INTEGRATING INFORMATION QUALITY IN VISUAL ANALYTICS

A Dissertation Submitted
to the Graduate School
University of Arkansas at Little Rock

in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in Integrated Computing

in the Department of Information Science
of the Donaghey College of Engineering and Information Technology

December 2011

Ahmed Abu Halimeh

B.Sc. Applied Science University, 2004
M.S. University of Arkansas Little Rock, 2008

This dissertation, “Integrating Information Quality in Visual Analytics”, by Ahmed Abu Halimeh, is approved by:

Dissertation Advisor: ______________________________

Mihail Tudoreanu
Associate Professor of Information Science

Dissertation Committee: ______________________________

John Talburt
Professor of Information Science

____

Elizabeth Pierce
Associate Professor of Information Science

____

Mariofanna Milanova
Professor of Computer Science

____

Ningning Wu
Professor of Information Science

Program coordinator: ______________________________

Kenji Yoshigoe
Associate Professor of Computer Science

Graduate Dean: ______________________________

Patrick J.Pellicane
Professor of Construction Management

Fair Use

Duplication

INTEGRATING INFORMATION QUALITY IN VISUAL ANALYTICS

by Ahmed Abu Halimeh, December 2011

ABSTRACT

Visualization and visual analytics rely heavily on rapid exploration, and a combination of various information dimensions and sources performed by an analyst. Integrating Information Quality involves combining Information Quality (IQ) residing in different sources and providing users with a unified view of this data quality. This process becomes significant in a variety of situations both commercial and scientific. Integrating information appears with increasing frequency as the volume and the need to share existing information grows. This research focus on a subset of IQ dimensions, which we term subjective IQ (SIQ). Visual analytics techniques rely on fast, continuous, and interactive user exploration and multiple changes the information being displayed, but they currently cannot convey Information Quality because no techniques exist for continuously updating the quality of the current information in a short period of time. Current approaches fail to provide awareness with regard to the quality and contribution of each of these sources to the combined information shown in the visualization. This research focuses on assessing the Subjective Information Quality dimensions in order to facilitate the integration of those metrics in visualizations, so that users can take the most of the advantage of visual analytics. It also focuses on the quality of combined information and how the combined Information Quality should be conveyed and presented to the users. The main deliverable of this research is to understand and establish rules/principles about how quality of information can be determined /assessed through visualizations.

ACKNOWLEDGEMENTS

I would like to acknowledge the advice and guidance of Dr. M.Eduard.Tudoreanu, committee chairman. I also thank the members of my committee for their guidance and suggestions, especially Dr.John Talburt for all his advice, encouragement, and financial assistance.

Special thanks go to my brother Thayer Abuhalimeh, without his guidance and assistance I would not have been successful.

I would like to thank my parents, my wife “Duaa”, and family members, for supporting and encouraging me to pursue this degree. Without their encouragement, I would not have finished the degree.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Visual analytics is the integration of interactive visualization with analysis techniques to answer a growing range of questions in science, business, and intelligence. It can address certain problems that vary in size, complexity, and need for closely coupled human and machine analysis that would have otherwise been intractable.

Visual representations translate information into a visible form that highlights important features, including commonalities and anomalies. These visual representations make it easy for users to perceive clear aspects of their information. One of these information aspects is Information Quality, or “the fitness for use of the information provided”. Providing an appropriate representation is the focus of visual analytics; it integrates new techniques and visual representations to facilitate human-information discourse and improve understanding of Information Quality.

Visualization and visual analytics rely heavily on rapid exploration and a combination of various information dimensions and sources performed by an analyst. Integrating Information Quality involves combining Information Quality residing in different sources and providing users with a unified view of this data quality. This process becomes significant in a variety of situations both commercial and scientific. Integrating information appears with increasing frequency as the volume and the need to share existing information grows.

Information Quality is not a simple scalar measure, but can be defined on multiple dimensions, with each dimension yielding different meanings to different information consumers and processes. Each dimension can be measured and assessed differently. Information Quality assessment implies providing a value for each dimension about how much of the dimension or quality feature is achieved in order to enable adequate understanding and management.

This research focuses on a subset of IQ dimensions, which we term *subjective IQ* (SIQ). These dimensions typically require a user's opinion and do not have a clear mathematical technique for finding their value. Note that most dimensions can be measured through multiple techniques, but the SIQ ones are most useful when the user's experience, opinion, or performance, are accounted for. The objective dimensions are the dimensions that can be largely and commonly assessed via mathematical or functional forms. It is possible for the same dimension to be measured by both subjective and objective means, depending on the context. For example, accuracy can be objectively computed when one checks the correctness of some data against the balance of a bank account. However, accuracy can also be subjectively employed, for example, the estimation of the amount of shade in a specific area.

The following table shows all IQ dimensions and the relative importance or usefulness of objective metrics versus subjective ones (L. Pipino, W. Lee and Y. Wang [10]).

Dimension	**Objective Assessment (width of gray area proportional to relevance of objective metrics)**	**Subjective Assessment (width of white area shows relevance for subjective)**
Believability (SIQ)	Apply a formula (integration only)	User’s opinion and experience determine~~s~~ whether they trust the data
Ease of Manipulation	Time to perform a computation	User’s experience or performance with the data
Interpretability	Whether some computation is successful	User can understand the data correctly
Relevancy	Can produce a valid result	Helps the user in their task
Reputation	Apply a formula (integration only)	The user can judge or assume the accuracy based on the result of the objective assessment
Value-Added	Can increase the value of data	The user can judge or assess the value added to the data
Timeliness	Can reflect how up-to-date the data is with respect to the task it is used for	User judges based on previous experience

Understandability	Can provide clear and simple data	User can understand the data easily
Concise Representation	The shortest representation is known; duplicates are counted	User judges based on previous experience
Appropriate Amount of Data	The needed amount is known	User expertise is required
Security	Against a standard metric.	Users experience or performance with the data
Accessibility	Against a standard metric	Based on user's experiences

Table 1: IQ dimensions and the relative usefulness of objective metrics versus subjective ones

This research focuses on the quality of information obtained from multiple sources and on how the combined Information Quality should be computed. The purpose of this research is to understand and establish rules/principles to enable users to estimate or determine the quality of the visual information. The rules can subsequently be used to estimate these quality dimensions, which can be presented to the users in near real time to help them perceive the subjective quality of the combined information and the major components of that data. As a secondary concern, this work will provide simple means of presenting the Subjective Information Quality dimensions through visualizations, so that users can better take advantage of visual analytics.

1.2 Current Information Quality Framework

In the current Information Quality framework, data is being captured from various sources and stored in a database. An Information Quality specialist performs an assessment of the quality of each data source. End users can rely on that assessment when using data from a single source, but there are no known methods, especially for SIQ, to assess the quality of information that needs to be integrated data from multiple databases.

Our approach is to investigate rules/principles to enable an IQ specialist to derive formulas to estimate the value of SIQ for integrated data sources. These formulas will be customized by the IQ specialist to their particular data source, and could be derived by sampling a small number of integration cases. The end-users will benefit from the SIQ estimation in near real time to help them perceive the subjective quality of the combined information and the major components of that data.

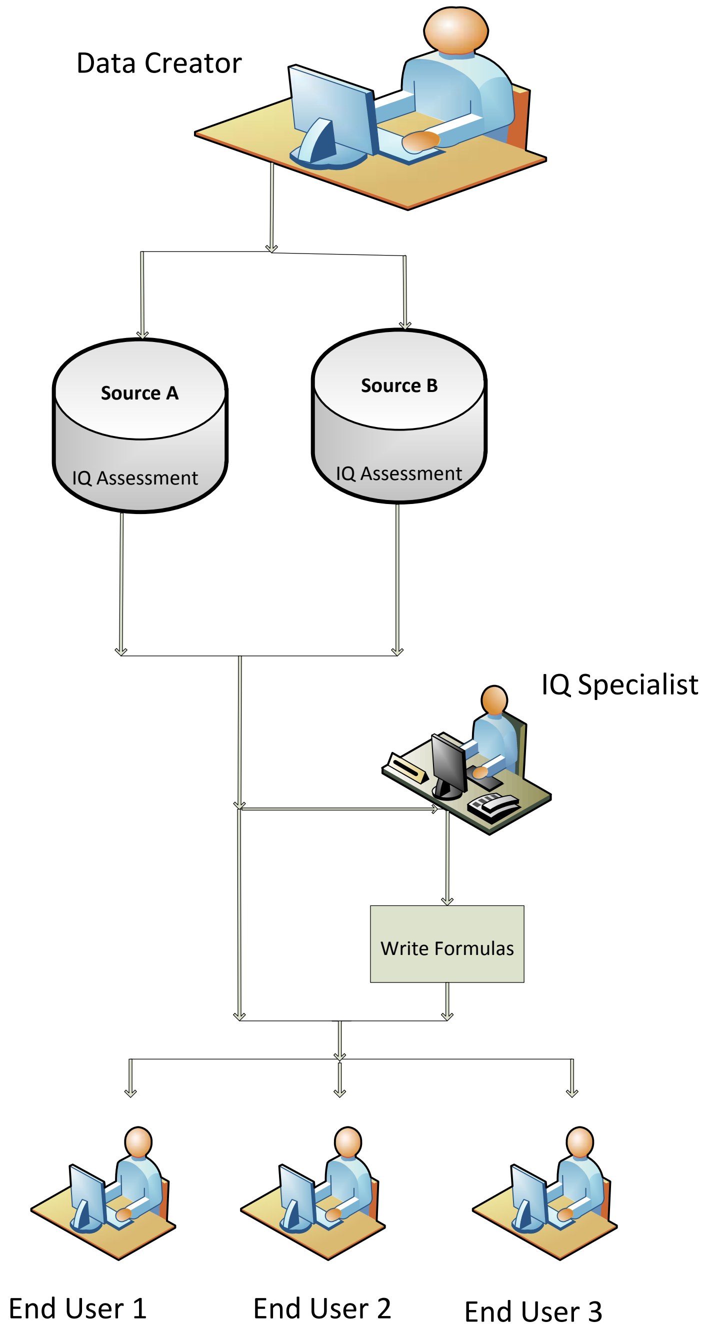


Figure 1: Proposed IQ Framework with the specialist deriving SIQ integration formulas

1.3 Problem

Visual analytics techniques rely on fast, continuous, and interactive user exploration and multiple changes the information being displayed, but they currently cannot convey

SIQ because no techniques exist for continuously updating the quality of the current information in a short period of time. Current approaches, such as those proposed by Tekusová, Knuth, Schreck and Kohlhammer [5], Pang [6], or MacEachren [7], fail to provide awareness with regard to the quality and contribution of each of these sources to the combined information shown in the visualization. Combining information from different information sets using a mixture of several different operations is crucial for visual analytics, but may also lead to several Information Quality problems, such as duplication, inconsistency and missing values. Other subjective Information Quality issues such as believability, value-added, and trust measurements are also difficult and slow to estimate. The propagation of the quality is not a simple additive model; it is entirely possible for two pieces of information with high-quality ratings to result in low quality information, if, for example, the two contradict each other. The opposite is also true, in that two pieces of low quality data can complement each other with a net boost in quality. Current integration techniques treat all IQ dimensions as mathematical entities and combine them through statistical methods (L. Pipino, W. Lee, and Y. Wang [10] Lin and Hua [11], Caballero, Verbo, Calero and Piattini [13] Peralta, Ruggia, Kedad and Bouzeghoub [14], Motro and Rakov [15], P. Ballou, InduShobha N., Chengalur-Smit and Y.Wang [16]). However, most people do not treat information in a statistical manner and may find value in seemingly low quality information or fault in information that is statistically of high quality. This becomes more important for the SIQ dimensions, in which the user's opinion is the most important measure. There are no methods for determining/estimating SIQ from multiple sources that have been verified or derived from actual human observations. As such, data quality of subjective dimensions is not appropriately conveyed for information obtained through a combination of multiple information sources. This holds true even if the quality of the individual sources is known.

Current techniques for determining Subjective Information Quality (SIQ) are time consuming and effort-intensive because they require surveys, sampling, and statistical analysis. Analyzing and assessing the quality takes a long time in the order of days, if not weeks. Such methods are not appropriate for visual analytics, where data combination and refinement must take place in a short period of time.

Another problem resides in a lack of visualization tools that present the quality of the information together with the information itself, and thus the full fitness of the information for any particular user is not realized. Some properties of the information may become apparent through visual examination, without the need to perform extensive computations in the background. Using visualizations would be the best technique to display the growth of the information along with its quality, and also show how the different SIQ dimensions would affect each other through the information propagation process.

CHAPTER 2

RELATED WORK

2.1Visual analytics and visualization aspects related work

- **Georgia Tech – Information Visualization & Visual Analytics**

Georgia Tech [1] is the NSF/DHS FODAVA-Lead, where FODAVA stands for Foundations of Visual Analytics. Their research goal is to develop new, interactive visualization techniques and systems that provide multiple and flexible perspectives on the data being examined. The data being examined may range from quantitative business information, stored in spreadsheets and databases, to textual documents such as news reports and articles. Often, the data is a heterogeneous collection of items drawn from different sources. Common to all these types of data is our user's need to draw information out that is hidden. What is the right course of action? Which option should the user choose? What is the best way to accomplish a goal? The answer is buried somewhere, and users just need a way to look for it. These systems help people and organizations to browse, explore and analyze data that is important to them. Fundamentally, these interactive visualizations are tools for sense-making; they assist users in understanding data by presenting it in a form that can be organized, queried and explored in order to gain new perspectives and insights about it.

- **University of Maryland HCI Lab Visualization Projects**

SEMVAST (Scientific Evaluation Methods for Visual Analytics Science and Technology) SEMVAST project/contest [2], aims to improve the evaluation of visual analytics technology. Their project has focused on two activities: 1) making benchmark data sets available, and 2) seeding an infrastructure for evaluation. They developed automated metrics for some aspects of Visual Analytics systems, and guidelines will help researchers/aspects assess the subjective aspects of the visual analytics environment. Their research focuses on the Accuracy of the Benchmark.

- **Developing Qualitative Metrics for Visual Analytic Environments**

Scholtz project [3], examined reviews for the entries to the 2009 Visual Analytics Science and Technology (VAST) Symposium Challenge. By analyzing these reviews, the authors gained a better understanding of what is important to the reviewer's visualization researchers and professional analysts. This is a bottom-up approach to the development of heuristics to use in the evaluation of visual analytic environments. Their project focus on the usefulness, efficiency, and intuitiveness of the visualizations presented to the participant; it evaluates the visualizations in the contest.

- **Data Quality Visualization for Multivariate Hierarchic Data**

Tekusova, Knuth, Schreck and Kolmar research [5], presents a brief survey of currently available uncertainty visualization techniques. Then, the authors present experimental results obtained from use of several techniques for visualization of multidimensional data quality information, applied to multivariate hierarchical data used in an economic data analysis scenario. Methods for visualizing error and uncertainty are presented in several surveys. Available techniques include:

- Usage of free graphical variables, including color, size, saturation of color, position, angle, clarity, fuzziness, transparency, edge crispness.
- Integration of additional graphical objects, such as uncertainty glyphs, labels, isosurfaces, textures.

- Usage of animation: speed, duration, blinking, motion blur.
- Interactive representation: clickable maps, difference images, mouse-over effects, magic lenses.
- Addressing other human senses: acoustic or haptic senses (e.g. sound or vibration).
- A user study of the methods for spatial data indicates that blinking, adjacency, and overlay are among the most useful techniques. At the same time, animation and saturation of color were deemed least useful. Interestingly, none of these techniques supports a combination of qualitative and quantitative uncertainty information.

They also implemented a prototype for testing various techniques to map data certainty attributes to visualization for hierarchically structured data. However, they only focus one information quality dimension, namely “uncertainty”.

- **Visualizing Uncertainty in Geo-spatial Data**

Pang [6], focuses on how computer graphics and visualization can help users access and understand the increasing volume of geospatial data. In particular, his research highlights some of the visualization challenges in visualizing uncertainty associated with geo-spatial data. Uncertainty comes in a variety of forms and representations, and requires different techniques for presentation together with the underlying data. In general, treating the uncertainty values as additional variables of a multivariate dataset is not always the best approach. His research presented some possible approaches and further challenges using two illustrative application domains. As the previous work, that the focus is only on the dimension of uncertainty, while our work is to include more IQ dimensions

- **Visualizing Uncertain Information**

MacEachren research [7], addresses the difference between data quality and uncertainty, the application of Bertin's graphic variables to the representation of uncertainty, conceptual models of spatial uncertainty as they relate to kinds of cartographic symbolization, and categories of user interfaces suited to presenting data and uncertainty about that data. Also touched on is the issue of how we might evaluate our attempts to depict uncertain information on maps. Again, they only show one quality dimension that of “uncertainty”. Our research focuses on subjective IQ dimensions and provides visual representation for the process of propagating quality; in order to asses not only the final result of algorithms, but also to convey the information quality to the user.

