories	A scale in which distances are stated with reference to a rational zero
Permissible Operations	counting	*greater than* or *less than* comparisons	addition and subtraction	addition, subtraction multiplication and division

Example	gender, marital status	movie ratings (PG, R)	degrees F; attitude	degrees K; distance
Analysis	cross-tabulation	frequencies	mean	coefficient of variation

One of the most important lessons in statistics is that *correlation is not causation*. This statement captures the belief that term “causation” carries a burden of proof that goes far beyond what is required to ascertain a simple, though persistent association, aka, correlation. Hence, what makes certain types of data “causal”? The answer begins with an up-close look at the basic tenets of the concept of causation.

Ascertaining Causation

Given its central role in knowledge creation, the notion of *causality* or *causation* has been a subject of centuries-long debate among scientists and philosophers alike. At the core of the debate has been the line of demarcation separating cause-and-effect from just simple concurrence-based relationships. Is factor A causing B, or do the two merely coincide? According to Hunt,[14] an explanation has to meet four separate criteria before it can be classified as causal. Those are shown in Table 6.4.

As pointed out above, temporal sequentiality and associative variation have a very simple meaning and application, in spite of the somewhat foreboding names. Occurring in sequence (A followed by B) and doing so persistently is both intuitively obvious and relatively easy to demonstrate. For instance, if certain marketing actions (such as a direct mail campaign) precede, time-wise, observed business outcomes and that relationship is recurring, that would generally constitute sufficient basis to attest to temporal sequentiality and associative variation.

Table 6.4 Causality Criteria

Criterion	*Description*
Temporal Sequentiality	Changes in factor A to be used to causally explain factor B must precede in time the occurrence of changes in B. Thus an attributed action must precede the observed outcome before it can be considered a cause of it. It is an intuitively obvious and an easily established requirement in practical business analysis.
Associative Variation	Changes in factor A must be systematically associated with changes in factor B. In other words, in order to conclude that certain types of promotions lead to higher sales, one must be able to observe systematic sales increases associated with those specific marketing actions. Again, a logical and usually relatively easy to meet requirement.
Non-Spurious Association	If A causes B, then there must be no factor C which, if introduced into the explanation would make the systematic A–B association vanish. Thus if a particular direct mail campaign is indeed one of the causes of sales gains, factoring out another metric, such as online advertising, should not nullify the *direct mail campaign–sales increase* relationship. Unlike the previous two requirements, this is a far more difficult

	condition to meet, primarily because data is not always available to make that determination and even when it is, its quality can vary considerably among sources.
Theoretical Support	If A causes B, is it consistent with an established theory X? If the aforementioned direct mail campaign indeed leads to higher sales, what is the theory that explains that dependence? Since practical business analyses are typically not concerned with abstract theoretical explanations, this particular requirement is rarely satisfied, though it is worthwhile to keep in mind.

It is less so with the remaining two causality thresholds: non-spurious association and demonstrated theoretical support. Even though many behaviors tend to be somewhat repetitive, the mere fact that they tend to be a part of a larger set of behavioral interdependencies makes the requirement of proving their non-spurious nature a difficult one. In addition, many of these activities are highly pragmatic—i.e., they do not espouse to adhere to specific general theories and their roots are often in fact, spurious ideas and decisions. Frankly, the pursuit of competitive advantage demands that firms take steps that are uniquely more advantageous to them than their competitors, rather than pursuing strategies that have been proven to work the same way for everyone else.

So what is the conclusion? Establishing *causality* in business analyses should be held to a somewhat different, frankly lower standard than in theoretical research. This may sound almost blasphemous to some, but let's consider some of the hallmark differences between practical and scientific endeavors. First and foremost, business analyses are typically concerned with uncovering unique though sustainable sources of competitive advantage, in contrast to scientific investigations which almost always are focused on formulating and testing generalizable, i.e., universally true, knowledge claims. This means that business and theoretical analyses both share in the requirement of ascertaining temporal sequentiality and the associative variation of the potentially causal relationships, but it also means that in contrast to the theoretical pursuits, business analyses can conclude that the relationships of interest are causal without conclusively demonstrating a clear theoretical support or the non-spurious nature of the said association. In other words, a particular cause–effect relationship does not need to be universally true (meaning, equally valid for all other organizations) in order to be a source of competitively advantageous decisions by a particular organization—it only needs to be persistent, or hold up across time.

Second, scientific worthiness of findings is demonstrated through their generalizability and only implicitly through longitudinal stability, while the value of practical business analyses is demonstrated almost exclusively through longitudinal persistence of results. To a firm, it matters little whether particular dependencies can be generalized to other firms or industries (frankly, as suggested above, the pursuit of a competitive edge would argue the opposite), but it is tremendously important that these relationship hold as expected when resources are invested into future business initiatives, whether these are promotional programs or other capital expenditures. In other words, while theoretical research aims to create universally applicable knowledge, business research strives to uncover the few dimensions of knowledge that are uniquely applicable—i.e., advantageous—to a particular firm.

Third, marketing analytics are contextualized by focus and data. Even the largest organizations only compete in a subset of all industries and their data and data analyses are a reflection of the scope of their operations. In other words, the focus of business analyses is the world in which a given organization operates, while the focus of theoretical investigations naturally transcends any idiosyncratic industry or a set of industries. Thus the resultant knowledge claims—including causality—should be evaluated in the proper context.

Therefore, as it is used in this text, the term “causal” will apply to data that can be used as the basis for establishing or validating enduring relationships demonstrating temporal sequentiality and associative variation. The same standard will be applied to ascertaining causality in the context of the database analytical process.

A Closer Look

In the context of marketing analytics, *causal data* are any data elements that systematically contribute to explaining the variability in the core data, most notably, in frequency and severity of loss-causing events. Causal data allow us to look past *what* happened to trying to understand *why* it happened, which is a crucial step in the development of predictive analytical models.

Overall, causal data encapsulate entity traits, preferences and propensities. *Entity traits* are the basic descriptors such as demographics or location. *Preferences* are either demonstrated or attributed choices, exemplified by psychographics, lifestyles or preferred product categories. And lastly, *propensities* represent modeled probabilities of engaging in a particular course of action, such as the likelihood of a particular consumer or a household responding to a specific marketing incentive.

Content wise, causal data can be broadly categorized as either business- or individual-focused.[15] At the center of the *causal business data* is a collection of basic individual or household characteristics collectively called *demographics*, which includes the basic descriptive traits such as age, income, education, etc. A second important grouping of business causal data falls under the umbrella of *demonstrated propensities*, which include the willingness to catalog-shop, buy online, respond to direct mail solicitations, etc.

In terms of its *source*, causal data can be either observed or derived. *Observed data* is factual and usually it is reported directly by entities, such as individuals or businesses—examples include the demographic information supplied by consumers through product registration, or the lifestyle groupings into which individuals self-select. *Derived data* is probabilistic and it typically is a result of sample-to-population or geographic area-based generalizations, as illustrated by promotion response propensities or geodemographics.

The *level of aggregation* reflects the granularity of the data, which can be individual, a group of individuals or a particular area. Individual-level causal data reflects the characteristics of a single individual, usually the customer in the context of marketing. A group of individuals, such as household, reflects the characteristics of the group that is expected to engage in a shared decision making. Lastly, area- or geography-based causal data represents location-derived estimates of individual characteristics, such as geodemographics, which are block-level (typically about 20–30 households, on average) estimates of the key descriptors derived from the more granular U.S. Census data.[16]

The considerable differences in data granularity across the three levels of aggregation suggest a number of tradeoffs, most notably between accuracy and coverage of the individual-level causal data, as shown in Figure 6.9.

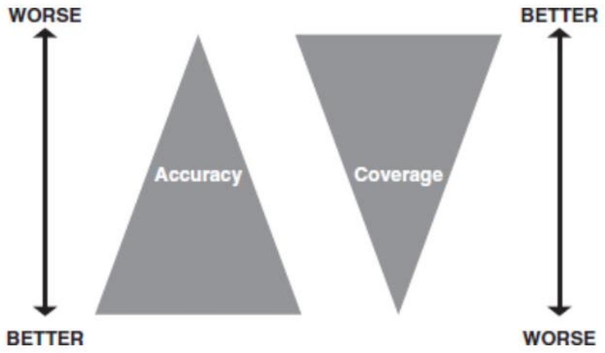

Figure 6.9 Accuracy vs. Coverage Tradeoff

As depicted above, there is an inverse relationship between accuracy and coverage: As one increases, the other one decreases. The most disaggregate consumer data is available at the individual level—obviously, it is the most accurate information because it is factual, insofar as it depicts the actual characteristics of those engaging in behaviors of interest. However, it is also the scarcest, i.e., it is only available for a relatively small fraction of all individuals, and at times it may not be accessible, for legal and other reasons. Group-level data is less accurate because it is probabilistic, i.e., it expresses the most likely, rather than actual characteristics. But at the same time, it offers a better coverage since it is not limited to observed outcomes. Lastly, geography-based causal data is least accurate since its values represent estimates derived for large groups of households, but it offers the most complete coverage because it is derived from the U.S. Census, which covers the entire country.

