Springer Proceedings in Complexity

Lorna Uden
Leon S. L. Wang
Tzung-Pei Hong
Hsin-Chang Yang
I-Hsien Ting *Editors*

# The 3rd International Workshop on Intelligent Data Analysis and Management

Springer

# Springer Proceedings in Complexity

For further volumes:
http://www.springer.com/series/11637

Lorna Uden · Leon S. L. Wang
Tzung-Pei Hong · Hsin-Chang Yang
I-Hsien Ting
Editors

# The 3rd International Workshop on Intelligent Data Analysis and Management

Springer

*Editors*
Lorna Uden
School of Computing
Staffordshire University
Stafford
UK

Leon S. L. Wang
College of Management
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

Tzung-Pei Hong
Department of Computer Science and Information Engineering
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

Hsin-Chang Yang
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

I-Hsien Ting
Department of Information Management
National University of Kaohsiung
Kaohsiung
Taiwan, R.O.C.

ISSN 2213-8684 ISSN 2213-8692 (electronic)
ISBN 978-94-007-7292-2 ISBN 978-94-007-7293-9 (eBook)
DOI 10.1007/978-94-007-7293-9
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013944207

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

Data analysis is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver, and enhance the value of data and information assets. Data analysis and data management both have multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. Intelligent Data Analysis and Management (IDAM) examines issues related to the research and applications of Artificial Intelligence techniques in data analysis and management across a variety of disciplines. It is an interdisciplinary research field involving academic researchers in information technologies, computer science, public policy, bioinformatics, medical informatics, and social and behavior studies, etc. The techniques studied include (but are not limited to): Data visualization, data pre-processing, data engineering, database mining techniques, tools and applications, evolutionary algorithms, machine learning, neural nets, fuzzy logic, statistical pattern recognition, knowledge filtering, and post-processing, etc.

On June 8, 2012 the IDAM was first held in College of Management, National University of Kaohsiung and the second IDAM was held on May 17, 2013. The third IDAM gathers people from previously disparate communities to provide a stimulating forum for exchange of ideas and results. We invite academic researchers (in information technologies, computer science, business and organizational studies, and social studies), as well as information technology companies, industry consultants, and practitioners in the fields involved.

The IDAM proceedings consist of 15 papers covering different aspects of Intelligent Data Analysis and Management. Authors of the papers come from many different countries such as Australia, India, Korea, Singapore, and Taiwan.

We would like to thank our authors, reviewers, and program committee for their contributions and the National University of Kaohsiung for hosting the conference.

Without their efforts, there would be no conference or proceedings.

Kaohsiung, Taiwan, September 2013

Lorna Uden
Leon S. L. Wang
Tzung-Pei Hong
Hsin-Chang Yang
I-Hsien Ting

# Organizations

## Conference Chair

Prof. Lorna Uden—Staffordshire University, UK

## Program Chairs

Prof. Leon S. L. Wang—National University of Kaohsiung, Taiwan
Prof. Tzung-Pei Hong—National University of Kaohsiung, Taiwan

## Local Chairs

Prof. Hsin-Chang Yang—National University of Kaohsiung, Taiwan
Prof. I-Hsien Ting—National University of Kaohsiung, Taiwan

## Organization Committee

Prof. Chian-Hsueng Chao—National University of Kaohsiung, Taiwan
Prof. Han-Wei Hsiao—National University of Kaohsiung, Taiwan
Prof. Ying-Feng Kuo—National University of Kaohsiung, Taiwan
Prof. Hsing-Tzu Lin—National University of Kaohsiung, Taiwan
Prof. Yu-Hui Tao—National University of Kaohsiung, Taiwan
Prof. Kai Wang—National University of Kaohsiung, Taiwan
Prof. Chen-Hsing Wu—National University of Kaohsiung, Taiwan
Prof. Shu-Cheng Yang—National University of Kaohsiung, Taiwan
Ms. Ming-Jun Chen—National University of Kaohsiung, Taiwan