- **Stanford Entity Resolution Framework**

The goal of the SERF project [8] is to develop a generic infrastructure for Entity Resolution (ER). ER (also known as reduplication or record linkage) is an important information integration problem. For instance, two records on the same person may provide different name spellings, and addresses may differ. The goal of ER is to "resolve" entities, by identifying the records that represent the same entity and reconciling them to obtain one record per entity. Our research focuses on a full view of the IQ dimensions including the subjective dimensions.

- **Stanford Trio Project**

Trio [9], is a new kind of database management system (DBMS), one in which data dimensions, uncertainty of the data, and data lineage are all first-class citizens. Combining data, uncertainty, and lineage yields a data management platform that is useful for data integration, data cleaning, and information extraction systems. They also look at data combination through regular SQL operations, and are more interested in objective IQ, not SIQ.

Our research focuses on the quality of the information being presented to the user through the visualizations. Our research may benefit in the future from the benchmarks in the way in which empirical evaluation is performed in SEMVAST [2], but one limitation of those benchmarks is that they do not cover Information Quality, especially SIQ.

Our research aims to develop principles to help people understand and assess the quality of the data showed to them through visualization, in addition to the broad goals of FODAVA stated above. The focus is on the quality of combined data, how the combined data quality should be conveyed and presented to the users, and ultimately how people perceive subjective dimensions, such as believability, value-added and accuracy, our research will provide visual representation of the quality along with its quality.

2.2 Information Quality research in subjective assessment.

- **Data Quality Assessment**

Pipino, Lee, and Wang article [10], describes the subjective and objective assessments of data quality, and presents three functional forms for developing objective data quality metrics. The article presents approaches that combine the subjective and objective assessments of data quality, and illustrate how it has been used in practice. Data and information are often used synonymously. In practice, managers differentiate information from data intuitively, and describe information as data that has been processed. Their research does not provide visual techniques to present the subjective quality, and it takes time to determine the quality of data. Also, this does not provide the estimation option to the users. Their work focuses on the data coming from one data source.

- **A Method for measuring data quality in Data Integration**

Lin, and Hue research [11] reports a method for measuring data quality in data integration. The article focuses on believability, a major aspect of quality. The authors present an approach for computing believability based on metadata. In their method, they make explicit use of lineage-based measurements and develop a precise approach to measuring data quality. Believability is itself divided into sub-dimensions: believability of source, believability compared to internal common-sense standard, and believability based on temporality of data. They present metrics for assessing the believability of data resulting from one process run; the believability of a data value is computed based on the lineage of this value. The next steps of their research will concentrate on the refinement of the proposed metrics in conjunction with further testing on real case studies, and the development of a tool to capture metadata. Their research only focuses on the believability, and does not include the human interactions with quality. Believability is one of the subjective data quality dimensions that we discussed in this research.

- **Beyond accuracy; what data quality means to data consumers**

Wang and Strong, [12] developed a framework that captures the aspects of data quality that are important to data consumers. A two-stage survey and a two-phase sorting study were conducted to develop a hierarchical framework for organizing data quality dimensions. Their framework captures dimensions of data quality that are important to data consumers. Intrinsic DQ denotes that data have quality in their own right. Contextual DQ highlights the requirement that data quality must be considered within the context of the task at hand. Representational DQ and accessibility DQ emphasize the importance of the role of systems. These findings are consistent with our understanding that high-quality data should be intrinsically good, contextually appropriate for the task, clearly represented, and accessible to the data consumer.

- **A data quality measurement information model**

Caballero, Verbo, Calero, and Piattini [13] propose a Data Quality Measurement Information Model (DQMIM), which provides a standardization of the referred terms by following ISO/IEC 15939 as a basis. They deal with the concepts implied in the measurement process, and not with the measures themselves. In order to make operative the DQMIM, we have also designed a XML schema, which can be used to outline Data Quality Measurement Plans.

- **A Framework for Data Quality Evaluation in a Data Integration System**

Peralta, Ruggia, Kedad, and Bouzeghoub project [14], addresses the problem of data quality evaluation in data integration systems. They present a framework, which is a first attempt to formalize the evaluation of data quality. It is based on a graph model of the data integration system, which allows them to define evaluation methods as graph properties. In their project they only focus on data freshness and currency dimension.

- **Estimating the Quality of Data in Relational Databases**

Amihai and Igor [15] propose a standard for rating information sources with respect to their quality. An important consideration is that the quality of information sources often varies considerably when specific areas within these sources are considered. They describe an approach that uses dual quality measures to gauge the distance of the information in a database from the truth, and propose to combine manual verification with statistical methods to arrive at useful estimates of the quality of databases. Amihai and Igor show how to derive quality estimates for individual queries from such quality specifications. In their project they focus only on two dimensions: accuracy and completeness.

- **Sample-Based Quality Estimation of Query Results in Relational Database Environments**

The approach in [16] provides a basis for the systematic analysis of the quality of information products (IPs). Their research uses the relational algebra framework to develop estimates for the quality of query results based on the quality estimates of samples taken from the base tables. Their procedure requires an initial sample from the base tables; these samples are then used for all possible information IPs. Each specific query governs the quality assessment of the relevant samples. By using the same sample repeatedly, their approach is relatively cost effective. They examine various, relevant statistical issues, including how to deal with the impact on quality of missing rows in base tables. They do not discuss any data quality dimensions; instead they measure the quality based on a specified condition, whether acceptable or not acceptable. The work assumes the quality of the data is not known.

- **Examining Data Quality**

Tyi, and Ballou research [17] discusses the term data quality, or “fitness for use,” which implies the concept of data quality is relative. Thus, data with quality considered appropriate for one use may not possess sufficient quality for another use. The trend toward multiple uses of data, exemplified by the popularity of data warehouses, has highlighted the need to address data quality concerns. Furthermore, fitness for use implies that one needs to look beyond traditional concerns with the accuracy of the data. Data found in accounting-type systems may be accurate but unfit for use if that data is not sufficiently timely. Also, personnel databases situated in different divisions of a company may be correct but unfit for use if the desire is to combine the two and they have incompatible formats. A related problem with multiple users of data is that of semantics. The data gatherer and initial user may be fully aware of the nuances regarding the meaning of the various data items, but that will not be true for all of the other users. Thus, although the value may be correct, it can easily be misinterpreted. Also, the capability of judging the reasonableness of the data is lost when users have no responsibility for the data's integrity and when they are removed from the data creators. Such problems are becoming increasingly critical as organizations implement data warehouses

- **Anchoring Data Quality Dimensions Ontological Foundations**

Yair, and Wang's project [4] addresses the impacts of poor data quality on the overall effectiveness of organizations. In a world where people are moving to total quality management, one of the critical areas is data. The quality of a product depends on the process by which the product is designed and produced. Likewise, the quality of data depends on the design and production processes involved in generating the data. To design for better quality, it is necessary first to understand what quality means and how it is measured. Data quality, as presented in their work, is a multidimensional concept. Frequently mentioned dimensions are accuracy, completeness, consistency, and timeliness. The choice of these dimensions is primarily based on intuitive understanding, industrial experience, or literature review. However, a literature review shows that there is no general agreement on data quality dimensions.

- **Measuring Data Believability: A Provenance Approach**

Nicolas.P and Madnick.S.E [21] presents the main concepts of a model for representing and storing data provenance, which includes ontology of the sub-dimensions of data believability. They use aggregation operators to compute believability across the sub-dimensions of data believability. Our work focus on subjective evaluation of believability for visual presentation and aims to help in developing a method that will enable users to better estimate the quality of the data coming from different sources. Our results may be better at determining SIQ measures than the statistical methods employed for their data provenance model.

- **The Credibility of Online Information**

The research of Huerta, Esperanza and Ryan [4] examines the factors affecting the credibility of online information. It uses the Elaboration Likelihood Model (Petty and Cacioppo [24]) as a theoretical framework, proposing a comprehensive model that includes factors from traditional means of communication and the Web. A field experiment was conducted that manipulated quality of content, reputation of the Web site owner, attractiveness, modality of exposure, and simulation. Out of these factors, quality of content and reputation of the Web site owner show statistical significance in the expected direction.

- **A Methodology for Information Quality Assessment**

Lee, Strong, Kahn, and Wang [23] developed a methodology, called AIM quality (AIMQ), to form a basis for IQ assessment and benchmarking. The methodology was illustrated through its application to five major organizations. The methodology encompasses a model of IQ, a questionnaire to measure IQ, and analysis techniques for interpreting the IQ measures. They developed and validated the questionnaire and used it to collect data on the status of organizational IQ. These pieces of data were used to assess and benchmark IQ for four quadrants of the model, which rely on questionnaires to find IQ scores. Our study uses different data from multiple sources, and we aim to help in developing a method that will enable users/organizations to better estimate the SIQ dimensions of the data coming from different sources.

- **A homogeneous framework to measure data quality**

Bobrowski, Marre and Yankelevich [25] presented a methodology to measure data quality within organizations. First, a list of IQ criteria must be set up. These IQ criteria are divided into directly and indirectly assessed criteria. Scores for the indirectly assessed IQ criteria are computed from the directly assessed IQ criteria. In order to assess the direct criteria, traditional software metrics techniques were applied. These techniques measured data quality following the goal-question-metric methodology: For each directly assessed criterion, a question is set up that characterizes the criterion, and then a metric is derived to answer that question, giving a precise evaluation of the quality. From these metrics a user questionnaire is set up, which is based on samples of the database.

- **The market for "LEMONS": Quality uncertainty and the market mechanism**

George Akerlof [46] studies the relationship between quality and uncertainty. His work includes interesting and important problems for the theory of markets. On the one hand, the interaction of quality differences and uncertainty may explain important institutions of the labor market.

- **Internet Data Quality: Perceptions of Graduate and Undergraduate Business Students**

Klein [54] examines user perceptions of the quality of information found on the Internet using surveys of graduate and undergraduate students taking specific courses. The Wang and Strong [12], framework was applied in the study as a tool for measuring data quality. The objective of the study is to further improve the understanding of users' evaluations of internet information quality by comparing perceptions of graduate and undergraduate students. The study was built on prior research aimed at understanding the dimensions of data quality. The study has several limitations. First, the sample size was relatively small. Second, all of the surveyed users were students taking specific courses. The findings of this study suggest that additional research in this area would be beneficial. Future research should focus on surveying a larger, less narrowly focused user population and on surveying users about their perceptions of the data quality of specific web sites.

- **User Perceptions of Data Quality: Internet and Traditional Text Sources**

Klein [55] examines user perceptions of the quality of information found on the Internet and in traditional text sources .This study uses a survey based on the Wang and Strong [12] framework. Her focus, as ours, was on the consumers (users) of data and information.

The objective of the study reported is to improve the understanding of users evaluations of Internet information quality. The study was built on prior research aimed at understanding the dimensions of data quality.

Several differences from our work and Klein's study are noted. First, the sample size was relatively small. Second, all of the surveyed users were students taking an MBA courses. Third, respondents were asked questions about the quality of Internet and traditional text sources in general rather than being asked questions about specific Internet sites and text sources. Forth, our work examines how users perceived data stemming from multiple sources.

Our research addresses the quality of combined visual data. We propose a set of principles of estimating data quality. The principles can subsequently be used to estimate these quality dimensions and present them to a user. Our studies researches various types data sources, and focus on how people perceive subjective dimensions such as believability, accuracy, and value-added. Our research assumes the data source is already assessed and the quality is known for the subjective IQ dimensions

2.3 Additional Contributing Resources

- **Signaling Theory**

Signaling theory was presented by Michael Spence in 1973 [29], [30], and [31], as a solution for Asymmetric information [32], [47], [50]. Spence used a basic job market model as an example to explain his theory; Spence assumes that generally, employers are willing to pay higher wages to employ better workers. While the individual may know his or her own level of ability, the hiring firm is not (usually) able to observe such an intangible trait thus there is an asymmetry of information between the two parties. Education credentials can be used as a signal to the firm, indicating a certain level of ability that the individual may possess; thereby narrowing the informational gap. This is beneficial to both parties as long as the signal indicates a desirable attribute .There are, number of problems that these parties would immediately run into.

- How much time, energy, or money should the sender (Information Source) spend on sending the signal (information)?
- How can the receiver trust the signal to be an honest declaration of information?
- Assuming there is a signaling equilibrium under which the sender signals honestly and the receiver trusts that information, under what circumstances will that equilibrium break down?

- **Information asymmetry**

Information asymmetry [32], [47], [50], deals with the study of decisions in transactions where one party has more or better information than the other. This creates an imbalance of power in transactions, which can sometimes cause the transactions to go away, a kind of market failure, in the worst case. An example of this problem is the adverse selection.

Adverse selection "*is a term used in economics, insurance, statistics, and risk management. It refers to a market process in which "bad" results occur when buyers and sellers have asymmetric information (i.e. access to different information): the "bad" products or services are more likely to be selected*". [49], [50], [52],

Our research will help to answer some of the introductory questions of the signaling theory. Our research will provide set of principles help filling the gaps in the problem of adverse selection [49], [52], presented by George Akerlof's "The Market for Lemons" [45], [46], which brought informational issues at the forefront of economic theory.

Information Asymmetry will fit perfectly in our research since it carries different information and different quality and our research addresses the quality of combined data from different resources.

- **Cognitive psychology**

Cognitive psychology [34], [35], "*the study of how people perceive, remember, think, speak, and solve problems*". (Wikipedia, [34], [35], [41], [42])

The term "cognition" refers to all processes by which the sensory input is transformed, reduced, elaborated, stored, recovered, and used. In other words the "cognition" is involved in everything a human being might do.

Cognitive bias is a pattern of poor judgment, often triggered by a particular situation. Identifying "poor judgment," or more precisely, a "deviation in judgment," requires a standard for comparison, i.e. "good judgment." In scientific investigations of cognitive bias, the source of "good judgment" is that of people outside the situation hypothesized to cause the poor judgment, or, if possible, a set of independently verifiable facts. The existence of most of the particular cognitive biases listed below has been verified empirically in psychology experiments.

"*Cognitive bias is a general term that is used to describe many distortions in the human mind that are difficult to eliminate and that lead to perceptual distortion, inaccurate judgment, or illogical interpretation*". (Wikipedia, [36], [37], [38]) .Many of these biases are studied for how they affect belief formation, business decisions, and scientific research. The main biases types that apply and illustrate our results are:

- **Anchoring**

Anchoring [38] is one of the cognitive biases that occurs when people tend to rely too heavily, on one piece of information when making decisions. During normal decision making. For example, as a person looks to buy a used car, he or she may focus excessively on the odometer reading and model year of the car, and use those criteria as a basis for evaluating the value of the car, rather than considering how well the engine or the transmission is maintained.

- **Attentional bias**

Attentional bias [38], [39], [40], many types of cognitive bias may occur due to an attentional bias. One example is when a person does not examine all possible outcomes when making a judgment about a correlation or association. They may focus on one or two possibilities, while ignoring the rest.

- **Confirmation bias (also called confirmatory bias or my side bias)**

Confirmation bias "*is a tendency for people to favor information that confirms their preconceptions or hypotheses regardless of whether the information is true. As a result, people gather evidence and recall information from memory selectively, and interpret it in a biased way. The biases appear, in particular, for emotionally significant issues and for established beliefs. For example, in reading about gun control, people usually prefer sources that affirm their existing attitudes. They also tend to interpret ambiguous evidence as supporting their existing position*". (Wikipedia, [38])

Confirmation biases contribute to overconfidence in personal beliefs and can maintain or strengthen beliefs in the face of contrary evidence. Hence, they can lead to poor decisions, especially in organizational, scientific, military, political and social contexts.

• Selective perception

Selective perception, "*it may refer to any number of cognitive biases in psychology related to the way expectations affect perception. For instance, several studies have shown that students who were told they were consuming alcoholic beverages (which, in fact, were non-alcoholic) perceived themselves as being "drunk," exhibited fewer physiological symptoms of social stress, and drove a simulated car similarly to other subjects who had actually consumed alcohol. The result is somewhat similar to the placebo effect*". (Wikipedia, [38], [43])

• Subjective validation

Subjective Validation [38], [43],[44], sometimes called personal validation effect is a cognitive bias by which a person will consider a statement or another piece of information to be correct if it has any personal meaning or significance to them. In other words, a person whose opinion is affected by subjective validation will perceive two unrelated events (i.e., a coincidence) to be related because their personal belief demands that they be related, subjective validation is an important element in cold reading. It is considered to be the main reason behind most reports of paranormal phenom.

The contribution of the psychology analysis will help the development of effective principles to better the awareness and estimation of the Subjective Information Quality. The principles provided by this research will help the signaling theory to discover the quality of the signal veracity.

