# Truth, Lies, and Data: Credibility Representation in Data Analysis

James Schaffer*, Tarek Abdelzaher†, Debra Jones§,
Tobias Höllerer*, Cleotilde Gonzalez‡, Jason Harman‡ and John O'Donovan*
*Department of Computer Science, University of California Santa Barbara, Santa Barbara, California 93106
Email: {james_schaffer, holl, jod}@cs.ucsb.edu
†Department of Computer Science, University of Illinois at Urbana Champaign, Urbana, IL 61801
Email: zaher@cs.uiuc.edu
‡Department of Social and Decision Sciences
Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
Email: {coty, jharman}@cmu.edu
§SA Technologies Inc, 751 Hebron Parkway Suite 310, Lewisville, Texas 75057
Email: debra@satechnologies.com

***Abstract*—The web has evolved in a scale free manner, with available information about different entities developing in different forms, different locations, and at massive scales. This paper addresses the cognitive limitations that information analysts typically experience as they approach the boundaries where automated analysis algorithms are sorely needed. An experiment is conducted to explore information analysts' interactions with recommendations from an automated fact-finder algorithm during the task of answering questions in a fictional humanitarian aid delivery scenario. An experiment (N=285) is performed using three increasingly complex user interfaces, with and without the presence of the automated recommendations. Results show that in the best performing group, interaction with the fact-finder recommendations was 47 percent greater than the worst performing group.**

## I. Introduction

Current web technology provides rapid generation and collection of information from diverse sources. Accordingly, the amount of information available to decision makers has become too large to be efficiently and effectively analyzed without the help of automated tools. Finding the most relevant, reliable, and *credible* information on which to base a decision can be a daunting, time consuming task even with automated approaches. An expert analyst is often the best judge of assessing which data is relevant in an information seeking task, especially when the incoming information is noisy or unexpected. Interactive visual interfaces have been employed to help decision makers establish the parameters that effectively tailor information to a set of criteria, but practical limitations of an analyst's attention and the increasing size of available data reinforce the need for automation. Automatically generated recommendations about credibility (corroboration with ground truth) of individual information reports and reliability (propensity to produce credible information) of information sources has the potential to improve the analysis process. Unfortunately, automated processing of data often leaves the user out of the loop and inhibits data understanding due to the complexity of data mining algorithms. Finding optimal combinations of automated assessments of information credibility and human assessments of information relevance remains a key challenge in large scale data analysis.[15][16][7], human-computer interaction [4][5], and visual data mining [6]. Figure 1 shows an overview of the recommendation pipeline and associated cognitive evaluation framework that was used in our experiments.

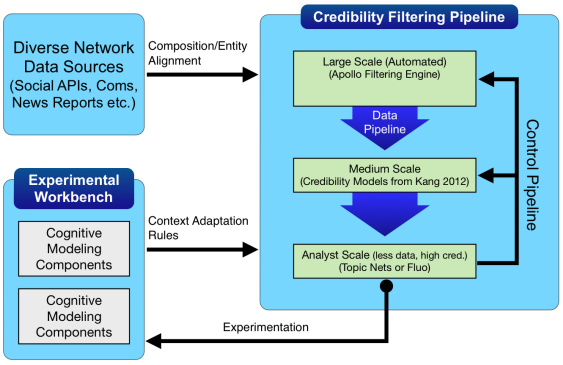


Fig. 1. Overview of a credibility-based recommender system with a supporting cognitive modeling component.

### A. Research Questions

- *What are the scientific challenges that arise from modeling credibility in networks of different sizes?*
- *What are the cognitive limitations of human analysts that can inform where automated algorithms should take over?*
- *How can we leverage the theoretical and practical boundaries of different types of credibility modeling to improve/optimize a credibility filtering pipeline?*
- *How do we leverage cognitive and human-factor models to discover rules that help analysts to better adapt to specific contexts/missions?*


Fig. 2. Fluo is a configurable user interface for interacting and visualizing relational data, such as those found in most modern databases. In this experiment, data is hierarchically and semantically organized into different lists (A) and related objects (e.g. by foreign key) are connected by pipes (B). Queries are performed by specifying relevance to upstream nodes which flows to connected downstream nodes. Each list sorts its contents by this flow of relevance.