## Program Committee

Prof. Ajith Abraham—Machine Intelligence Research Labs, USA
Prof. Damminda Alahakoon—Monash University, Australia
Prof. Chien-Chung Chan—University of Akron, USA
Prof. Bo-Rong Chang—National University of Kaohsiung, Taiwan
Prof. Ping-Tsai Chung—Long Island University, USA
Prof. Seng-Cho Chou—National Taiwan University, Taiwan
Prof. Saman Halgamuge—University of Melbourne, Australia
Prof. Liang-Cheng James Huang—Academia Sinica, Taiwan
Prof. Han-Wei Hsiao—National University of Kaohsiung, Taiwan
Prof. Hui-Yin Hsu—New York Institute of Technology, USA
Prof. Yada Katsutoshi—Kansei University, Japan
Prof. Tsau-Young Lin—San Jose State University, USA
Prof. Wen-Yang Lin—National University of Kaohsiung, Taiwan
Prof. Pawan Lingras—Saint Marys University, Canada
Prof. Victor Lu—St. John's University, USA
Prof. Javier Bajo Perez—Polytechnic University of Salamanca, Spain
Prof. James Tan—SIM University, Singapore
Prof. Ted Teng—Stony Brook University, USA
Prof. Shusaku Tsumoto—Shimane University, Japan
Prof. Da-Jin Wang—Montclair State University, USA
Prof. Chen-Shu Wang—National Taipei University of Technology, Taiwan
Prof. Shiang-Kwei Wang—New York Institute of Technology, USA
Prof. Jierui (Jerry) Xie—Oracle, USA
Prof. Hsin-Chang Yang—National University of Kaohsiung, Taiwan
Prof. Qiangfu Zhao—University of Aizu, Japan
Prof. Haibin Zhu—Nipissing University, Canada
Prof. Mihai Horia Zaharia—Gheorghe Asachi Technical University, Romania

# Contents

# Chapter 1
# 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 Ref. [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.

G. Shmueli (✉)
Srini Raju Centre for IT and the Networked Economy,
Indian School of Business, Hyderabad, 500032India
e-mail: galit_shmueli@isb.edu

G. Shmueli
Rigsum Institute of IT and Management, Thimphu, Bhutan

R. Kenett
KPA Ltd., Raanana Israel Dept of Statistics & Applied Mathematics,
University of Torino, Turin, Italy

R. Kenett
Center for Finance and Risk Engineering, NYU-Poly, Brooklyn, NY 11201, USA

L. Uden et al. (eds.), *The 3rd International Workshop on Intelligent Data Analysis and Management*, Springer Proceedings in Complexity,
DOI: 10.1007/978-94-007-7293-9_1,

**Keywords** Data analytics · Data mining · Statistical modeling · Study goal · Empirical study evaluation quality

## 1.1 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 Fig. 1.1.

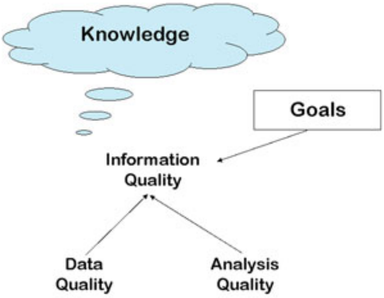

**Fig. 1.1** InfoQ depends on data quality and analysis quality, conditional on the goal at hand

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: Sect. 1.2 introduces InfoQ, its components, terminology and formal definition. In Sect. 1.3 we describe eight dimensions of InfoQ that are useful for assessing InfoQ in practice. Section 1.4 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 Sect. 1.5.

## 1.2 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.

### *1.2.1 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 predicting the output value for new observations given a set of input variables. Descriptive 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.

### *1.2.2 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:

$$InfoQ(f, X, g) = U(f(X \mid g)) \tag{1.1}$$

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 $U$.

### *1.2.3 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 versus 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.