Chapter 2
Data Collection

Sudhir Voleti

1 Introduction

Collecting data is the first step towards analyzing it. In order to understand and solve business problems, data scientists must have a strong grasp of the characteristics of the data in question. How do we collect data? What kinds of data exist? Where is it coming from? Before beginning to analyze data, analysts must know how to answer these questions. In doing so, we build the base upon which the rest of our examination follows. This chapter aims to introduce and explain the nuances of data collection, so that we understand the methods we can use to analyze it.

2 The Value of Data: A Motivating Example

In 2017, video-streaming company Netflix Inc. was worth more than $80 billion, more than 100 times its value when it listed in 2002. The company's current position as the market leader in the online-streaming sector is a far cry from its humble beginning as a DVD rental-by-mail service founded in 1997. So, what had driven Netflix's incredible success? What helped its shares, priced at $15 each on their initial public offering in May 2002, rise to nearly $190 in July 2017? It is well known that a firm's [market] valuation is the sum total in today's money, or the net present value (NPV) of all the profits the firm will earn over its lifetime. So investors reckon that Netflix is worth tens of billions of dollars in profits over its lifetime. Why might this be the case? After all, companies had been creating television and

S. Voleti (✉)
Indian School of Business, Hyderabad, Telangana, India
e-mail: sudhir_voleti@isb.edu

B. Pochiraju, S. Seshadri (eds.), *Essentials of Business Analytics*, International Series in Operations Research & Management Science 264,
https://doi.org/10.1007/978-3-319-68837-4_2

cinematic content for decades before Netflix came along, and Netflix did not start its own online business until 2007. Why is Netflix different from traditional cable companies that offer shows on their own channels?

Moreover, the vast majority of Netflix's content is actually owned by its competitors. Though the streaming company invests in original programming, the lion's share of the material available on Netflix is produced by cable companies across the world. Yet Netflix has access to one key asset that helps it to predict where its audience will go and understand their every quirk: data.

Netflix can track every action that a customer makes on its website—what they watch, how long they watch it for, when they tune out, and most importantly, what they might be looking for next. This data is invaluable to its business—it allows the company to target specific niches of the market with unerring accuracy.

On February 1, 2013, Netflix debuted House of Cards—a political thriller starring Kevin Spacey. The show was a hit, propelling Netflix's viewership and proving that its online strategy could work. A few months later, Spacey applauded Netflix's approach and cited its use of data for its ability to take a risk on a project that every other major television studio network had declined. Casey said in *Edinburgh*, at the *Guardian Edinburgh International Television Festival*[1] on August 22: "Netflix was the only company that said, 'We believe in you. We have run our data, and it tells us our audience would watch this series.'"

Netflix's data-oriented approach is key not just to its ability to pick winning television shows, but to its global reach and power. Though competitors are springing up the world over, Netflix remains at the top of the pack, and so long as it is able to exploit its knowledge of how its viewers behave and what they prefer to watch, it will remain there.

Let us take another example. The technology "cab" company Uber has taken the world by storm in the past 5 years. In 2014, Uber's valuation was a mammoth 40 billion USD, which by 2015 jumped another 50% to reach 60 billion USD. This fact begs the question: what makes Uber so special? What competitive advantage, strategic asset, and/or enabling platform accounts for Uber's valuation numbers? The investors reckon that Uber is worth tens of billions of dollars in profits over its lifetime. Why might this be the case? Uber is after all known as a ride-sharing business—and there are other cab companies available in every city.

We know that Uber is "asset-light," in the sense that it does not own the cab fleet or have drivers of the cabs on its direct payroll as employees. It employs a franchise model wherein drivers bring their own vehicles and sign up for Uber. Yet Uber does have one key asset that it actually owns, one that lies at the heart of its profit projections: data. Uber owns all rights to every bit of data from every passenger, every driver, every ride and every route on its network. Curious as to how much data are we talking about? Consider this. Uber took 6 years to reach one billion

[1]Guardian Edinburgh International Television Festival, 2017 (https://www.ibtimes.com/kevin-spacey-speech-why-netflix-model-can-save-television-video-full-transcript-1401970) accessed on Sep 13, 2018.

rides (Dec 2015). Six months later, it had reached the two billion mark. That is one billion rides in 180 days, or 5.5 million rides/day. How did having consumer data play a factor in the exponential growth of a company such as Uber? Moreover, how does data connect to analytics and, finally, to market value?

Data is a valuable asset that helps build sustainable competitive advantage. It enables what economists would call "supernormal profits" and thereby plausibly justify some of those wonderful valuation numbers we saw earlier. Uber had help, of course. The nature of demand for its product (contractual personal transportation), the ubiquity of its enabling platform (location-enabled mobile devices), and the profile of its typical customers (the smartphone-owning, convenience-seeking segment) has all contributed to its success. However, that does not take away from the central point being motivated here—the value contained in data, and the need to collect and corral this valuable resource into a strategic asset.

3 Data Collection Preliminaries

A well-known management adage goes, "We can only manage what we can measure." But why is measurement considered so critical? Measurement is important because it precedes *analysis*, which in turn precedes *modeling*. And more often than not, it is *modeling* that enables *prediction*. Without *prediction* (determination of the values an outcome or entity will take under specific conditions), there can be no *optimization*. And without *optimization*, there is no *management*. The quantity that gets measured is reflected in our records as "data." The word data comes from the Latin root *datum* for "given." Thus, data (datum in plural) becomes facts which are given or known to be true. In what follows, we will explore some preliminary conceptions about data, types of data, basic measurement scales, and the implications therein.