### B. *Preview of Study*

This research study evaluates several methods for how to visually incorporate results from data mining algorithms into a visual tool and explores the limitations, potential synergies, and other theoretical boundaries between automated credibility analysis algorithms and credibility assessments made by human analysts. Amazon Mechanical Turk is used to collect experiment results. Online workers are presented with an interactive visual interface of varying complexity and tasked with answering questions about a hypothetical relief scenario dataset. For half of the participants, the interface incorporated automated results from a highly scalable credibility analysis engine based on the Apollo system [18]. This is an automated fact-finder tool that assess credibility of information claims from different sources at large scales. The hypothesis is that there is a 'sweet spot' between complexity of visual and interactive features and the inclusion of recommendations from complex automated tools.

## II. Related Work

The *Fluo* experimental workbench integrates components from multiple research areas. Our discussion of related work covers user interface research, cognitive modeling, recommender systems and relevant research on mining information credibility.

### A. *Interactive Interfaces for Data Analysis*

Effective user interface design can overcome limitations in the user's attention and working memory [3][2]. Additionally, by increasing visual explanation of computational processes, user interface design can facilitate the correct perception of trust and data provenance in scientific information analysis [9]. Developers and interface designers must deal with the challenge of combining and representing large amounts of data at the right time in the right format. Unfortunately, there are finite limits to an analyst's ability to efficiently and effectively summarize large amounts of data, especially when tasks are time-sensitive.

### B. *Cognitive Models*

In Intelligence tasks humans often go through a detective process of "search and relate" to determine who is doing what to and with whom. Intelligence officers often gather information by field observation or consulting public records and confidential information sources. Information helps them make connections and finally answer important questions and make decisions. Cognitive models of a user can help improve computation, filtering, visualization and comprehension of credible information. Understanding of how to filter and visualize network data matched to individuals cognitive states is necessary in order to improve inspectability, control, and situation awareness. Figure 1 shows how cognitive models are leveraged in our experimental data analysis framework. Cognitive models are representations of human behavior that rely on mathematical and computational mechanisms. These mechanisms represent cognitive processes and effects that influence human behavior, such as memory retrieval, forgetting, recognition, judgment and decision making. In military

intelligence tasks cognitive models play an important role in explaining, predicting, and supporting the collection of intelligence that can serve a mission.

### C. Recommender Systems

There have been numerous studies that have addressed optimizing the synergy between human analysts and expert agents. The visual analytics community has leveraged automated algorithms to intelligently limit the subset of displayed information when viewing multivariate data [17]. Recent studies in recommender systems have noted the importance of user-recommender trust in improving satisfactions with recommendations [12][10]. The importance of system transparency and explanation of recommendation algorithms has also been shown to increase the effectiveness of user adoption of recommendations [7].

### D. Presenting Credibility Recommendations

A large number of studies spanning multiple disciplines have focused on information credibility in networks. However, there has been a lack of focus on the importance of the interface in communicating credibility information to end users. In many cases the problem of data sparsity is unavoidable and there is simply not enough information available to reliably estimate credibility of information in a network. Insights include the following: in online settings, the window of data available to assess credibility of a piece of information is small compared with real world scenarios [12]. In information networks, credibility models can focus on node content, node connectivity, information flow around a node, or some combination of these. Inspectability and control both independently improve the perceived credibility of information in a network [7]. Many visualization tools lack a careful consideration of cognitive phenomena. Cognitive Models can be applied to model interaction behavior with the information system [8], and resulting insights can be used to refine a system design.