Cognitive bias in all its forms, anchoring, subjective validation, and so on, have a heavy effect on most SIQ metrics. The most affected may be believability, where people tend to be biased toward certain information. Others such as subjective accuracy and value-added may not be as heavily influenced by X, Y, or Z, but an IQ specialist still needs to understand the effect of these biases. Samples of measurements can be taken to various data types and sources to identify the most common biases.

CHAPTER 3

APPROACH

3.1 Proposed Approach

The problems of unclear and slow estimation/evaluation of SIQ measures and poor visual representation for information quality will be addressed through three tracks: 1) user-centered evaluation of Subjective Information Quality (evaluation track, Chapters 4 - 6); 2) deriving rules and principles for estimating and evaluating SIQ (principles track, Chapter 7); and 3) presenting the SIQ to users (visualization track, part of Chapter 8).

In this approach, we will focus on three quality dimensions: accuracy, believability, and value-added dimension. Accuracy is defined as the extent to which information is correct and reliable [10], [12]. Believability is defined as the extent to which information is regarded as true and credible [10], [12], and the Value-added dimension is defined as the extent to which information is beneficial and provides advantages for its use [10], [12]. The reason for using these three dimensions is that accuracy can be accessed through objective methods, in addition to subjective methods. Believability is mainly a subjective dimension, although there are potential mathematical models for belief, which may or may not be applicable to believability. Value-added tends to be mostly a SIQ dimension with little mathematical background in computing a metric. This will provide a wide range of different dimension options to help develop an understanding of how the SIQ dimensions interact with each other. These three dimensions are interrelated due to the nature of Subjective Information Quality, which resides at the intersection between human and computers.

3.2 Limitation and methods

This research focuses on information integration operations that can be easily understood by a human and delimits those computer functions such as SQL queries that are too complex to be understood and visually performed by people. The operations considered in our work are intended for use with visual representations, and as such, the complexity of these operations is limited by what people can perceive. Very complex operations, such as those that can only be handled by a machine, are beyond the scope of this work, primarily because the quality of the information integration may be highly nondeterministic and difficult to interpret and estimate by a human. The three tracks of the proposed work on integrating SIQ into visual analytics are detailed below.

3.2.1 Evaluation Track

To determine how people perceive the SIQ dimension and its characteristics, different information sets from multiple information sources are assessed through empirical evaluations of how SIQ is regarded by users. We rely on information sets with already known quality, such as average temperatures in the United States, lab data on gold nanoparticles infused in bone cells, and economics information such as budget data. The information used is available and easily estimated by an average person living in the US for at least one SIQ dimension. The information sets are modified in different percentages and scenarios and presented to the user to determine how people can perceive small information quality variations. Also, the idea of how human actors discover different characteristics of the SIQ dimensions will be evaluated. The work of this track will help in understanding the characteristics of the SIQ, the sensitivity of people's perception of Information Quality, and possible Subjective Information Quality thresholds.

3.2.2 Principles Track

To enable an IQ professional to encode how the quality of information is propagating from the source to the current visual display, we developed a set of principles to based on the lessons learned in the evaluation track. We can implement rules and principles on how different types of information with different SIQ might be perceived by a human audience. The IQ professional can perform a small sample information integration and SIQ evaluation, and produce a set of simple formulas for estimating SIQ on-the-fly. The work of this track provides a partial solution to the problem of determining a method for fast estimation of SIQ during the information integration process.

3.2.3 Visualization Track

We present a number of simple visualization that can convey numerical estimates of IQ dimensions. These IQ visualizations can be linked with "main" visualizations of data, and can be updated on the fly as the main view is being explored. We employed Many Eyes [20] and Microsoft Excel to create stand-alone representations. Many other visualizations could be created, and IQ could be integrated in the main view, but that is very dependent on particular tasks, data, and tools used, and thus outside the scope of this thesis.

CHAPTER 4

SUBJECTIVE EVALUATION OF PERCEPTION OF ACCURACY IN VISUALIZATION OF DATA

4.1 Synopsis

This chapter focuses on the human perception of information quality and describes the results of a study on how accuracy is estimated for data shown through a visual representation. The subjective assessment of quality appears to be non-linear in relation to the actual degree of errors in the dataset. Users are sometimes unable to distinguish between datasets with different quality, and their ability to estimate is better for certain quality levels than for others. The study also shows that adding complementary information does not always help users to better assess the accuracy of the visualization, and thus of the data. The implication of these results is that, for subjective measures of quality, traditional statistical methods of assessing quality may need to be extended with additional methods to account for the non-linearity and the behavior of data integration.

4.2 Introduction

In many applications the quality of information cannot be determined through algorithmic means and can only be assessed through subjective judgment. Some quality dimensions, such as believability, value-added or reputation, are intrinsically dependent on the human actor, while others may in certain situations become subjective. This may be the case for accuracy, a precise dimension when the exact value can be computed, and a subjective one when an exact value is not available or cannot be computed. Exact accuracy, for example, can be achieved when one checks the accuracy of some data against the balance of a bank account, and subjective accuracy is employed in the estimation of the amount of oil being spilled from a deep-water oil rig.

The term we introduce to describe quality measures that cannot be determined by a computer alone is subjective information quality or SIQ. Subjective information quality may not necessarily behave the same as the precisely computed measurements because they involve human factors and human psychology. Assessment of SIQ may be more application- and situation-specific, and rules for determining such quality may be different than statistical calculations. For example, people may find faults in data that is very accurate, and may find the combination of two poor data sources to be more than the sum or average of the parts.

The aspects of subjective information quality covered in this chapter include how subjective rating varies with actual data quality, and how additional information supports better assessment. An orthogonal issue is data integration. Decision makers, experts, and regular users often have to combine information from different sources to get a unified view of the information, or to help guide decisions on larger amounts of information from various sources. This process has become significant in a variety of situations both commercial and scientific. Combining data has become increasingly important as organizations strive to integrate an increasing quantity of internal and external information. Users must combine data for a variety of reasons. Some of those reasons are:

- To have more attributes;
- To get more detailed information for an attribute or item for different purposes and cases; or
- Users cannot find an answer or a solution from a single dataset or data source.

Our results are based on a study on the perceived accuracy of weather data in the United States. The study employed data that can be easily judged by an average person living in the US. The data was obtained from the National Oceanic and Atmospheric Administration's (NOAA) website [19]. The investigators introduced a controlled amount of error in each visualization. Participants did not know the exact percentage of error introduced and were asked to estimate the visualization's overall accuracy. The study also examined the effect of data integration by including visualizations with two panels, each conveying complementary weather data for winter and summer. The same amount of controlled error was thus presented to participants both as a single-panel and as double-panel visualizations. The additional panel itself had also the accuracy controlled and varied from completely accurate to fifty percent errors.

The study results showed that estimated accuracy is non-linear as a function of the actual accuracy, and that data integration may not always help users. Participants did not make constant estimation errors, nor did their estimation error increased or decreased with the actual accuracy. Multiple peaks and valleys are apparent, which suggests that people may not be able to distinguish between certain levels of accuracy, and that certain thresholds make accuracy estimation easier for a given number of actual error levels. With regard to data integration, the study found that introducing additional, error-free data, such as temperatures for summer in addition to those for winter, resulted in worse accuracy assessment than additional data with error. For this weather data set, we found the counterintuitive result that single datasets (for example winter only) are better estimated than datasets with double the amount of information (winter and summer).

The remainder of the chapter is organized as follows: the next section discusses related work, followed by the experiment description and the results. The chapter concludes with a discussion and future work.

4.3 Experiment

4.3.1 *Participants*

The study was web-based, and was advertised to specific student groups in the information-related disciplines at the University of Arkansas at Little Rock and to colleagues of the authors. The study was open for about two weeks. 15 complete responses from 15 participants were identified. 3 responses were excluded because participants had selected the same answer for all questions. Participation was anonymous, and no information we stored could have been traced back to the participant. No incentives were offered.

4.3.2 *Materials*

4.3.2.1 Data

Data was obtained from NOAA [19] and included average temperatures for all US states broken down by season. Table 2 shows part of the dataset of the average temperature rates by state used in the study. Only winter and summer were included in our study.

STATE	WINTER	SPRING	SUMMER	FALL
ALABAMA	42.6	61.3	80.2	62.9
ALASKA	15.8	36.3	58.4	34.1
ARIZONA	51.7	60.8	86.5	70.5
ARKANSAS	40.1	61.4	82.4	63.3

CALIFORNIA	57.1	60.8	69.3	66.9

Table 2: *Part of the weather data set employed in the study. All states were shown to Participants*

4.3.2.2 Equipment and software:

The computer hosting the web study was an Intel Xeon dual-core workstation running at 3.06 GHz with 3 GB of RAM and Windows 7 Professional (32-bit). The web pages were dynamically generated using the ASP .NET v4.0 framework, running on top of Internet Information Services 7.0. The web pages could be viewed on any network connection computer on campus, running the browser of the participant's choosing.

4.3.3 *Methodology*

The survey was broken down in six different pages, and participants could move to the next one by pressing a button. The first two pages were always presented in the same order, while the last four were presented in a random order determined in real-time by the web-server and our software every time a new browsing session (a new user) was establish. The first page was a landing page with short instructions about the study. The second page, the warm-up, allowed the users to explore the features of the visualization type employed in the study. All visualizations in the study were created using the Many Eyes software [18] and consisted of a map(s) with the states within the US and an average temperature for each state. Many Eyes [18] is “an IBM research project and website whose stated goal is to enable data analysis by making it easy for laypeople to create, edit, share and discuss information visualizations”. Some visualizations contained one map for one season (winter), and others two panels for two seasons (winter and summer).

The other four pages contained the actual visualizations that needed to be rated by the user. Table [3] shows the different pages presented to the participants, the maps shown on each, and the percentage each map was modified. For the Green Page the additional data was summer and was presented in the second panel, while for the Yellow and Blue Pages, the additional dataset was winter and shown first.

Page	**Season(s) and % Modified**
Red	Winter 6%
	Winter 12%
	Winter 25%
	Winter 50%
Green	Winter 6% and Summer 0%
	Winter 12% and Summer 0%
	Winter 25% and Summer 0%
	Winter 50% and Summer 0%
Yellow	Winter 25% and Summer 6%
	Winter 25% and Summer 12%
	Winter 25% and Summer 25%
	Winter 25% and Summer 50%
Blue	Winter 50% and Summer 6%
	Winter 50% and Summer 12%
	Winter 50% and Summer 25%

	Winter 50% and Summer 50%

Table 3: *Pages shown to the user and the visualizations included in each. Note that the percentage shows the amount of error, which is the inverse of accuracy relative to 100%.*

The datasets and the visualizations were generated before the study took place. A separate dataset was generated for each map/panel of the visualization by modifying the temperature in a certain percentage of the states for a given season. Special purpose software was created for this task. The states were randomly chosen, and the temperatures were randomly generated within the minimum and maximum temperatures for that season that existed in the original data set. As such, all the modified temperatures still fall within some reasonable limits.

It is important to note that no dataset, except the 100% accurate (the original) set, was shown to the participants in more than one map/panel. Even for the same season and percentage of modification multiple datasets were generated with different states and different values. After each visualization such as the one in Figure 2, participants were asked to assess the accuracy on a five level rating scale. For visualization composed of two panels (that is two maps), the instructions asked to rate the overall accuracy. The scale consisted in (1) Very Accurate (100% - 80%), (2) Accurate (80% - 60%), (3) Fairly Accurate (60% - 40%), (4) Inaccurate (40% - 20%), and (5) Very Inaccurate (20% - 0%).

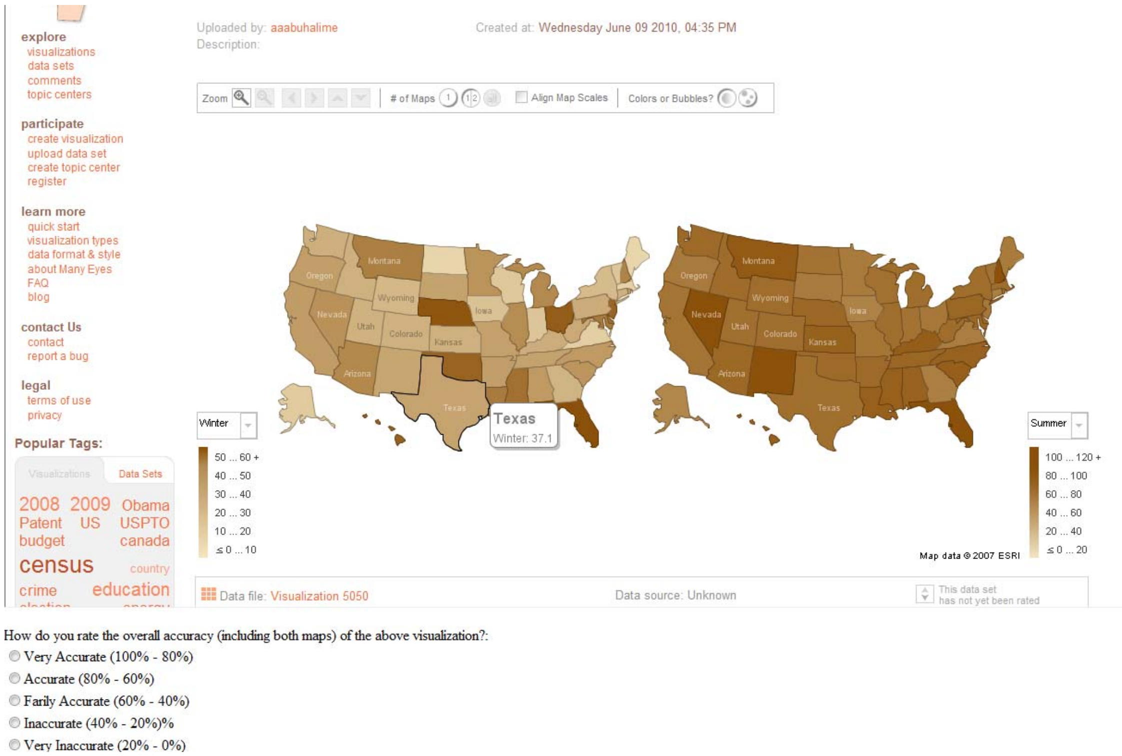


Figure 2: *Snapshot of a webpage showing a visualization and question.*

4.3.4 *Hypotheses*

The following hypotheses were considered:

A. Both the answers entered by the user and the amount of estimation error are dependent on the actual accuracy of the data shown in the visualization.

B. User answers and estimation error does not vary in a linear fashion with the actual accuracy of the data.

C. (C1) Adding additional information, such as a second season, changes both the answers of the participants and the amount of estimation error when compared to single-season data sets, and (C2) the more accurate the additional data is the more the overall subjective assessment of accuracy is improved.

4.3.5 *Design*

Two independent variables were considered: basic_accuracy, and additional accuracy. The basic accuracy is one of the 94%, 88%, 75%, or 50%, and represents the quality of the data presented in at least one of the panels of each visualization. A webpage of the survey has four visualizations, one for each accuracy level.

Additional accuracy captures the quality of the data added in double panel visual representations. The additional accuracy is a constant within each webpage, but it varies from webpage to webpage. Possible values are -1 for single panel visualizations, and 100%, 75%, and 50% for double visualizations. For simplicity, the results will be reported using the either the accuracy of the single panel for simple visualizations (one of 94%, 88%, 75%, or 50%), or the average of the basic and additional accuracy for visualization composed of double panels (one of 97%, 94%, 87.5%, 84.5%, 81.5%, 75%, 72%, 69%, 62.5%, and 50%).

The dependent variable is user_estimation and it is one selection on a five level scale. The participants can choose one of the five, equal-size intervals that divide 0% through 100% accuracy. User’s answer, as a measure, is independent of the actual accuracy, and the same answer for an accuracy of 94% can be significantly worse than for an actual accuracy of 50%. In order to quantify how exact participant’s assessment was, we derived a metric from the user_estimation and average accuracy. The new metric, error is the difference between the average accuracy of visualization and the closest edge of the interval answered by a participant. When the interval contains the average accuracy, the error is zero.

4.3.6 ***Results***

The survey produced 15 complete answers, but the results only considered 12 because the other three appeared to resemble test submissions containing the same answer for all 16 questions. Overall, the analysis included 192 answered questions.

An ANOVA was performed for both the user_estimation and error. The average accuracy was found to be statistically significant factors: $F_{10, 110}$=2.01, p=0.0385 for user estimation, and $F_{10, 110}$=2.26, p=0.0192 for error. The same holds true for additional accuracy: $F_{3, 33}$=9.17, p=0.0001 for user estimation, and $F_{3, 33}$=7.89, p=0.0004 for error.