Naturally, the quality of data plays a significant role in the validity and reliability of analytical insights. Causal data plays a particularly important role because it is suggestive of actions that can enhance the performance of marketing initiatives. Furthermore, in a strictly technical sense, it also has a pronounced impact on the quality of statistical models it helps to power, particularly as it relates to what is known as "model specification" (fully discussed later). In essence, a correctly specified model will have a relatively small number of statistically and substantively potent predictors, which is a logical as well as practical consequence of the *principle of parsimony* discussed in the opening chapter. Poor quality causal data will translate into poor explanatory power on one hand, while also contributing to an unnecessary inflation in the number of predictors (while offering little additional explanation).

It Is Not About the Digits

For all their complexity and the often hefty price tags, databases are fundamentally nothing more than large collections of digitally encoded facts, which require considerable effort and skill before they are of any value to decision makers. Although the investment in the database infrastructure is manifestly about the creation and dissemination of decision-aiding information, the amount of effort and resources put into knowledge creation activities pales by comparison to how much is spent on hardware and related infrastructure. Consider that on average about 85% of all database-related expenditures are consumed by hardware (i.e., storage), about 10% is spent on software and only as little as 5% on the analysis of stored data. As a result, a typical scenario looks something like this: A mid-size or a large organization spends millions of dollars (many millions, in case of the latter) to erect complex and expensive data storage and management facilities, but stops short of committing comparable resources to harvesting the raw material locked away in various databases. Nor is our hypothetical organization overly concerned with the lack of compelling evidence to suggest that these expensive databases are worth the investment… Once again, virtually all organizations have data, yet only a small subset of them are able to turn it into a source of competitive advantage.

Yet overtly, the reason organizations make database investments is to outwit their competitors, or at the very least, maintain competitive parity. Hence the database paradox: *The larger, more complex and comprehensive a database, the less likely it is to give rise to competitive advantage*. It is another way of saying that organizations tend to "choke" on the amount of data at their disposal. Quite often, large data volumes give rise to an even larger volume and array of reports, which tends to amount to nothing more than color-coding of raw data, which is the practical consequence of reporting on outcomes without clear cause delineation (which almost always requires more involved analytics). Untold man-hours are spent pouring over disparate pieces of information, but ultimately very little competitive-edge-producing knowledge is created, all while decisions continue to be driven more by intuition than by facts. Under this (common) scenario, building and maintaining of large corporate data reservoirs becomes a goal in itself, with the creation of competitively advantageous knowledge getting lost in the shuffle.

Obviously, not every organization falls into that trap. As mentioned earlier, Walmart's database, believed to be the largest of its kind at an estimated 600 terabytes (and growing) helped to propel it to the elite group of the most dominant (as measured by revenue) and most influential (as measured by market power) organizations in the world. Walmart's efficacy at persistently extracting knowledge out of the otherwise overwhelming quantities of raw data clearly illustrates the power of systematic and skillful data exploration. It also illustrates that merely translating digits into text—i.e., converting data into information disseminated as reports—is not enough to gain competitive advantage. To truly reap the benefits of its database investments, an organization must develop complementary knowledge-extraction capabilities.

The Data–Information–Knowledge Continuum

Utility-wise, databases exist for two basic reasons: First, they enable an ongoing capture and storage of facts; second, they serve as platforms for inferential knowledge creation. When both reasons are combined, databases become conduits for transforming data into information, some of which may lead to decision-aiding knowledge, ultimately giving rise to competitive advantage. Figure 6.10 below summarizes the *Data → Information → Knowledge* progression, shown in the context of each of the step's incremental value to users, interpretational challenges and the level of benefit.

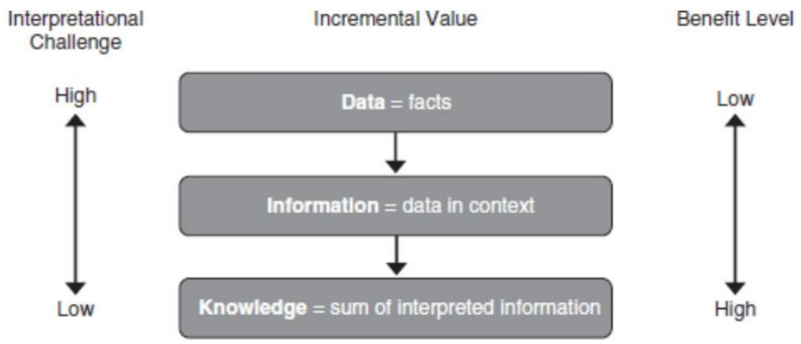
Interpretational Challenge
Incremental Value
Benefit Level
High
Low
Data = facts
Information = data in context
Knowledge = sum of interpreted information
Low
High

Figure 6.10 The Data–Information–Knowledge Continuum

From the standpoint of users, raw data, which is simply a repository of digitally coded (in the case of business databases) facts, represents the lowest level of utility. It is both the most interpretationally challenging (obviously, thousands or millions of seemingly random digits mean very little) and it embodies the least amount of user benefit, precisely because it does not clearly communicate anything … Regardless of its format (i.e., root vs. derived), source (i.e., transactions vs. surveys) or organizational model (i.e., relational or other), raw data will just about always present a considerable interpretational challenge because of its sheer volume, cryptic nature and the lack of self-evident differentiation between important and trivial facts.

However, once the raw data are converted into information, the resultant interpretation becomes easier and it is of more benefit to users. Its value increases as a result, but it still might be hampered by limited actionability. For example, translating individual purchases (raw data) into period or store-level summaries (information) certainly increases the benefit and lightens the interpretational challenge, but the resultant information is still of little benefit to decision makers as, in this case, it says little about the potential driver of the observed outcomes. In other words, while useful, information still needs additional “refinement” before the value of the database investment is truly maximized. In the example used here, that point is reached when the purchase-detailing information is combined with attitudinal details, promotional history and demographics as well as other individual difference variables, such as past purchase history or promotional response propensity, to ultimately give rise to *knowledge* hinting at the most effective marketing strategies. In a more general sense, information represents extracting findings out of raw data, while knowledge corresponds to application of these findings in the decision making process, a progression first described in the first chapter.

Hence the value progression implied in Figure 6.10 can form a foundation for thinking about data analyses. The *Data “Information” Knowledge* value continuum can be used to illustrate the most fundamental difference between database reporting and database analytics: The former enables the conversion of raw data into information, while the latter facilitates the creation of knowledge, which is the ultimate expression of the value of data. Figure 6.11 illustrates this distinction.

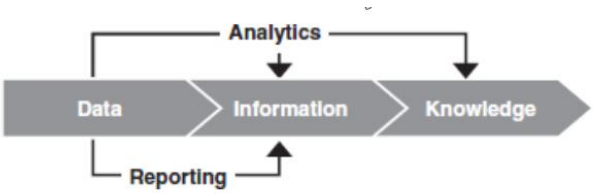

Figure 6.11 Data Value-Added Progression

The informational value creation process outlined above carries important usage considerations surrounding large, corporate databases. At the most rudimentary level, databases can be used to support ongoing business performance reporting through performance "dashboards," usually built around a pre-selected set of metrics of most importance to the organization. Such reports are of particular interest to line managers with vested interest in keeping an eye on operational aspects of business. The format of these reports, as well as their content and frequency, tend to be shaped by factors such as data availability, industry characteristics and the organization-specific needs. In general, database reporting tends to be data type specific (e.g., point-of-sales data is used to create brand-, product-, period- or location-specific sales summaries, while direct mail details are used to create promotional impact reports), making it difficult to cross-reference key pieces of information, and altogether impossible to draw cause (promotion)–effect (sales) conclusions. In essence, basic data reporting provides important though generic information, which means it is not likely to give rise to sustainable competitive advantage. However, many of these reports or dashboards can be produced with highly automated business intelligence software, thus requiring little-to-no advanced data analytical capabilities, which is one of the reasons behind their popularity.

Going beyond the mere status quo reporting, a more robust analytical set of processes can help in translating the often disparate pieces of information into higher level inferences, or knowledge. Of course, it is not quite as easy as it may sound. As pointed out earlier, the ability to distill large volumes of markedly dissimilar information into specific, competitive-advantage-producing insights hinges on a combination of a forward-looking informational vision and a robust analytical skill set. Converting raw data into (competitively) generic information can be handled, for the most part, with the help of highly automated, commercially available database tools. Funneling the often still voluminous and almost always inconclusive information into unique knowledge cannot be handled by standardized, off-the-shelf database reporting applications for reasons ranging from extensive data preparation requirements to the intricacies surrounding multivariate analyses (discussed in later chapters). Furthermore, the creation of competitively unique knowledge quite often necessitates the amalgamation of multiple (and otherwise disconnected) data sources into a single, yet multidimensional causal chain, which in turn requires the establishment of cross-factor correlations, cause– effect relationships as well as the more technically obtuse interaction and nonlinear effects.

It follows that it is considerably more difficult to develop robust *information-to-knowledge* conversion processes than it is to put in place robust database reporting, or data-to-information conversion capabilities. Of course, the higher level of difficulty carries with it more substantial benefits. As exemplified by Walmart, Capital One or Harrah's, it enables a migration from just generating competitive parity type of insights and toward the establishment of a source of sustainable competitive advantage. In a financial sense, the ability to systematically translate generic information into unique knowledge vastly increases the implied return on database-related investments.