3.1 Primary Versus Secondary Dichotomy

Data collection for research and analytics can broadly be divided into two major types: primary data and secondary data. Consider a project or a business task that requires certain data. Primary data would be data that is collected "at source" (hence, primary in form) and specifically for the research at hand. The data source could be individuals, groups, organizations, etc. and data from them would be actively elicited or passively observed and collected. Thus, surveys, interviews, and focus groups all fall under the ambit of primary data. The main advantage of primary data is that it is tailored specifically to the questions posed by the research project. The disadvantages are cost and time.

On the other hand, secondary data is that which has been previously collected for a purpose that is *not* specific to the research at hand. For example, sales records,

industry reports, and interview transcripts from past research are data that would continue to exist whether or not the project at hand had come to fruition. A good example of a means to obtain secondary data that is rapidly expanding is the API (Application Programming Interface)—an interface that is used by developers to securely query external systems and obtain a myriad of information.

In this chapter, we concentrate on data available in published sources and websites (often called secondary data sources) as these are the most commonly used data sources in business today.

4 Data Collection Methods

In this section, we describe various methods of data collection based on sources, structure, type, etc. There are basically two methods of data collection: (1) data generation through a designed experiment and (2) collecting data that already exists. A brief description of these methods is given below.

4.1 Designed Experiment

Suppose an agricultural scientist wants to compare the effects of five different fertilizers, A, B, C, D, and E, on the yield of a crop. The yield depends not only on the fertilizer but also on the fertility of the soil. The consultant considers a few relevant types of soil, for example, clay, silt, and sandy soil. In order to compare the fertilizer effect one has to control for the soil effect. For each soil type, the experimenter may choose ten representative plots of equal size and assign the five fertilizers to the ten plots at random in such a way that each fertilizer is assigned to two plots. He then observes the yield in each plot. This is the design of the experiment. Once the experiment is conducted as per this design, the yields in different plots are observed. This is the data collection procedure. As we notice, the data is not readily available to the scientist. He designs an experiment and generates the data. This method of data collection is possible when we can control different factors precisely while studying the effect of an important variable on the outcome. This is quite common in the manufacturing industry (while studying the effect of machines on output or various settings on the yield of a process), psychology, agriculture, etc. For well-designed experiments, determination of the causal effects is easy. However, in social sciences and business where human beings often are the instruments or subjects, experimentation is not easy and in fact may not even be feasible. Despite the limitations, there has been tremendous interest in behavioral experiments in disciplines such as finance, economics, marketing, and operations management. For a recent account on design of experiments, please refer to Montgomery (2017).

4.2 Collection of Data That Already Exists

Household income, expenditure, wealth, and demographic information are examples of data that already exists. Collection of such data is usually done in three possible ways: (1) *complete enumeration*, (2) *sample survey*, and (3) through available sources where the data was collected possibly for a different purpose and is available in different published sources. *Complete enumeration* is collecting data on all items/individuals/firms. Such data, say, on households, may be on consumption of essential commodities, the family income, births and deaths, education of each member of the household, etc. This data is already available with the households but needs to be collected by the investigator. The census is an example of complete enumeration. This method will give information on the whole population. It may appear to be the best way but is expensive both in terms of time and money. Also, it may involve several investigators and investigator bias can creep in (in ways that may not be easy to account for). Such errors are known as *non-sampling errors*. So often, a *sample survey* is employed. In a sample survey, the data is not collected on the entire population, but on a representative sample. Based on the data collected from the sample, inferences are drawn on the population. Since data is not collected on the entire population, there is bound to be an error in the inferences drawn. This error is known as the *sampling error*. The inferences through a sample survey can be made precise with error bounds. It is commonly employed in market research, social sciences, public administration, etc. A good account on sample surveys is available in Blair and Blair (2015).

Secondary data can be collected from two sources: internal or external. *Internal data* is collected by the company or its agents on behalf of the company. The defining characteristic of the internal data is its proprietary nature; the company has control over the data collection process and also has exclusive access to the data and thus the insights drawn on it. Although it is costlier than external data, the exclusivity of access to the data can offer competitive advantage to the company. *The external data*, on the other hand, can be collected by either third-party data providers (such as IRI, AC Nielsen) or government agencies. In addition, recently another source of external secondary data has come into existence in the form of social media/blogs/review websites/search engines where users themselves generate a lot of data through C2B or C2C interactions. Secondary data can also be classified on the nature of the data along the dimension of structure. Broadly, there are three types of data: *structured*, *semi-structured (hybrid)*, and *unstructured* data. Some examples of *structured data* are sales records, financial reports, customer records such as purchase history, etc. A typical example of *unstructured data* is in the form of free-flow text, images, audio, and videos, which are difficult to store in a traditional database. Usually, in reality, data is somewhere in between structured and unstructured and thus is called *semi-structured* or *hybrid* data. For example, a product web page will have product details (structured) and user reviews (unstructured).

The data and its analysis can also be classified on the basis of whether a single unit is observed over multiple time points (*time-series data*), many units observed once (*cross-sectional data*), or many units are observed over multiple time periods (*panel data*). The insights that can be drawn from the data depend on the nature of data, with the richest insights available from panel data. The panel could be *balanced* (all units are observed over all time periods) or *unbalanced* (observations on a few units are missing for a few time points either by design or by accident). If the data is not missing excessively, it can be accounted for using the methods described in Chap. 8.

5 Data Types

In programming, we primarily classify the data into three types—*numerals*, *alphabets*, and *special characters* and the computer converts any data type into binary code for further processing. However, the data collected through various sources can be of types such as numbers, text, image, video, voice, and biometrics.