A Tukey pairwise analysis of the contribution of each additional type of visualization was performed. In the case of user estimation, significant differences were found between single-map (additional accuracy = -1) and 100% accuracy additional panels (Adj. p < 0.0001). A weak statistical significance was found between single-panel and 75% (Adj. p = 0.0692). Another difference was found between 100% additional and 50% accurate additional maps (Adj. p = 0.0440). For error, single-panels (additional accuracy = -1) were significantly different than 100% additional views (Adj. p = 0.0002). Visualization containing 100% accurate additional panels were also found to be different than 50% and 75% ones, with Adj. p = 0.0080 and Adj. p = 0.0489, respectively. The absolute values for user_estimation the error are depicted graphically. For simplicity, all graphs use the convention that the higher the bar the higher the error. Figure 3 shows the user answers and error recorded for various actual accuracy levels. Single visualizations were better estimated in term of error than double-panel ones (Figure 4). Figure 5 conveys the same in more detail and split by each level of accuracy for the additional panel. Figure 5 presents how both single and double-panel visualizations were assessed relative to the actual fault level in the additional data.

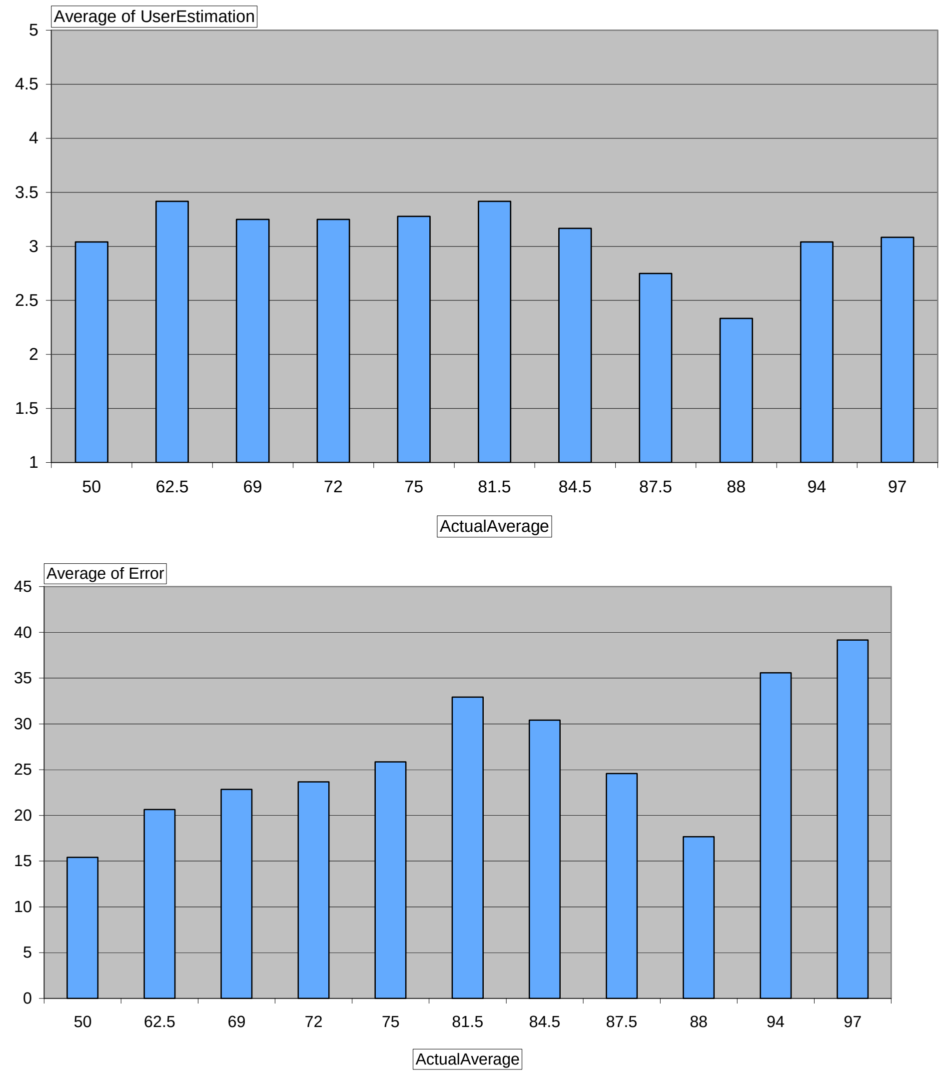


Figure 3: *Average of user estimation (top) and error (bottom) per actual accuracy level. Note that, in the top panel, a value of 1 for user estimation means "Very Accurate (100%-80%)", and 5 represents "Very Inaccurate (20% - 0%)".*

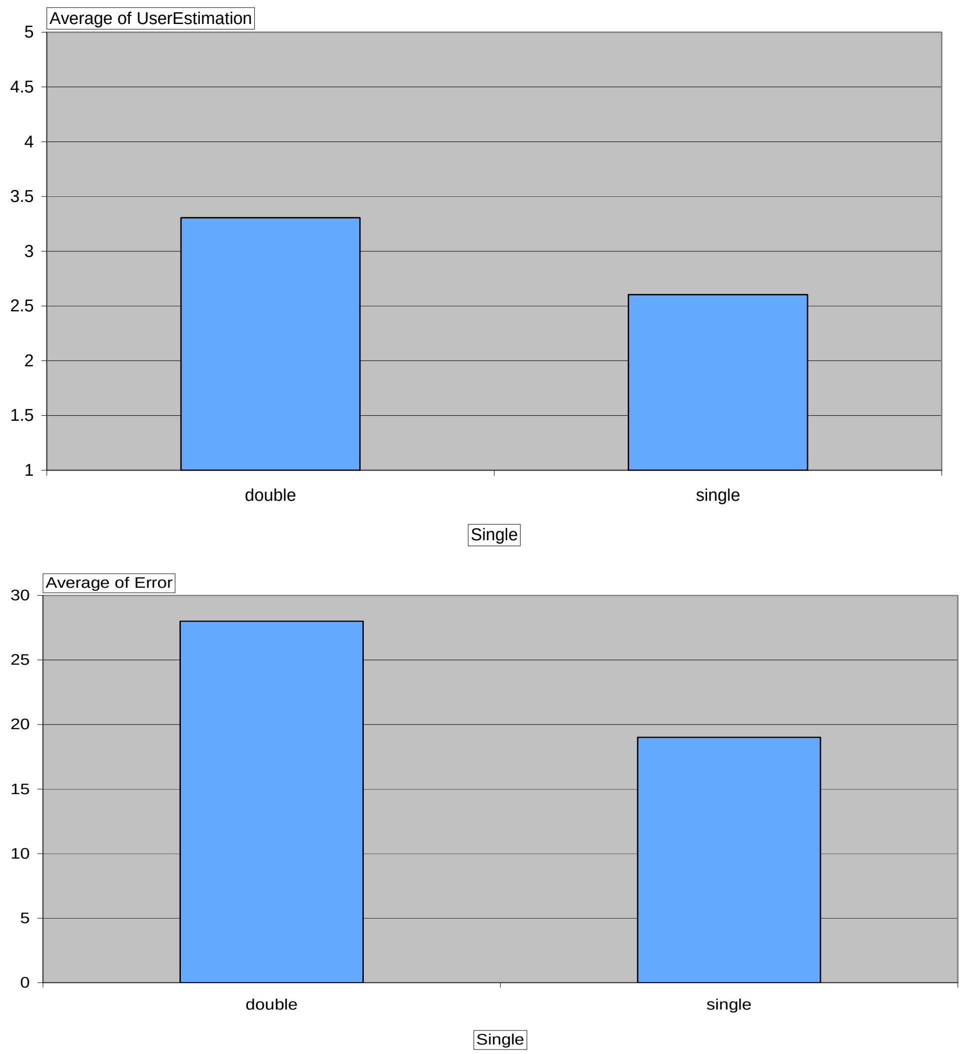
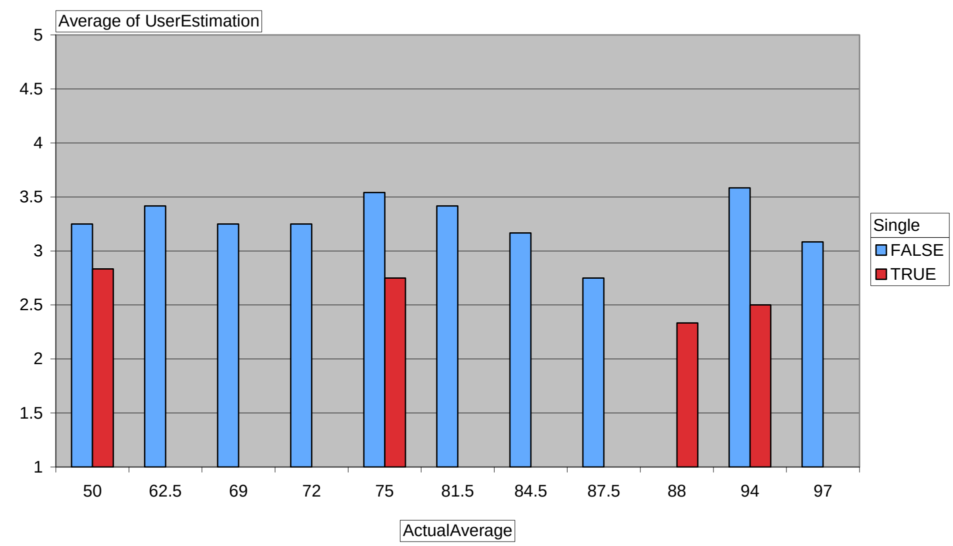


Figure 4: *User Estimation (top) and error (bottom) for double- and single-map visualizations. Note in the top panel that "1" for user estimation means "Very Accurate (100%-80%)" and "5" represents "Very Inaccurate (20% - 0%)".*



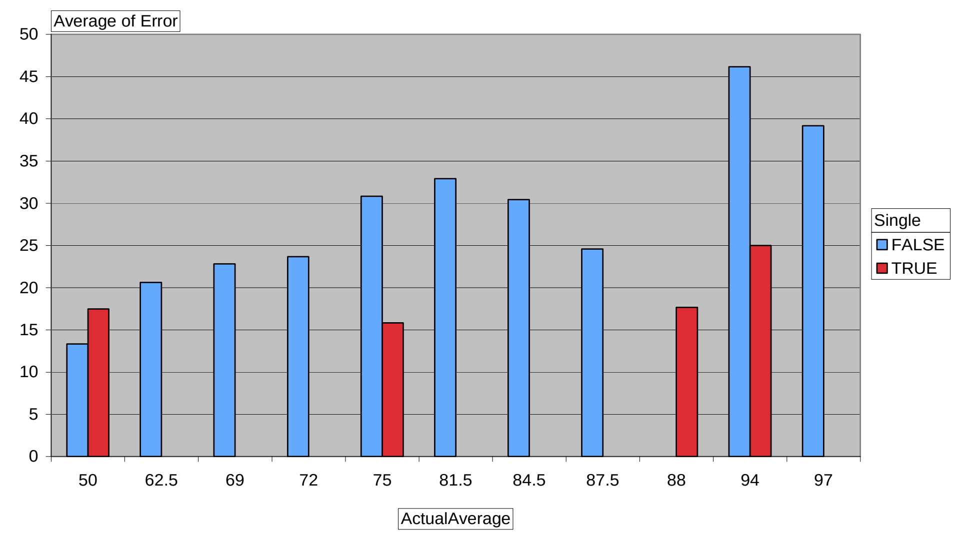
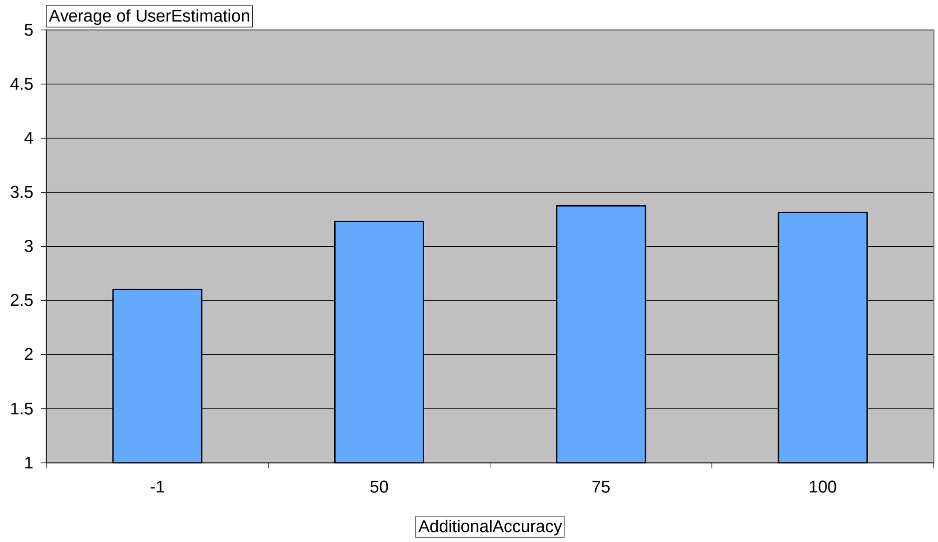


Figure 5: *User Estimation (top) and error (bottom) for double- and single-map visualizations and divided per actual accuracy average. Note that 1 for user estimation means "Very Accurate (100%-80%)" and 5 represents "Very Inaccurate (20% - 0%)."*



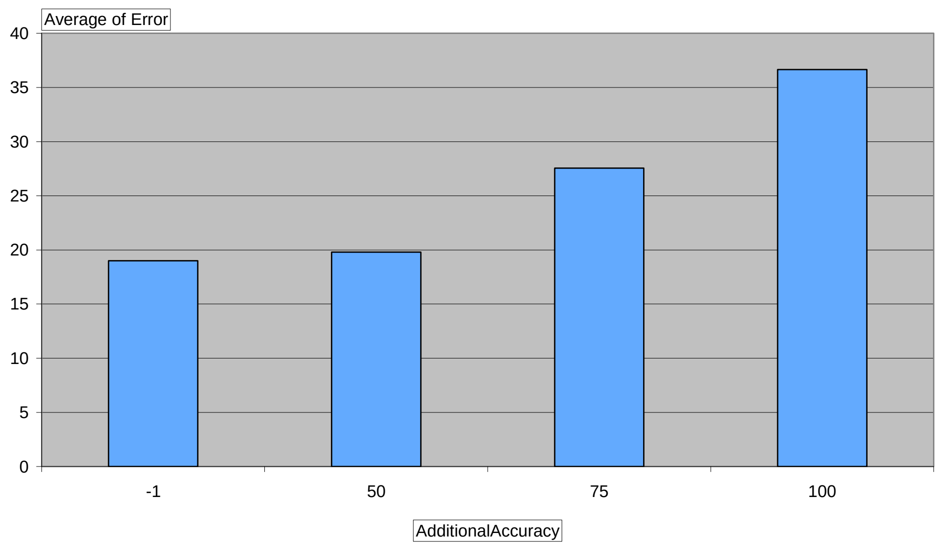


Figure 6: *User Estimation (top) and error (bottom) as a function of the accuracy of the additional map (-1 denotes single visualizations). Note that 1 for user estimation means "Very Accurate (100%-80%)" and 5 represents "Very Inaccurate (20% - 0%)".*

4.4 Discussion

Hypotheses A, B and C1 are confirmed by the experiment, but C2, the addition of accurate data helps subjective assessment, is not supported by the results. One explanation for the failure to observe C2 is that more information contributes to a task overload and participants performed worse. Initially, we thought that the addition of a winter or summer season would be able to provide more information about the minimum and maximum temperatures of each state. Then, if the one of the seasons was altered, the other season's temperatures would help in identifying a too hot or cold temperature. Another way in which the additional season could help is by providing a visual pattern of temperature variations for a whole region of the US. Wide differences between the summer and winter pattern would be a warning that information in that region is not accurate. In the end, none of these suppositions were correct as shown by the failure to prove C2.

The results show that visualizations with single maps behave better than other types of visualizations (Figures 4-6), although no significant difference was found, on any of the dependent variables, between single-map and double maps in which one of the panels is 50% accurate. It appears that for this task user perform better with additional inaccurate information than with completely accurate one. This behavior would be difficult to incorporate in a statistical model for data integration. It may be that various thresholds may exist for when users are able to best use additional information.

There does not seem to be a linear dependence between the user assessment and the actual accuracy. The assessment is best around 88% and 50% (Figure 3), but it becomes worse around 100% and 80% (Figure 3). While "bad" estimation at 100% and "good" at 50% can be explained by its distance from the average of a completely random answer (that would be 50%); however there is no explanation for the behavior at 88% and 80%. Moreover, all of 80%, 88%, and 100% belong to a single answer interval: "Very accurate (100%-80%)". It is unclear why people think that almost accurate visualizations are worse than the ones that have 12% errors. It may be possible that people have thresholds of how they perceive visualizations, and that they also suspect that an error has been introduced even for very accurate visualizations.

