An Information Quality (InfoQ) Framework for Ex-Ante and Ex-Post Evaluation of Empirical Studies

Galit Shmueli and Ron Kenett

Abstract Numbers are not data and data analysis does not necessarily produce information and knowledge. Statistics, data mining, and artificial intelligence are disciplines focused on extracting knowledge from data. They provide tools for testing hypotheses, predicting new observations, quantifying population effects, and summarizing data efficiently. In these fields, measurable data is used to derive knowledge. However, a clean, exact and complete dataset, which is analyzed professionally, might contain no useful information for the problem under investigation. The term *Information Quality* (InfoQ) was coined by [15] as the potential of a dataset to achieve a specific (scientific or practical) goal using a given data analysis method. InfoQ is a function of goal, data, data analysis, and utility. Eight dimensions that relate to these components help assess InfoQ: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Generalizability, Chronology of Data and Goal, Construct Operationalization, and Communication. The eight dimensions can be used for developing streamlined evaluation metrics of InfoQ. We describe two studies where InfoQ was integrated into research methods courses, guiding students in evaluating InfoQ of prospective and retrospective studies. The results and feedback indicate the importance and usefulness of InfoQ and its eight dimensions for evaluating empirical studies.

Galit Shmueli
Srini Raju Centre for IT and the Networked Economy, Indian School of Business, Hyderabad 500 032, India, and Rigsum Institute of IT & Management, Thimphu, Bhutan.
e-mail: galit_shmueli@isb.edu

Ron Kenett
KPA Ltd., Raanana, Israel Dept of Statistics & Applied Mathematics, University of Torino, Torino, Italy, and Center for Finance and Risk Engineering, NYU-Poly, NY 11201, USA.
e-mail: ron@kpa-group.com

1 Introduction and Motivation

The term Intelligent Data Analysis (IDA) implies an expectation that data analysis will yield insights and knowledge. Research and academic environments focus on developing intelligent tools for extracting information from data. Statistics education is typically aimed at teaching analysis quality. Godfrey [10] describes low quality of analysis as "poor models and poor analysis techniques, or even analyzing the data in a totally incorrect way". The book *Guide To Intelligent Data Analysis* [4] has the subtitle *How to Intelligently Make Sense of Real Data* and is focused on pitfalls that lead to wrong or insufficient analysis of results. In other words, *intelligent* most often refers to the analysis quality.

While analysis quality is critically important, another key component of IDA is the usefulness of a particular dataset for the problem at hand. The same data can contain high-quality information for one purpose and low-quality information for another purpose. An important question that arises both in scientific research and in practical applications is therefore: what is the potential of a dataset to achieve a particular goal of interest? This is related to the Zeroth Problem, coined by Mallows [18], which is the general question of "how do the data relate to the problem, and what other data might be relevant?" Hand [11] notes, "statisticians working in a research environment... may well have to explain that the data are inadequate to answer a particular question". Patzer [19] comments: "data may be of little or no value, or even negative value, if they misinform".

There is therefore a need to formalize these important aspects of IDA that have thus far not been formalized. Recently, Kenett and Shmueli coined the term *Information Quality*, or InfoQ, to define *the potential of a dataset to achieve a specific (scientific or practical) goal using a given data analysis method* ([15] with discussion and rejoinder). InfoQ lies on the interface of data, goal, and analyst and is tightly coupled with the analysis context. This is schematically illustrated in Figure 1.

The focus of this paper is on integrating InfoQ into the thought process of data analysts, while conducting an active empirical study as well as for ex-post evaluation of empirical studies. We proceed as follows: Section 2 introduces InfoQ, its components, terminology and formal definition. In Section 3 we describe eight dimensions of InfoQ that are useful for assessing InfoQ in practice. Section 4 discusses an evaluation methodology, and then describes two studies. The first study describes the integration of InfoQ into a graduate-level research methods course at Ljubljana University. The second study describes an InfoQ assignment designed for ex-post evaluation of empirical studies, and its implementation in a Masters in Statistical Practice program at Carnegie Mellon University. We conclude and offer future directions in Section 5.

