Editorial: In Search of Novel Ideas and Solutions with a Broader Context of Data Quality in Mind

Welcome to the latest installment of the ACM JDIQ. In this issue, four well prepared research articles are showcased. The authors in this issue question the impact of input data on the output data, such as (1) results of analysis, (2) decisions, (3) expected performance in search, and (4) in security.

These articles provide new theoretical advancement and solutions, keeping data quality perspective in mind. We hope that the insights gleaned from the four articles can be adapted to provide extended next steps for deepening and advancing theories by researchers and crafting useful ways for advancing solutions for data and information problems by practitioners.

The questions and findings raised in this issue set the groundwork for future research to investigate more complex problems encompassing the entire information manufacturing process, including collection, storage, and use environment. Data quality is not experienced in a vacuum and the same set of data can be evaluated differently based on evaluation contexts and broader environment, such as when it was used, by whom, and for what purpose.

We believe that these articles have risen to the challenge of understanding data quality in a broader context while still recognizing the data stored in a database system.

Articles 8 and 9 question the association between input data and output data, such as that given by analysis results. Article 10 provides an updated solution for reconciling ambiguous names. Article 11 cautions the use of biometrics-based authentification.

Article 8

In "The Effects and Interactions of Data Quality and Problem Complexity on Classification", Roger H. Blake and Paul Mangiameli explore the impact of data quality and problem complexity on the outcome of classification. The authors treat data quality as an independent variable and classification as a dependent variable. They use the generated dataset and ANOVAs for hypothesis testing. The authors then apply their findings to 1,000 customer datapoints from a telecommunications company for predicting customer retention with different categories. Both the generated data and the example application produced similar results to support their general hypothesis that poor data quality and problem complexity negatively impact classification results.

Researchers are encouraged to extend the results of this study to evaluate additional areas of decision-making and data-mining issues. Researchers can also apply the results to study additional models and applications of data quality for other tasks, such as economic, cognitive, and social aspects of using data and information. For practitioners, the results of this study can be applied to develop a pragmatic method in deciding their own threshold for efforts in quality of data and problem complexity towards producing a desired outcome, such as customer retention for their marketing campaign in various organizations and industries.

Article 8 has been revised four times by the authors since the submission of the original manuscript. The earlier version was presented at the International Conference on Information Quality (ICIQ) and received The Madnick Best Paper Award in 2008.

Article 9

"GIGO or not GIGO: The Accuracy of Multi-Criteria Satisficing Decisions" by Irit Askira Gelman questions the commonly believed notion that poor-quality input data produces poor-quality output data, which is dubbed colloquially as Garbage In Garbage Out (GIGO). She posits a more varied association between input data and output data (information) based on her investigation on the effect of input errors on the outcomes of disjunctive and conjunctive decisions. She specifically explains the negative association between the two, employing mathematical modeling.

This article is important to researchers who are interested in gaining more insight into associations between input data and the output data (information) errors. Specifically, researchers can continue the research and validate the results through additional simulations and empirical work. Practitioners can pay attention to various mechanisms in the data environment where input data can be influenced and impacted for output data errors.

This article has been revised three times by the authors since the submission of the original manuscript.

Article 10

"On Graph-Based Name Disambiguation" by Xiaoming Fan, Jianyong Wang, Xu Pu, Lizhu Zhou, and Bing Lv provides a graph-based model, called "GHOST," to help solve the name disambiguation problem. The solution, which uses a 5-step approach, is demonstrated using publication records with same or similar names as an example.

Researchers who are interested in various techniques and aspects of semantic reconciliation, such as entity resolution, record linkage, duplicate detection, and reference reconciliation and object matching are encouraged to extend this study by adapting this approach to accommodate evolving entities, attributes, and the adaptive techniques in different problem areas. Practitioners are encouraged to experiment with the approach and to resolve ambiguities in semantics used in critical business areas.

This article has been revised five times by the authors since the submission of the original manuscript.

Article 11

In "Typing Biometrics: Impact of Human Learning on Performance Quality", Benjamin Ngugi, Beverly Kahn, and Marilyn Tremaine investigate human-typing behaviors to determine whether or not behavioral biometrics can be an effective way to provide authentification. The authors found that there are critical data quality issues that need to be addressed in order for behavioral biometrics to be a sufficient solution. They found that learning and typing key placement has a significant impact on the quality of authentication performance.

Researchers can expand on this study by examining biometric template updating and the choice of uncorrelated PIN combinations. Practitioners can also benefit by gaining insight into how behavioral biometrics can be used and applied in their organization and industry.

The article has been revised twice by the authors since the submission of the original manuscript.

We continue to welcome new and innovative research papers in a broader context of data and information quality. We look forward to publishing articles that ask new and exciting questions. We also welcome research that applies theories from diverse disciplinary areas, and we do not discriminate against any particular research method.

We publish research work that uses both quantitative and qualitative research methods.

Enjoy the read, and we look forward to providing future articles with breakthrough ideas and useful insights.

—Stuart E. Madnick and Yang W. Lee
Editors-in-Chief
http://jdiq.acm.org