More on Reporting vs. Analytics: Tools and Applications

The simplest way to extract information out of a database is to query it. *Querying* involves relating individual data elements to create specific information. For example, to get at the “cost by region” information, the DBMS used to run the queries needs to divide all available sales into the appropriate regional “buckets,” create region-by-region summaries and return appropriately formatted information. The speed and the agility of database querying capabilities vary across the type of data model used, with the entity–relationship model offering the lowest levels of querying speed and agility, while the relational model tends to deliver the highest levels of performance in that regard.

Database querying is a somewhat manual process, requiring some level of technical proficiency. For example, accessing of a relational database, which is arguably the most dominant data storage mechanism in business today, requires familiarity with querying protocols of the particular relational database type, such as Microsoft Access, Oracle, IBM DB2 or MySQL. Combined with the repetitively ongoing nature of many of the informational needs, much of the ad hoc *database querying* is usually replaced with standard, automated *database reporting*. Business intelligence tools, which commonly provide the automated reporting functionality, rely on standard templates and predefined process to repeatedly generate the same set of reports, often in fixed time intervals. Thus rather than querying the database about period-by-period and/or region-by-region sales manually, an automated sales report is generated without the need for manual querying.

Unfortunately, such generic, standard database reporting processes are frequently confused with database analytics, so much so that in the eyes of many, analytics is synonymous with reporting. Although there are certainly similarities between the functions—both entail manipulating and translating raw data into the more meaningful information—there are sharp differences separating these two types of endeavors. Perhaps the most important is the type of data processing. Reporting relies primarily on summarization, tabulation and contrasting, all with the goal of generating basic *descriptive* conclusions about the underlying data. More analytically advanced data exploration often starts with basic descriptive analyses as well, though its ultimate goal typically entails forward-looking extrapolations and predictions. In other words, database analytics goes far beyond the status quo reporting, by offering causal explanations and making decision-guiding predictions. In that sense, basic database reporting offers descriptive summaries of past events, while the inferential knowledge-focused database analytics supports *probabilistic interpretation* of data, as illustrated by Figure 6.12.

Different missions call for different tools. Database querying and reporting typically utilize the database's own DBMS capabilities, usually in the form of outside database reporting tools known as “business intelligence solutions,” or BI tools for short (technical fields love acronyms!). These tools, exemplified by Business Objects, MicroStrategy or Cognos, help to automate report generation while also deepening database exploration. It is important to note, however, these tools effectively become a part of the DBMS and as such, operate within the database itself. In addition, the resultant reports are externally consumed and are not used to enrich the informational value of the database itself.

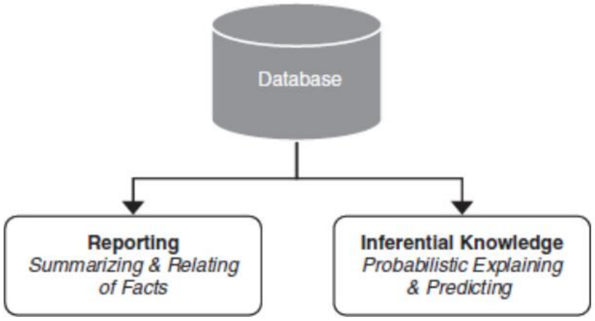
Database
Reporting
Summarizing & Relating of Facts
Inferential Knowledge
Probabilistic Explaining & Predicting

Figure 6.12 Factual vs. Probabilistic Data Exploration

The necessary coupling of hardware and software applications that are required to support an operational database environment gave rise to a new class of database computational devices known as *database appliances*. The fundamental difference between a database appliance and a traditional stand-alone hardware and software approach is that the former is a turnkey solution integrating the hardware and software components into a single unit, an appliance. The single unit performs all the functions of a database, a server and a storage device considerably faster[17] and thus might be preferred in situations where large volumes of standard reports need to be generated under tight time constraints. From the data analysis standpoint, the fundamental underlying assumption of this class of database devices is that it is desirable to analyze the contents of the entire database (i.e., to use all available records), and it is in the context of such "full database queries" that the greater speed of database appliances is most appreciated. However, as will be shown in the subsequent chapters, the database analytical process of translating data into competitive-edge-producing insights makes a heavy use of appropriately selected samples, the drawing of which can be quite cumbersome in a database appliance.

Although the specific characteristics of the database management may impact the speed and the ease of performing certain operations, data analysis and modeling are ultimately impacted far more by the power and the efficacy of the statistical analysis applications. The three most dominant and widely used systems—SAS, SPSS and R—are functionally independent of the DBMS and are in effect, open-ended methods of addressing data-related issues, rather than means of funneling data into pre-defined templates (although all three can be incorporated in most database designs, including the aforementioned appliance and used for basic reporting). What differentiates SAS, SPSS and R from lesser-known applications is their depth and comprehensiveness, in particular as it relates to data processing, management and manipulation, statistical analysis and modeling, database scoring and output management.

From the standpoint of the resultant information, all data management systems, inclusive of the earlier discussed business intelligence as well as the data analysis and modeling applications, can be used for either *descriptive* or *predictive* purposes. The former is focused primarily on retrospective outcome summarization and tabulation, while the latter is typically tasked with forward-looking decision support. Although both tend to be labeled "analytics," it is more correct to refer to retrospective outcome summarization as "reporting" and forward-looking decision support as "inferential knowledge" As depicted in Figure 6.13 below, reporting functionality tends to be database-resident, as it is quite conducive to automation.

Hence the business intelligence solutions mentioned earlier—e.g., MicroStrategy or Business Objects—are embedded in the database itself. On the other hand, the statistical data analysis systems—e.g., SAS and SPSS—tend to be stand-alone applications, primarily because the knowledge creation process they enable is not conducive to automation.

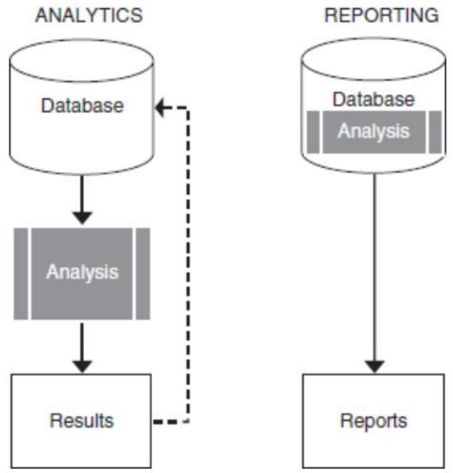

Figure 6.13 Database Reporting vs. Analysis

The Case of Data Mining

One particular aspect of database analysis called "data mining" does not fit in with the rationale outlined above. Broadly defined, *data mining* is an information extraction activity designed to uncover hidden facts contained in the data with the help of an automated system combining statistical analysis, machine learning and database technology (it should be noted that data mining can certainly be performed manually, but in the eyes of many, the term itself became synonymous with automated data pattern identification capabilities). So although data mining resembles database reporting in terms of its degree of automation, its methodological engine aligns it more closely with inferential knowledge discovery. At the same time, its goal of "roaming the database to find noteworthy relationships" differentiates it sharply from conventional database analytics focused on testing specific hypothesis in support the organizational strategy-driven informational needs. Overall, data mining is distinctly different from the goal-driven database analytical process described in this book, yet it is an alternative inferential knowledge creation avenue, as shown in Figure 6.14 below.

I should point out that data mining received a lot of attention over the past decade or so, particularly in the area of "enterprise data management." Leading software business applications developers, including the two largest statistical analysis software vendors, SAS and SPSS (the latter now a part of IBM), invested in complex systems built expressly to sift through large volumes of data in search of (statistically) significant relationships. However, there are a couple of key hurdles that, to the best of my knowledge, have not yet been overcome.

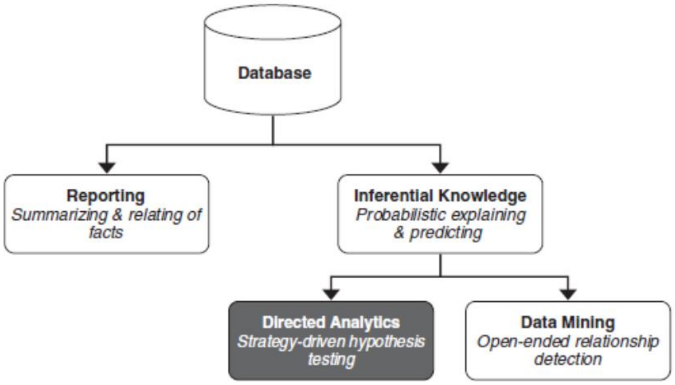
Database
Reporting
Summarizing & relating of facts
Inferential Knowledge
Probabilistic explaining & predicting
Directed Analytics
Strategy-driven hypothesis testing
Data Mining
Open-ended relationship detection

Figure 6.14 Database Exploration Venues

The first one is methodological in nature. In order to differentiate between "significant" and "not significant" relationships, data mining applications employ established statistical significance tests (of which there are different types, tied to data characteristics), which as a group are a set of techniques for determining if a given relationship is spurious (i.e., it represents a chance occurrence) or persistent (those tests are discussed in more detail in subsequent chapters). As fully discussed later, the reliability of statistical significance tests is negatively impacted by sample size, which leads to practically trivial relationships (e.g., extremely small correlations or differences) being deemed as "significant," in a statistical sense. Given the "pass vs. fail" nature of statistical significance tests, all "significant" relationships are deemed to be important, which will lead to lumping together of the truly important and trivial, or even spurious findings. In the context of a large database, for which data mining was intended, the resultant listing of statistically significant findings may be so extensive that finding the truly important ones may prove to be quite taxing.

