1

Introduction to information quality

1.1 Introduction

Suppose you are conducting a study on online auctions and consider purchasing a dataset from eBay, the online auction platform, for the purpose of your study. The data vendor offers you four options that are within your budget:

1. Data on all the online auctions that took place in January 2012
2. Data on all the online auctions, for cameras only, that took place in 2012
3. Data on all the online auctions, for cameras only, that will take place in the next year
4. Data on a random sample of online auctions that took place in 2012

Which option would you choose? Perhaps none of these options are of value? Of course, the answer depends on the goal of the study. But it also depends on other considerations such as the analysis methods and tools that you will be using, the quality of the data, and the utility that you are trying to derive from the analysis. In the words of David Hand (2008):

> Statisticians working in a research environment… may well have to explain that the data are inadequate to answer a particular question.

COPYRIGHTED MATERIAL

Information Quality: The Potential of Data and Analytics to Generate Knowledge,
First Edition. Ron S. Kenett and Galit Shmueli.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
Companion website: www.wiley.com/go/information_quality

While those experienced with data analysis will find this dilemma familiar, the statistics and related literature do not provide guidance on how to approach this question in a methodical fashion and how to evaluate the value of a dataset in such a scenario.

Statistics, data mining, econometrics, and related areas are disciplines that are focused on extracting knowledge from data. They provide a toolkit for testing hypotheses of interest, predicting new observations, quantifying population effects, and summarizing data efficiently. In these empirical fields, measurable data is used to derive knowledge. Yet, a clean, exact, and complete dataset, which is analyzed professionally, might contain no useful information for the problem under investigation. In contrast, a very "dirty" dataset, with missing values and incomplete coverage, can contain useful information for some goals. In some cases, available data can even be misleading (Patzer, 1995, p. 14):

> Data may be of little or no value, or even negative value, if they misinform.

The focus of this book is on assessing the potential of a particular dataset for achieving a given analysis goal by employing data analysis methods and considering a given utility. We call this concept **information quality** (InfoQ). We propose a formal definition of InfoQ and provide guidelines for its assessment. Our objective is to offer a general framework that applies to empirical research. Such element has not received much attention in the body of knowledge of the statistics profession and can be considered a contribution to both the theory and the practice of applied statistics (Kenett, 2015).

A framework for assessing InfoQ is needed both when designing a study to produce findings of high InfoQ as well as at the postdesign stage, after the data has been collected. Questions regarding the value of data to be collected, or that have already been collected, have important implications both in academic research and in practice. With this motivation in mind, we construct the concept of InfoQ and then operationalize it so that it can be implemented in practice.

In this book, we address and tackle a high-level issue at the core of any data analysis. Rather than concentrate on a specific set of methods or applications, we consider a general concept that underlies any empirical analysis. The InfoQ framework therefore contributes to the literature on statistical strategy, also known as metastatistics (see Hand, 1994).

1.2 Components of InfoQ

Our definition of InfoQ involves four major components that are present in every data analysis: an analysis goal, a dataset, an analysis method, and a utility (Kenett and Shmueli, 2014). The discussion and assessment of InfoQ require examining and considering the complete set of its components as well as the relationships between the components. In such an evaluation we also consider eight dimensions that deconstruct the InfoQ concept. These dimensions are presented in Chapter 3. We start our introduction of InfoQ by defining each of its components.

Before describing each of the four InfoQ components, we introduce the following notation and definitions to help avoid confusion:

- g denotes a specific analysis goal.
- X denotes the available dataset.
- f is an empirical analysis method.
- U is a utility measure.

We use subscript indices to indicate alternatives. For example, to convey K different analysis goals, we use $g_1, g_2,\ldots, g_K$; J different methods of analysis are denoted $f_1, f_2,\ldots, f_J$.

Following Hand's (2008) definition of statistics as "the technology of extracting meaning from data," we can think of the InfoQ framework as one for evaluating the application of a technology (data analysis) to a resource (data) for a given purpose.

1.2.1 Goal (g)

Data analysis is used for a variety of purposes in research and in industry. The term "goal" can refer to two goals: the high-level goal of the study (the "domain goal") and the empirical goal (the "analysis goal"). One starts from the domain goal and then converts it into an analysis goal. A classic example is translating a hypothesis driven by a theory into a set of statistical hypotheses.