CHAPTER 5

EMPIRICAL EVALUATION OF BELIEVABILITY IN VISUALIZATION OF DATA

5.1 Synopsis

Believability is one of the major information quality dimensions that play a role in the operational fitness and sound decision making. This chapter presents an empirical evaluation of how people perceive believability of data shown through visual and textual representations. Integration of text and images is also studied with respect to believability. The subjective assessment exhibits variation for different types of data sources: textual, image, and both. The manner in which believability varies appears to be heavily dependent on task. Some tasks are more believable when text is integrated with images, others do not benefit from the combination. Scientific data collected in the process of incubation of the bone cells with gold nanoparticles is selected for the study because it alleviates the effect of the accuracy dimension on the assessment of believability. The implication of these results is that, for subjective measures of believability, traditional statistical methods of assessing quality may need to be extended with additional methods to account for the non-linearity and the behavior of data integration.

5.2 Introduction

Believability, defined as the extent to which information is true and credible [10, 12], is one of the information quality (IQ) dimensions that can be best determined/assessed through subjective assessment of the information users rather than through algorithmic means. Some other quality dimensions, such as value-added or reputation, are also intrinsically dependent on the human actor, while others may become subjective in certain situations.

The term we introduce to describe quality measures that cannot be determined by a computer alone is subjective information quality or SIQ [2]. Subjective information quality may not necessarily behave the same as the precisely computed measurements because they involve human factors and human psychology. Assessment of SIQ may be more application and situation specific, and rules for determining such quality may be different than statistical calculations. For example, people may find faults in data that is very true and credible, and may find the combination of two poor data sources to be more than the sum or average of the parts.

The aspects of subjective information quality covered in this chapter include how subjective rating varies with different pieces of information displayed, and how additional information influences the assessment of believability. Our results are based on a study on the perceived believability of the concentration of gold particles added to the bone cells. The study employed data that can be easily judged by an average person. The data was obtained from the Nanotechnology Center at University of Arkansas at Little Rock. The investigators introduced two types of data: lab notes (textual) and microscopy (image). The study assessed the believability of each type of data as well as the believability of the integration of text and image.

Believability was assessed with regard to two different concentrations of gold particles. Participants were asked to express their belief that the presented data was of a given concentration. The design of the study surreptitiously forced participants to provide their opinion based solely on belief because scientists could not determine any difference in results between the two concentrations. Thus, believability was measured alone without interference from other IQ dimensions, more importantly accuracy, which can skew the results for believability when people tend to not believe data that appears inaccurate. Note that choosing experts for the study would result in them validating accuracy rather than believability because experts already use and trust these visual tools and lab notes.

The study revealed that believability varies with the type of data source, image, text, or both, and that it behaves differently for each task. Users assessed themselves as being neutral to confident in their results, with the text data source scoring the lowest, and the image scoring the highest.

The remainder of the chapter is organized as follows: the next section discusses related work, followed by a description of study and the results. The chapter concludes with a discussion and future work plans.

5.3 EXPERIMENT

5.3.1 *Participants*

The study was web-based, and was conducted through Amazon's Mechanical Turk. The study was open for about one week. 161 complete responses from 200 answers were identified. Some responses were excluded because participants had selected random numbers not within the two options provided, and all the answers where work time was less than 30 seconds were excluded. Participation was anonymous, and no information we stored could have been traced back to the participant. Each answer was paid $0.25.

5.3.2 *Materials*

5.3.2.1 Data

Data was obtained from scientists in the Nanotechnology Center at the University of Arkansas at Little Rock , and a sketch of the processed of incubation of the bone cells with gold nanoparticles is shown in Figure 7. The cells were sliced and visualized under a transmission electron microscope (TEM), where gold nanoparticles appear as black dots. The gold nanoparticles deposited on upper surface of cell plasma membrane, which triggers arms forming a round the gold nanoparticles (Endocytosis). Some pictures used for the experiment show the arms in the process of Endocytosis.Two different concentrations of gold nanoparticles are used 10 µg/ml and 160 µg/ml, but the end result of the incubation of the cells is the same regardless of the concentration.

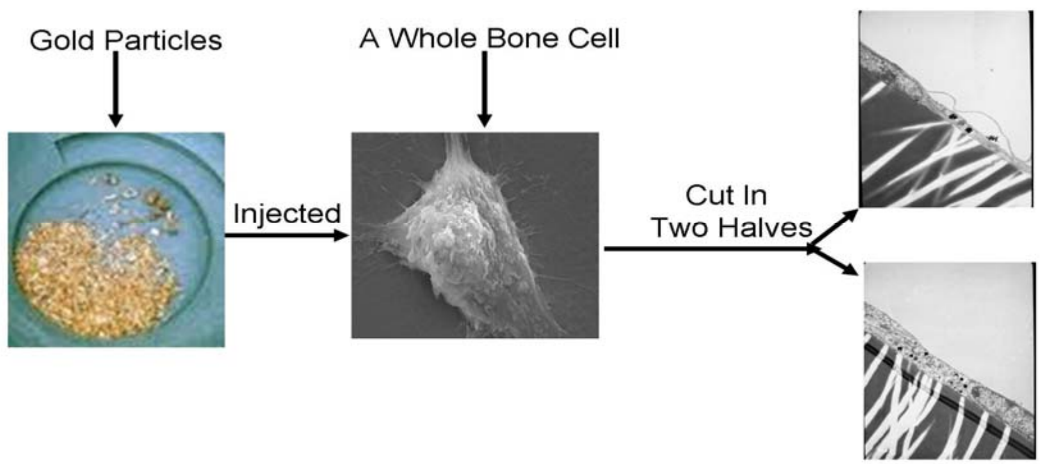


Figure 7: *Diagram describing the experimental process of incubation of the gold nanoparticles with the bone cells. This image was provided to the participants in the study*.

5.3.2.2 Equipment and software:

The software and environment used to perform the study is Amazon Mechanical Turk [20], a marketplace in which people use their innate human intelligence to solve various tasks. The Mechanical Turk web service enables companies to programmatically access this marketplace, which is supported by a diverse, on-demand workforce. Mechanical Turk aims to make accessing human intelligence simple, scalable, and cost-effective. Businesses or developers that have tasks that cannot be solved by a machine can create small pieces of work, called Human Intelligence Tasks or "HITs", via the Mechanical Turk APIs. Workers registered with the Mechanical Turk, then perform the tasks. Upon verifying the results, businesses and developers direct Mechanical Turk to pay the workers. We employed Mechanical Turk as a way to distribute questions about the gold-doped bone cells and to estimate the level of believability in the two gold concentrations from the professional workers registered with the Mechanical Turk.

5.3.3 *Methodology*

The study was designed in such a way to not be dependent on accuracy. We achieved this goal by choosing a task based on the resulting cell configurations, which appears the same regardless of the gold concentration. The scientists discovered that the end-result of gold nanoparticles incubation is the same for both concentrations. However, scientists and experts were excluded from taking the study to avoid introducing bias towards accuracy in the results, since they would be familiar with materials and the images. This will not help the main goal of the study.

The first section of each HIT starts with short instructions about the HIT. The image shown in Figure 7 provides an overview of the whole process of adding the gold particles to the bone cells, and another image (Figure 8) provides a sample image with description of important features to allow the participants to familiarize themselves with the data types employed in the study. The second section of the HIT includes a textual description of the process of incubating the gold nanoparticles in the bone cells, and a sample of the two concentrations. The last section of the HIT describes the task the user needs to perform, and it is captured in Figure 3.

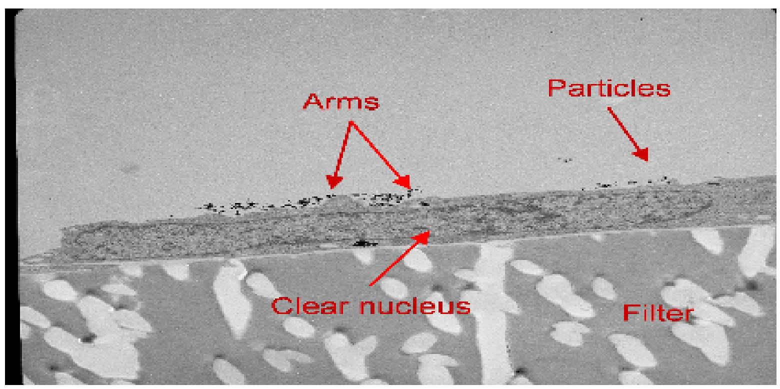


Figure 8: *Snapshot showing the contents of images*

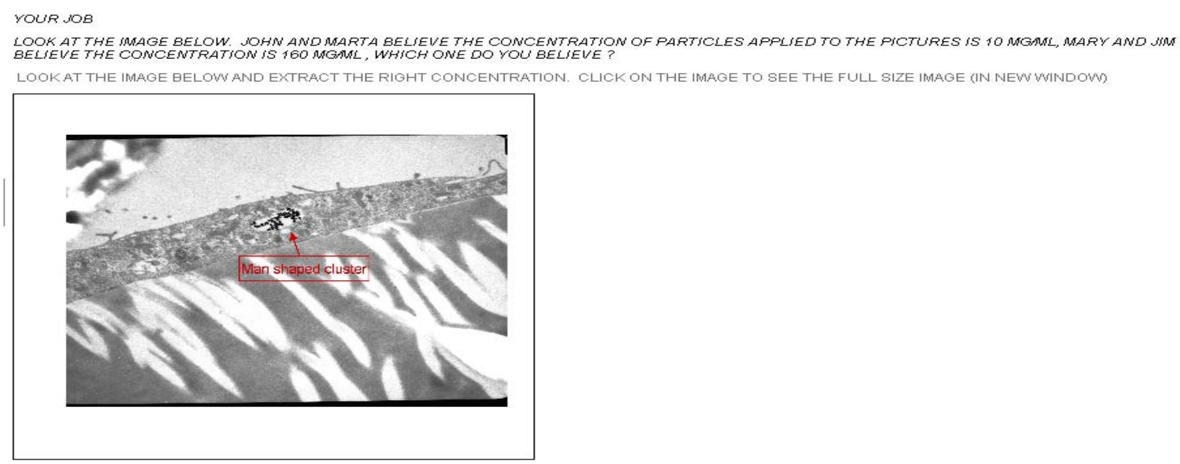


***Figure 9**: Snapshot showing a task that integrated images with text*

The study was broken down into nine different tasks (HITs). The first three HITs we designed included questions based on only images of bone cells doped with either 160 µg/ml and 10 µg/ml gold, while the next three HITs included text description of cells with each of the two concentrations. The final three HITs included both image and text integrated as in Figure 8. Different cells were presented in each HIT. All the HITs were published in a random order and at different times.

In each HIT, the following scenario was included “John and Marta believe the concentration of particles applied to the pictures is 10 µg/ml, Mary and Jim believe the concentration is 160 µg/ml“as captured in Figure 9. Participants were asked to provide their answer whether they agree with John and Marta or Mary and Jim, and also to assess how confident they are in their answers on a five level rating scale. The scale presented the users with the following five choices: Very confident (5), Confident (4), Neutral (3), Not confident (2), and Not confident at all (1). The time allotted per assignment was two minutes, and ten unique workers were allowed to work on each HIT. Only Mechanical

Turk workers over 18 were allowed to work on the HITs. The payment for each assignment was $0.25.

5.3.4 *Hypotheses*

The following hypotheses were considered:

D. User's answers and believability does vary when showing image, text, or a combination.

E. Showing more pieces of information, combined information, improves the overall subjective assessment of believability.

5.3.5 *Design*

The independent variable in the experiment was source of data whose possible values are image, text, or both, and refers to the medium through which the participants in the study are getting their information. The textual information was extracted from the pictures in such a way to be similar to lab notes which present the features present in the observations (images). Note that actual concentration was another independent variable, but experts believe that it is not distinguishable in the images or text, and we do not consider the actual concentration as part of the model.

Two dependent variables were measured during the study believed_concentration and confidence. The believed concentration provides an objective assessment of the participant's believability and can be either 10 µg/ml or 160 µg/ml. The confidence is a self assessment from the user on a five level scale.

5.3.6 *Results*

The study was open for about one week, and 161 complete responses from 200 participants were identified. Some responses were excluded because participants had selected random numbers not within the two options provided, and all the answers where work time was less than 30 seconds were excluded.

An ANOVA revealed that source is a significant factor for believed_concentration ($F_{2,\,160}$ = 3.02, p = 0.0516).A Tukey pairwise comparison found significant differences between *image* and *both* (p = 0.0398). For confidence, the presentation medium is a marginal factor ($F_{2,\,159}$ = 2.64, p = 0.0748). Pairwise, *image* and *text* sources appear the most statistically different for confidence (p = 0.0597). Note that as expected, actual concentration is not a statistically significant factor.

Figures 10-12 illustrate the number of answers who believed either the 10 µg/ml or 160 µg/ml tasks. The self-assessment of the user confidence in their answers is given in Figure 13.

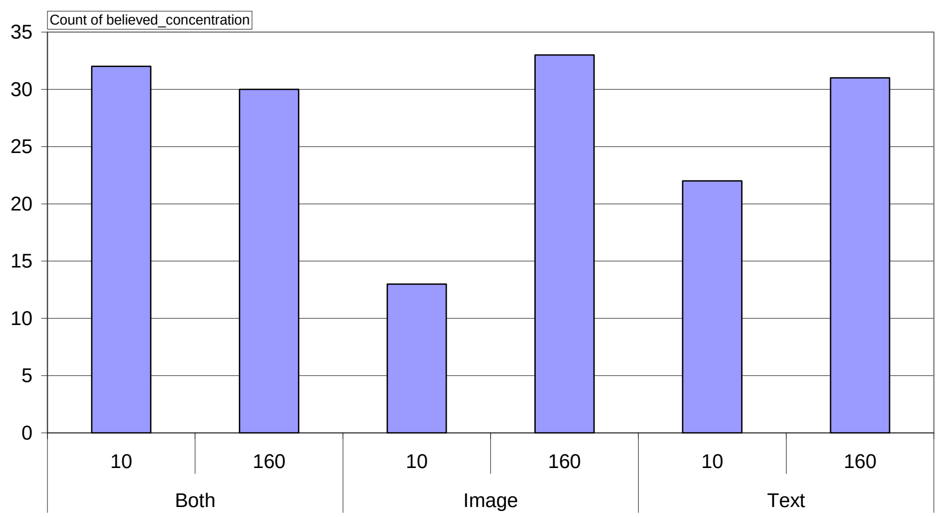


Figure 10: *User believability in the two gold particle concentrations. The information is broken down by data source type and believed concentration. The y-axis shows the number of answers who believed in a given concentration.*

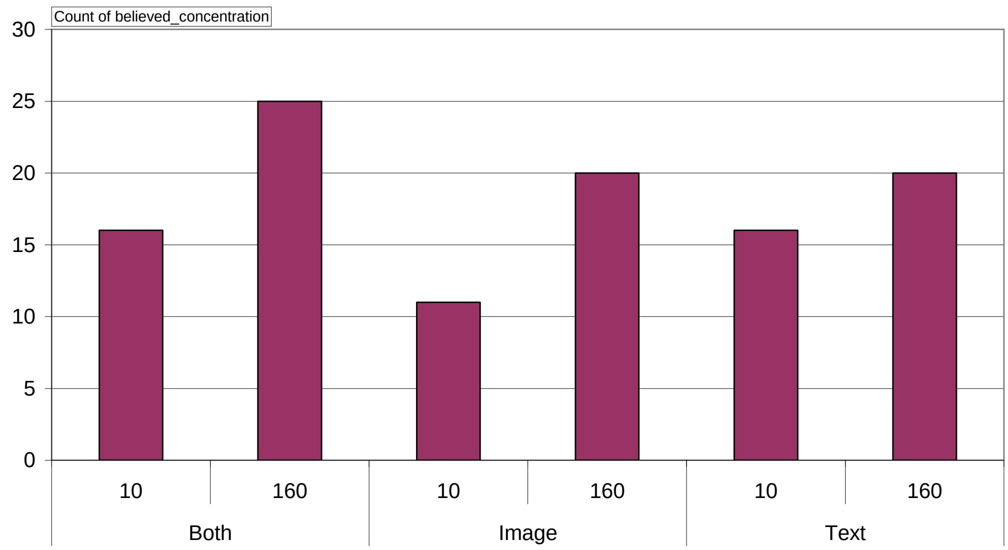


Figure 11*: Actual concentration of 160 µg/ml: user believability in the two concentrations by data type and believed concentration.*