The second key shortcoming of automated data mining is of more pragmatic nature, reflecting the relative inability of such systems to deliver sustainable informational—and ultimately, competitive—advantage. Recall the data and application commoditization discussion. The bulk of the transactional and related data is not organization-unique, just as electronic transaction processing systems generating the data are not unique. Similarly, data mining applications are commercially available applications. Combining two relatively generic entities will lead to generic outcomes, just about guaranteeing the lack of unique, competitively advantageous insights.

The limitations of machine search-based systems notwithstanding, these technologies proved to be tremendously important insofar as they paved the way to the exploration of the largest, though historically inaccessible source of data: text. The next section offers a closer look at what has come to be known as "text mining."

Textual Information

An important, though often largely inaccessible (for large scale analyses) source of data is *text*, defined here as words treated as data. In contrast to numerical data, the bulk of which is a product of machine-to-machine communication, textual data is a result of human communication, which means it is ubiquitous and analytically challenging. There are numerous reasons for why textual data is analytically challenging, but they can be reduced to two primary factors, which are *structural variability* and *volume*. The former encapsulates the intricacy of human communication—the interweaving of explicit and implicit elements, structured vs. unstructured modes, the multiplicity of meanings associated with many commonly used terms, to name just a few. The latter is self-evident, though perhaps it might be instructive to more explicitly define that quality: The World Wide Web contains more than 7 billion indexed primarily text containing pages[18]; Google, the widely recognized online search leader (accounting for about half of all internet searches) grows its search database by more than 100 million searches per day; the Library of Congress, the world's largest library (boasting more than 130 million items, such as books, maps, photographs, etc.) contains an estimated 20 terabytes[19] of textual data. Yet perhaps the most telling illustration of the volume of textual data is social media, where Facebook alone sits atop of more than 30 petabytes of data—more than 3,000 times the size of the Library of Congress.

Textually encoded information, as exemplified by social networking data, is important to marketing analytics as it contains "unfiltered" expressions of consumer motivational factors that play a key role in determining consumer marketplace behaviors, especially as it relates to brand and other choices. In fact, within the confines of consumer products, it could be argued that the broadly defined *consumer voice*, which encompasses all electronically captured consumer communications, is one of the strongest predictors of the future performance of consumer brands. Historically, consumer voice was qualitative and non-generalizable in nature—the former referring to the subjectivity of its analysis, while the latter underscoring non-representativeness of the resultant conclusions.[20] That said, the relatively recent emergence of wide scale *social networking*,[21] coupled with advances in text mining analytic technologies is beginning to reshape how we think about—and how we analytically approach—non-numeric, or qualitative consumer data.

Social Communication and Marketing Analytics

Social networking, as a field of academic inquiry, is built around an idea (really, an axiom) that the key to understanding social interactions is to focus on properties of relations among network participants, rather than on the participants themselves. However, when trying to develop an understanding of social-network-encapsulated interactions in the context of a brand, product type or company, it is myopic to focus exclusively in social relations. To a business entity (a company, a brand, etc.), the "who" of social networking is as important as the "what" because the former contextualizes the latter. For example, the management of a consumer brand would want to differentiate between qualitatively expressed opinions of the brand's high-value customers and those of its current non-purchasers or casual (i.e., low-value) buyers for a number of reasons, such as to get a better understanding of the brand's "perception gap." This is suggestive of two somewhat distinct outcomes of analyses of text data: 1. qualitative-research-like extraction of the "emergent themes"; and 2. digitization of key text-encoded metrics. The former can be considered an end outcome in and of itself, while the latter is an input into broader analysis, where insights gleaned from text analysis are related to those garnered from analyses of consumer behaviors. In more operational terms, it means that textually and numerically encoded data need to be—ultimately—amalgamated into a single analytic dataset. Under most circumstances, doing so is a fairly tall order—to appreciate why that is the case, it is instructive to take a closer look at some of the fundamental analytical properties of text data.

Structured vs. Unstructured

Much like a typical numeric data file, text data files are two-dimensional matrices, where columns demark individual variables and rows delimit successive cases. Given that, a key consideration in machine-aided analysis of text is the organizational composition of data files, with a particular emphasis on the relationship between successive rows and individual columns in the data matrix. Within the confines of the successive rows vs. individual columns layout, text files can be categorized as either structured or unstructured.

Structured data follows a repeatable pattern where each successive data record (row) mimics the layout of the previous one and columns demark distinct variables. Form-based information, such as warranty cards that accompany household appliance purchases are a good example of this category of text—each purchaser completes the same type of information captured in the same manner. Naturally, given the fixed and repeatable format of structured data files, it stands to reason that so-formatted files are far more amenable to machine processing, though at the same time, the informational value of outcomes of those analyses might be somewhat limited. Why? Structured data files' computer processing malleability is rooted in their layout—as mentioned earlier, to be considered "structured," a text data file has to follow a basic two-dimensional matrix layout where columns contain distinct fields and rows enumerate successive data records. Given that, the only truly meaningful difference between structured text and numeric data is that the individual data fields (i.e., variables) comprising the former contain text rather than digits. Although somewhat less obvious, it also means that the informational content of those fields tends to be skewed toward categorical, rather than syntactical substance[22] (think of the warranty cards that usually accompany household appliance purchases). As a result, the output of the analysis of structured text tends to be limited to the identification of patterns and/or communalities across records, which is obviously worthwhile, though at the same time only represents a fairly small subset of the informational content of text-encoded communications. Hence, the task of reliable machine-aided analysis of structured text data is a fairly manageable one, but the resultant insights tend to be fairly limited.

Unstructured data , which are by far the most common form of text files,[23] do not adhere to a recognizable layout schema. The layout of individual records comprising unstructured data can be characterized as being amorphous and lacking discernible organizational schema—it is non-repeatable in the sense that each successive record is independent of the previous one and any similarity is coincidental; furthermore, individual columns do not delimit distinct and repeatable variables. Examples of unstructured or free-flowing text abound, as most of the written human communication, ranging from books, newspapers and the like to contents of billions of web pages and social communications tend to fall within that category. The lack of structure coupled with the inherent ambiguity of "natural" (meaning, used in human communication) language makes unstructured data quite difficult to deal with algorithmically, primarily because, in principle, each data record is different from all other records, both in terms of the layout as well as content.[24] The upside to the absence of a rigid organizational schema is that unstructured text tends be syntactically rich, which means that although it is possible to focus the mining efforts on the identification of cross-record patterns and communalities (as is the case with structured text), the true value in analyzing unstructured text often lies in surmising the "deeper meaning" hidden in what and how is being communicated. However, it can be a daunting task, primarily because natural language depends heavily on "common sense knowledge," which is the understanding of the implied meaning of expressions that goes beyond technical definitions of individual terms. This is one area where precision-demanding machine processing is—in spite of over-promising terms such as "artificial intelligence"—far inferior to the human mind. Thus, unlike the relatively manageable task of categorizing structured text, unstructured textual data is exceptionally hard to encode logarithmically, which means it does not succumb to traditional data analytic techniques.

Given the considerable differences separating structured and unstructured text, it follows that there are a number of different methods that can be used to machine-process large volumes of text data. Broadly referred to as "text mining," these techniques are summarized next.

Text Mining Approaches

Text data are very common—by some estimates, a typical business organization has anywhere between two and ten times more textual than numeric data. Analyses of textual data, which are overwhelmingly exploratory (as opposed to hypothesis testing) in nature, are commonly referred to as *text mining*, which itself is considered a subset of a broader category of *data mining*.[25] As applied business analytic notions, both data and text mining tend to be associated with machine-aided processing, largely because of the computer applications first introduced in the 1990s by companies like SAS, SPSS[26] or IBM; though in principle, those terms apply to all exploratory analyses, whether or not aided by machines. Thus, an analyst using a computer-based application to explore a large repository of text data, such as customer comments posted on a brand's website, or a researcher reading a transcript of a focus group are both examples of text mining. Naturally, "manual"—reading and summarizing by humans—mining of text or other data is only operationally feasible when the quantity of data is fairly small.