There are various classifications of study goals; some classifications span both the domain and analysis goals, while other classification systems focus on describing different analysis goals.

One classification approach divides the domain and analysis goals into three general classes: *causal explanation*, *empirical prediction*, and *description* (see Shmueli, 2010; Shmueli and Koppius, 2011). Causal explanation is concerned with establishing and quantifying the causal relationship between inputs and outcomes of interest. Lab experiments in the life sciences are often intended to establish causal relationships. Academic research in the social sciences is typically focused on causal explanation. In the social science context, the causality structure is based on a theoretical model that establishes the causal effect of some constructs (abstract concepts) on other constructs. The data collection stage is therefore preceded by a *construct operationalization* stage, where the researcher establishes which measurable variables can represent the constructs of interest. An example is investigating the causal effect of parents' intelligence on their children's intelligence. The construct "intelligence" can be measured in various ways, such as via IQ tests. The goal of empirical prediction differs from causal explanation. Examples include forecasting future values of a time series and predicting the output value for new observations given a set of input variables. Examples include recommendation systems on various websites, which are aimed at predicting services or products that the user is most likely to be interested in. Predictions of the economy are another type of predictive goal, with forecasts of particular

economic measures or indices being of interest. Finally, descriptive goals include quantifying and testing for population effects by using data summaries, graphical visualizations, statistical models, and statistical tests.

A different, but related goal classification approach (Deming, 1953) introduces the distinction between *enumerative studies*, aimed at answering the question "how many?," and *analytic studies*, aimed at answering the question "why?"

A third classification (Tukey, 1977) classifies studies into exploratory and confirmatory data analysis.

Our use of the term "goal" includes all these different types of goals and goal classifications. For examples of such goals in the context of customer satisfaction surveys, see Chapter 7 and Kenett and Salini (2012).

1.2.2 Data (X)

Data is a broadly defined term that includes any type of data intended to be used in the empirical analysis. Data can arise from different collection instruments: surveys, laboratory tests, field experiments, computer experiments, simulations, web searches, mobile recordings, observational studies, and more. Data can be primary, collected specifically for the purpose of the study, or secondary, collected for a different reason. Data can be univariate or multivariate, discrete, continuous, or mixed. Data can contain semantic unstructured information in the form of text, images, audio, and video. Data can have various structures, including cross-sectional data, time series, panel data, networked data, geographic data, and more. Data can include information from a single source or from multiple sources. Data can be of any size (from a single observation in case studies to "big data" with zettabytes) and any dimension.

1.2.3 Analysis (f)

We use the general term *data analysis* to encompass any empirical analysis applied to data. This includes statistical models and methods (parametric, semiparametric, nonparametric, Bayesian and classical, etc.), data mining algorithms, econometric models, graphical methods, and operations research methods (such as simplex optimization). Methods can be as simple as summary statistics or complex multilayer models, computationally simple or computationally intensive.

1.2.4 Utility (U)

The extent to which the analysis goal is achieved is typically measured by some performance measure. We call this measure "utility." As with the study goal, utility refers to two dimensions: the utility from the domain point of view and the operationalized measurable utility measure. As with the goal, the linkage between the domain utility and the analysis utility measure should be properly established so that the analysis utility can be used to infer about the domain utility.

In predictive studies, popular utility measures are predictive accuracy, lift, and expected cost per prediction. In descriptive studies, utility is often assessed based on

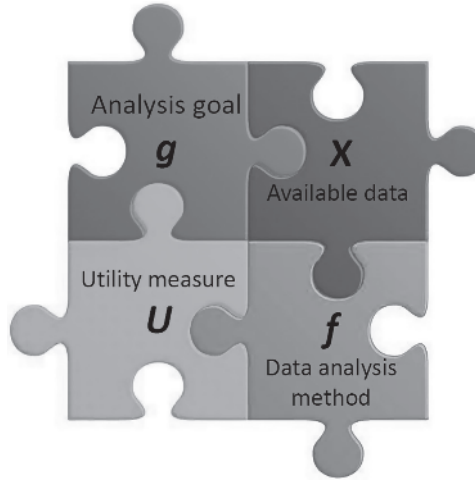

Figure 1.1 The four InfoQ components.

goodness-of-fit measures. In causal explanatory modeling, statistical significance, statistical power, and strength-of-fit measures (e.g., R^2) are common.