## III. System Description

This section describes the scalable fact-finder algorithm, Apollo [18] and the Fluo research framework which were used in the experimental setup. Apollo is a fact-finding tool designed to jointly assess both the credibility of information and the reliability of sources. The underlying algorithm performs maximum-likelihood estimation. Given a set of sources and their claims (e.g., statements, tweets, blogs, or other assertions), Apollo iteratively computes the credibility of those claims given their degree of corroboration, and the credibility of sources given credibility of their claims. Apollo also considers non-independence relations between sources to discount rumors that are corroborated only within one social group. Once credibility values are computed, Apollo can rank the information based on credibility.

Fluo (Figure 2) is a configurable web-based user interface developed at UC Santa Barbara. The system is intended as a research tool for isolating design ideas in user interfaces, and will eventually synthesize a suite of commonly used network-analysis tools with a consistent visual and interaction style. A prototype of the framework was used to configure and execute the experiment described here.

Here, Fluo is used to model different semantic schema as connected nodes that are organized into multiple sort-able lists (similar to a traditional spreadsheet). The inspiration for the semantic organization of network data and corresponding visual metaphor used in the experiment builds on [7]. The lists were placed serially (creating an upstream/downstream relationship) or in parallel based on experimental condition.

Participants were presented with one of three configurations of the interface (Figure 3), which varied in the diversity and complexity of interaction methods and data presentation. The simplest user interface presented to users, 'spreadsheet' Fluo, mimics the functionality and visual layout of the familiar spreadsheet by organizing the scenario data in a single list. The spreadsheet condition offered only two methods for interaction: sort (by schema column) and scan (via a scroll bar), but the data in this condition is highly organized and readable. The Apollo credibility ranking of each message was presented in column format and sort-able by clicking the column header. The second interface, 'limited' Fluo, increases visual and interactive complexity by organizing categorical properties (such as region) of each message into upstream lists, but deliberately does not give users the querying tool (scoring) necessary to overcome the organizational complexity. Despite this, 'limited' Fluo does allow a user to get on-demand details of all messages that match a certain category (e.g. messages that originated in the region of Brickland), where 'spreadsheet' required sorting and then scanning. Apollo ranking of messages was still displayed as text on each cell and items were sort-able, as in 'spreadsheet'. Finally, 'advanced' Fluo expands on 'limited' primarily by allowing users to assign a relevance score to each node. The relevance is then automatically propagated to connected downstream nodes and the appropriate lists are re-sorted. Users were also given the option to filter messages interactively by specifying a numeric range (right side of Figure 2) in addition to the sorting that was available in 'limited' and 'spreadsheet'.

Obtaining the correct answer to each analysis question in the experimental task was always possible regardless of the interface configuration, but the minimum level of time and interaction required to perform the same query varied from interface to interface. For instance, comparing messages from different regions in the 'spreadsheet' condition requires a single click for sorting, followed by scrolling back and forth between the two message groups from each region of interest; in the 'limited' and 'advanced' conditions, the relevant region is first selected and then scrolling is used to scroll through messages. 'Advanced' could also be used to more quickly answer complex queries. For instance, if users want to compare messages between two regions that have high Apollo ranking and match a set of key terms, they can boost the score of the regions and key terms of interest, then adjust the Apollo ranking filter to match the query.


Fig. 3. The 'spreadsheet' condition (A) and the 'limited' condition(B). 'Advanced' is shown in Figure 2. The answer to the question 'What was the occupation of the person who died when entering a leaning NGO building?' is highlighted in each interface.

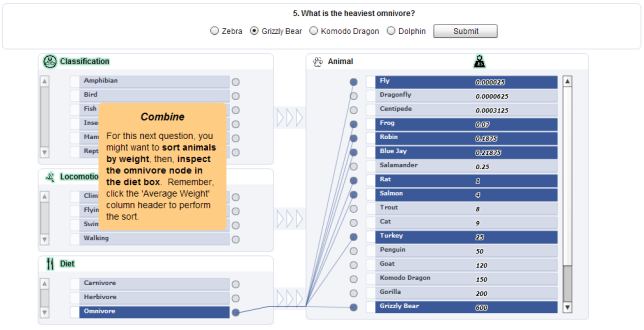


Fig. 4. During the training session, users are required to correctly answer a series of questions designed to verify their understanding of the interface. An easily understood animal dataset is used tso help ease users into views and interaction modalities that may be unfamiliar to them. Here, 'limited' is shown answering a query related to the diet and weight of the animals present in the dataset.