In many regards, business organizations (and others) tend not to think of text as explorable data, primarily because systematic and wide-scale analyses of large volumes of text-encoded data have been largely out of reach. For the vast majority of companies, the only meaningful exposure to non-numeric data analysis usually came in the form of qualitative marketing research, as exemplified by focus groups or other forms of *ethnographic research*.[27] From the standpoint of decision-guiding knowledge, the very attribute (i.e., data quantity that is small enough to be manually analyzable) that makes qualitative marketing research studies analytically manageable also diminishes the utility of is informational outcomes. For instance, an average focus group engages a group of 8–12 people in a guided discussion spanning about 90–120 minutes, which yields the amount of data that is small enough to allow "manual" processing, which usually means a properly trained researcher or a team of researchers reading and summarizing participants' text-expressed opinions and comments with the goal of arriving at a set of findings, commonly referred to as "emergent themes." Of course, even with a fairly small dataset it is nonetheless a time-consuming and a laborious process—even more importantly, however, such analyses are also permeated by readers' subjectivity, which is an inescapable consequence of the stated goal of qualitative research: to "interpret" the comments/opinions contained therein. In practice, there is no meaningful method of quantifying, much less eliminating researcher bias, which means it is a problem without a clear solution. Hence, even if it was operationally feasible to "manually mine" large text databases, such as the totality of consumer comments relating to a particular brand, the potential reviewer bias, or more specifically, inter-reviewer variability (given that such endeavor would require many raters in order to be completed within a reasonable amount of time) would clearly cast doubt on the validity of conclusions.

The above considerations underscore that to be operationally feasible as largely free of *rater bias*, analyses of large reservoirs of text (and numeric) data require machine processing. Thus, although as pointed out earlier any exploratory analysis of text data can be characterized as "text mining," within the confines of the marketing database analytics process detailed in this book it refers to the discovery of trends and/or patterns in (typically) large quantities of textual data with the help of automated or semiautomated means.

Process-wise, mining of textual data entails four distinct steps: 1. retrieval, 2. summarization, 3. structural mining and 4. digitization. *Retrieval* pertains to searching for records of interest that are hidden in large repositories, such general or topical databases; the outcome of the retrieval efforts will usually take the form of an extract. *Summarization*, on the other hand, involves the condensing of otherwise large quantities of (textual) data; its outcome typically takes the form of an abstract or a synopsis. S *tructural mining* is probably the broadest, as well as most complex aspect of text mining as it involves converting voluminous text data into (statistically) analyzable, categorized metadata.[28] Depending on the combination of purpose and the type of data (discussed below), structural mining can mean searching for predetermined expressions or attempting to surmise the deeper meaning hidden in the syntactic structure of text. Lastly, *digitization* refers to the process of number-coding of metadata, or converting nonnumeric data values into numeric ones. The ultimate goal of that conversion is to facilitate amalgamation of form-dissimilar (i.e., text vs. numeric), but complementary-in-meaning text-mining-derived insights and already digitally coded numeric data, with the goal of enabling *multi-source analytics*.[29]

Function-wise, text mining can be performed with several different outcomes in mind, such as: 1. *summarization*: identifying co-occurrences of themes in a body of text; 2. *categorization*: reducing documents' content to pre-defined categories; 3. *clustering*: reducing documents' content to emergent (i.e., based on documents' content, rather than being pre-defined) categories; 4. *visualization*: re-casting textually expressed information into graphics; 5. *filtering*: selecting subsets of all information, based on predetermined logic. It is important to note that, though quite dissimilar, the individual text mining functions can be viewed as complements, as each delivers a distinctly different end-user informational utility.

Method-wise, mining of text data can take the path of either frequency count and tabulation or natural language processing. The *frequency count and tabulation* approach itself can take one of two paths: 1. "tagging" of a priori identified expressions, or searching a body of data for specific expressions or terms that have been spelled out in advance of the search; or 2. "term funneling," where instead of using a priori lists, the starting point of the analysis is the generation of comprehensive frequency counts of all recurring terms. The former—*tagging*—requires a substantial amount of knowledge on the part of the analyst, to the degree to which specific expressions or terms have to be identified as important ahead of the search; as such, it is deductive in nature, which is to say it is focused on answering specific questions stemming from the hypothesis formed at the outset of analyses. Furthermore, simply searching for terms that have been identified as important beforehand is not conducive to uncovering new "truths," as the focused mechanics of deductive search make it difficult, if not practically impossible, to notice unexpected results. The latter of the two frequency and tabulation data mining approaches—*term funneling*—requires no prior knowledge hence in contrast to tagging it is inductive in nature, but it can produce overwhelmingly large quantities of output (tens of thousands of terms and/or expressions), which in turn will demand a substantial amount of post-extraction processing. It follows that it is not only time-consuming, but also likely to infuse potentially large amounts of the earlier discussed *rater bias*, effectively reducing the objectivity of findings. Overall, their differences notwithstanding, tagging and term funneling are focused strictly on pinpointing of terms without considering the context or the way in which those terms are used. In other words, using the frequency count and tabulation method, it might be difficult, if not outright impossible to distinguish between positively and negatively valenced lexical items.[30]

The second broadly defined approach to text mining—*natural language processing* (NLP)—attempts to remedy the limitations of the frequency count and tabulation methodology by attempting to extract insights from the semantic[31] structure of text. NLP is an outgrowth of computational linguistics (itself a part of a broader domain of artificial intelligence), which is statistical and/or rule-based modeling of natural language. The goal of NLP is to capture the meaning of written communications in the form of tabular metadata amenable to statistical analysis—as such, it represents an inductive approach to knowledge creation, well adept at uncovering new "truths." Given the significantly more ambitious goal of natural language processing, namely, objectively summarizing and extracting the meaning of nuanced human communications, the level of difficulty associated with this endeavor is significantly higher, which means that the reliability of findings will typically be proportionately lower.

Although NLP clearly offers a potentially deeper set of insights, it is also fraught with difficulties that directly impact the validity and reliability of the resultant findings—on the other hand, the comparatively more superficial frequency count and tabulation is straightforward to implement and can deliver a fairly consistent—keeping its limitations in mind—set of insights. All considered, both text mining approaches have merit and in order to gain a better understanding of the applicability of both methods, a more in-depth overview of presented next.

Frequency Count and Tabulation

As outlined earlier, the goal of frequency count and tabulation approach to text mining is to identify, tag and tabulate, in a given body (known as "corpora") of text, predetermined terms and/or expressions. Conceptually, it can be thought of as a confirmatory tool as it is focused on finding concepts that are already (i.e., prior to search) believed to be important, rather than identifying new ones—i.e., pinpointing concepts that were not previously believed to be important. As a result, the efficacy of the frequency count and tabulation approach is highly dependent on prior knowledge, as manifested by the completeness of the a priori created external categories.

However, even the most complete external schemas in no way assure robust outcomes as the search or the mining process itself can produce incomplete findings due to ambiguity stemming from wording or phrasing variability and the potential impact of synonyms and homographs. The *word* or *phrase variability* is a syntax (principles and rules used in sentence construction) problem stemming from the fact that the same term or an idea can oftentimes be written in somewhat different ways. A common approach to addressing word- or phrase-related variability is to use the so-called "stemming algorithms" which reduce words to their Latin or Greek stems with the goal of recognizing communalities. *Synonyms*, which are yet another potential source of search ambiguity, are terms that have a different spelling but the same or similar meaning, while *homographs* are terms that have the same spelling but different meaning. The most common approach to addressing those sources of possible confusion or ambiguity is to create external reference categories, which are de facto libraries of terms delineating all known synonyms and homographs for all a priori identified search terms. Clearly, a rather substantial undertaking…

Process-wise, the frequency count and tabulation approach to mining textual data makes use of *text transformations*, defined here as the process of translating words into digits, where each word is an attribute defined by its frequency of occurrence.[32] This process is built around a notion of "bag-of-words"—a transformation-simplifying assumption which posits that a text can be viewed as an unordered collection of words, where grammar and even the word order are disregarded. Considered in the context of the two types of textual data discussed earlier—structured and unstructured—text transformation can take on somewhat different operational meanings: Transforming structured data typically involves supplanting textual fields with numerically coded and a priori delineated categories, which effectively translates textual expressions into numeric values. In that context, it is a relatively straightforward process because structured textual data tends to exhibit relatively little ambiguity, due to heavy reliance on predetermined lexical categories.[33] At the same time, the process of transforming unstructured data is considerably more involved because of an open-ended nature of that source of data—in principle, terms appearing in unstructured data can be thought of as unconstrained choices, as opposed to those representing selections from a predetermined (i.e., closed) menu of options, as is often the case with structured text. A direct consequence of unconstrained nature of unstructured text is a far greater need for *disambiguation*, which is resolving of potential conflicts that tend to arise when a single term or an expression can take on multiple meanings. Conceptually, disambiguation is somewhat similar to the earlier discussed notion of term/phrase variability, but operationally it is quite a bit more complex as it necessitates anticipating the potential conflicts ahead of knowing what terms can appear in what context.

Overall, the frequency count and tabulation approach to text mining entails a considerable amount of analyst input, while at the same time its primary focus is on the identification and re-coding (from text into digits) of identifiable and definable terms and expressions, all while skipping over any deeper meaning that might be hidden in the semantic structure of text. As noted earlier, this broadly defined approach offers, in principle, no clear way of contextualizing or otherwise qualifying search-identified terms, beyond merely pinpointing their occurrence and counting the subsequent recurrences. Looked at as a sum of advantages and disadvantages, frequency count and tabulation method can be an effective approach to extracting numerically analyzable details out of the otherwise inaccessible textually expressed data, but overall it is not an effective mean of discovering new knowledge.