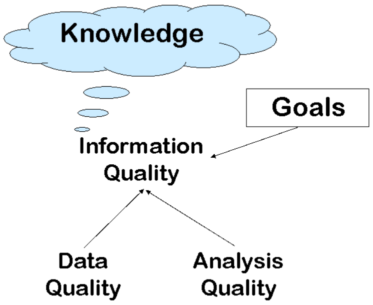


Fig. 1 InfoQ depends on data quality and analysis quality, conditional on the goal at hand.

2 Information Quality: Terminology and Definition

InfoQ is a function of several components: data, analysis goal, data analysis method, and the anticipated utility from the analysis. We describe each of these four components and then define the InfoQ function.

2.1 InfoQ components

Analysis Goal (g): Data analysis is used for variety of purposes. Three general classes of goals are causal explanation, prediction, and description [21, 22]. Causal explanation includes questions such as Which factors cause the out-come?" Prediction goals include forecasting future values of a time series and pre-dicting the output value for new observations given a set of input variables. De-scriptive goals include quantifying and testing for population effects using data summaries, graphical visualizations, statistical models, and statistical tests. Deming [6] introduced the distinction between enumerative studies, aimed at answering the question "how many?" and analytic studies, aimed at answering the question why? Later, Tukey [23] proposed a classification of exploratory and confirmatory data analysis. Our use of the term goal generalizes all of these different types of goals and goal classifications.

Data (X): The term data includes any type of data to which empirical analysis can be applied. Data can arise from different collection tools: surveys, laboratory tests, field and computer experiments, simulations, web searches, observational studies and more. Data can be univariate or multivariate (one or more variables) and of any size (from a single observation in case studies to many observations). It can also contain semantic, unstructured information in the form of text or images with or without a dynamic time dimension. Data is the foundation of any application of empirical analysis.

Data Analysis Method (f): We use the term data analysis to refer to statistical analysis and data mining. This includes statistical models and methods (parametric, semiparametric, nonparametric), data mining algorithms, and graphical methods. Operations research methods, such as simplex optimization, where problems are modeled and parametrized, fall into this category as well.

Utility (U): The extent to which the analysis goal is achieved is typically measured by some performance measure. We call this measure utility. For example, in studies with a predictive goal a popular performance measure is predictive accuracy. In descriptive studies, common utility measures are goodness-of-fit measures. In explanatory models, statistical power and strength-of-fit measures are common utility measures.

2.2 Information Quality (InfoQ): Definition

Following Hands definition of statistics as The technology of extracting meaning from data [11], 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 empirical analysis method? To formalize this question of interest, we define the concept of Information Quality (InfoQ) as:

$$InfoQ(f,X,g) = U(f(X \mid g)) \tag{1}$$

InfoQ is affected 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 X, f, g and U.

2.3 Example: Online Auctions

Some of the large online auction websites, such as eBay, provide data on closed and ongoing auctions, triggering a growing body of research in academia and in practice. A few popular analysis goals have been:

- Determining factors affecting the final price of an auction [17]
- Predicting the final price of an auction [8]
- Descriptive characterization of bidding strategies [2, 5]
- Comparing behavioral characteristics of auction winners vs. fixed-price buyers [1]
- Building descriptive statistical models of bid arrivals or bidder arrivals [5]

Given the diverse goals, it is intuitive that one dataset of eBay auctions would hold different value (InfoQ) in terms of its potential to derive insights.

Let us consider a particular goal for illustrating the components of InfoQ. Econometricians are interested in determining factors affecting the final price of an online auction. While game theory provides an underlying theoretical causal model of price in offline auctions, the online environment differs in substantial ways. We consider a study by Katkar and Reiley [13] who investigated the effect of two types of reserve prices on the final auction price. A reserve price is a value 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 (bidders only see that there is a reserve price but do not know its value). InfoQ, in the context of this study, consists of asking the question: "Given the data collected on a set of auctions, what is their potential to allow quantifying the difference between secret and public reserve prices using regression modeling?"

Study Goal (g) : Quantify the effect of using a secret vs. public reserve price on the final price of an auction.