The data type helps analyst to evaluate which operations can be performed to analyze the data in a meaningful way. The data can limit or enhance the complexity and quality of analysis.

Table 2.1 lists a few examples of data categorized by type, source, and uses. You can read more about them following the links (all accessed on Aug 10, 2017).

5.1 Four Data Types and Primary Scales

Generally, there are four types of data associated with four primary scales, namely, *nominal*, *ordinal*, *interval*, and *ratio*. Nominal scale is used to describe categories in which there is no specific order while the ordinal scale is used to describe categories in which there is an inherent order. For example, green, yellow, and red are three colors that in general are not bound by an inherent order. In such a case, a nominal scale is appropriate. However, if we are using the same colors in connection with the traffic light signals there is clear order. In this case, these categories carry an ordinal scale. Typical examples of the ordinal scale are (1) sick, recovering, healthy; (2) lower income, middle income, higher income; (3) illiterate, primary school pass, higher school pass, graduate or higher, and so on. In the ordinal scale, the differences in the categories are not of the same magnitude (or even of measurable magnitude). Interval scale is used to convey relative magnitude information such as temperature. The term "Interval" comes about because rulers (and rating scales) have intervals of uniform lengths. Example: "I rate A as a 7 and B as a 4 on a scale of 10." In this case, we not only know that A is preferred to B, but we also have some idea of how much more A is preferred to B. Ratio scales convey information on an absolute scale. Example: "I paid $11 for A and $12 for B." The 11 and 12

here are termed "absolute" measures because the corresponding zero point ($0) is understood in the same way by different people (i.e., the measure is independent of subject).

Another set of examples for the four data types, this time from the world of sports, could be as follows. The numbers assigned to runners are of nominal data type, whereas the rank order of winners is of the ordinal data type. Note in the latter case that while we may know who came first and who came second, we would not know by how much based on the rank order alone. A performance rating on a 0–10

Table 2.1 A description of data and their types, sources, and examples

Category	Examples	Type	Sources[a]
Internal data			
Transaction data	Sales (POS/online) transactions, stock market orders and trades, customer IP and geolocation data	Numbers, text	http://times.cs.uiuc.edu/~wang296/Data/ https://www.quandl.com/ https://www.nyse.com/data/transactions-statistics-data-library https://www.sec.gov/answers/shortsalevolume.htm
Customer preference data	Website click stream, cookies, shopping cart, wish list, preorder	Numbers, text	C:\Users\username\App Data\Roaming\Microsoft \Windows\Cookies, Nearbuy.com (advance coupon sold)
Experimental data	Simulation games, clinical trials, live experiments	Text, number, image, audio, video	https://www.clinicaltrialsregister.eu/ https://www.novctrd.com/ http://ctri.nic.in/
Customer relationship data	Demographics, purchase history, loyalty rewards data, phone book	Text, number, image, biometrics	
External data			
Survey data	Census, national sample survey, annual survey of industries, geographical survey, land registry	Text, number, image, audio, video	http://www.census.gov/data.html http://www.mospi.gov.in/ http://www.csoisw.gov.in/ https://www.gsi.gov.in/ http://landrecords.mp.gov.in/
Biometric data (fingerprint, retina, pupil, palm, face)	Immigration data, social security identity, Aadhar card (UID)	Number, text, image, biometric	http://www.migrationpolicy.org/programs/migration-data-hub https://www.dhs.gov/immigration-statistics

(continued)

Table 2.1 (continued)

Category	Examples	Type	Sources[a]
Third party data	RenTrak, A. C. Nielsen, IRI, MIDT (Market Information Data Tapes) in airline industry, people finder, associations, NGOs, database vendors, Google Trends, Google Public Data	All possible data types	http://aws.amazon.com/datasets https://www.worldwildlife.org/pages/conservation-science-data-and-tools http://www.whitepages.com/ https://pipl.com/ https://www.bloomberg.com/ https://in.reuters.com/ http://www.imdb.com/ http://datacatalogs.org/ http://www.google.com/trends/explore https://www.google.com/publicdata/directory
Govt and quasi govt agencies	Federal governments, regulators—Telecom, BFSI, etc., World Bank, IMF, credit reports, climate and weather reports, agriculture production, benchmark indicators—GDP, etc., electoral roll, driver and vehicle licenses, health statistics, judicial records	All possible data types	http://data.gov/ https://data.gov.in/ http://data.gov.uk/ http://open-data.europa.eu/en/data/ http://www.imf.org/en/Data https://www.rbi.org.in/Scripts/Statistics.aspx https://www.healthdata.gov/ https://www.cibil.com/ http://eci.nic.in/ http://data.worldbank.org/
Social sites data, user-generated data	Twitter, Facebook, YouTube, Instagram, Pinterest Wikipedia, YouTube videos, blogs, articles, reviews, comments	All possible data types	https://dev.twitter.com/streaming/overview https://developers.facebook.com/docs/graph-api https://en.wikipedia.org/ https://www.youtube.com/ https://snap.stanford.edu/data/web-Amazon.html http://www.cs.cornell.edu/people/pabo/movie-review-data/

[a]All the sources are last accessed on Aug 10, 2017

scale would be an example of an interval scale. We see this used in certain sports ratings (i.e., gymnastics) wherein judges assign points based on certain metrics. Finally, in track and field events, the time to finish in seconds is an example of ratio data. The reference point of zero seconds is well understood by all observers.