Natural Language Processing

Perhaps the most obvious limitation of the frequency count and tabulation approach to text mining is the embrace of the bag-of-words idea, which leads to lumping together of stem-related but differently valenced terms (i.e., the same term used in the positive vs. negative context). The stripping away of grammar linking together individual—i.e., merely counting the recurrence of terms without regard for their order or a broader context—inescapably leads to loss of information, which in turn diminishes the quantity and the quality of the resultant insights. Hence an obvious (and challenging) path to substantially enriching the depth and the breadth of newly discovered knowledge is to set aside the limiting bag-of-words assumption and to expressly consider the syntactical structure of text, which means to process what is known as “natural language.”

Broadly conceived, *natural language* is human communication that is distinct and different from “constructed languages,” best exemplified by computer programming or mathematical logic. In general, natural language can be spoken, written or signed, though text mining is obviously only concerned with the written aspect of it—more specifically, unpremeditated descriptions of states, outcomes and/or phenomena that comprise textual data. Formally defined, *natural language processing* (NLP) is an exploratory process of extracting meaningful insights out of the semantic structure of a body of text. Approach-wise, NLP can take one of two broadly defined forms: 1. supervised machine learning-based automated classification, or 2. unsupervised mining. At their core, both types of methodologies expressly consider words, word order and grammar in trying to discern generalizable rules that can be applied to distilling large quantities of text into manageable sets of summarized findings; however, they are quite different in terms of operational mechanics.

The first of the two NLP approaches, *automated classification*, can be thought of as a “pseudo-exploratory” methodology as it is a type of supervised learning where an algorithm (a decision rule) is "trained" using a previously classified text. The training task is essentially that of establishing generalizable rules for assigning topical labels to content, where the efficacy of the resultant algorithm is, to a large degree, dependent on the balancing of two important, though somewhat contradictory notions of accuracy and simplicity.[34] There are two distinct schools of thought as it regards the development of automated text classification systems : 1. *knowledge-based*, which relies on codification of expert-derived classification rules; and 2. *learning-based*, where experts supply classified examples rather than classification rules. Within the realm of the marketing database analytics process detailed in this book (as well as the broader context of marketing analytics), knowledge-based systems are generally deemed more workable, primarily because the requisite training inputs are more obtainable[35] (under most circumstances, it is prohibitively difficult to compile adequately representative classification samples that are required by learning-based systems). It is important to note that to be effective as a classification tool, the initial algorithmic learning cannot be overly tailored to idiosyncrasies of the training file (a condition known as "overfitting"), as doing so will lead to poor generalizability—yet, it needs to exhibit adequate classificatory accuracy (hence the need to balance the two somewhat contradictory notions of accuracy and simplicity).

Unsupervised mining, the second of the two general types of machine learning techniques, can be thought of as a "purely exploratory" text mining approach (in contrast to the above-described pseudo-exploratory automatic classification). However, even the purely exploratory text mining mechanisms do not represent, in a strict sense of the word, truly independent machine-based processing of human communications. These methods leverage similarity/difference heuristics—such as hierarchical clustering techniques—that are informed by text records' content which are emergent from data, rather than being based on learning from already classified text. That said, it is important to note that although the mining itself is unsupervised, the general rules within which it is conducted are governed by explicit vocabulary control, which typically takes the form of an a priori constructed (by human experts) thesaurus. The individual terms commonly comprising a particular thesaurus are noun phrases (i.e., content words), with the meaning of those terms restricted to that most effective for the purpose of a particular thesaurus, which in turn is shaped by the context of the search (e.g., a search of a database of customer comments might be focused on pinpointing drivers of brand users' satisfaction and/or dissatisfaction). In terms of *lexical inference*, or deriving meaning from grammar-connected word combinations, the human-expert-provided thesaurus also needs to expressly define three main types of cross-term relationships: equivalence, which is critical to resolving synonym-related ambiguity, as well as hierarchy and association, both of which play important roles in extracting semantic insights by imbuing meaning to multi-term expressions.

A choice of a specific approach notwithstanding, the goal of natural language processing is to go beyond enumerating categories by extracting kernels of syntactical meaning out of textual data. Hence, in the vast majority of cases NLP is used with semantically rich unstructured text, although in principle it could also be employed with structured text. In practice, however, the frequency count and tabulation text mining techniques discussed earlier are more appropriate—and probably more effective, all considered—to use with the categorization-friendly structured text.

Machine processing of natural human communications is a lofty goal and the NLP approaches briefly outlined above should be viewed as a probabilistic science yielding a moderate degree of success. In general, the quality of outcomes of syntactical analyses is highly dependent on the nature of data and the desired utility of findings: In general, the higher the specificity of data and the expected utility of results—the better the efficacy (in terms of validity and reliability) of outcomes. The reason for that dependency is that, under most circumstances, greater specificity translates into lower variability between the inputs (e.g., classification rules, examples or thesauri) provided by human experts and the raw data used in the analysis. Overall, in contrast to numerically encoded data, which can be analyzed with the expectation of highly valid and reliable results, analyses of the often highly nuanced and context-dependent text data cannot be expected to yield comparable (to the aforementioned numeric analyses) levels of result efficacy. In fact, the very task of objectively assessing the validity of text mining results is fraught with difficulties, not the least of which is the scarcity of evaluative benchmarks. More specifically, whereas the results of numeric analyses can be cross-referenced in a manner suggesting possible inconsistencies, the use of such evaluative methods is rarely feasible within the confines of text mining.

Still, its limitations notwithstanding, machine processing of written human communication offers an opportunity to peer into previously inaccessible domains of data, ultimately further enhancing the efficacy of the decision making process. And as suggested by the history of technological progress, the text mining technologies will continue to improve, which means that the validity and the reliability of machine-created insights will continue to get better.

Single- vs. Multi-Source Analytics

The explosive growth in the volume and diversity of data brought to light yet another important data-related consideration: single-source vs. multi-source analyses, briefly discussed earlier. Although rarely receiving much more than a cursory mention, this is a tremendously important consideration from the standpoint of creating competitively advantageous knowledge. Let's take a close look at the nature of multi-sourced analytics.

To start, it is important to ask what constitutes a data source. For the most part, it is *uniqueness* and *homogeneity*. To constitute its own source, data needs a unique point of origin. The so-called UPC (Universal Product Code) scanner data owes its name to its unique origin, which is a barcode-reading scanning device. It is also homogenous in the sense that individual records have fundamentally the same informational content. Hence the UPC data constitutes a unique source-based and informationally homogenous category of data. A more recent (UPC scanners have been in use since the mid-1970s) variation of technology-spurred data is exemplified by the radio frequency identification tags, or RFID tags for short. An RFID tag is an object that can be applied to or incorporated into a product, animal, or person for the purpose of identification and tracking using radio waves. (Some tags can be read from several meters away and beyond the line of sight of the reader; active RFID tags contain a battery to supply power for ongoing transmission, while passive RFID tags do not contain a battery.)

Similar to scanner data, *demographics* can be thought of as a unique and homogenous data source, to the degree to which individual (demographic) metrics are manifestations of different physical characteristics of a population; in a similar manner, *psychographics* represent another unique, though complementary data source, capturing the intangible characteristics of a population, such as values, attitudes or lifestyles. Yet another example of a distinct data source is consumer creditworthiness information, which is a reflection of consumers' borrowing and bill-paying behaviors collected and aggregated by consumer reporting agencies known in the United States as "credit bureaus."

All of these individual data sources—buying behaviors, demographics, psycho-graphics or creditworthiness—are naturally self-delimiting, hence their distinctiveness is fairly obvious. That is not necessarily the case with the textual data discussed in the previous section, which can also be considered a distinct data source to the degree to which it provides otherwise unavailable measures. Of course, “text data” is a general designation as there are numerous distinct sources of text-encoded metrics, hence there numerous unique and homogenous text data venues, such as the now-ubiquitous consumer product reviews or other sources of open-ended commentaries.