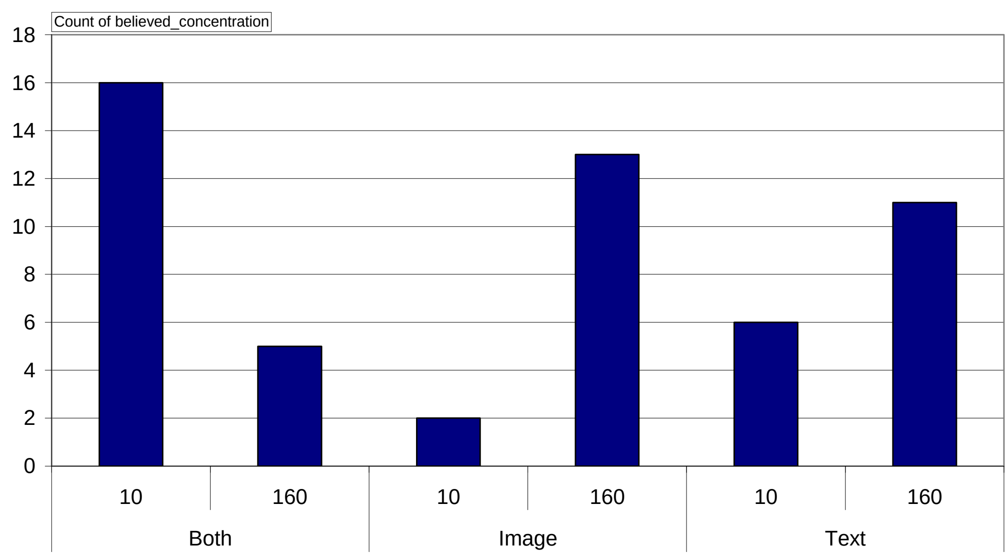


Figure 12: *Actual concentration of 10 µg/ml: user believability in the two concentrations by data type and believed concentration.*

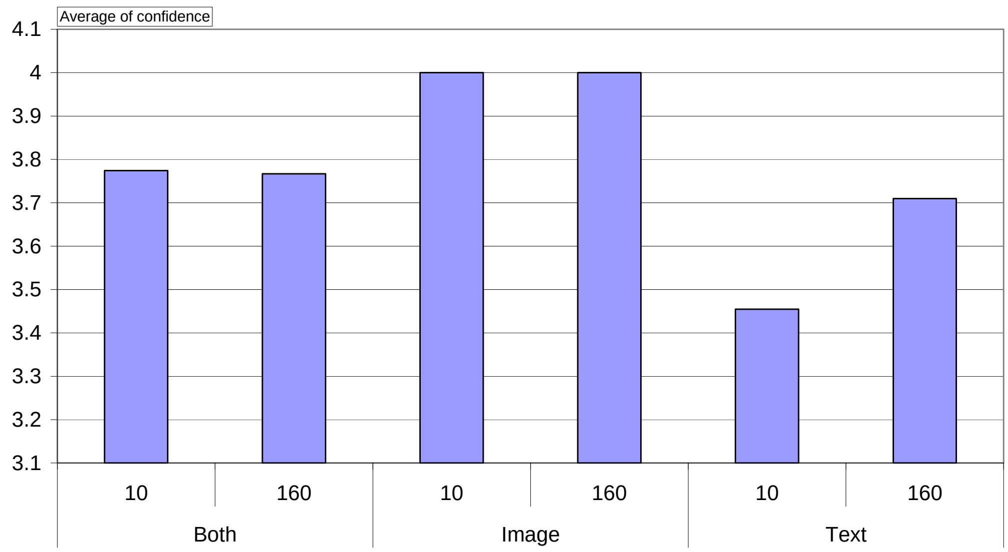


Figure 13*: Average rating of users' confidence in their answers broken down by source type and believed concentration. Note than 1 for user rating means "Not confident at all", 2 means "Somewhat not confident", 3 "Neutral", 4 represents "Confident", and 5 represents "Very Confident".*

5.4 Discussion

Hypothesis (A) holds for both tasks for which the users were assessed, that is for both believing in 10 µg/ml and believing in 160 µg/ml. Hypothesis (B) holds only for the believability of 10 µg/ml task, as shown in Figure 4 more answers selected 10 for the both condition than for image or text alone. The believability of the 160 µg/ml is the lowest when users were presented both image and text combined, and thus hypothesis (B) does not hold.

The results show a user preference (or bias) for the 160 µg/ml task, and consequently a bias against the 10 µg/ml. Figures 10-12 show that most people and under most conditions believed the concentration to be 160 µg/ml more than 10 µg/ml (except for the both case in Figure 12). Further research is needed to confirm the existence of this kind of biased.

Believability was task dependent in our experiment, which may make automated estimation of this dimension a complicated endeavor. A different behavior of believability assessment is observed for the two tasks, none of them being simple averaging. The combination of the two datasets seems to affect slightly negatively the combination of text and images for the preferred task (160 µg/ml), while for the biased-against task (10 µg/ml) the combination improves the level of believability when compared to either image or text.

Most people are confident in their answers, which translates into them being confident in their belief. Image seems to inspire more confidence then text. For confidence, the combination of image and text produces a result that is about the average of the individual confidence levels as shown in Figure 13.

CHAPTER 6

SUBJECTIVE EVALUATION OF VALUE-ADDED IN VISUALIZATION OF DATA

6.1 Synopsis

Value-added is one of the contextual information quality dimensions that depends on the nature of task and plays a role in the operational fitness and the goals to be achieved. This chapter presents an empirical evaluation of how people perceive the value-added of data shown through visual representations. Single and composite visualizations are shown to users and they are asked to provide value-added for a given task. The manner in which value-added varies appears to be heavily dependent on task. Some tasks are more believable and valuable when a single visualization is displayed than when two views are combined. Economics data is selected for the study because it alleviates the effect of the accuracy dimension on the assessment of value-added. The implication of these results is that, for subjective measures of value-added, traditional statistical methods of assessing quality may need to be extended with additional methods to account for the non-linearity and the behavior of data integration

6.2 Introduction

Value added refers to "extra" feature(s) of an item of interest (product, service, person, etc.) that go beyond the standard expectations and provide something "more" while adding little or nothing to its cost [27]. Each industry has its own definition of value-added. The definition of value-added varies in accounting, economics, marketing or education. The value-added of information is defined as the extent to which information is beneficial and provides advantage from its use [10], [12], and it is one of the information quality (IQ) dimensions that can be best assessed through subjective, empirical means rather than through algorithmic means. Some other quality dimensions, such as believability or reputation, are also intrinsically dependent on the human actor, while others, such as accuracy, may become subjective in certain situations [Appendix A].

The term we introduce to describe quality measures that cannot be determined by a computer alone is subjective information quality or SIQ [Appendix A]. Subjective information quality may not necessarily behave the same as the precisely computed measurements because they involve human factors and human psychology. Assessment of SIQ may be more application and situation specific, and rules for determining such quality may be different than statistical calculations. For example, people may find faults in data that is very true and credible, and may find the combination of two poor data sources to be more than the sum or average of the parts.

The aspects of subjective information quality covered in this chapter include how subjective rating varies with different parts of data/categories displayed, and how additional information influences the assessment of value-added. Our results are based on a study on the perceived value-added of restructuring images of two different visualization. The visualizations represent competing U.S. budgets, with no party affiliation replaced by two neutral names: John and Mary. The study employed data that can be easily judged by an average person. The data was obtained from the office of management and budget [26], and another web resource [28]. The investigators introduced different parts of budgets information through visualizations. The study assessed the value-added of each image part as well as the value-added of the integration of the different parts. The value-added is determined by amount of dollars the participants would pay for the selected image part.

The study revealed that value-added varies with the information parts displayed, and that it behaves differently for each task. Users assessed themselves as being neutral to confident in their results, with the combined data scoring the lowest, and single data (John's budget) scoring the highest.

The remainder of the chapter is organized as follows: the next section discusses related work, followed by a description of study and the results. The chapter concludes with a discussion and future work plans.

6.3 Experiment

6.3.1 Participants

The study was web-based, and was conducted through Amazon's Mechanical Turk. The study was open for about one week. 89 complete responses from 120 answers were identified. Some responses were excluded because participants had selected random numbers not within the two options provided ($74, $63), and all the answers where work time was less than 30 seconds were excluded. Participation was anonymous, and no information we stored could have been traced back to the participant. Each answer was paid $0.25.

6.3.2 *Materials*

6.3.2.1 Data

Data was obtained from The Office of Management and Budget, the White House website [26], and from Paul Ryan's web resource [28]. The data used are the budget proposals of President Barack Obama and Senator Paul Ryan. The budget categories were distributed in two groups and visualized in two charts that complement each other. Parts of the budgets charts used are shown in Figure 14.

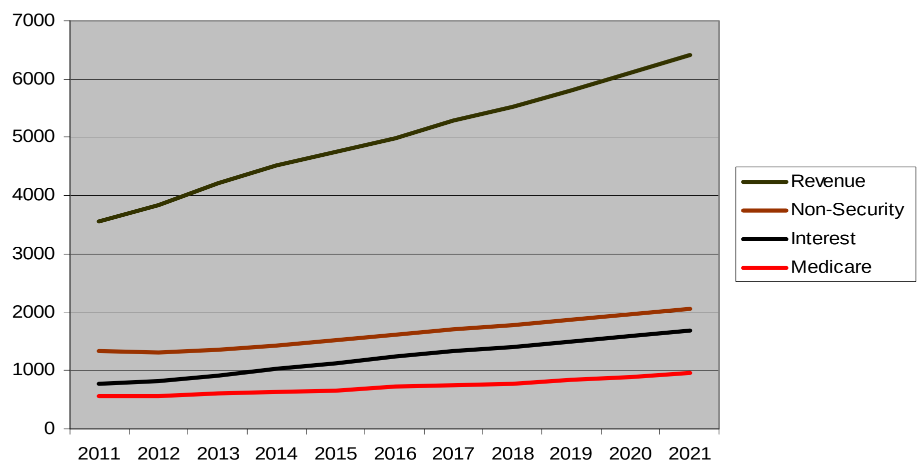


Figure 14: Chart showing the distribution of John's Budget. This image was displayed to the participants in the study as the first part of John's Budget.

6.3.2.2 Equipment and software

The software and environment used to perform the study is Amazon Mechanical Turk [20], a marketplace in which people use their innate human intelligence to solve various tasks. The Mechanical Turk web service enables companies to programmatically access this marketplace, which is supported by a diverse, on-demand workforce. Mechanical Turk aims to make accessing human intelligence simple, scalable, and cost-effective.

Businesses or developers that have tasks that cannot be solved by a machine can create small parts of work, called Human Intelligence Tasks or “HITs”, via the Mechanical Turk APIs. Workers registered with the Mechanical Turk, then perform the tasks. Upon verifying the results, businesses and developers direct Mechanical Turk to pay the workers. We employed Mechanical Turk as a way to distribute questions about the budget visualizations and to estimate the level of value-added in the two visualization-based tasks from workers registered with the Mechanical Turk.

6.3.3 *Methodology*

The study was designed in such a way to determine the how users assess the value-added dimension in a task involving matching the categories from the two budgets. Also the study was designed in such a way to not be dependent on accuracy. We achieved this goal by choosing the task and type of data used, which could be performed by an average person. In the first three HITs, the budget data were distributed in two groups and visualized in two charts complement each other using two different aliases, John and Mary, (Figure 15). The last three HITs of the study, presented similar questions, but with a different break-out of categories.

The first section of each HIT starts with short instructions and present either John's budget, or Mary's budget, or both (Figure 15). The workers are then asked to match either Budget A or Budget B, with complementary categories (Figure 16), to one of the target budgets (John's or Mary's). Finally, the workers were asked to assign a set amount of money to each of the Budget A and Budget B, and to provide a self-assessment of their confidence level.

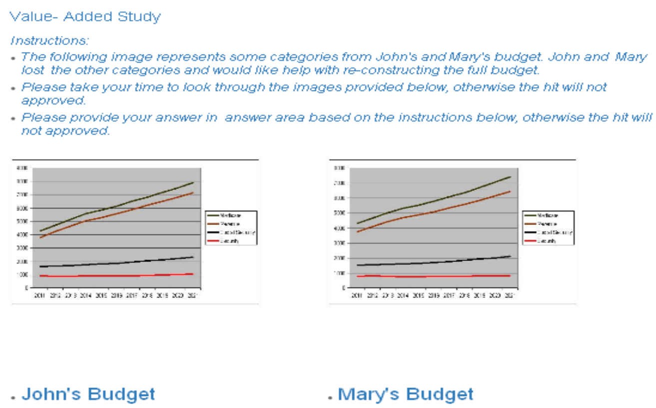


Figure 15: Snapshot showing the distribution of John's & Mary's Budgets after modification.

The study was broken down into six different tasks (HITs). The first two Hits we designed included questions based on only one budget chart (John's or Mary's), the third one include a combination of the two budgets. HITs four and five included only one budget chart (John's or Mary's) using a different break-down of the categories, while the last HIT include a combination of the two modified budget data. All the HITs were published in a random order and at different times.

In each HIT, a scenario like the following was included: "Determine which image (A or B), best fits Mary's budget. Both images below have the same categories. Both images below have the same categories, all of which are different from the categories presented above. If you would have to pay for the two images below, how much would you spend on image A and how much on image B given that John gave you $63. Assume that all funds need to be spent and receipts given back to Mary".

Figure 16: Snapshot showing the complementary part of John's and Mary's Budget

Participants were asked to provide their answer whether they select image A or image B and provide the amount they would pay for each image, two different amounts were provided $74 and $63, and also to assess how confident they are in their answers on a five level rating scale. The scale presented the users with the following five choices: Very confident (5), Confident (4), Neutral (3), Not confident (2), and Not confident at all (1). The time allotted per assignment was 10 minutes, and 20 unique workers were allowed to work on each HIT. Only Mechanical Turk workers over 18 were allowed to work on the HITs. The payment for each assignment was $0.25.

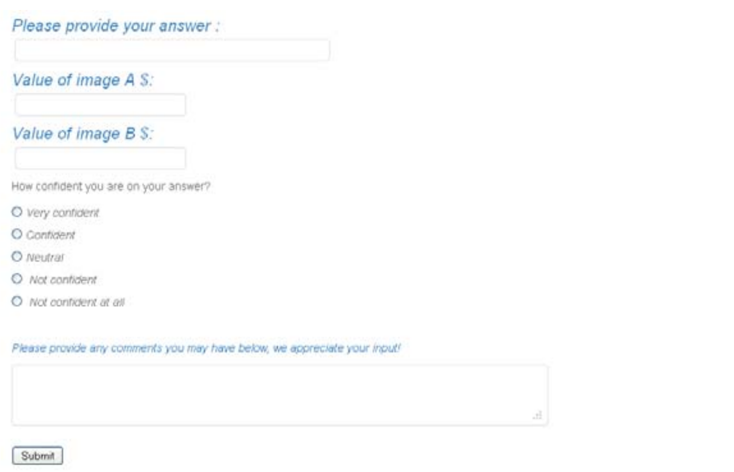


Figure 17: Snapshot showing the lat section of the HIT

6.3.4 Hypotheses

The following hypotheses were considered:

A. User's answers and value-added does vary when showing single budget chart (John's or Mary's), or a combination.

B. Showing more parts of information, combined information, improves the overall subjective assessment of value-added.

6.3.5 *Design*

The independent variable in the experiment was source of data whose possible values are John's budget, Mary's budget, or both. Two dependent variables were measured during the study value-added and confidence. The value-added provides an objective assessment of the participant's average value-added, and we converted the amount of money the workers would spend into a value between 0 and 1. The confidence is a self assessment from the user on a five level scale.

6.3.6 *Results*

The study was open for about one week, and 89 complete responses from 120 participants were identified. Some responses were excluded because participants had selected random numbers not within the two options provided, and all the answers where work time was less than 30 seconds were excluded.

Figures 18, illustrate the number of answers Users selections/answers of Budget A or Budget B within each data type displayed. The average value-added of Budget A and B is broken down by data type displayed in Figures 19 and 20. The self-assessment of the user confidence in their answers is given in Figures 21 and 22.

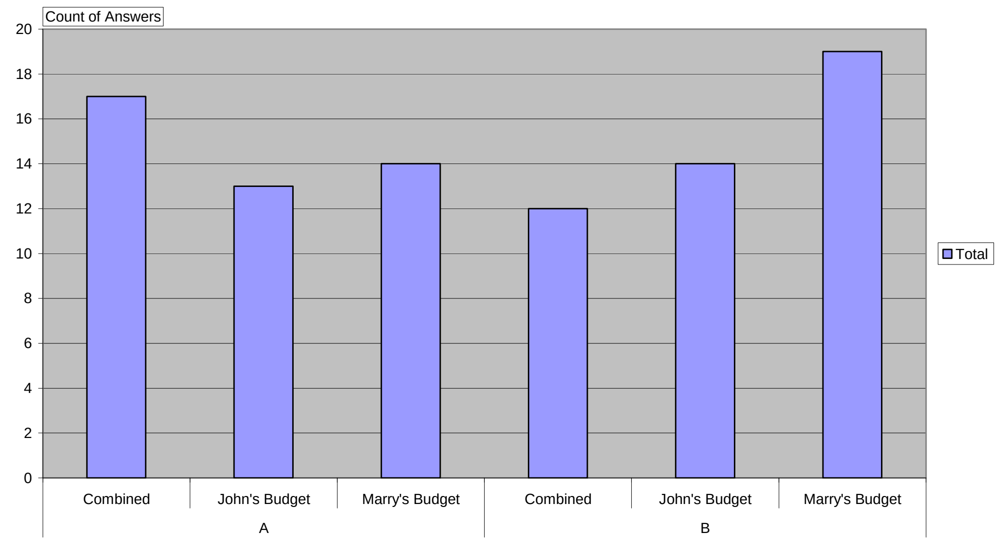


Figure 18: Users selections/answers of Budget A or Budget B. The information is broken down by whose budget was shown (John's, Mary's, or both).