1.3 Definition of information quality

Following Hand's (2008) definition of statistics as "the technology of extracting meaning from data," we consider the utility of applying a technology f to a resource X for a given purpose g. In particular, we focus on the question: What is the potential of a particular dataset to achieve a particular goal using a given data analysis method and utility? To formalize this question, we define the concept of InfoQ as

$$\text{InfoQ}\left(g, X, f, U\right) = U\left\{f\left(X \mid g\right)\right\}$$

The quality of information, InfoQ, is determined by the quality of its components g ("quality of goal definition"), X ("data quality"), f ("analysis quality"), and U ("quality of utility measure") as well as by the relationships between them. (See Figure 1.1 for a visual representation of InfoQ components.)

1.4 Examples from online auction studies

Let us recall the four options of eBay datasets we described at the beginning of the chapter. In order to evaluate the InfoQ of each of these datasets, we would have to specify the study goal, the intended data analysis, and the utility measure.

To better illustrate the role that the different components play, let us examine four studies in the field of online auctions, each using data to address a particular goal.

Case study 1 Determining factors affecting the final price of an auction

Econometricians are interested in determining factors that affect the final price of an online auction. Although game theory provides an underlying theoretical causal model of price in offline auctions, the online environment differs in substantial ways. Online auction platforms such as eBay.com have lowered the entry barrier for sellers and buyers to participate in auctions. Auction rules and settings can differ from classic on-ground auctions, and so can dynamics between bidders.

Let us examine the study "Public versus Secret Reserve Prices in eBay Auctions: Results from a Pokémon Field Experiment" (Katkar and Reiley, 2006) which investigated the effect of two types of reserve prices on the final auction price. A reserve price is a value that is set by the seller at the start of the auction. If the final price does not exceed the reserve price, the auction does not transact. On eBay, sellers can choose to place a public reserve price that is visible to bidders or an invisible secret reserve price, where bidders see only that there is a reserve price but do not know its value.

STUDY GOAL (*g*)

The researchers' goal is stated as follows:

> We ask, empirically, whether the seller is made better or worse off by setting a secret reserve above a low minimum bid, versus the option of making the reserve public by using it as the minimum bid level.

This question is then converted into the statistical goal (g) of testing a hypothesis "that secret reserve prices actually do produce higher expected revenues."

DATA (*X*)

The researchers proceed by setting up auctions for Pokémon cards[1] on eBay.com and auctioning off 50 matched pairs of Pokémon cards, half with secret reserves and half with equivalently high public minimum bids. The resulting dataset included information about bids,

[1] The Pokémon trading card game was one of the largest collectible toy crazes of 1999 and 2000. Introduced in early 1999, Pokémon game cards appeal both to game players and to collectors. Source: Katkar and Reiley (2006). © National Bureau of Economic Research.

bidders, and the final price in each of the 100 auctions, as well as whether the auction had a secret or public reserve price. The dataset also included information about the sellers' choices, such as the start and close time of each auction, the shipping costs, etc. This dataset constitutes X.

Data analysis (f)

The researchers decided to "measure the effects of a secret reserve price (relative to an equivalent public reserve) on three different independent variables: the probability of the auction resulting in a sale, the number of bids received, and the price received for the card in the auction." This was done via linear regression models (f). For example, the sale/no sale outcome was regressed on the type of reserve (public/private) and other control variables, and the statistical significance of the reserve variable was examined.

Utility (U)

The authors conclude "The average drop in the probability of sale when using a secret reserve is statistically significant." Using another linear regression model with price as the dependent variable, statistical significance (the p-value) of the regression coefficient was used to test the presence of an effect for a private or public reserve price, and the regression coefficient value was used to quantify the magnitude of the effect, concluding that "a secret-reserve auction will generate a price $0.63 lower, on average, than will a public-reserve auction." Hence, the utility (U) in this study relies mostly on statistical significance and p-values as well as the practical interpretation of the magnitude of a regression coefficient.