## IV. Scenario Description

Before the analysis session, participants were presented with a brief description of a scenario and a map of the fictional Brickland-Gordania region. The scenario is presented to users as a collection of 400+ messages broadcast from various fictional regions over different fictional media networks. The hypothetical task was to search through these messages to dispense advice on humanitarian efforts during a crisis situation following an earthquake and involving riots by three insurgent factions.

Participants were asked to inspect the messages using the interface provided and answer the four Priority Intelligence Requirement questions (PIR), as listed below:

1) *Which insurgent/militia cell is encouraging the most violence against Brickland and the Gordanian government?*
2) *Of the three cities, where are search and rescue efforts most needed?*
3) *Where will protests against Brickland and Gordania most likely occur to disrupt relief operations?*
4) *What degree of risk exists to NGO elements operating in the cities and towns around the cities?*


Fig. 5. Elements contributing to the "Brickland" Humanitarian Aid Scenario Data and Metadata.

Figure 5 shows a graph of the different components of the Brickland humanitarian aid delivery scenario. In addition to the core messages, entities and ground truths in the scenario, multiple augmentations were provided to improve the data set. A collection of related social media (mainly Twitter) messages were appended to the scenario to add realism, and to provide more scale to test the Apollo fact-finder tool. Each message was appended with a credibility score from Apollo. This data is revealed to participants in some of our experimental conditions, as detailed in the next section.

## V. Experimental Setup

The study described in this paper explored how well parameters that allow the systematic manipulation of data based on criteria related to data credibility and source reliability support the decision maker in efficiently exploring a large data set. Specifically, this study addressed the following hypothesis.

- H1: Increased control over automated credibility filtering mechanisms (i.e., a control pipeline that is regulating the data pipeline) improve human analysts information requirement research process, thereby improving the speed and quality of decision making.

A 3x2 between-subjects design was utilized. The conditions varied the level of functionality available in the user interface, along with presence of credibility information in order to assess which manipulations improve decision speed and quality.

The experimental system was deployed on Amazon Mechanical Turk and data was collected from AMT workers. The AMT web service is attractive for researchers who require large participant pools and inexpensive overhead for their experiments, however, there is valid concern that data collected online may be of low quality and require robust methods

| *Experimental Conditions* | | |
|---|---|---|
| *Condition Number* | *Apollo* | *Fluo* |
| 1 | No | Spreadsheet |
| 2 | Yes | Spreadsheet |
| 3 | No | Limited |
| 4 | Yes | Limited |
| 5 | No | Advanced |
| 6 | Yes | Advanced |

for validation. Numerous experiments have been conducted, notably [1] and [14], that have attempted to show the validity of using the service for the collection of data intended for academic and applied research. These studies have generally found that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled.

After accessing the experimental system online, participants were presented with a pre-study questionnaire that collects some basic demographic and expertise information, and required the user to answer three screening questions to test their attention. They were then directed to an interactive training session where the interface was explained while operating on a simple animals (taxonomy) dataset. After the training, participants were prompted to either continue or to re-take the training session. If they were ready to continue, the relief scenario data was loaded into the interface and participants were allowed to explore the data and required to answer some basic comprehension questions before metrics were collected. When all analysis questions were answered and metric collection was complete, the users were directed to a post-study questionnaire where they provided feedback on the user interface and scenario data.