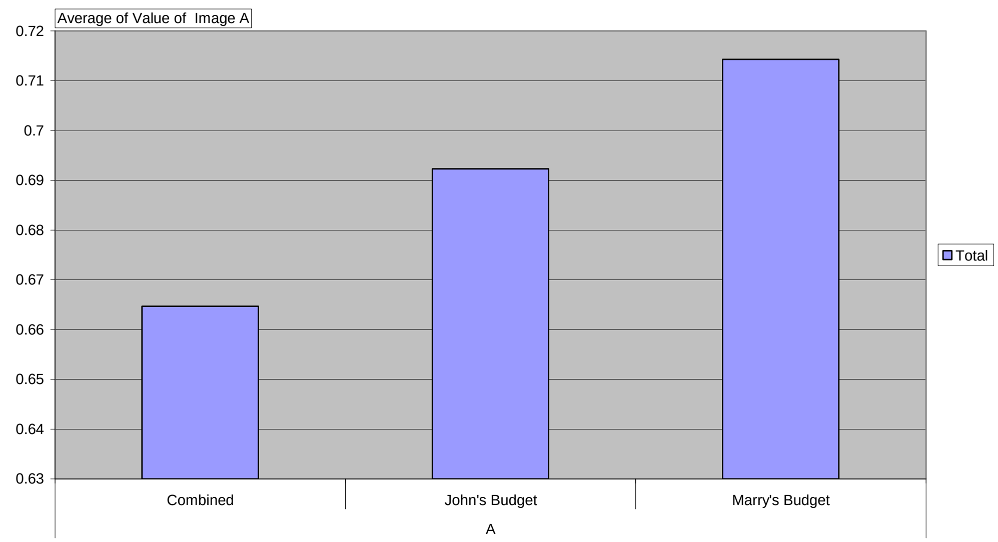


Figure 19: Average value-added of Budget A broken down by budget originator.

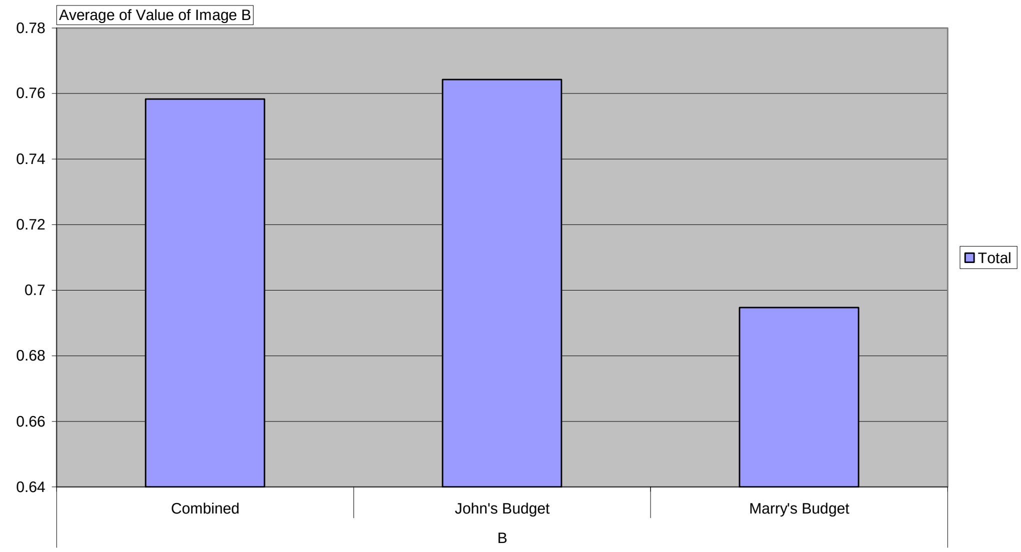


Figure 20: Average value-added of Budget B broken down by budget originator.

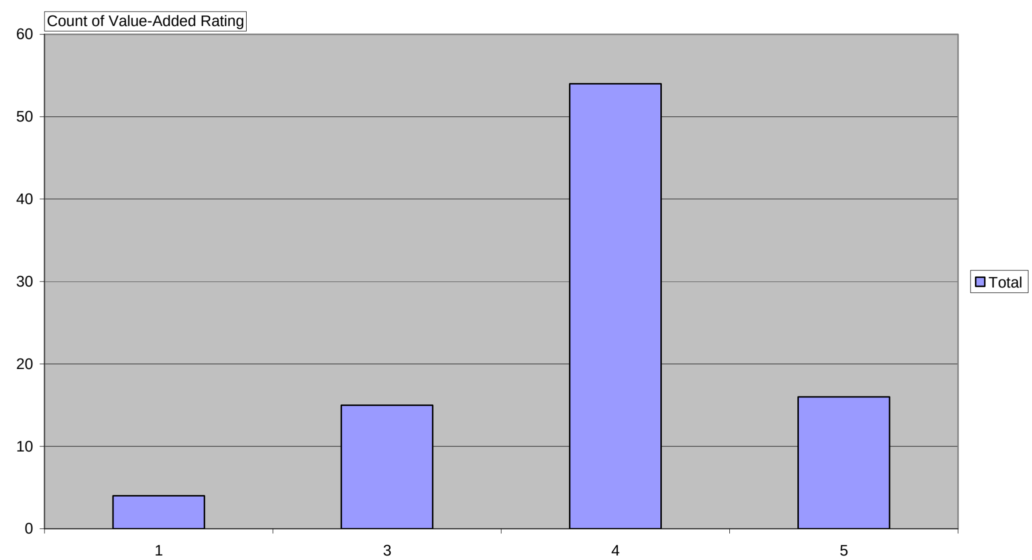


Figure 21: Count rating of users' confidence in their answers broken down by rating. Note than 1 for user rating means "Not confident at all", 2 means "Somewhat not confident", 3 "Neutral", 4 represents "Confident", and 5 represents "Very Confident".

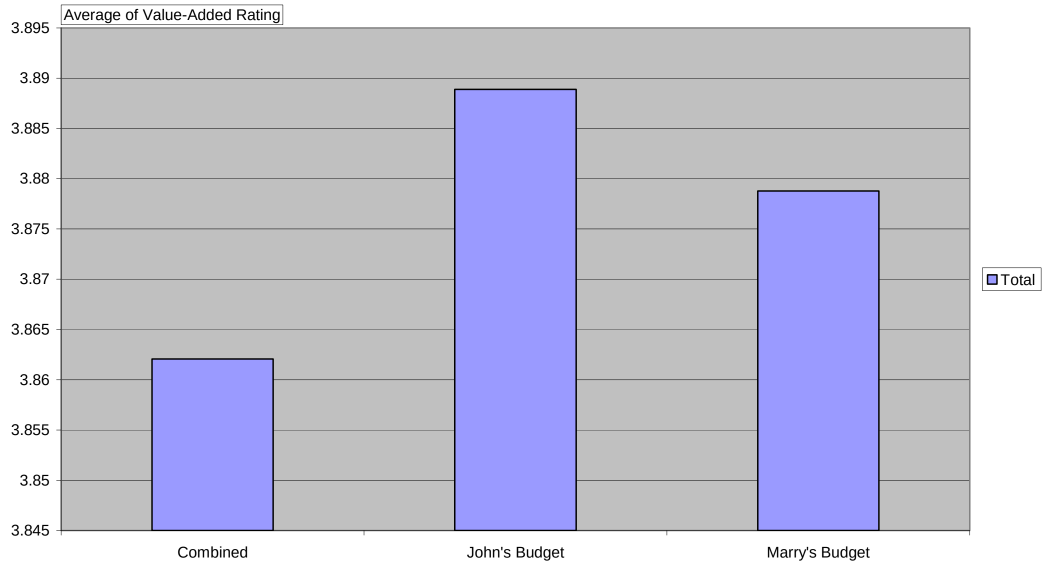


Figure 22: Average rating of users' confidence in their answers broken down by budget originator. Note than 1 for user rating means "Not confident at all", 2 means "Somewhat not confident", 3 "Neutral", 4 represents "Confident", and 5 represents "Very Confident".

6.4 Discussion

Hypothesis (A) was confirmed for all tasks for which the users were assessed, that is for both single budget data (John's or Mary's) and combination. Hypothesis (B) failed perhaps because the more information displayed the more the participants get overloaded.

As shown in Figures 19-21, more answers selected Budget A for the combined condition than for John's budget or Mary's budget alone. The value-added of Budget A is the lowest when participants were presented with combined visualizations. More answers selected Budget B for Mary's budget condition than for John's budget or combined. The value-added of Budget B is high in the combined and John's budget condition, while the value-added of Budget B is the lowest in Mary's budget condition, and thus hypothesis (B) does not hold.

The results show a user preference (or bias) for Budget A in the combined condition. The bias with Budget B is for Mary's budget. Further research is needed to better explain the existence of this kind of biases.

Value-added was task dependent in our experiment, which may make automated estimation of this dimension a complicated endeavor. A different behavior of value-added assessment is observed for all tasks, though none of them is simple averaging. The combination of the two datasets/images seems to affect slightly negatively the combination of data/images for the preferred Budget B; while for Budget A the combination negatively affect the level of value-added when compared to either John's budget or Mary's budget.

Most people are confident in their answers. Single budget condition seems to inspire more confidence then combination. For confidence in value–added ratings the combination of John's and Mary's budget produces a result that is the lowest of the individual confidence levels as shown in Figures 21 and 22.

CHAPTER 7

DISCOVERY

7.1 SIQ Assessment Principles

Following similar fields of science that depend on human behavior, such as psychology or human-computer interface, we distill our findings into a set of principles. For each task, an IQ professional would need to determine how these principles can be incorporated into the SIQ assessment particular to the data types and end-users.

"*People rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations*" (Tversky and Kahneman, [41], [42]).

A. Information integration does not necessarily improve the perceived accuracy.

A.1. Adding high quality information to existing data does not help people estimate the overall accuracy better.

A.2 Low-quality information integrated with target data helps people better estimate accuracy.

Our findings show that, for the temperature map integration, visualizations with single maps behave better than other types of visualizations, although no significant difference was found, on any of the dependent variables, between single-map and double maps in which one of the panels is 50% accurate. Moreover, users seem to perform better with additional inaccurate information than with completely accurate one.

In Chapter 6, the combination of the two datasets/images seems to affect slightly negatively the combination of data/images for the preferred Budget B; while for Budget

A the combination negatively affects the level of value added when compared to either John's budget or Mary's budget in the value-added study chapter 5.

B. Believability can suffer from an intrinsic bias.

The results show a user preference (or bias) for the 160 µg/ml task in the believability perception study (Chapter 5), and consequently, a bias against the 10 µg/ml. Most people and under most conditions believed the concentration to be 160 µg/ml more than 10 µg/ml.

Cognitive psychology and cognitive bias in psychology will help with the assessment of SIQ. Consider the answer to the following questions:

Why people think that almost accurate visualizations are worse than the ones that have 12% errors?

Why most people are confident in their answers?

Why the image seems to inspire more confidence then text?

"Cognitive psychology is the study of how people perceive, remember, think, speak, and solve problems". (Wikipedia, [34], [35], [41], [42])

"*Cognitive bias is a general term that is used to describe many distortions in the human mind that are difficult to eliminate and that lead to perceptual distortion, inaccurate judgment, or illogical interpretation*".(Wikipedia, [36],[37],[38])

Many of these biases are studied for how they affect belief formation, business decisions, and scientific research. The main biases types that apply and illustrate our results are: anchoring, attentional bias, confirmation bias, selective perception, and subjective validation.

Anchoring [38] is one of the cognitive bias types, that describes the common human tendency to rely too heavily, on one side or piece of information when making decisions. During normal decision-making, anchoring occurs when individuals overly rely on a specific piece of information to govern their thought-process.

Attentional bias [38], [39], [40] is an other type of cognitive bias, occurs due to an attentional bias. One example is when a person does not examine all possible outcomes when making a judgment about a correlation or association. They may focus on one or two possibilities, while ignoring the rest.

Confirmation bias [38] is the cognitive bias type that shows the people's tendency to favor information that confirms their preconceptions or hypotheses regardless of whether the information is true.

Selective perception "*it may refer to any number of cognitive biases in psychology related to the way expectations affect perception. For instance, several studies have shown that students who were told they were consuming alcoholic beverages (which, in fact, were non-alcoholic) perceived themselves as being "drunk," exhibited fewer physiological symptoms of social stress, and drove a simulated car similarly to other subjects who had actually consumed alcohol. The result is somewhat similar to the placebo effect*". (Wikipedia, [38], [43])

Subjective Validation [38], [43], [44]. This type of bias also called personal validation, and it is one of cognitive bias types by which a person will consider a statement or another piece of information to be correct if it has any personal meaning or importance to them. In other words, it is affected by personal beliefs and opinions.

C. Accuracy, believability, and Value added are not always interrelated dimensions.

It is entirely possible for some pieces of information to be believable but not necessary to be accurate or add value to a specific task. When the information is accurate and adds value to the targeted task, then the information becomes trusted and believed and the interrelationship between these dimensions is achieved and become significant.

D. Subjective Information quality is task and data dependent.

This principle stems from the nature of the human and problem factor, and it suggests that an IQ professional will have to sample some typical data integration tasks that are common with their data consumers. Based on the findings regarding bias and perception of accuracy, the IQ professional can develop a simple linear and threshold- based formula to estimate SIQ. This formula can then be automatically computed when users demand data to be integrated from various sources.

7.2 Other useful principles for improving IQ

E. It's better to be approximately right than exactly wrong (Carveth Read, [53])

F. Know your task and goals

G. Do not judge data based on your personal beliefs.

"*Many decisions based on beliefs*" (Tversky and Kahneman, [41], [42]).It is not always true that the believable information to be accurate or add value to a specific task.

H. The accurate information for the right task increases the value of the information and increases the believability specific task.

I. Use your logic effectively

If the visualization is overloaded, to simplify understanding and estimating the quality of what you are seeing:

- Start with the points you know and proceed to the points that you don't know in logical order.
- Look at the visualization as whole to part and back to a whole, link a particular point the general points available logically.
- Always starts with the simple points and proceeds to the more complex points this will help building your assumption gradually.

J. Divide task

To make better estimate the data quality, you should specify the dimension to be assessed, in case of the overall quality dimension you should assess each dimension individually.

Include all related factors/dimensions, and activities when making estimates

- Comparaisons (variables, data resources, etc.).
- Ask questions.

K. Look at the combined data

Look at the combined data and check the following variables or ask the following questions:

- Is the amount of information sufficient for the specified task?
- If the answer is no, then add more data, and look again

- Is the data is relevant to the specified task?
- Is the data meaningful?
- Is the data logically accepted?
- Is the data true and reliable for the specified task?
- Is the data true and credible for the specified task?
- Is the data beneficial for the specified task?

L. Revise assessment and refine results as you combine data

CHAPTER 8

CONCLUSIONS

8.1 Goals achieved

Our work on SIQ resulted in experimental results on how people perceive IQ and combinations of information with various levels of quality. We also distilled a set of rules/principles/techniques to allow simple, better estimation of SIQ. The rules can subsequently be used to estimate these quality dimensions and present them to a user.

The computed SIQ measure can be represented through visual representations in order to convey the Information Quality to the user. We could present the quality in its own window for both the individual data sources and the integrated data. Figures 23 and 24 illustrated a Treemap and bar chart of quality.