InfoQ components evaluation

What is the quality of the information contained in this study's dataset for testing the effect of private versus public reserve price on the final price, using regression models and statistical significance? The authors compare the advantages of their experimental design for answering their question of interest with designs of previous studies using observational data:

> With enough [observational] data and enough identifying econometric assumptions, one could conceivably tease out an empirical measurement of the reserve price effect from eBay field data… Such structural models make strong identifying assumptions in order to recover economic unobservables (such as bidders' private information about the item's value)… In contrast, our research project is much less ambitious, for we focus only on the effect of secret reserve prices relative to public reserve prices (starting bids). Our experiment allows us to carry out this measurement in a manner that is as simple, direct, and assumption-free as possible.

In other words, with a simple two-level experiment, the authors aim to answer a specific research question (g_1) in a robust manner, rather than build an extensive theoretical economic model (g_2) that is based on many assumptions.

Interestingly, when comparing their conclusions against prior literature on the effect of reserve prices in a study that used observational data, the authors mention that they find an opposite effect:

> Our results are somewhat inconsistent with those of Bajari and Hortaçsu…. Perhaps Bajari and Hortaçsu have made an inaccurate modeling assumption, or perhaps there is some important difference between bidding for coin sets and bidding for Pokémon cards.

This discrepancy even leads the researchers to propose a new dataset that can help tackle the original goal with less confounding:

> A new experiment, auctioning one hundred items each in the $100 range, for example could shed some important light on this question.

This means that the InfoQ of the Pokémon card auction dataset is considered lower than that of a more expensive item.

Case study 2 Predicting the final price of an auction at the start of the auction

On any given day, thousands of auctions take place online. Forecasting the price of ongoing auctions is beneficial to buyers, sellers, auction houses, and third parties. For potential bidders, price forecasts can be used for deciding if, when, and how much to bid. For sellers, price forecasts can help decide whether and when to post another item for sale. For auction houses and third parties, services such as seller insurance can be offered with adjustable rates. Hence, there are different possible goals for empirical studies where price is the outcome variable, which translate into different InfoQ of a dataset. We describe in the succeeding text one particular study.

STUDY GOAL (*g*)

In a study by Ghani and Simmons (2004), the researchers collected historical auction data from eBay and used machine learning algorithms to predict end prices of auction items. Their question (g) was whether end prices of online auctions can be predicted accurately using machine learning methods. This is

a predictive forward-looking goal, and the results of the study can improve scientific knowledge about predictability of online auction prices as well as serve as the basis for practical applications.

Data (X)

The data collected for each closed auction included information about the seller, the item, the auction format, and "temporal features" (price statistics: starting bid, shipping price, and end price) of other auctions that closed recently. Note that all this information is available at the start of an auction of interest and therefore can be used as predictors for its final price. In terms of the outcome variable of interest—price—the data included the numerical end price (in USD). However, the authors considered two versions of this variable: the raw continuous variable and a multiclass categorical price variable where the numerical price is binned into $5 intervals.

Data analysis (f)

In this study, several predictive algorithms (f) were used: for the numerical price, they used linear regression (and "polynomial regression with degrees 2 and 3"). For the categorical price, they used classification trees and neural networks.

Utility (U)

Because the authors' goal focused on predictive accuracy, their performance measures (U) were computed from a holdout set (RMSE for numerical price and accuracy % for categorical price). This set consisted of 400 auctions that were not used when building ("training") the models. They benchmarked their performance against a naive prediction—the average price (for numerical price) or most common price bin (for categorical price). The authors concluded:

> All of the methods we use[d] are effective at predicting the end-price of auctions. Regression results are not as promising as the ones for classification, mainly because the task is harder since an exact price is being predicted as opposed to a price range. In the future, we plan to narrow the bins for the price range and experiment with using classification algorithms to achieve more fine-grained results.

InfoQ components evaluation

For the purpose of their research goal, the dataset proved to be of high InfoQ. Moreover, they were able to assert the difference in InfoQ between two versions of their data (numerical and categorical price). Following their results,

the authors proposed two applications where predicting price intervals of an auction might be useful:

> **Price Insurance**: Knowing the end-price before an auction starts provides an opportunity for a third-party to offer price insurance to sellers….
>
> **Listing Optimizer**: The model of the end price based on the input attributes of the auction can also be used to help sellers optimize the selling price of their items.