Data was collected from 297 participants, 12 of which were marked as erroneous after closer inspection for a total of 285. Satisficing elimination was done similar to [13] - participants were required to undergo 3 Instructional Manipulation Checks (IMCs) in the pre-study and were called out if they answered the questions incorrectly. Furthermore, an extra level of satisficing detection was added by requiring participants to manually type their answers to the analysis questions which were inspected by hand. Participants were removed either because their answers to the analysis questions were unintelligible or due to timing glitches in our experimental system. Of the 285 use-able data points, participant age ranged from 18 to 67 with an average of 31.42 and a median of 29. 59% (168) of participants were male while 39% (117). were female.

Participants could not be observed as they undertook the study, so the system itself logged detailed results for nearly every possible interaction, including time taken, node clicks, filter manipulations, sorts, and score manipulations.

## VI. Results

Table I shows a list of the primary metrics recorded in the experiment. These can be classified into accuracy, Interaction and perception based metrics. To better understand the 'sweet-spot' between analyst and information system described in the introduction, the following questions were posed about the experimental data:

1) How did the level of control and presented metaphors in the user interface impact on accuracy of PIR questions? What about analyst score (response time and accuracy?
2) How did presence of Apollo impact on accuracy or speed of PIR questions?
3) Did participants who interacted with Apollo (when it was available) do better (correctness) overall than people who did not?

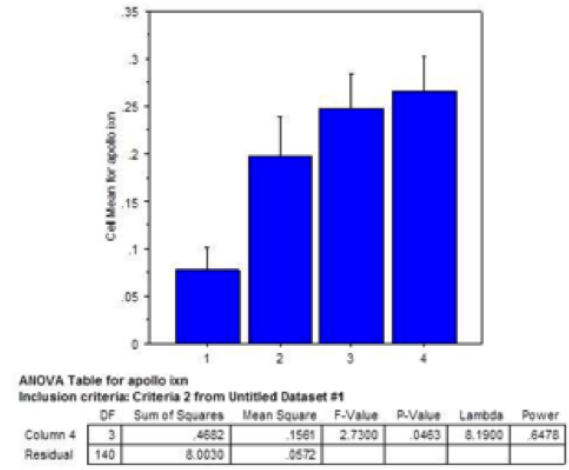


Fig. 6. Number of correct PIRs v/s Apollo Interaction Level. Mean interaction with Apollo was 47.1% higher for correct answer group than incorrect group (p=.0463, ANOVA)

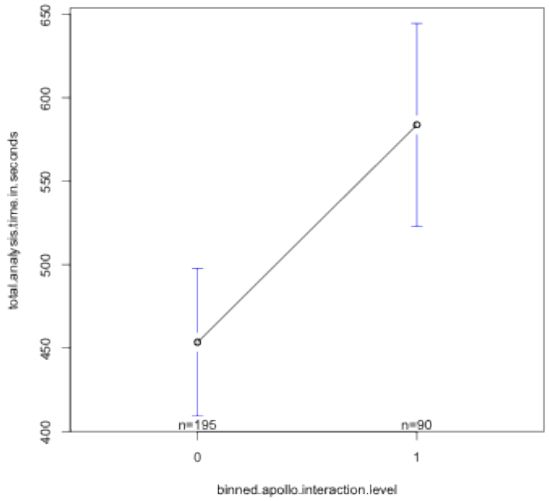


Fig. 7. Response time v/s Apollo Interaction Level (bins: high N=195, low N=90)

TABLE I
AN OVERVIEW OF THE CRITICAL METRICS RECORDED FOR EACH PARTICIPANT.

| *Metrics* | | |
|---|---|---|
| *Type of Metric* | *Specific Metric* | *Measurement Method* |
| PIR Response | Speed to answer PIR | Time from beginning of scenario until PIR answered. |
| PIR Response | Accuracy of PIR answer | Compare participant response to ground truth. |
| Semantic Entity Usage | Quantity/quality of terms used | Specific terms and number of times revisited |
| Semantic Entity Usage | Nature of term usage | Order of usage, frequency of search, frequency of modification. |
| Interface Interaction | Utilization of each control | Frequency of manipulation. |
| Interace Interaction | Patterns of Interaction | Ordering of control usage. |
| Usability | Usefulness/value of features | Survey administered at the end of the scenario. |

4) What is the performance difference between participants who showed a high level of interaction overall and participants who showed a high level of interaction with Apollo? This was to validate that observed performance improvements were a result of Apollo interaction, rather than an increase in general interaction with the system.
5) For the three previous questions, did the effect size change based on the user-interface (i.e. did participants who used 'limited' benefit more from the recommendations)?