Data (X) : The authors conducted a field experiment by selling Pokemon cards on eBay. They auctioned 25 identical pairs of Pokemon cards in week-long auctions during a two-week period in April 2000, where each card was auctioned twice: once with a public reserve price and once with a secret reserve price. The resulting data included the complete information on all 50 auctions.

Data Analysis Method (f) : The authors used linear regression to test for the effect of private/public reserve price on the final auction price and to quantify it.

Utility (U) : The authors used statistical significance (p-value) of the regression coefficient to assess the presence of an effect for private/public reserve price. They used the regression coefficient value for quantifying the magnitude of the effect (they conclude: "a secret-reserve auction will generate a price $0.63 lower, on average, than will a public-reserve auction".)

3 Eight Dimensions of InfoQ

Quality of Statistical Data is a concept developed and used in European official statistics and international organizations such as the International Monetary Fund (IMF) and The Organisation for Economic Cooperation and Development (OECD). This concept refers to the usefulness of summary statistics produced by national statistics agencies and other producers of official statistics. This is a special case of InfoQ, where the data analysis method (f) is the computation of summary statistics. Although this operation might seem very simple, it is nonetheless considered analysis, because it is in fact estimation. Hence, InfoQ is more general. Quality of statistical data is evaluated in terms of the usefulness of the statistics for a particular goal. The OECD uses seven dimensions for quality assessment: relevance, accuracy, timeliness and punctuality, accessability, interpretability, coherence, and credibility (Chap 5 in [9]). We use a similar framework to determine InfoQ dimensions.

Taking an approach similar to data quality assessment, we define eight dimensions for assessing InfoQ that consider and affect not only the data and goal, but also the analysis method and utility function. With this approach we provide a decomposition of InfoQ that can be used for assessing and improving research initiatives or ex-post evaluations.

3.1 Data Resolution

Data resolution refers to the measurement scale and aggregation level of X. The measurement scale of the data should be carefully evaluated in terms of its suitability to the goal, the analysis methods to be used, and the required resolution of U. Given the original recorded scale, the researcher should evaluate its adequacy. It is usually easy to produce a more aggregated scale (e.g., two income categories instead of ten), but not a finer scale. Data might be recorded by multiple instruments or by multiple sources. To choose among the multiple measurements, supplemental information about the reliability and precision of the measuring devices or data sources is useful. A finer measurement scale is often associated with more noise; hence the choice of scale can affect the empirical analysis directly. The data aggregation level must also be evaluated relative to the goal.

3.2 Data Structure

Data structure relates to the type(s) of data and data characteristics such as corrupted and missing values due to the study design or data collection mechanism. Data types include structured numerical data in different forms (e.g., cross-sectional, time series, network data) as well as unstructured, non-numerical data (e.g., text, text with hyperlinks, audio, video, and semantic data). The InfoQ level of a certain data type depends on the goal at hand.

3.3 Data Integration

With the variety of data source and data types, there is often a need to integrate multiple sources and/or types. Often, the integration of multiple data types creates new knowledge regarding the goal at hand, thereby increasing InfoQ. For example, in online auction research, the integration of temporal bid sequences with cross-sectional auction and seller information has led to more precise predictions of final prices (see Chapter 4 in [12]) as well as to an ability to quantify the effects of different factors on the price process [3].

3.4 Temporal Relevance

The process of deriving knowledge from data can be put on a time line that includes the data collection, data analysis, and study deployment periods as well as the temporal gaps between the data collection, the data analysis, and the study deployment stages (see Figure 2). These different durations and gaps can each affect InfoQ. The data collection duration can increase or decrease InfoQ, depending on the study goal, e.g., studying longitudinal effects vs. a cross-sectional goal. Similarly, if the collection period includes uncontrollable transitions, this can be useful or disruptive, depending on the study goal.

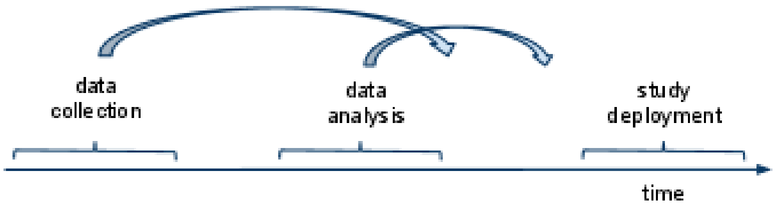


Fig. 2 Temporal durations and gaps that affect InfoQ. Feedback arrows indicate the cyclic process of data collection and analysis.

