December 2006

Uncertainty in the Information Supply Chain: Integration of Multiple Data Sources

Monica Tremblay
University of South Florida

Uncertainty in the Information Supply Chain: Integration of Multiple Data Sources

Monica Chiarini Tremblay
Information Systems and Decision Sciences
University of South Florida
mtrembla@coba.usf.edu

ABSTRACT

Knowledge workers often deal with multiple sources of data when acquiring information for decision making. Multiple data sources are valuable for knowledge creation, but deciding how to integrate and analyze these different data sources is difficult. The data acquisition process and the task of correctly combining and manipulating data these data in the "information supply chain" have challenges: data are unbounded, have different data definitions, and there is no guarantee of its quality. This study utilizes the design science guidelines proposed by Hevner et al. (2004) to design, develop, and evaluate an agent-based information system to aid effective knowledge creation from the information supply chain in the context of health planning.

Keywords

Knowledge management, business intelligence, agent-based modeling, healthcare planning, decision support systems

INTRODUCTION AND MOTIVATION

Knowledge workers draw on a set of pre-existing tools when acquiring data from multiple data sources available from the information supply chain (Berndt 2006). The data acquisition process and the task of correctly combining and manipulating data from multiple data sources in the information supply chain have challenges: data are unbounded, have different data definitions, and there is no guarantee of its quality. In most cases, knowledge workers make decisions with available information and use "gut instinct" or experience to choose the correct course of action when data sources conflict or do not match expectations. Yet, decisions must be made with the data available. These challenges are made even more complicated by the knowledge worker's own judgment biases. Existing tools can aid knowledge workers, yet the lack of integration among these tools (Davenport, Jarvenpaa and Beers 1996) aggravate cognitive and behavioral biases and result in missed opportunities for knowledge creation.

This study identifies a knowledge management gap between the desktop inventory available to knowledge workers and the demands of interacting with multiple sources of data. The design science guidelines identified by Hevner et al. (2004) are utilized to design, develop and evaluate an agent-based information system (ABIS) that supports the information supply chain. The ABIS will have multiple functionalities: it will compare and select relevant data sources, filter important information, and provide the following data quality metrics: volatility and completeness. The ABIS will expose the impact of judgment biases, in particular: biases due to insensitivity to sample size, biases due to imaginability and anchoring and adjusting.

The context for this research emerges from a longitudinal study conducted by the researcher on the use of a BI tool by a health planning agency in the state of Florida (Tremblay, Fuller, Berndt and Studnicki 2006). Knowledge workers in organizations that collect data used in creating public policies deal with extremely unstructured tasks and the outcome of this data collection can have an important impact on society. This knowledge work is data intensive and involves the information supply chain that is the focus of this study.

The remainder of the paper is structured as follows: a brief review of the foundational theories that will inform the design of the ABIS are outlined, followed by the research questions for this thesis, a description of the research methodology and concludes with the anticipated contributions.

FOUNDATION THEORIES

Dealing with multiple data sources on the desktop is comparable to dealing with data streams, in that methods for comparing data are fundamental operations for knowledge workers: data are unbounded, have different data definitions, and have no guarantee of its quality. Ongoing research by computer scientists in the area of data streams and data quality help guide the design of the ABIS proposed by this study.

Data Streams

Today's data often arrives in streams: network traffic data, click stream data from websites, transactions from retail chains, and ATM card operations are a few examples. Data streams are defined as a possibly unbounded sequence of data items. Often businesses will need to analyze stream data sooner than is possible with the current model of data warehouse storage and off-line analysis (Chaudry, Shaw and Abdelguerfi 2004). This poses architectural and functional challenges when querying this type of data. Two important challenges Chaudry (2004) outlines are:

- Queries need to include a notion of time since the data is continuous.
- Since there is no quality guarantee, the queries must compute answers based on unreliable data.