Scenario-based experimentation has high variance in participant responses. The experimental subjects were 285 AMT users, all with different propensity and abilities for scenario-based data analysis. Despite the comprehensive metrics recorded, only a few interesting effects were found through statistical power analysis. Notably, no sizable effect for speed and accuracy was revealed by varying user interface richness. This was most likely due to large individual differences in the analysis capabilities of participants and their individual receptions of the presented metaphors. To address this, a followup study is currently in progress utilizing a within-subjects design, fewer differences between interface configurations, and allowing users more control over presented metaphors and query styles.

To answer questions 2 and 3, performance and time were assessed across the conditions with and without the Apollo credibility data. As not every participant was forced to interact with Apollo, there was predictably no significant performance difference in correctness. Despite this, the best performers (those who answered 3 or 4 of the PIR questions correctly), showed a relative increase of 47 % more interaction with Apollo's credibility assessments compared to those who correctly answered 0-2 PIR questions. This result reinforces the idea that credibility information is important in the process of data analysis, and also that *analyst interaction* (e.g.: sorting and filtering of messages) with automated credibility information leads to better performance. Figure 6 shows the results of the experiment along with an ANOVA table (p=0.463).

Results show that Apollo credibility data can lead to increased accuracy, but this does come at a cost. Figure 7 shows a comparison of the mean response times for participants with low and high Apollo interaction scores. Response time was recorded as the number of seconds required to answer all four of the PIR questions (note that the y-axis begins at 400 seconds). Participants in the high-use bin took 27% longer to answer the PIR questions than those in the low-use bin. When considering this result, the limited training time available to participants in this experiment should be kept in mind. In a real-world scenario, where an analyst is well versed in the system functionality, it may not be the case that use of credibility data leads to longer analysis times. This question will be addressed in a follow-up that will better account for training time and agreement of user-interface tools and the analyst's mental model.

## VII. DISCUSSION AND FUTURE WORK

This section provides the authors' perspective on four key lessons learned from the experiment. First, participants reported the most frustrations in the 'limited' user interface condition. This is reflected by task completion time (qualitatively assessed from the sentiment of free-response feedback). There were also noticeably more frustrations for participants using 'limited' over 'spreadsheet'. Additionally, participants in the 'limited' and 'advanced' also reported more reliance (average 3.87 and 3.74) on the automated credibility assessments than in 'spreadsheet' (3.24). This might suggest that frustration with a user interface or task leads to a strong reliance on automation, but how and when users relinquish control is not well understood.

Second, results from this experiment may have been confounded by too many small differences between the three user interfaces and not enough differences at the level of ideas. For instance, a decrease in efficiency in the 'limited' condition was most likely due to the large amount of scanning that was necessary to complete the task, which could have been remedied with well-established tools like keyword search. Meanwhile, the 'spreadsheet' and 'advanced' conditions differed only in the particular *way* that users were sorting and filtering, but did not try to compare the information filtering paradigm with other established methodologies. It might be more useful simply to provide consistent designs for common data interaction intents (as described in [19]) and observe an analyst's use patterns.

Third, the task questions presented were fairly prescriptive and could be eventually solved just by persistent scanning, mitigating the need for a dynamic interface. One of the major advantages of manual information browsing over automated analysis is that a human analyst may gain unexpected insight

into the data that may help with future tasks and overall data understanding, but an analyst may not always have the leisure of time. Trying to understand how to design interfaces to quickly develop insight (and how to measure this in an experimental setting) is an open problem [11]. Using a longitudinal experimental methodology with less prescriptive questions and forcing participants to work from memory for some questions may help isolate promising interface ideas.