5.2 Common Analysis Types with the Four Primary Scales

The reason why it matters what primary scale was used to collect data is that downstream analysis is constrained by data type. For instance, with nominal data, all we can compute are the mode, some frequencies and percentages. Nothing beyond this is possible due to the nature of the data. With ordinal data, we can compute the median and some rank order statistics in addition to whatever is possible with nominal data. This is because ordinal data retains all the properties of the nominal data type. When we proceed further to interval data and then on to ratio data, we encounter a qualitative leap over what was possible before. Now, suddenly, the arithmetic mean and the variance become meaningful. Hence, most statistical analysis and parametric statistical tests (and associated inference procedures) all become available. With ratio data, in addition to everything that is possible with interval data, ratios of quantities also make sense.

The multiple-choice examples that follow are meant to concretize the understanding of the four primary scales and corresponding data types.

6 Problem Formulation Preliminaries

Even before data collection can begin, the purpose for which the data collection is being conducted must be clarified. Enter, problem formulation. The importance of problem formulation cannot be overstated—it comes first in any research project, ideally speaking. Moreover, even small deviations from the intended path at the very beginning of a project's trajectory can lead to a vastly different destination than was intended. That said, problem formulation can often be a tricky issue to get right. To see why, consider the musings of a decision-maker and country head for XYZ Inc.

> Sales fell short last year. But sales would've approached target except for 6 territories in 2 regions where results were poor. Of course, we implemented a price increase across-the-board last year, so our profit margin goals were just met, even though sales revenue fell short. Yet, 2 of our competitors saw above-trend sales increases last year. Still, another competitor seems to be struggling, and word on the street is that they have been slashing prices to close deals. Of course, the economy was pretty uneven across our geographies last year and the 2 regions in question, weak anyway, were particularly so last year. Then there was that mess with the new salesforce compensation policy coming into effect last year. 1 of the 2 weak regions saw much salesforce turnover last year ...

These are everyday musings in the lives of business executives and are far from unusual. Depending on the identification of the problem, data collection strategies,

resources, and approaches will differ. The difficulty in being able to readily pinpoint any one cause or a combination of causes as specific problem highlights the issues that crop up in problem formulation. Four important points jump out from the above example. First, that reality is messy. Unlike textbook examples of problems, wherein irrelevant information is filtered out a priori and only that which is required to solve "the" identified problem exactly is retained, life seldom simplifies issues in such a clear-cut manner. Second, borrowing from a medical analogy, there are symptoms—observable manifestations of an underlying problem or ailment—and then there is the cause or ailment itself. Symptoms could be a fever or a cold and the causes could be bacterial or viral agents. However, curing the symptoms may not cure the ailment. Similarly, in the previous example from XYZ Inc., we see symptoms ("sales are falling") and hypothesize the existence of one or more underlying problems or causes. Third, note the pattern of connections between symptom(s) and potential causes. One symptom (falling sales) is assumed to be coming from one or more potential causes (product line, salesforce compensation, weak economy, competitors, etc.). This brings up the fourth point—How can we diagnose a problem (or cause)? One strategy would be to narrow the field of "ailments" by ruling out low-hanging fruits—ideally, as quickly and cheaply as feasible. It is not hard to see that the data required for this problem depends on what potential ailments we have shortlisted in the first place.

6.1 Towards a Problem Formulation Framework

For illustrative purposes, consider a list of three probable causes from the messy reality of the problem statement given above, namely, (1) product line is obsolete; (2) customer-connect is ineffective; and (3) product pricing is uncompetitive (say). Then, from this messy reality we can formulate decision problems (D.P.s) that correspond to the three identified probable causes:

- D.P. #1: "Should new product(s) be introduced?"
- D.P. #2: "Should advertising campaign be changed?"
- D.P. #3: "Should product prices be changed?"

Note what we are doing in mathematical terms—if messy reality is a large multidimensional object, then these D.P.s are small-dimensional subsets of that reality. This "reduces" a messy large-dimensional object to a relatively more manageable small-dimensional one.

The D.P., even though it is of small dimension, may not contain sufficient detail to map directly onto tools. Hence, another level of refinement called the research objective (R.O.) may be needed. While the D.P. is a small-dimensional object, the R.O. is (ideally) a one-dimensional object. Multiple R.O.s may be needed to completely "cover" or address a single D.P. Furthermore, because each R.O. is one-dimensional, it maps easily and directly onto one or more specific tools in the analytics toolbox. A one-dimensional problem formulation component better be

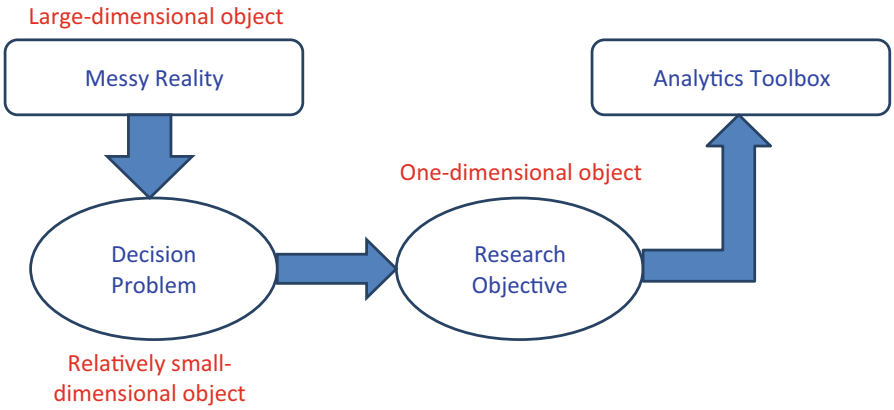

Fig. 2.1 A framework for problem formulation

well defined. The R.O. has three essential parts that together lend necessary clarity to its definition. R.O.s comprise of (a) an action verb and (b) an actionable object, and typically fit within one handwritten line (to enforce brevity). For instance, the active voice statement "Identify the real and perceived gaps in our product line vis-à-vis that of our main competitors" is an R.O. because its components action verb ("identify"), actionable object ("real and perceived gaps"), and brevity are satisfied.