Presently, computer scientists outline algorithms to handle data streams, with a focus on how to manage, process, and query these data sources. The most common approach is to attach a timestamp or sequence number to each arriving stream item (Arasu, Babcock, Babu, Cieslewicz, Datar, Ito, Motwani, Srivastava and Widom 2004). Another approach is to "window" the data streams into subsets of data (Chaudry 2004; Li, Maier, Papadimos, Tucker and Tufte 2004; Moon, Lopez and Immanuel 2003)

An additional challenge is identifying the fundamental operations for managing data streams. Several authors offer extensions to SQL or methods of "punctuating" the streams (Arasu et al. 2004; Cormode, Datar, Indyk and Muthukrishnan 2003; Tucker, Maier, Sheard and Fegaras 2003).

Data Quality

In the information supply chain, proprietary data assets can be seen as off the shelf *data products*. Just as a consumer purchasing an off-the-shelf product wishes to know information about the product (such as the ingredients, instructions for use, or date of expiration), data consumers should be informed about the quality of data products (Wang, Reddy and Gupta 1993).

Wang has described more then 179 data quality attributes, several whose definitions overlap. This study concentrates on two well known metrics of data quality: volatility and completeness. Volatility describes the data's propensity to change. Completeness describes how many missing data items we have.

Judgment under Uncertainty

The general heuristics and biases that people use in making judgments are well researched. Though this study is mainly interested in strategies of data retrieval and representation that minimize these biases, it is important to understand the heuristics knowledge workers may use for decision making, as well as the possible biases that could result from the use of these heuristics. Heuristics are based on past experience and generally give good results, but they can also lead to severe and systematic errors (Tversky and Kahneman 1982). Tversky and Kahneman identify three heuristics that are used to access probabilities of an event that lead to biases in decision making: representativeness, availability, and anchoring and adjusting. This study investigates one form of bias for each type of heuristic.

1. Representativeness (insensitivity to sample size) - If an event appears similar to a past experience or event it is judged to belong to that event. In some cases this may result in an accurate classification of an event; but often the decision maker overlooks factors that should be considered, for example sample size or sample distribution.
2. Availability (imaginability) - refers to the tendency to retrieve information that is plausible without regard to its probability. In health planning, this could lead to incorrectly inflating the probability of event due to their imaginability, and the adoption of a very conservative approach toward prevention even in the face of highly unlikely events (Tracey and Rounds 1999).
3. Anchoring and Adjustment - Individuals tend not to sway (adjust) too far from initial information or impressions (their anchor), even when presented with very different information (Tracey and Rounds 1999).

RESEARCH QUESTIONS

This study investigates the management of multiple data sources from the information supply chain. Two research questions are addressed: the first focuses on the design of the artifact, and the second its evaluation.

1. What is the design (functionality, interface and structure) of an agent-based information system that will automate and support effective integration of multiple data sources in the information supply chain?
2. How do we evaluate the utility, quality, and efficacy of this agent-based information system?

RESEARCH MODEL

Figure 1 outlines the research model for this study. This model is based on Hevner et al.'s (2004) framework for information systems research. Utilizing the design research cycle an artifact is *built* with the intention to solve an identified organizational problem and is *evaluated* in an appropriate context to both provide feedback to the design process and a better understanding of the process (Hevner, March, Park and Ram 2004).

