Volume xx (200y), Number z, pp. 1–5

Analytic Behavior and Trust Building in Visual Analytics

Dominik Sacha, Ina Boesecke, Johannes Fuchs, and Daniel A. Keim

University of Konstanz, Germany

Abstract

Visual Analytics (VA) is a collaborative process between human and computer, where analysts are performing numerous interactions and reasoning activities. This paper presents our current progress in developing a note taking environment (NTE) that can be plugged to any VA system. The NTE supports the analysis process on the one hand, and captures user interactions on the other hand. Our aim is to integrate human lower- (exploration) with higher- (verification) level analytic processes and to investigate those together related to further human factors, such as trust building. We conducted a user study to collect and investigate analytic provenance data. Our early results reveal that analysis strategies and trust building are very individual. However, we were able to identify significant correlations between trust levels and interactions of particular participants.

Categories and Subject Descriptors (according to ACM CCS): H.1.2 [User/Machine System]: Human factors—Human information processing H.5.2 [User Interfaces]: Evaluation/methodology—

1. Introduction

Visual Analytics (VA) allows analysts to generate knowledge from data through visual interactive interfaces. During this process, analysts have to perform numerous lower level interactions (with the VA system) and higher level reasoning activities in order to arrive at the desired and verified pieces of information [SSS*14]. In order to validate the different pieces of gathered information, these facts have to be reviewed, related, organized, and combined with the analysts prior knowledge and assumptions (also described as "connecting the dots" [SGL09]). However, we know little about all the human factors affecting this entire process. Human analytic activities may be very unstructured and unorganized due to interruptions or unexpected events, such as spotting unforeseen patterns in the data. To cope with these issues, many systems allow analysts to bookmark, annotate, and organize interesting visualizations (e.g., [WSP*06]). In addition, approaches for capturing the human analytic process (by means of interaction logging) have emerged, enabling researchers and analysts to review their analysis processes and to retrieve (or "jump back" to) specific states/visualizations [XAJ*15]. However, relating human lower and higher level activities remains an open research challenge. Recent research identifies very individual human factors, such as trust building and the awareness of uncertainties [SSK*16]. In fact, these factors are hard to measure and to investigate.

This paper presents our work in progress for combining and analyzing these diverse aspects together. Our approach is based on a note taking environment (NTE) that can be connected to VA systems, supporting analysts in organizing their findings and externalizing their thoughts. The NTE offers interfaces for external interaction capturing (e.g., for VA tools) with the aim to combine provenance information of different levels. In addition, the NTE enables analysts to apply trust ratings or textual input as a third source of captured information. We conducted a user study to capture all the different kinds of data in order to create an initial dataset to be investigated. Our results reveal different analysis behavior among the individual participants between exploration and verification activities. Furthermore, we found a significant positive correlation between local and global trust ratings for a subgroup of participants. However, we did not discover any positive correlation between trust and the analysis efforts (on a exploration, verification or total measure). In contrast, two exceptional cases show a significant negative correlation.

2. Related Works

Many theoretical works on human thinking, reasoning and sensemaking during data analysis exist. Pirolli and Card describe this process as foraging and sensemaking loops [PC05] iteratively traversing several analysis stages. Sacha et al. [SSS*14] propose that the knowledge generation process is assembled by three loops (*exploration*, *verification*, and *knowledge generation*). More recently, Sacha et al. [SSK*16] describe the role of uncertainties, their awareness and human trust building within this process. They propose guidelines for supporting human factors, such as 1) supporting analysts in uncertainty aware sensemaking, 2) enabling analysts to review the analysis process, or 3) to analyze human behavior in order to derive hints on problems. Interaction categorizations (e.g., [GZ08], [BM13]) offer a useful foundation for capturing human analytic processes. Nguyen et al. [NXW14] survey