Figure 2.1 depicts the problem formulation framework we just described in pictorial form. It is clear from the figure that as we impose preliminary structure, we effectively reduce problem dimensionality from large (messy reality) to somewhat small (D.P.) to the concise and the precise (R.O.).

6.2 Problem Clarity and Research Type

A quotation attributed to former US defense secretary Donald Rumsfeld in the run-up to the Iraq war goes as follows: "There are known-knowns. These are things we know that we know. There are known-unknowns. That is to say, there are things that we know we don't know. But there are also unknown-unknowns. There are things we don't know we don't know." This statement is useful in that it helps discern the differing degrees of the awareness of our ignorance about the true state of affairs.

To understand why the above statement might be relevant for problem formulation, consider that there are broadly three types of research that correspond to three levels of clarity in problem definition. The first is exploratory research wherein the problem is at best ambiguous. For instance, "Our sales are falling .... Why?" or "Our ad campaign isn't working. Don't know why." When identifying the problem is itself a problem, owing to unknown-unknowns, we take an exploratory approach to trace and list potential problem sources and then define what the problems

may be. The second type is descriptive research wherein the problem's identity is somewhat clear. For instance, "What kind of people buy our products?" or "Who is perceived as competition to us?" These are examples of known-unknowns. The third type is causal research wherein the problem is clearly defined. For instance, "Will changing this particular promotional campaign raise sales?" is a clearly identified known-unknown. Causal research (the cause in causal comes from the cause in because) tries to uncover the "why" behind phenomena of interest and its most powerful and practical tool is the experimentation method. It is not hard to see that the level of clarity in problem definition vastly affects the choices available in terms of data collection and downstream analysis.

7 Challenges in Data Collection

Data collection is about data and about collection. We have seen the value inherent in the right data in Sect. 1. In Sect. 3, we have seen the importance of clarity in problem formulation while determining what data to collect. Now it is time to turn to the "collection" piece of data collection. What challenges might a data scientist typically face in collecting data? There are various ways to list the challenges that arise. The approach taken here follows a logical sequence.

The first challenge is in knowing what data to collect. This often requires some familiarity with or knowledge of the problem domain. Second, after the data scientist knows what data to collect, the hunt for data sources can proceed apace. Third, having identified data sources (the next section features a lengthy listing of data sources in one domain as part of an illustrative example), the actual process of mining of raw data can follow. Fourth, once the raw data is mined, data quality assessment follows. This includes various data cleaning/wrangling, imputation, and other data "janitorial" work that consumes a major part of the typical data science project's time. Fifth, after assessing data quality, the data scientist must now judge the relevance of the data to the problem at hand. While considering the above, at each stage one has to take into consideration the cost and time constraints.

Consider a retailing context. What kinds of data would or could a grocery retail store collect? Of course, there would be point-of-sale data on items purchased, promotions availed, payment modes and prices paid in each market basket, captured by UPC scanner machines. Apart from that, retailers would likely be interested in (and can easily collect) data on a varied set of parameters. For example, that may include store traffic and footfalls by time of the day and day of the week, basic segmentation (e.g., demographic) of the store's clientele, past purchase history of customers (provided customers can be uniquely identified, that is, through a loyalty or bonus program), routes taken by the average customer when navigating the store, or time spent on an average by a customer in different aisles and product departments. Clearly, in the retail sector, the wide variety of data sources and capture points to data are typically large in the following three areas:

- Volume
- Variety (ranges from structured metric data on sales, inventory, and geo location to unstructured data types such as text, images, and audiovisual files)
- Velocity—(the speed at which data comes in and gets updated, i.e., sales or inventory data, social media monitoring data, clickstreams, RFIDs—Radio-frequency identification, etc.)

These fulfill the three attribute criteria that are required to being labeled "Big Data" (Diebold 2012). The next subsection dives into the retail sector as an illustrative example of data collection possibilities, opportunities, and challenges.

8 Data Collation, Validation, and Presentation

Collecting data from multiple sources will not result in rich insights unless the data is collated to retain its integrity. Data validity may be compromised if proper care is not taken during collation. One may face various challenges while trying to collate the data. Below, we describe a few challenges along with the approaches to handle them in the light of business problems.

- No common identifier: A challenge while collating data from multiple sources arises due to the absence of common identifiers across different sources. The analyst may seek a third identifier that can serve as a link between two data sources.
- Missing data, data entry error: Missing data can either be ignored, deleted, or imputed with relevant statistics (see Chap. 8).
- Different levels of granularity: The data could be aggregated at different levels. For example, primary data is collected at the individual level, while secondary data is usually available at the aggregate level. One can either aggregate the data in order to bring all the observations to the same level of granularity or can apportion the data using business logic.
- Change in data type over the period or across the samples: In financial and economic data, many a time the base period or multipliers are changed, which needs to be accounted for to achieve data consistency. Similarly, samples collected from different populations such as India and the USA may suffer from inconsistent definitions of time periods—the financial year in India is from April to March and in the USA, it is from January to December. One may require remapping of old versus new data types in order to bring the data to the same level for analysis.
- Validation and reliability: As the secondary data is collected by another user, the researcher may want to validate to check the correctness and reliability of the data to answer a particular research question.