Case study 3 Predicting the final price of an ongoing auction

We now consider a different study, also related to predicting end prices of online auctions, but in this case predictions will be generated during an ongoing auction. The model used by Ghani and Simmons (2004) for forecasting the price of an auction is a "static model" in the sense that it uses information that is available at the start of the auction, but not later. This must be the case if the price forecasting takes place at the start of the auction. Forecasting the price of an ongoing auction is different: in addition to information available at the start of the auction, we can take into account all the information available at the time of prediction, such as bids that were placed thus far.

Recent literature on online auctions has suggested such models that integrate dynamic information that changes during the auction. Wang et al. (2008) developed a dynamic forecasting model that accounts for the unequal spacing of bids, the changing dynamics of price and bids throughout the auction, as well as static information about the auction, seller, and product. Their model has been used for predicting auction end prices for a variety of products (electronics, contemporary art, etc.) and across different auction websites (see Jank and Shmueli, 2010, Chapter 4). In the following, we briefly describe the Wang et al. (2008) study in terms of the InfoQ components.

Study goal (*g*)

The goal (g) stated by Wang et al. (2008) is to develop a forecasting model that predicts end prices of an ongoing online auction more accurately than traditional models. This is a forward-looking, predictive goal, which aims to benchmark a new modeling approach against existing methods. In addition to the main forecasting goal, the authors also state a secondary goal, to "systematically describe the empirical regularities of auction dynamics."

DATA (X)

The researchers collected data on a set of 190 closed seven-day auctions of Microsoft Xbox gaming systems and *Harry Potter and the Half-Blood Prince* books sold on eBay.com in August–September 2005. For each auction, the data included the bid history (bid amounts, time stamps, and bidder identification) and information on the product characteristics, the auction parameters (e.g., the day of week on which the auction started), and bidder and seller. Bid history information, which includes the timings and amounts of bids placed during the auction, was also used as predictor information.

DATA ANALYSIS (f)

The forecasting model proposed by Wang et al. (2008) is based on representing the sequences of bids from each auction by a smooth curve (using *functional data analysis*). An example for four auctions is shown in Figure 1.2. Then, a regression model for the price at time t includes four types of predictors:

a. Static predictors (such as product characteristics)

b. Time-varying predictors (such as the number of bids by time t)

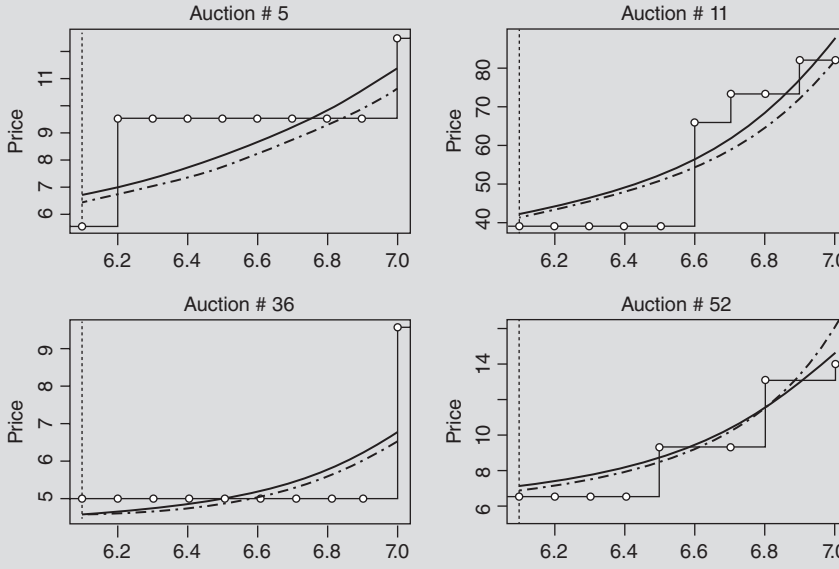

Figure 1.2 Price curves for the last day of four seven-day auctions (x-axis denotes day of auction). Current auction price (line with circles), functional price curve (smooth line) and forecasted price curve (broken line).