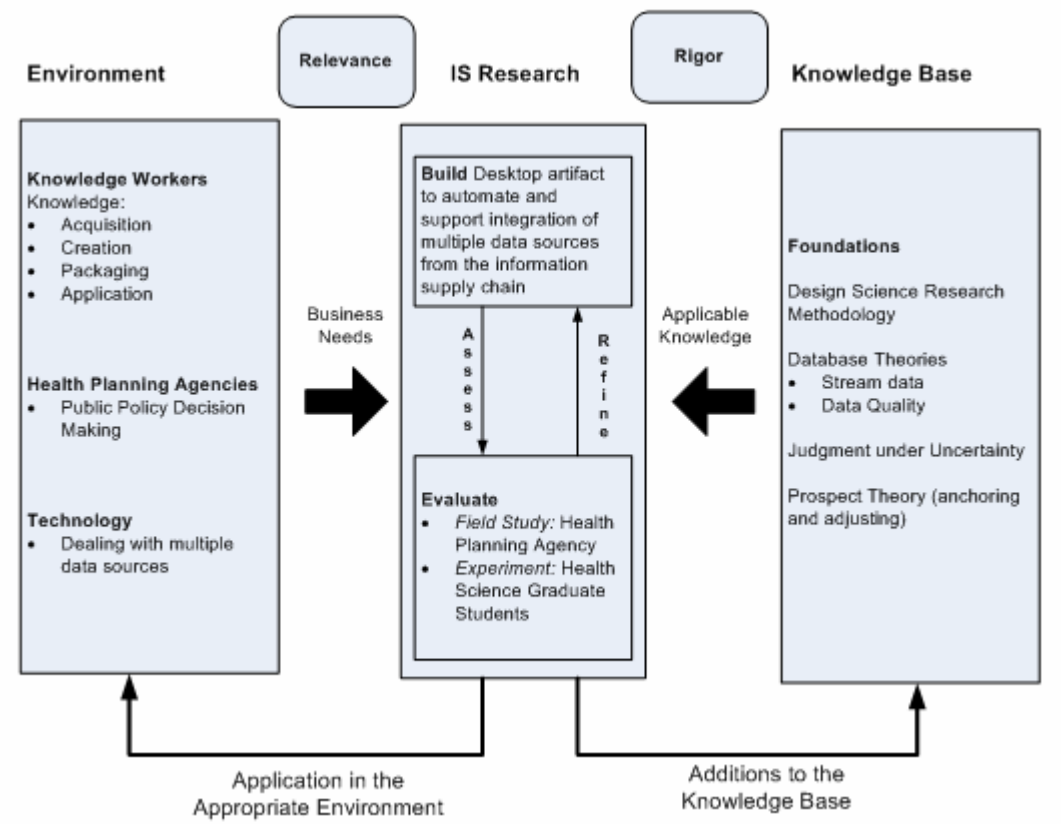

Figure 1 - Research Framework (adapted from Hevner et al. 2004)

The artifact is an *instantiation* (Hevner et al. 2004; Markus, Majchrzak and Gasser 2002) of an ABIS which will aid health planners in the process of acquisition and manipulation of data from the information supply chain. The design will be informed both by database theories and behavioral decision making theories. The ABIS will be evaluated in two phases.

Proposed Artifact Development

The ABIS will have multiple functionalities. In order to deal with multiple data sources the ABIS will compare and select relevant data sources, filter important information, and provide users with the following data quality metrics: volatility and completeness. The ABIS will also minimize the impact of judgment biases (in particular: biases due to insensitivity to sample size, biases due to imaginability and anchoring and adjusting).