Data presentation is also very important to understand the issues in the data. The basic presentation may include relevant charts such as scatter plots, histograms, and

pie charts or summary statistics such as the number of observations, mean, median, variance, minimum, and maximum. You will read more about data visualization in Chap. 5 and about basic inferences in Chap. 6.

9 Data Collection in the Retailing Industry: An Illustrative Example

Bradlow et al. (2017) provide a detailed framework to understand and classify the various data sources becoming popular with retailers in the era of Big Data and analytics. Figure 2.2, taken from Bradlow et al. (2017), "organizes (an admittedly incomplete) set of eight broad retail data sources into three primary groups, namely, (1) traditional enterprise data capture; (2) customer identity, characteristics, social graph and profile data capture; and (3) location-based data capture." The claim is that insight and possibilities lie at the intersection of these groups of diverse, contextual, and relevant data.

Traditional enterprise data capture (marked #1 in Fig. 2.2) from UPC scanners combined with inventory data from ERP or SCM software and syndicated databases (such as those from IRI or Nielsen) enable a host of analyses, including the following:

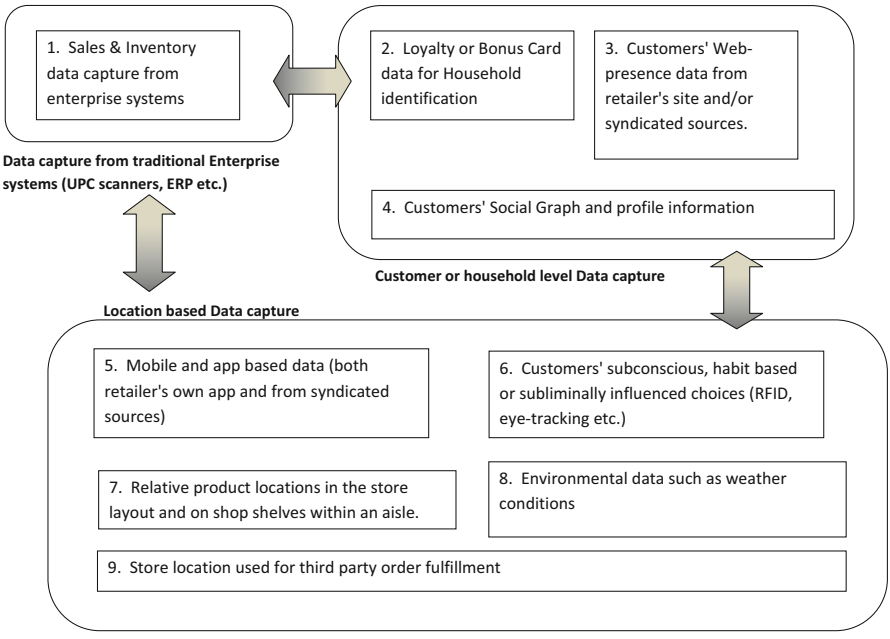

Fig. 2.2 Data sources in the modern retail sector

- Cross-sectional analysis of market baskets—item co-occurrences, complements and substitutes, cross-category dependence, etc. (e.g., Blattberg et al. 2008; Russell and Petersen 2000)
- Analysis of aggregate sales and inventory movement patterns by stock-keeping unit
- Computation of price or shelf-space elasticities at different levels of aggregation such as category, brand, and SKU (see Bijmolt et al. (2005) for a review of this literature)
- Assessment of aggregate effects of prices, promotions, and product attributes on sales

In other words, traditional enterprise data capture in a retailing context enables an overview of the four P's of Marketing (product, price, promotion, and place at the level of store, aisle, shelf, etc.).

Customer identity, characteristics, social graph, and profile data capture identify consumers and thereby make available a slew of consumer- or household-specific information such as demographics, purchase history, preferences and promotional response history, product returns history, and basic contacts such as email for email marketing campaigns and personalized flyers and promotions. Bradlow et al. (2017, p. 12) write:

> Such data capture adds not just a slew of columns (consumer characteristics) to the most detailed datasets retailers would have from previous data sources, but also rows in that household-purchase occasion becomes the new unit of analysis. A common data source for customer identification is loyalty or bonus card data (marked #2 in Fig. 2.2) that customers sign up for in return for discounts and promotional offers from retailers. The advent of household specific "panel" data enabled the estimation of household specific parameters in traditional choice models (e.g., Rossi and Allenby 1993; Rossi et al. 1996) and their use thereafter to better design household specific promotions, catalogs, email campaigns, flyers, etc. The use of household- or customer identity requires that a single customer ID be used as primary key to link together all relevant information about a customer across multiple data sources. Within this data capture type, another data source of interest (marked #3 in Fig. 2.2) is predicated on the retailer's web-presence and is relevant even for purely brick-and-mortar retailers. Any type of customer initiated online contact with the firm—think of an email click-through, online browser behavior and cookies, complaints or feedback via email, inquiries, etc. are captured and recorded, and linked to the customer's primary key. Data about customers' online behavior purchased from syndicated sources are also included here. This data source adds new data columns to retailer data on consumers' online search, products viewed (consideration set) but not necessarily bought, purchase and behavior patterns, which can be used to better infer consumer preferences, purchase contexts, promotional response propensities, etc.