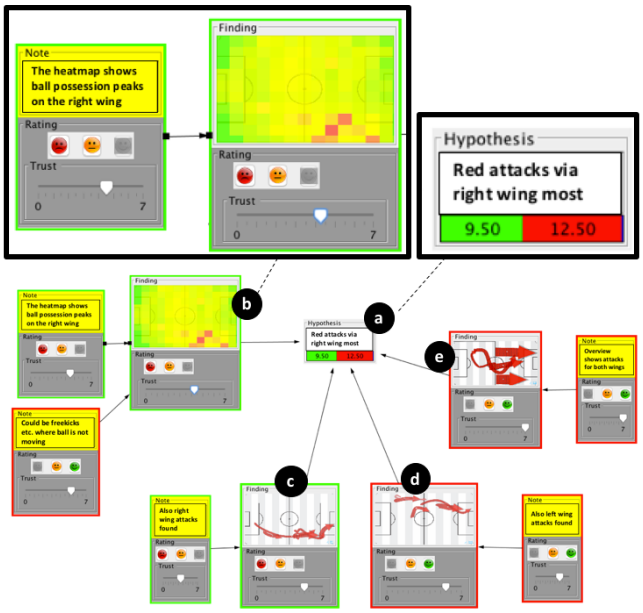

Figure 1: *Knowledge graph that has been built during the analysis process. Verifying as well as falsifying findings for confirming or rejecting the hypothesis have been collected. Additionally, the analyst added notes and trust ratings to the elements. The evidence bar at the hypothesis widget indicates that the hypothesis is rejected.*

several emerging approaches for analytic provenance and highlight that there are different stages of capturing, visualizing and utilizing provenance information.

We found several related systems in our literature review. Examples are Jigsaw's tablet view [LGK*10], the Sandbox for analysis [WSP*06], HARVEST [SGL09], Aruvi [SvW08], or VisTrails [SVK*07]. Typical components are the capturing of visualization states, history visualization, and the management of bookmarks, which is usually represented as a graph composed of entities, images, annotations and connections (from now on called a "knowledge graph"). Despite the fact that these tools share the same goal of supporting the analysis, we are able to identify some differences. Some focus on note taking capabilities (e.g., Sandbox [WSP*06]), whereas others focus on capturing system states (e.g., VisTrails [SVK*07]), or leverage the captured interactions for ranking or organizing bookmarks automatically (e.g., HARVEST [SGL09]).

Aruvi [SGL09] and Jigsaw [KGS11] have been used to study analytic behavior. Additionally, other user studies investigated further human factors. For example, the works by Harrison et al. [HDL*11] investigates user frustration and interaction. Dzindolet et al. [DPP*03] analyzed trust development between humans and automated decision aids. Observing errors caused a decrease in trust towards the system unless an explanation was provided. However, understanding the uncertainties caused in turn increasing trust towards the decision aid, even under uncertainty. Uggirala et al. [UGMT04] tried to find a way to measure trust in complex and dynamic systems and showed that an increase in uncertainty caused a decrease in trust towards the used system. Bass et al. [BBS13] investigated human judgment and proposed a method to measure and predict a humans understanding of automation. They offer a trust questionnaire for distinguishing high- from low-trusting participants. These studies are interesting illustrative examples for investigating further human factors that haven't been focused in the VA community so far.

In summary, we found very inspiring works on supporting, capturing, and analyzing analytic processes. However, we are not aware of a system that tracks interactions beyond system borders (either exploration or verification), enables knowledge management (by means of note taking) enriched with further capabilities to gather human inputs (such as trust ratings), with the goal to analyze these aspects together.

3. Note-Taking and Capturing Approach

The foundation of our approach is the capability to build a knowledge graph composed of gathered information and human assumptions. The note taking capability (verification) is smoothly integrated with the actual analysis within a VA system (exploration). Figure 1 shows a knowledge graph that has been built by a soccer analyst. At the very first, the analyst defined a hypothesis widget in the NTE interface (Figure 1-a). In our example, he had watched the game and assumed that "the red team mostly attacked via the right wing". In order to prove his assumption, he switched to the soccer tool and created a heat map for all ball movements of red team attacks. Subsequently, he imported the bookmarked finding including annotations to the NTE (Figure 1-b). The analyst marked the finding as a verifying piece of evidence (assigning a "falsifying" or "neutral" tag is also possible). However, by looking at the peaks within the heat map, he found out that the peaks may be caused by standard situations (e.g., free-kicks) where the ball is not moving. Consequently, this lowered the trust in this finding. Therefore, the soccer analyst adjusted the trust rating slider (value range 1-7). As a second step, the analyst visualized all right wing attacks, imported again the bookmarked visualization and applied again a note, a trust rating, and marked this finding as verifying (Figure 1-c). In the following, the analyst started to seek for counter evidences. He explored the left wing attacks and added them to the knowledge graph (Figure 1-d). Subsequently, the analyst produced a finding that clearly illustrates attacks for both wings (Figure 1-e). Finally, the analyst had collected enough pieces of evidence for rejecting his hypothesis. The evidence bar placed at the hypothesis supports this conclusion (Figure 1-a) by aggregating the trust inputs with verifying or falsifying information. During this process, the analyst switched between the two provided tools, revisiting and refining bookmarked visualizations.