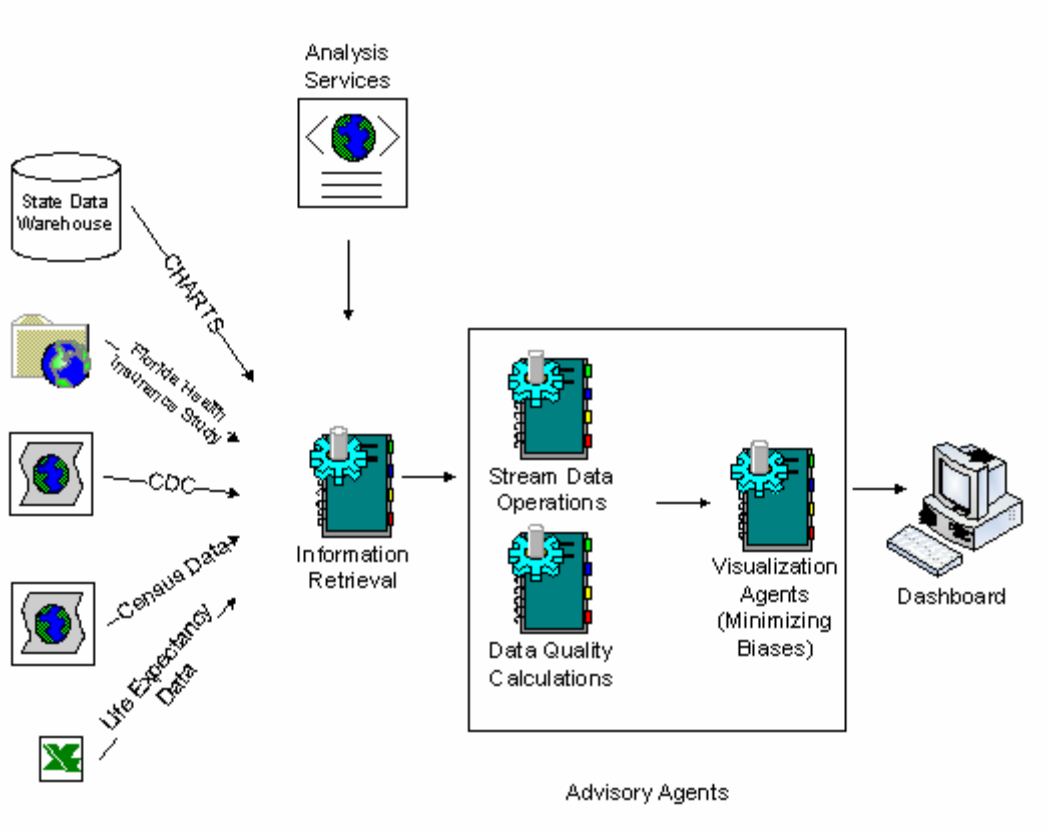

Figure 2 - Artifact Design

Nissen (2001, 2006) describes three classes of agents: information retrieval agents, advisory agents and performative agents. This study does not focus on performative agents, since the proposed ABIS will not make decisions. The ABIS (see Figure 2) will be composed of software agents that fall in two classes: *information retrieval agents*, and *advisory agents* (Nissen 2001; Nissen and Sengupta 2006). Information retrieval agents will focus on collecting all needed data available at the point of inquiry. They will guided by the knowledge workers as to what data to retrieve, and will form retrieval sets. Advisory agents aid users in forming consideration sets.

Design Evaluation

The first phase of evaluation will be several iterations with health planners at a local health planning agency. This will help the researcher better understand the problem and will provide feedback for improvement of the design of the artifact (Hevner et al. 2004; Markus et al. 2002). The artifact will then be evaluated using a controlled experiment.

The experiment is designed to be similar to the process health planners follow to collect data needed in the application of a grant. Subjects will be recruited from a graduate program in health sciences (the intention is to make this task part of class project in a public health planning class). Some participants will receive the complete tool, while some will receive a tool without advanced functionality. This task will utilize realistic data, but the data is "seeded" with potential pitfalls for the subject to make a sub-optimal decision (see Table 1 for examples). As is the case in health planning, the participants will be provided with several sources of data. The participant's experience and pre-existing biases regarding the task (self reported on survey) and cognitive ability (existing instrument) will be measured prior to the experiment. The experiment will consist of three tasks:

1. Prioritization Task: Subjects will be asked to rank the most important health issues in a specific geographical area (referred to as a community) based on the data available to them through the tool. An accurate ranking would reflect that the subject utilized the quality metrics and potential bias information.