Marked #4 in Fig. 2.2 is another potential data source—consumers' social graph information. This could be obtained either from syndicated means or by customers volunteering their social media identities to use as logins at various websites. Mapping the consumer's social graph opens the door to increased opportunities in psychographic and behavior-based targeting, personalization and hyper-segmentation, preference and latent need identification, selling, word of mouth, social influence, recommendation systems, etc. While the famous AIDA

framework in marketing has four conventional stages, namely, awareness, interest, desire, and action, it is clear that the "social" component's importance in data collection, analysis, modeling, and prediction is rising. Finally, the third type of data capture—location-based data capture—leverages customers' locations to infer customer preferences, purchase propensities, and design marketing interventions on that basis. The biggest change in recent years in location-based data capture and use has been enabled by customer's smartphones (e.g., Ghose and Han 2011, 2014). Figure 2.2 marks consumers' mobiles as data source #5. Data capture here involves mining location-based services data such as geo location, navigation, and usage data from those consumers who have installed and use the retailer's mobile shopping apps on their smartphones. Consumers' real-time locations within or around retail stores potentially provide a lot of context that can be exploited to make marketing messaging on deals, promotions, new offerings, etc. more relevant and impactful to consumer attention (see, e.g., Luo et al. 2014) and hence to behavior (including impulse behavior).

Another distinct data source, marked #6 in Fig. 2.2, draws upon habit patterns and subconscious consumer behaviors that consumers are unaware of at a conscious level and are hence unable to explain or articulate. Examples of such phenomena include eye-movement when examining a product or web-page (eye-tracking studies started with Wedel and Pieters 2000), the varied paths different shoppers take inside physical stores which can be tracked using RFID chips inside shopping carts (see, e.g., Larson et al. 2005) or inside virtual stores using clickstream data (e.g., Montgomery et al. 2004), the distribution of first-cut emotional responses to varied product and context stimuli which neuro-marketing researchers are trying to understand using functional magnetic resonance imaging (fMRI) studies (see, e.g., Lee et al. (2007) for a survey of the literature), etc.

Data source #7 in Fig. 2.1 draws on how retailers optimize their physical store spaces for meeting sales, share, or profit objectives. Different product arrangements on store shelves lead to differential visibility and salience. This results in a heightened awareness, recall, and inter-product comparison and therefore differential purchase propensity, sales, and share for any focal product. More generally, an optimization of store layouts and other situational factors both offline (e.g., Park et al. 1989) as well as online (e.g., Vrechopoulos et al. 2004) can be considered given the physical store data sources that are now available. Data source #8 pertains to environmental data that retailers routinely draw upon to make assortment, promotion, and/or inventory stocking decisions. For example, that weather data affects consumer spending propensities (e.g., Murray et al. 2010) and store sales has been known and studied for a long time (see, e.g., Steele 1951). Today, retailers can access a well-oiled data collection, collation, and analysis ecosystem that regularly takes in weather data feeds from weather monitoring system APIs and collates it into a format wherein a rules engine can apply, and thereafter output either recommendations or automatically trigger actions or interventions on the retailer's behalf.

Finally, data source #9 in Fig. 2.2 is pertinent largely to emerging markets and lets small, unorganized sector retailers (mom-and-pop stores) to leverage their physical location and act as fulfillment center franchisees for large retailers (Forbes 2015).

10 Summary and Conclusion

This chapter was an introduction to the important task of data collection, a process that precedes and heavily influences the success or failure of data science and analytics projects in meeting their objectives. We started with why data is such a big deal and used an illustrative example (Uber) to see the value inherent in the right kind of data. We followed up with some preliminaries on the four main types of data, their corresponding four primary scales, and the implications for analysis downstream. We then ventured into problem formulation, discussed why it is of such critical importance in determining what data to collect, and built a simple framework against which data scientists could check and validate their current problem formulation tasks. Finally, we walked through an extensive example of the various kinds of data sources available in just one business domain—retailing—and the implications thereof.

Exercises

Ex. 2.1 Prepare the movie release dataset of all the movies released in the last 5 years using IMDB.

(a) Find all movies that were released in the last 5 years.
(b) Generate a file containing URLs for the top 50 movies every year on IMDB.
(c) Read in the URL's IMDB page and scrape the following information:
Producer(s), Director(s), Star(s), Taglines, Genres, (Partial) Storyline, Box office budget, and Box office gross.
(d) Make a table out of these variables as columns with movie name being the first variable.
(e) Analyze the movie-count for every Genre. See if you can come up with some interesting hypotheses. For example, you could hypothesize that "Action Genres occur significantly more often than Drama in the top-250 list." or that "Action movies gross higher than Romance movies in the top-250 list."
(f) Write a markdown doc with your code and explanation. See if you can *storify* your hypotheses.

Note: You can web-scrape with the rvest package in R or use any platform that you are comfortable with.

Ex. 2.2 Download the movie reviews from IMDB for the list of movies.

(a) Go to www.imdb.com and examine the page.
(b) Scrape the page and tabulate the output into a data frame with columns "name, url, movie type, votes."
(c) Filter the data frame. Retain only those movies that got over 500 reviews. Let us call this Table 1.
(d) Now for each of the remaining movies, go to the movie's own web page on the IMDB, and extract the following information:
Duration of the movie, Genre, Release date, Total views, Commercial description from the top of the page.
(e) Add these fields to Table 1 in that movie's row.
(f) Now build a separate table for each movie in Table 1 from that movie's web page on IMDB. Extract the first five pages of reviews of that movie and in each review, scrape the following information:
Reviewer, Feedback, Likes, Overall, Review (text), Location (of the reviewer), Date of the review.
(g) Store the output in a table. Let us call it Table 2.
(h) Create a list (List 1) with as many elements as there are rows in Table 1. For the *i*th movie in Table 1, store Table 2 as the *i*th element of a second list, say, List 2.