It follows that there are multiple types of data available to businesses that could be used for broadly defined purposes of marketing analytics. For instance, individual-level metrics such as purchases, promotional responses, behavioral surveys, psychographics, lifestyles or credit bureau information can all be used to derive robust estimates of buyer behavior. In the vast majority of cases, the analysis of these and other data types takes place in a single-source context, simply because it is a lot more straightforward than concurrent analyses of multiple, diverse data types. In some instances, single-source analyses are indeed quite appropriate, namely, when informational needs can be adequately met in that fashion. However, there are many other instances where multi-source analyses would reveal insights that cannot be gleaned from single-source explorations. In fact, the goal of developing cost-effective and impactful marketing strategies—and enabling tactics—demands causal knowledge, which cannot be attained by delving into one data source at a time. In other words, to understand the totality of consumer behavior it is necessary to simultaneously consider all of the possible behavioral determinants, which points toward multi-source analytics. However, for many organizations, the idea of amalgamating source-dissimilar data into a singular source of insights continues to be just that—an elusive idea…

A theme that is repeated throughout this book is that single-source analyses—i.e., analyses of a unique and homogenous data type carried out separately from analyses of other unique and homogenous data types—facilitates conversion of data into information, while multi-source analyses make possible information-to-knowledge conversion. Consider Figure 6.15 below:

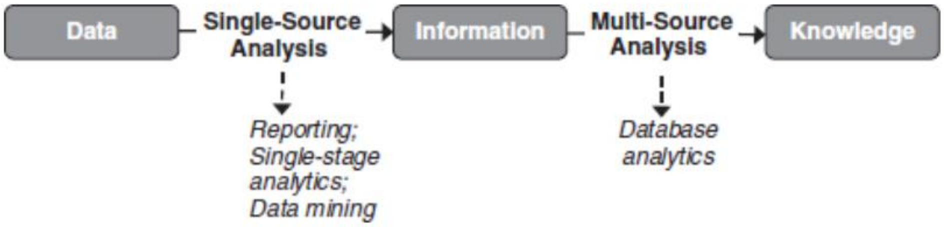

Figure 6.15 Multi-Source Analytics as the Source of Unique Knowledge

Conversion of raw data into usable information is the essence of single-source analytics—it is also the staple of business analytics. To the degree to which being in business is synonymous with selling of products and/or services, and the lion's share of sales transactions are carried out electronically, it should not be surprising that the vast majority of business organizations collect sales-related (i.e., single-source) data. Perhaps the most obvious informational utility of that data source is, broadly defined, *tracking and reporting* of outcomes of interest, such as sales by brand, product/service category or geography. The resultant time- or geography-based sales summaries have been a staple of business analytics for many years, and the more recent proliferation of data management and reporting technologies made the basic data reporting functionality essentially ubiquitous. A natural extension of the basic reporting functionality, *single-stage analytics*, entails looking beyond the “what is” basic reporting, typically focusing on estimation of responsiveness of outcomes, such as sales, to known (i.e., captured as data) stimuli, such as the various forms of sales incentives. In more methodological terms, single-stage analyses are usually confined to a predetermined set of outcomes and possible influencers of those outcomes, which is suggestive of a *hypothesis testing*. In other words, to the degree to which the analytic goal is to determine if any one or more of a predetermined set of factors has a measurable impact on the outcome or outcomes of interest,[36] the underlying analytical processes and methodologies can be characterized as *confirmatory*.

Still, analyses of single-source data are by no means limited to testing of a priori delineated hypotheses—an equally important and potentially informationally fruitful source of insights can be *data mining*, or an open-ended search for insights. In contrast to confirmatory hypothesis testing, mining of a set of data is purely *exploratory* in nature, as it is geared toward identifying previously unnoticed relationships with the goal of expanding current levels of understanding of a topic or a set of topics. An open-ended search for insights can be particularly effective when the amount of data (variable-wise) is large, as is usually the case with the earlier discussed UPC scanner data.[37] In such situations, an exploratory "consider-all" looks beyond the realm of already recognized interdependencies, in the hope of spotting new insights that could be suggestive of previously unconsidered marketing options (for example, exploratory data mining could identify new cross-product-type promotional opportunities by isolating previously unnoticed purchase interdependencies among seemingly unrelated products).

All considered, single-source analytics can certainly be a spring of worthwhile information—however, considering individual sources of data in isolation from one another invariably results in the forgoing of potentially important informational synergies, which is perhaps the most fundamental shortcoming of single-source analytics. As suggested in Figure 6.15, amalgamating unique data sources into a broader, source-dissimilar but meaning-interconnected informational reservoir is the first step in evolving the organization's data analytical capabilities from a function focused on pursuing of topical insights to one that is capable of creating competitively advantageous, decision-guiding knowledge.

As discussed in the opening chapter, the fundamental difference between information and knowledge is that the former sheds light on "what happened," while the latter is capable of explaining "why it happened." Naturally, it is immeasurably more beneficial to have the understanding of the underlying causes than to merely be well-informed regarding outcomes or events of interest, given that causal knowledge can inform future decisions, which is generally not the case with explanation-lacking outcome information. This is the essence of multi-source data analytics imperative—it offers the means of realizing the long-heralded but rarely realized value of data.

The notion of multi-source analytics is suggestive of yet another emergent business concept—a *marketing ecosystem*.[38] The following section offers a closer look.

Marketing Ecosystems

There is a long, though not always glorious, tradition of likening different aspects of business to natural phenomena, of which the idea of an *ecosystem* is one of the more recent examples. As a scientific concept, an ecosystem is the totality of living organisms and nonliving elements inhibiting a particular area; in a given ecosystems, organisms interact with each other and with nonliving elements. The natural—i.e., biological—definition of an ecosystem implies a certain degree of distinctiveness of the system as well as a certain degree of interconnectedness among the living and nonliving components of the system. If an ecosystem can be assumed to have a purpose, then it would most likely be to perpetuate itself and by extension, the purpose of its living elements can be assumed to be that of survival.

Given the basic outline of a natural ecosystem, it is easy to see its close parallels with business: “the totality of living organisms and nonliving elements inhabiting a particular area” can be likened to “the totality of firms competing in a particular industry”; the ecosystem's purpose of “self-perpetuation” is essentially the same as it is for any industry, as is the case with “survival” being the goal of individual ecosystem “members.” The application of the concept of ecosystem to a somewhat narrower context of marketing retains those similarities: A *marketing ecosystem* can be thought of as the totality of promotional means, encompassing platforms, channels and techniques, which individually and jointly contribute to the fundamental purpose of marketing—to win and retain a customer.[39] Those promotional means are striving for relevance as they are confronted with an array of adverse developments, including continual emergence of new alternatives, progressively more fragmented media usage and ever-rising targeting precision expectations. In a sense, like the organisms in a biological ecosystem, individual promotional elements are trying to survive as viable marketing ecosystem participants.

Within the realm of marketing, perhaps the most obvious application of the notion of ecosystems is to characterize the totality of the industry, as exemplified by a comprehensive Booz & Company research study, “Marketing & Media Ecosystem 2010.”[40] The study, which focuses on the changing dynamics of the marketing and media industry as a whole, likens the broadly scoped industry—encompassing service providers (advertising agencies, marketing service organizations and the like) and media content providers (MTV Networks, The Disney Company, Facebook and others)—to a marketing ecosystem, where individual “organisms,” or companies, compete and, to a lesser degree, collaborate with each other.

A somewhat different way of thinking about a marketing ecosystem is to consider the concept from the point of view of a brand. As suggested earlier, the goal of a brand is to win and retain profitable customers, with broadly defined marketing activities aiding in reaching that objective. The gradual but persistent shift toward “always on” promotional communication formats, coupled with the almost dizzying proliferation of promotional media and related options are effectively redefining the traditional conception of *marketing mix*. The persistently used, decades-old conceptualization of the four focal elements of marketing,[41] the *four Ps* of marketing (product, price, promotion and place) no longer offers an adequately clear way of thinking about the impact of marketing activities. There are several reasons for that: First, it reflects a producer-centric view, which is not in keeping with the reality of consumer-centric economy. Second, the conceptualization lumps together brand communication strategy (promotion) with what can be viewed as constraints (product, place and price) placed upon that strategy. In other words, the role of marketing, as a business function, is to devise maximally impactful brand promotional communications, subject to the offering's (a physical product or an intangible service) core attributes—or the “product” part of the four Ps—in addition to its price and distribution.

Perhaps most importantly, the four Ps conceptualization conveys a largely deterministic picture of the brand's world, one which is not in keeping with today's highly fluid business environment. More specifically, the four Ps framework implicitly assumes that, at a point in time, the bundle of attributes comprising the product of interest (or several varieties of that product), as well as the product's price and distribution are all effectively fixed. However, that is not today's reality: The advent of mass customization (the use of flexible, computerized manufacturing systems to produce output tailored to a specific order) is changing the way we think about products, as product attributes at a point in time can vary, reflecting individual buyers’ preferences. The rise of ecommerce has had a similar impact on the notion of place (distribution), which also lost a lot of its clarity, largely because to a consumer ordering a product on the internet from the comfort of his or her home, the distribution channel for that product is mostly invisible and practically irrelevant. Although more tacitly, the four Ps conceptualization also fails to capture the changing character of brands' communications: The past trend of brands both initiating and controlling promotional communications is also largely giving way to a more interactive model, where both brands and consumers initiate and control brand-related communications—the former by the way of consumer-targeted promotions and the latter through brand-related, largely online social interactions. Yet, even that distinctiveness is beginning to fade, as the traditionally "one-way" media, such as television, are moving in the direction of true interactivity.[42]

All things considered, internet-driven technological innovations are fundamentally reshaping the marketing paradigm—away from the narrowly defined, somewhat static and producer-centric *marketing mix* concept and toward a considerably more expansive *marketing ecosystem*. One of the more compelling depictions of a marketing ecosystem was proposed by Mullen, a Boston-based integrated advertising agency (see Figure 6.16).