2. Synthesis Task: Subjects will be asked to write an executive summary reporting on the health status of the community, sustaining claims with data provided by the tool. The summary will be coded by two independent raters/coders. The arguments provided by the subjects should enrich or help explain results of the first task.
3. Elaboration Task: Subjects will select a single health issue and elaborate on its importance, using supporting data. This qualitative data will also be coded by two independent raters/coders. The data they will use to elaborate on an issue will not have "seeded" pitfalls, but natural existing problems that are found in this type of data.

Issue	Factor	Pitfall
Biases	Insensitivity to sample size	There is natural variation in the rate of diseases, with some being more widespread, and some being rarer. For example, data sources may report occurrence rates for a community. A rate of 10 per 10,000 inhabitants should be more alarming for a rare infectious disease like tuberculosis than for a more common event such as heart disease.
Data Quality	Completeness	Hospital discharges occur continuously. Hospitals are continuously collecting data; but they may differ in their batching and transmission strategies. Data may arrive out of order (for example one hospital transmits once a year, while another may do so monthly), or some hospitals may send incomplete data, filling in information with later transmissions. Knowledge workers should be cognizant of quality metrics about this data.

Table 1 - Example of Pitfalls

For each task time spent (from system log), and the participant's confidence and satisfaction with the results (using established instruments) will be captured. The following hypotheses are proposed (see also Figure 3):

H1: *Knowledge workers, in particular, health care planners who use this agent-based information system will be more effective and efficient in health planning activities than those not using the system, in that:*

a) Their task outcome will be more accurate

b) They will spend less time on a task

c) They will be more confident with their results

d) They will be more satisfied with their results

H2: *The provision of support for potential biases, in particular biases due to insensitivity to sample size, biases due to imaginability and biases due to anchoring and adjusting will make knowledge workers, in particular, health care planners who use this agent-based information system more effective and efficient in health planning activities than those not using the system, in that:*

a) Their task outcome will be more accurate

b) They will be more confident with their results

c) They will be more satisfied with their results

H3: *The provision of the following data quality metrics: volatility and completeness will make knowledge workers, in particular, health care planners who use this agent-based information system more effective and efficient in health planning activities than those not using the system, in that:*