Ex. 2.3 Download the Twitter data through APIs.

(a) Read up on how to use the Twitter API (https://dev.twitter.com/overview/api). If required, make a twitter ID (if you do not already have one).
(b) There are three evaluation dimensions for a movie at IMDB, namely, Author, Feedback, and Likes. More than the dictionary meanings of these words, it is interesting how they are used in different contexts.
(c) Download 50 tweets each that contain these terms and 100 tweets for each movie.
(d) Analyze these tweets and classify what movie categories they typically refer to. Insights here could, for instance, be useful in designing promotional campaigns for the movies.

P.S.: R has a dedicated package twitteR (note capital R in the end). For additional functions, refer twitteR package manual.

Ex. 2.4 Prepare the beer dataset of all the beers that got over 500 reviews.

(a) Go to (https://www.ratebeer.com/beer/top-50/) and examine the page.
(b) Scrape the page and tabulate the output into a data frame with columns "name, url, count, style."
(c) Filter the data frame. Retain only those beers that got over 500 reviews. Let us call this Table 1.
(d) Now for each of the remaining beers, go to the beer's own web page on the ratebeer site, and scrape the following information:

"Brewed by, Weighted Avg, Seasonal, Est.Calories, ABV, commercial description" from the top of the page.

Add these fields to Table 1 in that beer's row.

(e) Now build a separate table for each beer in Table 1 from that beer's ratebeer web page. Scrape the first three pages of reviews of that beer and in each review, scrape the following info:

"rating, aroma, appearance, taste, palate, overall, review (text), location (of the reviewer), date of the review."

(f) Store the output in a dataframe, let us call it Table 2.

(g) Create a list (let us call it List 1) with as many elements as there are rows in Table 1. For the *i*th beer in Table 1, store Table 2 as the *i*th element List 2.

Ex. 2.5 Download the Twitter data through APIs.

(a) Read up on how to use the twitter API here (https://dev.twitter.com/overview/api). If required, make a twitter ID (if you do not already have one).

(b) Recall three evaluation dimensions for beer at ratebeer.com, viz., aroma, taste, and palate. More than the dictionary meanings of these words, what is interesting is how they are used in context.

So pull 50 tweets each containing these terms.

(c) Read through these tweets and note what product categories they typically refer to. Insights here could, for instance, be useful in designing promotional campaigns for the beers. We will do text analysis, etc. next visit.

P.S.: R has a dedicated package twitteR (note capital R in the end). For additional functions, refer twitteR package manual.

Ex. 2.6 WhatsApp Data collection.

(a) Form a WhatsApp group with few friends/colleagues/relatives.

(b) Whenever you travel or visit different places as part of your everyday work, share your location to the WhatsApp group.

For example, if you are visiting an ATM, your office, a grocery store, the local mall, etc., then send the WhatsApp group a message saying: "ATM, [share of location here]."

Ideally, you should share a handful of locations every day. Do this DC exercise for a week. It is possible you may repeat-share certain locations.

P.S.: We assume you have a smartphone with google maps enabled on it to share locations with.

(c) Once this exercise is completed export the WhatsApp chat history of DC group to a text file. To do this, see below:

Go to WhatsApp > Settings > Chat history > Email Chat > Select the chat you want to export.

(d) Your data file should look like this:

28/02/17, 7:17 pm—fname lname: location: https://maps.google.com/?q=17.463869,78.367403

28/02/17, 7:17 pm—fname lname: ATM

(e) Now compile this data in a tabular format. Your data should have these columns:

- Sender name
- Time
- Latitude
- Longitude
- Type of place

(f) Extract your locations from the chat history table and plot it on google maps. You can use the spatial DC code we used on this list of latitude and longitude co-ordinates or use leaflet() package in R to do the same. Remember to extract and map only your own locations not those of other group members.

(g) Analyze your own movements over a week *AND* record your observations about your travels as a story that connects these locations together.

References

Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research, 42*(2), 141–156.

Blair, E., & Blair, C. (2015). *Applied survey sampling*. Los Angeles: Sage Publications.

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). Market basket analysis. *Database Marketing: Analyzing and Managing Customers*, 339–351.

Bradlow, E., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing, 93*, 79–95.

Diebold, F. X. (2012). On the origin (s) and development of the term 'Big Data'.

Forbes. (2015). From Dabbawallas to Kirana stores, five unique E-commerce delivery innovations in India. Retrieved April 15, 2015, from http://tinyurl.com/j3eqb5f.

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

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

Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing, 22*(4), 395–414.

Lee, N., Broderick, A. J., & Chamberlain, L. (2007). What is 'neuromarketing'? A discussion and agenda for future research. *International Journal of Psychophysiology, 63*(2), 199–204.

Luo, X., Andrews, M., Fang, Z., & Phang, C. W. (2014). Mobile targeting. *Management Science, 60*(7), 1738–1756.

Montgomery, C. (2017). *Design and analysis of experiments* (9th ed.). New York: John Wiley and Sons.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science, 23*(4), 579–595.

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

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

Rossi, P. E., & Allenby, G. M. (1993). A Bayesian approach to estimating household parameters. *Journal of Marketing Research, 30*, 171–182.

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science, 15*(4), 321–340.

Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing, 76*(3), 367–392.

Steele, A. T. (1951). Weather's effect on the sales of a department store. *Journal of Marketing, 15*(4), 436–443.

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

Wedel, M., & Pieters, R. (2000). Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Science, 19*(4), 297–312.