3.5 Chronology of Data and Goal

The choice of variables to collect, the temporal relationship between them, and their meaning in the context of the goal at hand also affects InfoQ. For example, in the context of online auctions, classic auction theory dictates that the number of bidders is an important driver of auction price. Models based on this theory are useful for explaining the effect of the number of bidders on price. However, for the purpose of predicting the price of ongoing online auctions, where the number of bidders is unknown until the auction end, the variable “number of bidders”, even if available in the data, is useless. Hence, the level of InfoQ contained in number of bidders for models of auction price depends on the goal at hand.

3.6 Generalizability

The utility of $f(X \mid g)$ is dependent on the ability to generalize f to the appropriate population. Two types of generalizability are statistical and scientific generalizability. Statistical generalizability refers to inferring from a sample to a target population. Scientific generalizability refers to applying a model based on a particular target population to other populations. This can mean either generalizing an estimated population pattern/model f to other populations, or applying f estimated from one population to predict individual observations in other populations. Determining the level of generalizability requires careful characterization of g. For instance, for inferring about a population parameter, statistical generalizability and sampling bias are the focus, and the question of interest is "What population does the sample represent?" [24]. In contrast, for predicting the values of new observations, the question of interest is whether f captures associations in the training data X that are generalizable to the to-be-predicted data.

3.7 Construct Operationalization

Constructs are abstractions that describe a phenomenon of theoretical interest. Measurable data are an operationalization of underlying constructs. The relationship between the underlying construct and its operationalization can vary, and its level relative to the goal is another important aspect of InfoQ. The role of construct operationalization is dependent on the goal, and especially on whether the goal is explanatory, predictive, or descriptive. In explanatory models, based on underlying causal theories, multiple operationalizations might be acceptable for representing the construct of interest. As long as the data are assumed to measure the construct, the variable is considered adequate. In contrast, in a predictive task, where the goal is to create sufficiently accurate predictions of a certain measurable variable, the choice of operationalized variable is critical.

3.8 Communication

Effective communication of the analysis and its utility directly impacts InfoQ. There are plenty of examples where miscommunication of valid results has led to disasters, such as the NASA shuttle Challenger disaster [16]. Communication media are visual, textual, and verbal presentations and reports. Within research environments, communication focuses on written publications and conference presentations. Research mentoring and the refereeing process are aimed at improving communication and InfoQ within the research community.

4 Assessing Information Quality

The eight dimensions of InfoQ are intended to help operationalize the concept into actionable evaluation. Considering InfoQ and its dimensions can be useful during an ongoing empirical study (from study design, through data collection and analysis, to presentation) as well as in ex-post evaluations of completed studies. InfoQ evaluation can range from qualitative to quantitative. Qualitative evaluation consists of verbal or written assessments of a study on each dimension. Quantitative evaluation requires defining a metric for which we can evaluate each dimension. One such rating-based metric is described in the next section.

4.1 Rating-Based Evaluation

Similar to the use of "data quality" dimensions by statistical agencies for evaluating data quality, we evaluate the eight InfoQ dimensions to assess InfoQ. This evaluation integrates different aspects of a study and assigns an overall InfoQ score. The broad perspective of InfoQ dimensions is designed to help researchers enhance the added value of their studies.

Assessing InfoQ using quantitative metrics can be done in several ways. [15] presented a rating-based approach that examines a study report and scores each of the eight InfoQ dimensions. A coarse grained approach is to rate each dimension on a 1-5 scale, with "5" indicating "high" achievement on that dimension. The ratings (Y_i, i=1,...,8) can then be normalized into a desirability function (see [7]) for each dimension, ($0 \leq d(Y_i) \leq 1$), which are then combined to produce an overall InfoQ score using the geometric mean of the individual desirabilities:

$$\text{InfoQ Score} = [d_1(Y_1) \times d_2(Y_2) \times \ldots \times d_8(Y_8)]^{1/8} \qquad (2)$$

In the next sections we describe two studies where participants used InfoQ dimensions to evaluate an empirical study - either a proposed study or a completed one. The goal of the two studies was to introduce graduate students to the application of statistics in practice, to key questions that a statistician should ask, and to the link between goal, data, analysis and context. The InfoQ framework provides such an integrative view.