a) Their task outcome will be more accurate

b) They will be more confident with their results

c) They will be more satisfied with their results

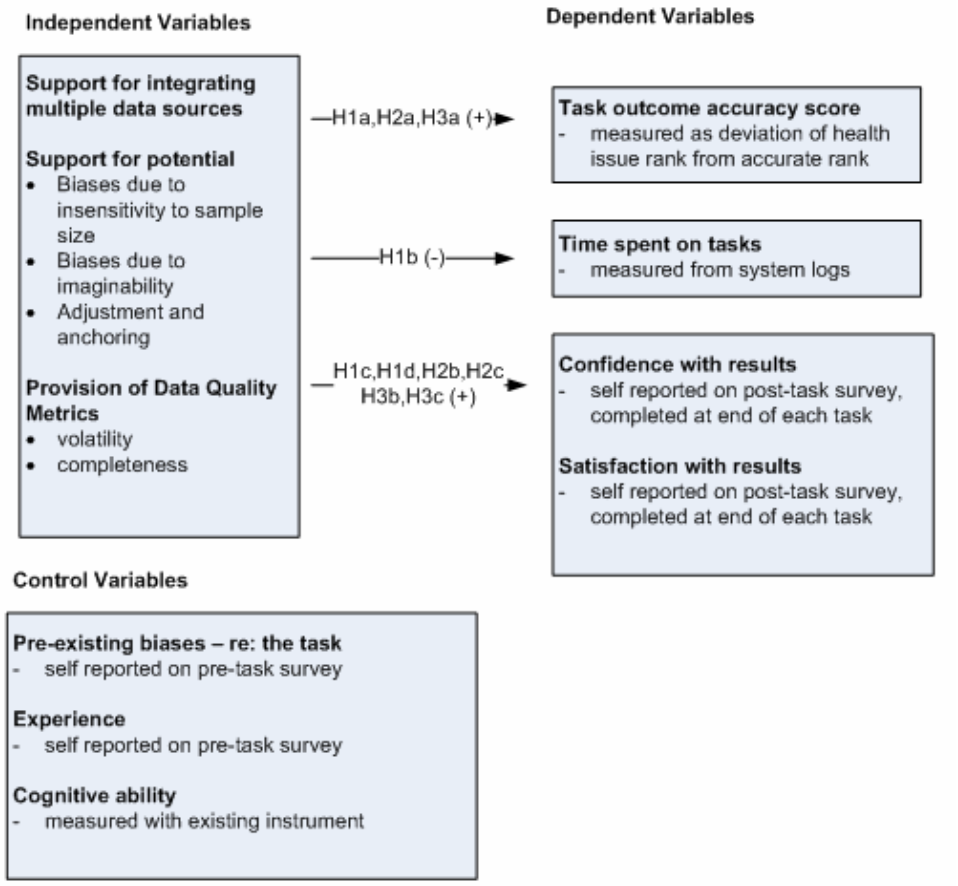

Figure 3 - Hypotheses –Phase 2 Evaluation of Artifact

ANTICIPATED CONTRIBUTIONS

Building and evaluating the artifact with the aid of a health planning agency allows the researcher to identify and refine how to deal with multiple data streams and minimize the selected judgment biases in a realistic setting, making this research relevant both to theory and practice. The application of algorithms from data stream theories to real-world, noisy data should uncover some new issues and potential resolutions which will extend the chosen algorithms. The ABIS will also provide two of the metrics outlined in Wang et al. (1993): volatility and completeness. Deciding how to calculate these data quality measures when dealing with data from multiple sources will be a contribution to both data stream and data quality theories. Studying the effect of the provision of these data quality metrics will inform the practitioner community.

REFERENCES

1. "http://hcecf.org/hcabout.html."
2. Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Datar, M., Ito, K., Motwani, R., Srivastava, U., and Widom, J. "STREAM: The Stanford Data Stream Management System," in: *Stream Data Management (Advances in Database Systems)*, N.A. Chaudry, K. Shaw and M. Abdelguerfi (eds.), Springer Science+Business Media, Inc, 2004.
3. Berndt, D. "Next-Generation Software Engineering: Challenges in Data and Knowledge Management," Next Generation Software Engineering Workshop at HICSS, Hawaii, 2006.
4. Chaudry, N.A. "Introduction to Stream Data Management," in: *Stream Data Management*, N.A. Chaudry, K. Shaw and M. Abdelguerfi (eds.), Springer Science+Business Media, Inc, 2004, pp. 1-13.
5. Chaudry, N.A., Shaw, K., and Abdelguerfi, M. "Preface," in: *Stream Data Management*, N.A. Chaudry, K. Shaw and M. Abdelguerfi (eds.), Springer Science+Business Media, Inc, 2004, pp. xiiv-xiv.