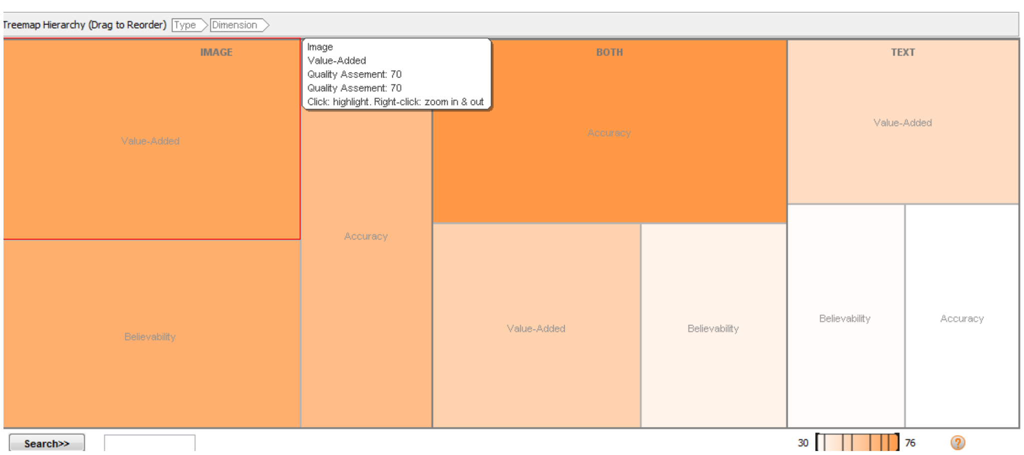


Figure 23:Treemap showing Information along with its quality

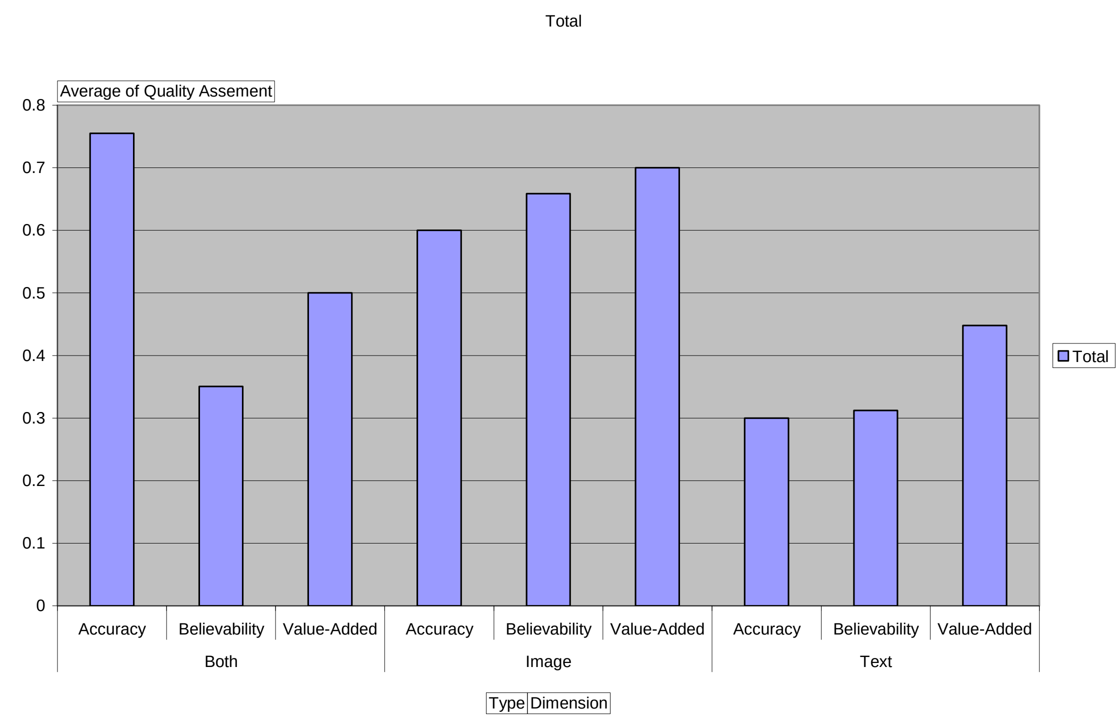


Figure 24 :Bar chart Information along with its quality

8.2 Research Findings Usage

This research would help and answer some of the introductory questions of the signaling theory by Michael Spence [29], [30], [31]. The signaling theory has been presented as a solution to the problem of adverse selection [49], [52], presented by George Akerlof's "The Market for Lemons" chapter [45], [46], which brought informational issues at the forefront of economic theory.

"Lemon market" effects have also been noted in other markets. The Information market is among those markets where Information Asymmetry is present, and data provenance can be from different sources with possibly unknown quality.

"Studies of signaling theory often examine the quality of the signaler (source of information). Future research would benefit from examining in more depth the various qualities signaled and more carefully linking the signals used to measure these qualities. At its essence, the link between a signal and the underlying quality represents a measurement issue that we call signal fit" (Connelly, BL, Certo, ST, Ireland, RD, & Reutzel, C. [31]).

Signaling theory examines the quality based on the reputation of the information source, so the quality is assumed to be known based on the signal, which is similar with this research because various aspects of information's quality are studied.

"The usefulness of a signal to the receiver depends on the extent to which the signal corresponds with the sought-after quality of the signaler (i.e., referred to as signal fit) and the extent to which signalers attempt to deceive (i.e., average honesty). Since both are required, some scholars define this combination as signal reliability, other scholars use the term credibility to describe the same notion, the extent to which the signaler is honest and the signal corresponds with signaler quality "(Davila, Foster, & Gupta, 2003 [33]).

There is common confusion between signal fit, honesty, reliability, and related terms, often using them interchangeably, but research into SIQ may help clarify their distinct underlying concepts of the signal fit as the three dimensions of SIQ: believability, accuracy, and value added. Our research could be linked to the signaling theory, especially if the signal are sent and received visually. The information provided through visualizations and the visualization itself will be the signal to a specific goal/task. The principles provided in this dissertation will help the receiver of the information visually to discover the quality of the signal veracity. Our work will help in understanding the characteristics of the SIQ of the signal (information), the sensitivity of people's perception of (Signal) Information Quality, and possible Subjective Information Quality thresholds. If the signals (information) provided are implausible or far from reality the signaling equilibrium will break down.

8.3 Future Work

Assessing information quality is not an easy task and requires knowledge and awareness of the subjective and objective information quality metrics. Further studies may focus on additional tasks to better understand the existence of preferences for and biases against tasks. Such investigations may also need to be determined for other data types, and data presentation methods.

Subjective assessment is not limited to accuracy, believability and value added, and our plans are to consider other SIQ dimensions and verify whether their behavior is similar to the subjective accuracy. Other dimensions that are inherently subjective, such as reputation, may lead to the development of a more complete theory of SIQ.

Any theory of SIQ may need to also consider the effect of data integration, an important topic in information and data quality. This research also showed that combining information/adding extra information is not always beneficial. Furthermore, for the cases when additional data is included, lower quality data may provide better support for subjective evaluation than higher quality data

REFERENCES

[1] Infovis, Georgia Institute of Technology. (2009, November 08).Retrieved November 08, 2009, from http://www.ic.gatech.edu/research/infovis, Georgia Institute of Technology

[2] Semvast, University of Maryland HCI Lab Visualization Projects. (2009, November 08).Retrieved November 08, 2009, from http://www.cs.umd.edu/hcil/semvast/

[3] Jean Scholtz, *Developing Qualitative Metrics for Visual Analytic Environments*, Pacific Northwest National Laboratory Richland, WA, http://www.beliv.org/chapters/

[4] Yair Wand and Richard Y.Wan g, *Anchoring Data Quality Dimensions Ontological Foundations,* 1996.

[5] Tatiana Tekusová , Martin Knuth, Tobias Schreck, Jörn Kohlhammer, *Data Quality Visualization for Multivariate Hierarchic Data*, 2008.

[6] Alex Pang, *Visualizing Uncertainty in Geo-spatial Data*, 2001.

[7] Alan M. MacEachren, *Visualizing Uncertain Information, published in Cartographic Perspective*, Number 13, fall, 1992, pp. 10-19.

[8] Stanford Entity Resolution Framework. , (2010, May 15).Retrieved May 15, 2010, from http://infolab.stanford.edu/serf/).

[9] Stanford Trio Project. (2010, May 20).Retrieved May 15, 2010, from (http://infolab.stanford.edu/Trio).

[10] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang, *Data Quality Assessment* Communications of the ACM, 2002.

[11] MO Lin, ZHENG Hua , *A method for measuring data quality in Data Integration, International Seminar on Future Information Technology and Management Engineering*, 2008.

[12] Wang, R. Y. and Strong, D. M. *Beyond accuracy: what data quality means to data consumers*. Journal of Management Information Systems 12, 4 (1996), 5-34.

[13] Ismael Caballero, Eugenio Verbo, Coral Calero, Mario Piattini, *A data quality measurement information model based on ISO/IEC 15939*, (Research-in-Progress).

[14] Verónika Peralta, Raúl Ruggia, Zoubida Kedad, Mokrane Bouzeghoub, *A Framework for Data Quality Evaluation in a Data Integration System*, 2004.

[15] Amihai.M, Igor.R, *Estimating the Quality of Data in Relational Databases*, 1996.

[16] Donald P. Ballou , InduShobha N. Chengalur-Smit , Richard Y. Wang, *Sample Based Quality Estimation of Query Results in Relational Database Environments*, IEEE Transactions on Knowledge and Data Engineering ,(May 2006) .

[17] Giri Kumar Tayi and Donald P. Ballou, Guest Editors, *Examining Data Quality*, 1998.

[18] Many_Eyes (2010 October 10). *In Wikipedia, the free encyclopedia* .Retrived October 10, 2010 from http://en.wikipedia.org/wiki/Many_Eyes.

[19] National Oceanic and Atmospheric Administration, (2010, July 08).Retrieved July 08, 2010, from www.nws.noaa.gov

[20] Amazon web services website, (2010, July 15).Retrieved July 15, 2010, from http://aws.amazon.com/mturk/

[21] Nicolas.P, Madnick.S.E, 2008. "*Measuring Data Believability: A Provenance Approach*," Working chapters 40086, Massachusetts Institute of Technology (MIT), Sloan School of Management

[22] Huerta, Esperanza and Ryan, Terry, "*The Credibility of Online Information*" (2003). AMCIS 2003 Proceedings. Chapter 279.http://aisel.aisnet.org/amcis2003/279

[23] Lee, Strong, Khan, and Wang, AIMQ: *A Methodology for Information Quality Assessment,” Information Management*, December2002, Volume 40, Issue 2, pp.133-146.

[24] Petty, R.E. & Cacioppo, J.T. (1986). *The Elaboration Likelihood Model of persuasion*. New York: Academic Press.

[25] Bobrowski M, Marre M, Yankelevich D. A homogeneous framework to measure data quality. In Proceedings of the International Conference on Information Quality (IQ), pages 115-124, Cambridge, MA, 1999.

[26] The Budget. (2011, June 10).Retrieved June 10, 2011, from Office of Management and Budget, http://www.whitehouse.gov/omb/

[27] Value added (2011, August 29). *In Wikipedia, the free encyclopedia* Retrieved August 29, 2011, from Wikipedia:http://en.wikipedia.org/wiki/Value_added

[28] Genuine, M. (2011, April 20). *Miller’s Genuine Draft*.Retrived April 20,2011, from http://millergd.blogspot.com/2011/04/interactive-chart-on-ryan-budget-vs.html

[29] Michael Spence (1973). "*Job Market Signaling*". Quarterly Journal of Economics (The Quarterly Journal of Economics, Vol. 87, and No. 3) 87 (3): 355–374. doi: 10.2307/1882010. JSTOR 1882010.

[30]Michael Spence (2002). "*Signaling in Retrospect and the Informational Structure of Markets*". [30] American Economic Review 92 (3): 434–459. doi: 10.1257/00028280260136200.

[31] Connelly, BL, Certo, ST, Ireland, RD, & Reutzel, C (2011). Signaling Theory: A Review and Assessment. Journal of Management 37(1):39-65.

[32] Cohen, B. D., & Dean, T. J. 2005. *Information asymmetry and investor valuation of IPOs: Top management team legitimacy as a capital market signal*. Strategic Management Journal, 26: 683-690.

[33] Davila, A., Foster, G., & Gupta, M. 2003. *Venture capital financing and the growth of startup firms*. Journal of Business Venturing, 18: 689-708.

[34] Cognitive Psychology (2011, October 15). *In Wikipedia, the free encyclopedia*.Retrived October 15, 2011, from http://en.wikipedia.org/wiki/Cognitive_psychology

[35] Neisser, U. (1967). *Cognitive psychology*. New York, NY: Meredith.

[36] Plous, Scott (1993), *the Psychology of Judgment and Decision Making*, New York: McGraw-Hill, ISBN 0-07-050477-6.

[37] Hardman, David (2009), *Judgment and decision making: psychological perspectives*, Wiley-Blackwell, ISBN 9781405123983

[38] Cognitive biases, (2011, October 15). *In Wikipedia, the free encyclopedia*.Retrived October 15, 2011, from http://en.wikipedia.org/wiki/List_of_cognitive_biases

[39] (Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M.J., & van IJzendoorn, M.H. (2007). *Threat-related attentional bias in anxious and non-anxious individuals*: A meta-analytic study. Psychological Bulletin.

[40] Schoth, D.E., & Liossi, C. (2010). *Attentional bias towards pictorial representations of pain in individuals with chronic headache.* The Clinical Journal of Pain. 26 (3): 244–250.

[41]Tversky, A., & Kahneman, D. (1973). *Availability: A heuristic for judging frequency and probability*. Cognitive Psychology, 5, 207-232.

[42] Kahneman, D., & Tversky, A. (1972). *Subjective probability: A judgment of representativeness*. Cognitive Psychology, 3, 430-454.

[43] Forer, B.R. (1949) "*The Fallacy of Personal Validation: A classroom Demonstration of Gullibility,*" Journal of Abnormal Psychology, 44, 118-121.

[44] Hastorf, A.H. & Cantril, H. (1954). *They saw a game: A case study*. Journal of Abnormal and Social Psychology, 49, 129-134

[45] Hoffer, George E.; Pratt, Michael D. (1987). "*Used vehicles, lemons markets, and Used Car Rules: Some empirical evidence*". Journal of Consumer Policy 10 (4): 409–414. doi: 10.1007/BF00411482.

[46] Akerlof, George A. (1970). "*The Market for 'Lemons': Quality Uncertainty and the Market Mechanism*". Quarterly Journal of Economics (The MIT Press) 84 (3): 488–500. doi: 10.2307/1879431. JSTOR1879431.

[47] Aboody, David; Lev, Baruch (2000). "*Information Asymmetry, R&D, and Insider Gains*". Journal of Finance 55 (6): 2747–2766. doi:10.1111/0022-1082.00305.

[48] Joseph Stiglitz .(2011,August). *In Wikipedia, the free encyclopedia*.Retrived August 2011, from http://en.wikipedia.org/wiki/Joseph_Stiglitz

[49] Polborn, M. K.; Hoy, M.; Sadanand, A. (2006). "*Advantageous Effects of Regulatory Adverse Selection in the Life Insurance Market*". Economic Journal 116 (508): 327–354. doi:10.1111/j.1468-0297.2006.01059.x.

[50] Cardon, J. H.; Hendel, I. (2001). "*Asymmetric information in health insurance: evidence from the National Medical Expenditure Survey*". Rand Journal of Economics 32 (3): 408–427. JSTOR 2696362.

[51] Busenitz, L. W., Fiet, J. O., & Moesel, D. D. 2005. *Signaling in venture capitalist–new venture team funding decisions: Does it indicate long-term venture outcomes? Entrepreneurship Theory and Practice*, 29: 1-12.

[52] Dionne, G.; Gouriéroux, C.; Vanasse, C. (2001). "*Testing for Evidence of Adverse Selection in the Automobile Insurance Market: A Comment"*. Journal of Political Economy 109 (2): 444–453. doi: 10.1086/319557.

[53] Read, C. (2006*). Logic Deductive and Inductive*. New York, NY: C. Susan Skinner. Retrieved from Project Gutenberg: `http://www.gutenberg.org`

[54] Klein, B.D.,"Internet Data Quality: Perceptions of Graduate and Undergraduate Business Student," *Journal of Business and Management*, vol.8, no. 4, 2002, pp.425-432.

[55] Klein, B.D.,"User Perception of Data Quality: Internet and Traditional Text Sources," *Journal of Computer Information Systems*, vol.41, no. 4, 2001, pp.9-15.

APPENDIX A

- Abuhalimeh, A., Tudoreanu, M.E. 2011. "Subjective Evaluation of Value-Added in Visualization of Data. To appear in SigIQ workshop in Shanghai, China.
- Abuhalimeh, A., Tudoreanu, M.E., and Mustafa, T. 2011. "How People Evaluate Believability in Visualization of Data. To appear in SigIQ workshop in Shanghai, China.
- Abuhalimeh, A., Tudoreanu, M.E., and Mustafa, T. 2011. "Subjective Evaluation of Perception of Believability in Visualization of Data". To appear in ICIQ Conference proceeding (Adelaide, Australia).
- Ahmed Abuhalimeh, M. Eduard Tudoreanu, and Erich A. Peterson, "How People Perceive The Accuracy of Information ", AIP Conf. Proc. 1326, 74 (2011), DOI:10.1063/1.3587460
- Abuhalimeh, A., Tudoreanu, M.E., and Peterson, E. 2010. "Subjective Evaluation of Perception of Accuracy in Visualization of Data". Proceedings ICIQ Conference (Little Rock, AR), pp. 112-123.