Perhaps the most important aspect of the ecosystem conceptualization is that a consumer is at the center of a multimedia swirl, where multiple communication channels both compete for attention and collaborate with each other to bring about the desired outcomes. It is a "yin and yang"[43] world: Individual promotional channels, such as print, direct mail, online, mobile or out-of-home (public place advertising, such as billboards) are both in competition, as each is hoping to attract the largest possible share of advertisers' budget, while at the same time they reinforce each other by offering multiple exposure points to advertisers' messages. Mullen's depiction captures what is nothing less than a staggering number of communication alternatives available to marketers; the graphic's life-resembling layout also hints of the fluidity of the system. Perhaps most importantly, the ecosystem representation draws attention to the essence, or the core of marketing's mission: To allow brands to communicate with their current and prospective customers.

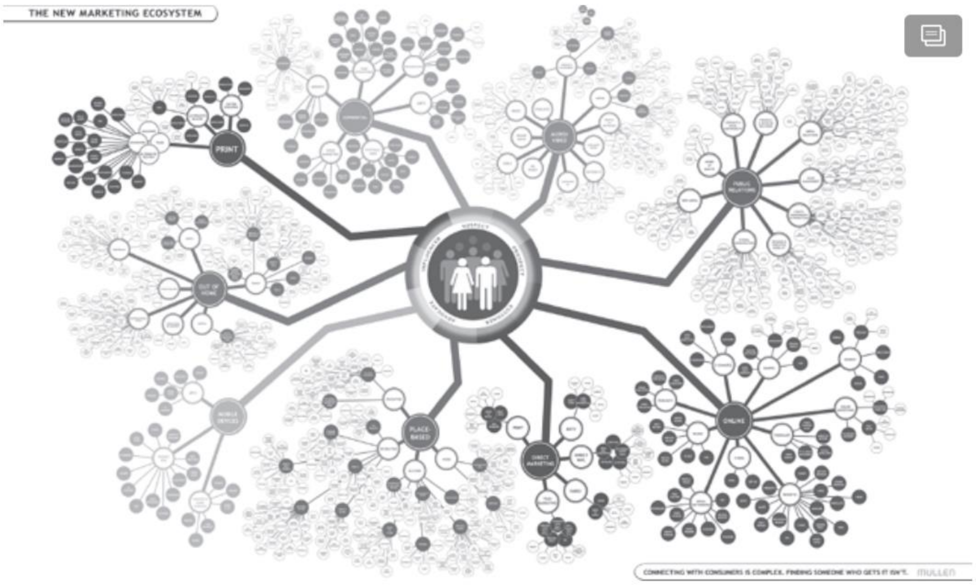
THE NEW MARKETING ECOSYSTEM
PRINT
OUT OF HOME
MOBILE DEVICES
PLACE-BASED
DIRECT MARKETING
ONLINE
PUBLIC RELATIONS
CONNECTING WITH CONSUMERS IS COMPLEX. FINDING SOMEONE WHO GETS IT ISN'T.
MULLEN

Figure 6.16 Mullen's New Marketing Ecosystem

Source: Mullen. Reprinted with permission. www.mullen.com

Given the sheer number of promotional options, how does one decide which vehicles to lean on and which ones to, possibly, de-emphasize? Aggregate trend-wise, there has been a gradual though persistent shift away from the "old" media—traditional TV, print, radio and to some degree, direct mail—and toward the "new" communication channels, most notably online and mobile.[44] However, aggregate trends are not necessarily indicative of which media vehicles are likely to work best for a particular brand—the answer ultimately lies in thorough assessment of the effectiveness of individual vehicles, individually as well as collectively.

Consider the marketing ecosystem in the context of earlier discussed *multi-source analytics*. Clearly, combining highly dissimilar data sources is far from easy, but not rising up to that challenge would be tantamount to forgoing a great deal of informational richness hidden in the interplay of the individual elements of the brand's marketing ecosystem. In fact, robust promotional effectiveness insights are at the core of creating competitively advantageous marketing knowledge.

Process-wise, combining of dissimilar data sources can follow one of two paths: 1. disaggregate, or 2. aggregate amalgamation. *Disaggregate amalgamation* entails merging two or more files into a single data reservoir using a common-to-all link, such as customer ID. The merger generally takes place at the most disaggregate level of data hierarchy (see the earlier discussion of *root* data) and requires the existence of a case-unique link, such as the aforementioned customer ID. Given that the goal of dis-aggregate amalgamation is increase the number of attributes (i.e., variables) that could be used to analyze cross-case relationships at the most granular level, ideally there should be high degree of case-overlap across all files that are to be merged together, to minimize potential data distortions.[45] Under most circumstances, cases that only appear in one file should be closely examined for possible elimination to keep the proportion of missing values in the post-merger dataset to a minimum. All considered, disaggregate amalgamation is the preferred approach to combining physically distinct data files as it offers the greatest amount of analytic flexibility—however, it is also most demanding in the sense that it requires communality of file organizational schemas and the presence of a common-to-all (files) unique case identifier.

Aggregate amalgamation involves combining files where cases represent summaries of more detailed records (see the earlier discussion of *derived* data). Under most circumstances, this type of a merger tends to be a product of necessity—for instance, if we were to combine sales and coupon redemption files and the former was organized by UPC but the latter by brand (i.e., each case represented an individual UPC or a brand, respectively), it would only be possible to combine these two files at the brand level (which means that the UPC-level records would need to be aggregated, or "rolled-up" to the higher aggregate of a brand). Hence, aggregate amalgamation should be undertaken at the most detailed level possible, which is usually the lowest level of aggregation that is common to all files that are to be combined. In fact, it is generally advantageous to combine data at lowest level of detail, simply because more detailed data can always be aggregated, but once aggregated, data cannot be disaggregated.

Mini-Case 6.1: Multi-Source Analytics

One of the U.S. wireless carriers recently embarked on a major customer relationship management (CRM) initiative to develop a "360-degree customer view," defined as *providing all organizational stakeholders with the consistent view of the customer*. At its core, the initiative entails amalgamating multiple customer-related, but source-dissimilar data sources in a manner where each set of data represents a different customer dimension—needless to say, it is a considerable informational undertaking. As is the case with other, large, multi-department organizations, the wireless carrier's approach to data capture is highly functional, to the degree to which much of the data capture is a passive consequence of electronic transaction processing. That, coupled with the fragmentation of data (i.e., different parts of the organization focusing on just certain subsets of all customer data) has led to a great deal of under-utilization of data as a decision-guiding resource. Among the undesirable consequences of that status quo is that the same customer might appear to be more or less attractive to the organization, depending on who in the organization is looking. In other words, since different parts of the organization are focused on different slices of the overall customer data, when taken out of the "total data" context, those individual data slices will likely tell a different story… Not only can such relativism lead to conflicting conclusions, it stands in the way of implementing more "enlightened" CRM approaches, such as tying loyalty rewards to customer lifetime profitability. Those are among the reasons compelling the wireless carrier to invest in the *360-degree customer view*.

One of the key challenges in the wireless telecommunication industry is the distinction between a customer—which could be an individual person, a household or a business entity—and a telephone number. At present, the total U.S. population is a bit under 314 million, while the total number of active phone numbers is about 1 billion. Even when non-private (i.e., business, government, etc.) phone numbers are set aside, there is still a one-to-many relationship between a wireless subscriber and phone numbers, because many of today's popular communication devices, such as tablet PCs or mobile broadband modems for laptops, have phone numbers embedded in them (e.g., a broadband-enabled tablet PC uses a built-in phone number to connect to internet). Hence, a wireless subscriber might have a phone number associated with his/her wireless phone, another one associated with his/her tablet PC and yet another one associated with his/her mobile broadband modem. Furthermore, that subscriber can also have still other, separate phone numbers, such as those used by his/her children… In a data sense, many of those will usually exist as separate accounts (since the individual devices tend to be acquired and activated at different points in time), which means that a single wireless subscriber can be seen as several different customers. On top of that, each of those individual "customers" will usually yield a considerably different ARPU (average revenue per user), as a smart phone voice + text + data subscription generates about $100 monthly, while the data-only wireless internet access for tablet or laptop generates between $20 and $40 monthly. But the informational proliferation only begins here: Multiple accounts will then spawn a number of account-related "data derivatives," such as billing records, account service and management, customer interaction and other records. Add to that departmental data fragmentation—i.e., smart phones vs. feature phones vs. tablets vs. wireless broadband modems—the number of single-source data grows exponentially, as does the fragmentation of the customer view.

The first step in the journey to create the *360-degree customer view* is to construct a controlling master identification system, which in the case of the wireless carrier is a combination of the name and address associated with individual accounts. Once created, the "master ID" forms the basic unit of analysis—i.e., each individual row in the data matrix will delimit a subscriber; to the degree to which a particular subscriber might have multiple phone numbers associated with his/her account, those numbers will become attributes (variables in the data matrix context), as will other characteristics, such as account payment history, promotional responses, etc. In doing so, multiple-data records, where a single subscriber appears as an attribute of multiple phone numbers, are reconfigured into a single record where the individual phone numbers (and all details associated with those numbers) are attributed to the controlling subscriber ID. Once the amalgamation has been completed, no matter who in the wireless subscriber organization is looking at the customer base, the resultant picture will be the same.