6. Chen, A.N.K., and Edgington, T.M. "Assessing Value in Organizational Knowledge Creation: Considerations for Knowledge Workers," *MIS Quarterly* (29:2), Jun 2005, p 279.
7. Cormode, G., Datar, M., Indyk, P., and Muthukrishnan, S. "Comparing Data Streams Using Hamming Norms (How to Zero In)," *IEEE Transactions on Knowledge and Data Engineering* (15:3), May/June 2003 2003, pp 529-540.
8. Davenport, T.H., Jarvenpaa, S.L., and Beers, M.C. "Improving knowledge work processes," *Sloan Management Review* (37:4), Summer 1996, p 53.
9. Derose, S.F., and Petitti, D.B. "Measuring Quality of Care and Performance from a Population Health Care Perspective," *Annual Review of Public Health* (24:1) 2003, pp 363-384.
10. Derose, S.F., Schuster, M.A., Fielding, J.E., and Asch, S.M. "Public Health Quality Measurement: Concepts and Challenges," *Annual Review of Public Health* (23:1) 2002, pp 1-21.
11. Einhorn, H.J., and Hogarth, R.M. "Behavioral Decision Theory: Processes of Judgment and Choice," *Annual Review of Psychology* (32:1) 1981, pp 53-88.
12. Friede, A., Blum, H.L., and McDonald, M. "Public Health Informatics: How Information-Age Technology Can Strengthen Public Health," *Annual Review of Public Health* (16:1) 1995, pp 239-252.
13. Guha, S., Meyerson, A., Mishra, N., Motwani, R., and O'Callahan, L. "Clustering Data Streams: Theory and Practice," *IEEE Transactions on Knowledge and Data Engineering* (15:3), May/June 2003, pp 515-528.
14. Hevner, A., March, S., Park, J., and Ram, S. "Design Science Research in Information Systems," *Management Information Systems Quarterly* (28:1), Mar 2004, pp 75-105.
15. Iezzoni, L.I. "Assessing Quality Using Administrative Data," *Ann Intern Med* (127:8_Part_2), October 15 1997, pp 666-674.
16. Li, J., Maier, D., Papadimos, V., Tucker, P., and Tufte, K. "Evaluating window aggregate queries over streams, Technical Report."
17. Maibach, E., and Holtgrave, D.R. "Advances in Public Health Communication," *Annual Review of Public Health* (16:1) 1995, pp 219-238.
18. Markus, M.L., Majchrzak, A., and Gasser, L. "A Design Theory for Systems that Support Emergent Knowledge Processes," *MIS Quarterly* (26:3), September 2002, pp 179-212.
19. Moon, B., Lopez, I.F.V., and Immanuel, V. "Efficient Algorithms for Large-Scale Temporal Aggregation," *IEEE Transactions on Knowledge and Data Engineering* (15:3), May/June 2003, pp 744-759.
20. Nissen, M.E. "Agent-Based Supply Chain Integration," *Information Technology and Management* (2:3), July 2001, pp 289-312.
21. Nissen, M.E., and Sengupta, K. "Incorporating Software Agents Into Supply Chains: Experimental Investigation With A Procurement Task," *MIS Quarterly* (30:1), March 2006, pp 145-165.
22. O'Connell, T.C., and Stearns, R.E. "Mechanism design for software agents with complete information," *Decision Support Systems* (39:2), Apr 2005, p 197.
23. Sikora, R., and Shaw, M.J. "A multi-agent framework for the coordination and integration of information systems," *Management Science* (44:11), Nov 1998, pp S65-S78.
24. Tracey, T.J., and Rounds, J. "Inference and Attribution Errors in Test Interpretation," in: *Test interpretation: Integrating science and practice*, R.K. Goodyear and J.W. Lichtenberg (eds.), Allyn & Bacon, Boston, 1999.
25. Tremblay, M.C., Fuller, R., Berndt, D., and Studnicki, J. "Doing More with More Information: Changing Healthcare Planning with OLAP Tools," *Decision Support Systems* (In Press) 2006.
26. Tucker, P.A., Maier, D., Sheard, T., and Fegaras, L. "Exploiting Punctuation Semantics in Continuous Data Streams," *IEEE Transactions on Knowledge and Data Engineering* (15:3), May/June 2003, pp 555-568.
27. Tversky, A., and Kahneman, D. "Judgment Under Uncertainty:Heuristics and Biases," in: *Judgment under uncertainty:Heuristics and biases*, D. Kahneman, P. Slovic and A. Tversky (eds.), Cambridge University Press, Cambridge, 1982.
28. Wang, R., Reddy, M.P., and Gupta, A. "An Object-Oriented Implementation of Quality Data Products," WITS-93, Orlando, Florida, 1993.
29. Whyte, W.F., and Whyte, K.K. *Learning from the field: a guide from experience* Sage Publications, Beverly Hills, 1984, p. 295.