This simple example illustrates that our approach smoothly supports and integrates the knowledge generation process. The NTE design is based on the knowledge generation model for VA [SSS*14] and offers different widget types according to the concepts: *Hypotheses*, *Actions*, *Findings*, and *Insights* which are part of the NTEs data model. *Actions* are captured automatically (in the NTE and the VA system) and additional *Notes* are created by the analyst. Note that *Findings* are the bookmarks that are imported from an external VA tool (also importing external images as *Findings* is possible). Beyond, the NTE incorporates functionality for externalizing and supporting the analysts' trust building

process by means of the guidelines (on human factors - G6, G7, and G8) provided in [SSK*16]. First, it enables the analyst to apply trust ratings and annotations to the graph elements. Second, it is capable to capture and visualize further measures that belong to a finding (e.g., uncertainty measures) and provides visual cues by aggregating them (evidence bar-Figure 1-a). Further, it is possible to map the measures (e.g., trust or uncertainty) to the elements of the knowledge graph (transparency). Third, it enables the analyst to review her/his analysis processes and to "jump back" to the specific system state once a finding in the NTE is clicked. Additionally, interaction sequence visualizations are provided (e.g., one line in Figure 2). Furthermore, it is possible to save and load the knowledge graph and captured interactions.

The NTE captures human behavior on different levels and spaces. On the one hand, all the exploration interactions in the VA system are captured (according to [BM13]) and low level operations (e.g., mouse clicks or moves) are counted. On the other hand, the NTE captures all the verification interactions in the knowledge graph and allows analysts to provide further inputs (such as the trust ratings). As a third source of information we consider individual aspects that can be captured by a questionnaire before and after the analysis process. This information provides hints on the different user characteristics (e.g., their experiences and attitude towards automated analysis systems [BBS13]). In sum, our capturing approach enables us to investigate different factors of the knowledge generation process.

Our process is realized as a note taking component (written in JAVA) providing an application programming interface (API) for integrating external VA systems. The API offers two different interfaces. One interface allows to send bookmarks (visualization images) including additional information (annotations, measures, or a callback function) to the NTE. The other interface provides an interaction logger, that offers different logging levels and types. Implementing the desired interaction logging and callback functions has to be done by the VA system developer. However, in this way the NTE is completely independent from the VA system. Please refer to our supplemental material for more technical details.

4. Experiment

We conducted a user study to prove our concept and to create an initial captured data set for investigation. We wanted to observe how the NTE is used and to gather feedback about possible areas of improvement. Our main goal was to measure the amount of actions per phase (exploration and verification) and to capture trust ratings on a global level (per task) and on a local level (per finding). Therefore, we formulated the following research questions (RQ):

RQ1: "Is it possible to identify different user groups based on their analysis strategies?" (e.g., performed analysis efforts between exploration and verification phase).

RQ2: "Is there a positive correlation between trust in findings and the overall trust in the system?" (e.g., a low trusted finding causes a general decrease of trust towards the system as a whole).

RQ3: "Is there a positive correlation between interaction activity and trust in findings?" (e.g., the more analysis effort is spent, the more trust should be built by the analyst).

Design and Procedure: *Participants:* We recruited 9 participants from the local student population (4 female, 5 male, age 23-29 (median: 23)). All participants reported normal or corrected to normal vision, had mixed experiences with soccer and only little background in using data analysis systems.

Apparatus: The study was conducted in a lab setting using two 24 inch screens. The only input device was a common computer mouse and a keyboard for textual input. The participants were seated approximately 50 cm away from the screen. The experimenter was present during the study for answering questions and introducing the study procedure. For recording the user input we plugged our NTE component to a soccer analysis tool [JSS*14] and implemented the API for importing the bookmarks and interaction logging.

Task and Procedure: We defined six analysis tasks that had to be solved using the soccer analysis tool. Each task comprised a given hypothesis (similar to the example at the beginning of Section 3) and soccer data with which the hypothesis had to be proven or rejected. After introducing the two systems and our intention of trust, participants had to work on all six tasks independently. The tasks were ordered according to their difficulty. Between two tasks participants were told to take a short break and assign a trust rating to their answer, as well as to the overall system. In order to enforce/influence a trust variation, we told the participants after the forth task that the soccer analysis tool might not work as accurate as expected due to interpolation operations. Exploration and verification interactions have been captured and counted for the particular phases. Additionally, participants had to answer a questionnaire (designed according to Bass et al. [BBS13]) at the end of the study.