c. Price dynamics (estimated from the price curve derivatives)

d. Price lags

Their model for the price at time t is given by

$$y(t) = \alpha + \sum_{i=1}^{Q} \beta_i x_i(t) + \sum_{j=1}^{J} \beta_j D^j y(t) + \sum_{l=1}^{L} \eta_L y(t-l),$$

where $x_1(t),\ldots, x_Q(t)$ is the set of static and time-varying predictors, $D^{(j)}y(t)$ denotes the jth derivative of price at time t, and $y(t-l)$ is the lth price lag. The h-step-ahead forecast, given information up to time T, is given by

$$\bar{y}(T+h|T) = \hat{\alpha} + \sum_{i=1}^{Q} \hat{\beta}_i x_i(T+h|T) + \sum_{j=1}^{J} \hat{\gamma}_j \bar{D}^{(j)} y(T+h|T) + \sum_{l=1}^{L} \hat{\eta}_L \bar{y}(T+h-1|T).$$

UTILITY (*U*)

As in case study 2, predictive accuracy on a holdout set of auctions was used for evaluating model performance. In this study, the authors looked at two types of errors: (i) comparing the functional price curve and the forecasted price curve and (ii) comparing the forecast curves with the actual current auction prices.

INFOQ COMPONENTS EVALUATION

The authors make use of information in online auction data that are typically not used in other studies forecasting end prices of auctions: the information that becomes available during the auction regarding bid amounts and timings. They show that this additional information, if integrated into the prediction model, can improve forecast accuracy. Hence, they show that the InfoQ is high by generating more accurate forecasts as well as by shedding more light on the relationship between different auction features and the resulting bid dynamics. The authors conclude:

> The model produces forecasts with low errors, and it outperforms standard forecasting methods, such as double exponential smoothing, that severely underpredict the price evolution. This also shows that online auction forecasting is not an easy task. Whereas traditional methods are hard to apply, they are also inaccurate because they do not take into account the dramatic change in auction dynamics. Our model, on the other hand, achieves high forecasting accuracy and accommodates the changing price dynamics well.

Case study 4 Quantifying consumer surplus in eBay auctions

Classic microeconomic theory uses the notion of consumer surplus as the welfare measure that quantifies benefits to a consumer from an exchange. Marshall (1920, p. 124) defined consumer surplus as "the excess of the price which he (a consumer) would be willing to pay rather than go without the thing, over that which he actually does pay …."

Despite the growing research interest in online auctions, little is known about quantifiable consumer surplus levels in such mechanisms. On eBay, the winner is the highest bidder, and she or he pays the second highest bid. Whereas bid histories are publicly available, eBay never reveals the highest bid. Bapna et al. (2008) set out to quantify consumer surplus on eBay by using a unique dataset which revealed the highest bids for a sample of almost 5000 auctions. They found that, under a certain assumption, "eBay's auctions generated at least $7.05 billion in total consumer surplus in 2003."

Study goal (*g*)

The researchers state the goal (*g*) as estimating the consumer surplus generated in eBay in 2003. This is a descriptive goal, and the purpose is to estimate this quantity with as much accuracy as possible.

Data (*X*)

Since eBay does not disclose the highest bid in an auction, the researchers used a large dataset from Cniper.com, a Web-based tool used at the time by many eBay users for placing a "last minute bid." Placing a bid very close to the auction close ("sniping") is a tactic for winning an auction by avoiding the placement of higher bids by competing bidders. The Cniper dataset contained the highest bid for all the winners. The authors then merged the Cniper information with the eBay data for those auctions and obtained a dataset of 4514 auctions that took place between January and April 2003. Their dataset was also unique in that it contained information on auctions in three different currencies and across all eBay product categories.

Empirical analysis (*f*)

The researchers computed the median surplus by using the sample median with a 95% bootstrap confidence interval. They examined various subsets of the data and used regression analysis to correct for possible biases and to evaluate robustness to various assumption violations. For example, they compared their sample with a random sample from eBay in terms of the various variables, to evaluate whether Cniper winners were savvier and hence derived a higher surplus.

UTILITY (*U*)

The precision of the estimated surplus value was measured via a confidence interval. The bias due to nonrepresentative sampling was quantified by calculating an upper bound.