4.2 Study 1: Evaluating Research Proposals

Graduate students at the Faculty of Economics of the University of Ljubljana undergo a 3-day workshop on research methods, to equip them with methodlogy for developing and presenting a research proposal that is the basis for their doctorate thesis. One of the first milestones for students is to defend their research proposal in front of a committee.

The class consisted of approximately 50 graduate students from a wide range of areas, including organizational behavior, operations research, marketing, and economics. In 2009, InfoQ was integrated into the research methods workshop. The goal was to help students (and their advisors) figure out whether their proposed research is properly defines as to potentially generate effective knowledge.

Students worked in small teams and discussed the InfoQ dimensions of their draft proposal. Each student then gave a 15 minute presentation to the whole team. Details of the workshop and pre-workshop assignments are availalbe at `goo.gl/f6bIA`. Students' grades in the workshop were derived from an InfoQ score of their proposal submission, which consisted of a PowerPoint presentation and a written document. This approach was designed to make their research journey more efficient and more effective. Feedback by students and faculty, has indicated that the InfoQ-based research methods workshop has indeed met this goal [14].

4.3 Study 2: Ex-Post Evaluation of Empirical Studies

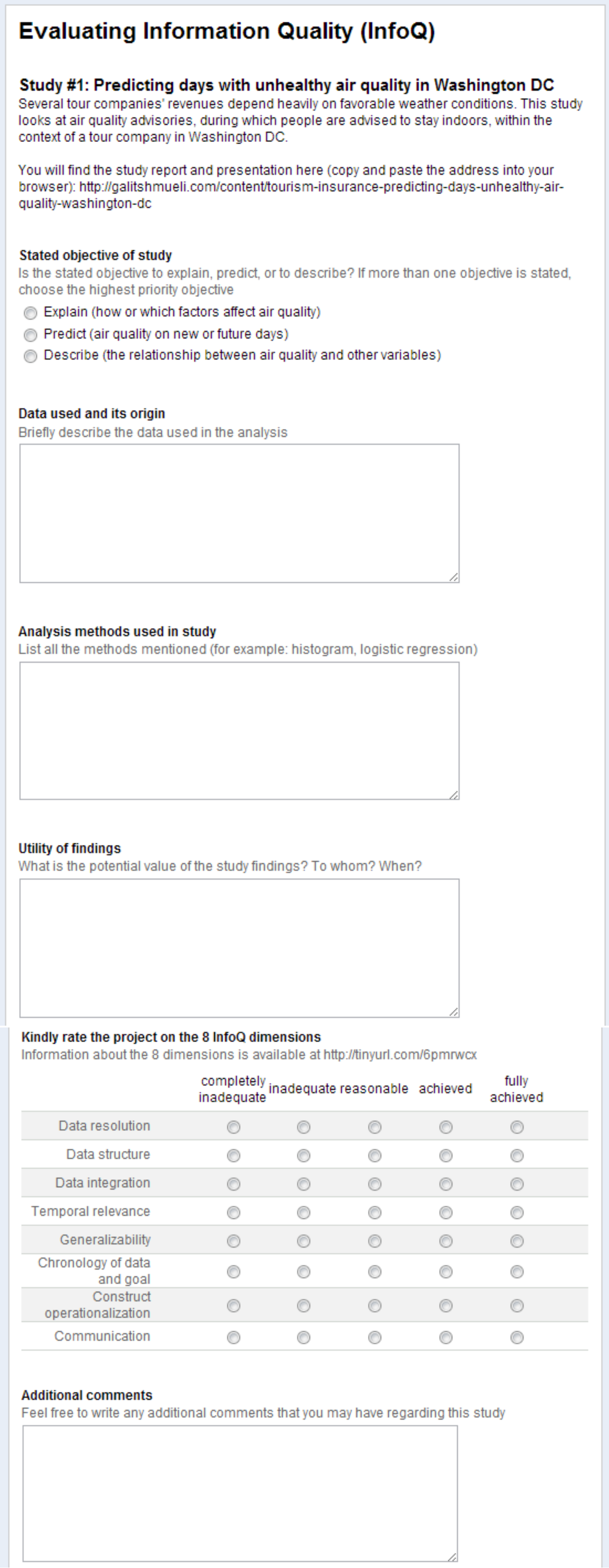
We designed an assignment that requires participants to evaluate five empirical studies based on written reports. Based on the reports and on a brief introduction to InfoQ, participants were asked to:

1. give a brief description of the goal, data, analysis, and utility measure for each study, and
2. rate the study on each of the eight InfoQ dimensions

The form with information on InfoQ, on the five studies, and the InfoQ questions and ratings are available at `goo.gl/erNPF`. Figure 4.3 shows the questions asked for one of the studies.

In 2012, the InfoQ assignment was integrated into a course in the Masters in Statistics Practice program at Carnegie Mellon Universitys Statistics Department. Each of the 16 students spent about 60-90 min reading the five studies and evaluating the eight dimensions for each study.

Comparing the responses of the 16 participants on each of the five studies revealed variability in respondents ratings of the InfoQ dimensions. This variability indicates a need to further streamline the process of quantifying each dimension rating. An important result of this study was the feedback regarding the value added by going through the evaluation process. Participants reported that using this approach helped them “sort out all of the information”, and several reported that they will adopt this evaluation approach for future studies.



Evaluating Information Quality (InfoQ)

Study #1: Predicting days with unhealthy air quality in Washington DC
Several tour companies' revenues depend heavily on favorable weather conditions. This study looks at air quality advisories, during which people are advised to stay indoors, within the context of a tour company in Washington DC.

You will find the study report and presentation here (copy and paste the address into your browser): http://galitshmueli.com/content/tourism-insurance-predicting-days-unhealthy-air-quality-washington-dc

Stated objective of study
Is the stated objective to explain, predict, or to describe? If more than one objective is stated, choose the highest priority objective

○ Explain (how or which factors affect air quality)
○ Predict (air quality on new or future days)
○ Describe (the relationship between air quality and other variables)

Data used and its origin
Briefly describe the data used in the analysis

Analysis methods used in study
List all the methods mentioned (for example: histogram, logistic regression)

Utility of findings
What is the potential value of the study findings? To whom? When?

Kindly rate the project on the 8 InfoQ dimensions
Information about the 8 dimensions is available at http://tinyurl.com/6pmrwcx

	completely inadequate	inadequate	reasonable	achieved	fully achieved
Data resolution	○	○	○	○	○
Data structure	○	○	○	○	○
Data integration	○	○	○	○	○
Temporal relevance	○	○	○	○	○
Generalizability	○	○	○	○	○
Chronology of data and goal	○	○	○	○	○
Construct operationalization	○	○	○	○	○
Communication	○	○	○	○	○

Additional comments
Feel free to write any additional comments that you may have regarding this study

Fig. 3 InfoQ evaluation form for an empirical study on air quality. The complete form with information and additional studies for evaluation are available at `goo.gl/erNPF`

5 Conclusions and Future Directions

Our discussion of InfoQ as a crucial component in the empirical analysis framework calls for further discussion and research in various directions. In assessing InfoQ, we proposed a rating-based approach. We describe two initial studies that attempt to gauge the effectiveness of using the InfoQ framework for developing an analysis plan and for evaluating an empirical study. Future research is needed on specific implementations such as investigating the reliability of ratings across raters.