Results: For the analysis of RQ1, we visualized the amounts of captured interactions per phase (exploration and verification). For the investigation of RQ2 and RQ3, we analyzed the captured trust values to calculate the pearson correlation index and report only on significant results.

RQ1-Exploration and Verification: Figure 2 shows the beginning of the analysis process of the participants (task 1). The visualization reveals that there are different analysis strategies. For example, P8 and P4 have very long exploration phases interrupted by brief verification phases. They started to collect and refine several findings before they switched to longer verification phases. In contrast, P7 and P3 directly put more efforts in verifying their findings. These user characteristics are also reflected in the respective derived phase measures (for all tasks), as shown on the right hand side of Figure 2. In general, exploration phases tend to be longer than verification phases. In contrast to the other participants, P7 and P6 invest less efforts in exploration and consequently start verifying earlier. This characteristic may be identified automatically by measuring the ratio between exploration and verification efforts. Since we identified different analysis strategies we are able to validate **RQ1**. An interesting observation is that the majority of participants only searched for findings verifying the given hypothesis (instead of collecting findings to reject them). Even when we raised their awareness on this issue, they did not start seeking for counter evidence.

RQ2-Global and Local Trust: In order to investigate our second RQ, we determined the finding with the lowest trust rating per task

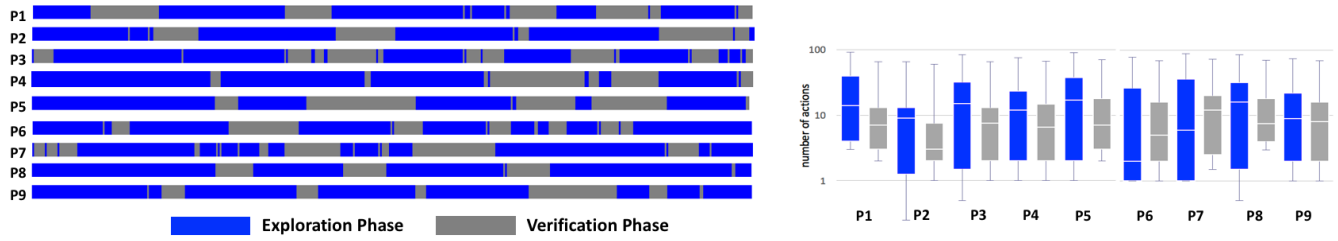

Figure 2: *Captured analysis phases of the participants solving the same tasks. Left: Phases are visualized for task 1, Right: Boxplots for the number of actions per phase (logarithmic scale) are shown for the whole user study (all tasks).*

(local) and compared it to the general trust in the analysis system after solving the task (global). We calculated pearson correlations for each individual participant. The results showed a significant positive correlation for three participants (P1: $r(4) = .82$, $p = .05$ - P6: $r(4) = .87$, $p = .02$ - P8: $r(4) = .84$, $p = .03$). On the one hand this means that if the trustworthiness of a finding was declared as low, also the trustworthiness of the general system was determined by this particular trust rating (independent from the presence of highly trusted findings). On the other hand, global trust increased when the trust of the (lowest) finding increased. Further, another group of three participants showed smaller positive correlations but included a slightly higher p value (P4: $r(4) = .79$, $p = .06$ - P5: $r(4) = .63$, $p = .18$ - P9: $r(4) = .78$, $p = .07$). The last group of participants had very low or even negative correlations and a very high p value, indicating that there was no relation between local and global trust (P2: $r(4) = .09$, $p = .86$ - P3: $r(4) = .31$, $p = .56$ - P7: $r(4) = -.08$, $p = .88$). In summary, these results showed that trust building is very individual, but we were able to identify different user groups based on our correlation analysis. Therefore, we are able to verify **RQ2** only partially for a particular sub group (P1, P6, P8).

Interestingly, our hint on potential faults/uncertainties after task 4 did not decrease the system's trustworthiness for each participant. For some participants the trust value already decreased before or even increased after task 4. A possible explanation could be the diverse background of the participants and previous experiences with the system before task 4 (e.g., user frustration/success).