Finally, there may have been a mismatch between the AMT community and the topic of the data set in the experimental task. Providing participants with a data set in a domain that is more likely to be familiar and trying to model their domain knowledge with a pre-study questionnaire will better isolate variables related to interface and recommendation design.

The follow-up experiment using the Fluo experimental toolbench will allow more pronounced paradigms between conditions (spreadsheet, graph views, and hybrid spreadsheet/graph) in terms of data model, visualization, and interaction, which may show more significant results. Additionally, user interaction with the system will be captured at a higher level which might better capture frustration and intent. A more open-ended design will be used, where participants will be allowed to explore the data set in advance of answering questions while others will force participants to work from memory. Finally, the authors believe that analysis skills and domain knowledge are a critical factor that should be adequately modeled, so AMT users will be presented with a data set related to traffic sensors and a deeper cognitive evaluation of analyst interactions will be performed using Instance Based Learning cognitive models [8].

## VIII. Conclusion

A crowdsourced experiment (N=285) was performed and evaluated to assess the cognitive limitations of human analysts in a variety of conditions, particularly with and without presence of credibility information from a large scale automated fact-finder tool. To achieve this goal, a novel experimental toolkit called "Fluo" was introduced. Fluo is a configurable network data analysis tool for use in multiple domains. Our experimental results showed that information analysts who answered both correct and fastest, *interacted* with the interface components related to credibility recommendations 47% more than other participants (p=0.463). However, the group with credibility information displayed to them did not show a significant accuracy improvement over those who had no credibility information available. Some participants reported that complexity of both UI and scenario data (including acronyms and other jargon) was confusing at times. The authors are conducting a followup experiment based around a more simplistic traffic analysis scenario. A followup paper will compare and contrast both studies to help provide generalizable insight for design of crowdsourced scenario-based analysis experiments.

## Acknowledgment

The authors would like to thank Laura Marusich at ARL for her valuable advice, and John Hyatt at SA Technologies for his role in constructing the Brickland dataset.

This work was partially funded by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

[1] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[2] Mica R Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.

[3] Mica R Endsley. *Designing for situation awareness: An approach to user-centered design*. Taylor & Francis US, 2003.

[4] Mica R Endsley and Esin O Kiris. The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(2):381–394, 1995.

[5] Christopher D Hundhausen, Sarah A Douglas, and John T Stasko. A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages & Computing*, 13(3):259–290, 2002.

[6] Daniel A Keim. Information visualization and visual data mining. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):1–8, 2002.

[7] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 43–50. ACM, 2012.

[8] Tomás Lejarraga, Varun Dutt, and Cleotilde Gonzalez. Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, 25(2):143–153, 2012.

[9] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.

[10] Paolo Massa and Paolo Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 17–24. ACM, 2007.

[11] Chris North. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE*, 26(3):6–9, 2006.

[12] John O'Donovan and Barry Smyth. Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 167–174. ACM, 2005.

[13] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4):867–872, 2009.

[14] Gabriele Paolacci, Jesse Chandler, and Panagiotis Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[15] Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 830–831. ACM, 2002.

[16] Kirsten Swearingen and Rashmi Sinha. Interaction design for recommender systems. In *Designing Interactive Systems*, volume 6, pages 312–334. Citeseer, 2002.

[17] Andrada Tatu, Georgia Albuquerque, Martin Eisemann, Jörn Schneidewind, Holger Theisel, M Magnork, and Daniel Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 59–66. IEEE, 2009.

[18] Dong Wang, Lance Kaplan, Hieu Le, and Tarek Abdelzaher. On truth discovery in social sensing: a maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, pages 233–244. ACM, 2012.

[19] Ji Soo Yi, Youn ah Kang, John T Stasko, and Julie A Jacko. Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1224–1231, 2007.