Although we discussed several dimensions of InfoQ and their relation to other quality concepts, there exist others that might be considered and new dimensions might evolve over time. For example, in todays environment, an important aspect of InfoQ is related to data privacy and confidentiality. In some areas, InfoQ could include a measure of risk in terms of confidentiality or human subjects. We also note the relationship with Big Data dimensions: the first five InfoQ dimensions relate to the three V's of Big Data [20]: *Volume* (data resolution; data structure), *Velocity* (temporal relevance; chronology of data and goal), *Variety* (data structure; data integration). Other proposed V's (value, veracity, viscosity, virality) can also be related to InfoQ dimensions, depending on their definitions.

Our studies have indicated the usefulness of considering InfoQ and its eight dimensions for evaluating prospective and retrospective studies. The InfoQ framework helps formalize and streamline the informal process that an experienced data analyst goes through. Future work will focus on streamlining rating-based and other InfoQ evaluation methods.

Acknowledgements We thank Professors Joel Greenhouse (Carnegie Mellon University), Shirley Coleman (Newcastle University), and Irena Ograjenek (University of Ljubljana) for their support of integrating InfoQ into graduate courses at CMU and University of Ljubljana, and helping assess its impact.

References

1. CM Angst, R Agarwal, and J Kuruzovich. Bid or buy? individual shopping traits as pre-dictors of strategic exit in on-line auctions. *International Journal of Electronic Com-merce*, 13:59–84, 2008.
2. R Bapna, P Goes, A Gupta, and Y Jin. User heterogeneity and its impact on electronic auction market design: An empirical exploration. *MIS Quarterly*, 28, 2004.
3. R Bapna, W Jank, and G Shmueli. Price formation and its dynamics in online auctions. *Decision Support Systems*, 44:641–656, 2008.
4. MR Berthold, C Borgelt, F Hoppner, and Klawonn F. *Guide To Intelligent Data Analysis*. Springer, 2010.
5. S Borle, P Boatwright, and JB Kadane. The timing of bid placement and extent of multiple bidding: An empirical investigation using ebay online auctions. *Statistical Science*, 21:194–205, 2006.
6. WE Deming. On the distinction between enumerative and analytic studies. *Journal of the American Statistical Association*, 48:244–255, 1953.

7. S Figini, RS Kenett, and Salini S. Integrating operational and financial risk assessments. *Quality and Reliability Engineering International*, 26, 2010.
8. R Ghani and H Simmons. Predicting the end-price of online auctions. Pisa, Italy, 2004.
9. E Giovanni. Understanding economic statistics. Technical report, 2008.
10. AB Godfrey. Eye on data quality. *Six Sigma Forum Magazine*, pages 5–6, 2008.
11. DJ Hand. *Statistics: A Very Short Introduction*. Oxford University Press, 2008.
12. W Jank and G Shmueli. *Modeling Online Auctions*. John Wiley & Sons, Hoboken, New Jersey, 2010.
13. R Katkar and DH Reiley. Public versus secret reserve prices in ebay auctions: Results from a pokémon field experiment. *Advances in Econc Analysis and Policy*, 6:Article 7, 2006.
14. RS Kenett, S Coleman, and I Ograjenek. On quality research: An application of infoq to the phd research process. In *Proceedings of the European Network for Business and Industrial Statistics (ENBIS) Tenth Annual Conference on Business and Industrial Statistics*, Antwerp, Belgium, September 2010.
15. RS Kenett and G Shmueli. On information quality. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, forthcoming, 2013.
16. RS Kenett and P Thyregod. Aspects of statistical consulting not taught by academia. *Statistica Neerlandica*, 60:396–412, 2006.
17. D Lucking-Reiley, D Bryan, N Prasad, and D Reeves. Pennies from ebay: The determinants of price in online auctions. *Journal of Industrial Economics*, 55:223–233, 2007.
18. C Mallows. The zeroth problem. *The American Statistician*, 52:1–9, 1998.
19. GL Patzer. *Using Secondary Data in Marketing Research*. Praeger, 2005.
20. P Russom. Big data analytics. Technical report, Q4, 2011.
21. G Shmueli. To explain or to predict? *Statistical Science*, 25:289–310, 2010.
22. G Shmueli and OR Koppius. Predictive analytics in information systems research. *Management Information Systems Quarterly*, 35:553–572, 2011.
23. JW Tukey. *Exploratory Data Analysis*. Addison Wesley, 1977.