INFOQ COMPONENTS EVALUATION

The unique dataset available to the researchers allowed them to compute a metric that is otherwise unavailable from publicly available information on eBay.com. The researchers conducted special analyses to correct for various biases and arrived at the estimate of interest with conservative bounds. The InfoQ of this dataset is therefore high for the purpose of the study.

1.5 InfoQ and study quality

We defined InfoQ as a framework for answering the question: What is the potential of a particular dataset to achieve a particular goal using a given data analysis method and utility? In each of the four studies in Section 1.4, we examined the four InfoQ components and then evaluated the InfoQ based on examining the components. In Chapter 3 we introduce an InfoQ assessment approach, which is based on eight dimensions of InfoQ. Examining each of the eight dimensions assists researchers and analysts in evaluating the InfoQ of a dataset and its associated study.

In addition to using the InfoQ framework for evaluating the potential of a dataset to generate information of quality, the InfoQ framework can be used for retrospective evaluation of an empirical study. By identifying the four InfoQ components and assessing the eight InfoQ dimensions introduced in Chapter 3, one can determine the usefulness of a study in achieving its stated goal. In part II of the book, we take this approach and examine multiple studies in various domains. Chapter 12 in part III describes how the InfoQ framework can provide a more guided process for authors, reviewers and editors of scientific journals and publications.

1.6 Summary

In this chapter we introduced the concept of InfoQ and its four components. In the following chapters, we discuss how InfoQ differs from the common concepts of data quality and analysis quality. Moving from a concept to a framework that can be applied in practice requires a methodology for assessing InfoQ. In Chapter 3, we break down InfoQ into eight dimensions, to facilitate quantitative assessment of InfoQ. The final chapters (Chapters 4 and 5) in part I examine existing statistical methodology aimed at increasing InfoQ at the study design stage and at the postdata collection stage. Structuring and examining various statistical approaches through the InfoQ lens creates a clearer picture of the role of different statistical approaches

and methods, often taught in different courses or used in separate fields. In summary, InfoQ is about assessing and improving the potential of a dataset to achieve a particular goal using a given data analysis method and utility. This book is about structuring and consolidating such an approach.

References

Bapna, R., Jank, W. and Shmueli, G. (2008) Consumer surplus in online auctions. *Information Systems Research*, 19, pp. 400–416.

Deming, W.E. (1953) On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48, pp. 244–255.

Ghani, R. and Simmons, H. (2004) Predicting the End-Price of Online Auctions. *International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management*, Pisa, Italy.

Hand, D.J. (1994) Deconstructing statistical questions (with discussion). *Journal of the Royal Statistical Society, Series A*, 157(3), pp. 317–356.

Hand, D.J. (2008) *Statistics: A Very Short Introduction*. Oxford University Press, Oxford.

Jank, W. and Shmueli, G. (2010) *Modeling Online Auctions*. John Wiley & Sons, Inc., Hoboken.

Katkar, R. and Reiley, D.H. (2006) Public versus secret reserve prices in eBay auctions: results from a Pokemon field experiment. *Advances in Economic Analysis and Policy*, 6(2), article 7.

Kenett, R.S. (2015) Statistics: a life cycle view (with discussion). *Quality Engineering*, 27(1), pp. 111–129.

Kenett, R.S. and Salini, S. (2012) Modern analysis of customer surveys: comparison of models and integrated analysis (with discussion). *Applied Stochastic Models in Business and Industry*, 27, pp. 465–475.

Kenett, R.S. and Shmueli, G. (2014) On information quality (with discussion). *Journal of the Royal Statistical Society, Series A*, 177(1), pp. 3–38.

Marshall, A. (1920) *Principles of Economics*, 8th edition. MacMillan, London.

Patzer, G.L. (1995) *Using Secondary Data in Marketing Research*. Praeger, Westport, CT.

Shmueli, G. (2010) To explain or to predict? *Statistical Science*, 25, pp. 289–310.

Shmueli, G. and Koppius, O.R. (2011) Predictive analytics in information systems research. *Management Information Systems Quarterly*, 35, pp. 553–572.

Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, PA.

Wang, S., Jank, W. and Shmueli, G. (2008) Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economics Statistics*, 26, pp. 144–160.