RQ3-Trust and Interaction: We analyzed for each participant if a positive correlation between the assigned trust ratings and the analysis effort exists (for all the findings that have been collected and rated). Therefore, we measured the amount of exploration and verification interactions as well as their sum. These measures have been calculated for each finding and correlations have been calculated for each participant. Among all users we did not find a general hint on significant positive correlations to the trust value ($p > .07$). Consequently, we are not able to prove **RQ3**. Interestingly, we observe two exceptions. P1 shows a significant negative correlation to the total analysis efforts ($r(15) = -.57$, $p = .02$) and to his explorations ($r(15) = -.52$, $p = .03$). Furthermore, for P3 we found a significant negative correlation for his verification efforts ($r(16) = -.60$, $p = .01$). In these cases, the participants assigned higher trust ratings to findings with less analysis efforts (however, on different levels). This could mean that these participants trusted their findings and therefore analyzed less. These results indicate the opposite of our initial assumption (that trusted items are investigated more intensively), however, for a small set of our participants.

Qualitative Feedback: After analyzing the answers in the questionnaire the participants stated that they felt on a medium to high level annoyed by adjusting the trust value (range: 4-7, median: 5), although, it was easy to learn how to do it. In addition to the 9 participants we invited a soccer expert to work with our prototype. He also stated that adjusting the trust value was annoying (7) but useful. He reported that while assigning a trust value he reflected about his own work. Furthermore, he thought that the NTE is a useful extension, as it provides an overview over the current analysis.

5. Conclusion and Future Work

In this paper, we present our general approach and early results for integrating, capturing, and analyzing exploration, verification, and trust building activities. We are able to mention interesting findings: 1) Different analysis strategies/behavior can be analyzed and identified (based on exploration and verification capturing), 2) a positive correlation between local and global trust exists (for a particular user group), and 3) there is no significant positive correlation between trust and the amount of interaction (in two exceptional cases there is a *negative* correlation). However, so far we just reported on little first steps of our work in progress and the gathered results are hard to generalize and need to be verified by more focused and extensive investigations (e.g., more participants). In addition, our derived interaction measures could be defined in more detail for specific activities (such as navigation, configuration, recording or annotation) to further distinguish user characteristics with respect to trust building activities. This would also enable us to analyze specific interaction sequences and patterns. Additionally, we found out that especially assigning trust ratings delivers a very subjective measure that has to be analyzed individually (or the trust has to be normalized among participants). Furthermore, we assume that human trust building is highly influenced by the analysis case and the impact of the decisions that have to be made during the analysis process. Therefore, we want to conduct similar studies in other (more critical) domains, such as crime analysis, or crisis management. Our vision is to investigate methods that enable us to identify different user characteristics automatically with the final goal to support the analysts adaptively according to their needs.

Acknowledgments

This work was partially supported by the EU project Visual Analytics for Sensemaking in Criminal Intelligence Analysis (VALCRI) under grant number FP7-SEC-2013-608142 and by the German Research Foundation (DFG) within the project "Visual Analysis of Language Change and Use Patterns."

References

[BBS13] BASS E. J., BAUMGART L. A., SHEPLEY K. K.: The effect of information analysis automation display content on human judgment performance in noisy environments. *Journal of cognitive engineering and decision making 7*, 1 (2013), 49–65. 2, 3

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2376–2385. doi:10.1109/TVCG.2013.124. 1, 3

[DPP*03] DZINDOLET M. T., PETERSON S. A., POMRANKY R. A., PIERCE L. G., BECK H. P.: The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud. 58*, 6 (2003), 697–718. doi:10.1016/S1071-5819(03)00038-7. 2

[GZ08] GOTZ D., ZHOU M. X.: Characterizing users' visual analytic activity for insight provenance. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2008, Columbus, Ohio, USA, 19-24 October 2008* (2008), pp. 123–130. doi:10.1109/VAST.2008.4677365. 1

[HDL*11] HARRISON L., DOU W., LU A., RIBARSKY W., WANG X.: Analysts aren't machines: Inferring frustration through visualization interaction. In *2011 IEEE Conference on Visual Analytics Science and Technology, VAST 2011, Providence, Rhode Island, USA, October 23-28, 2011* (2011), pp. 279–280. doi:10.1109/VAST.2011.6102473. 2

[JSS*14] JANETZKO H., SACHA D., STEIN M., SCHRECK T., KEIM D. A., DEUSSEN O.: Feature-driven visual analytics of soccer data. In *2014 IEEE Conference on Visual Analytics Science and Technology, VAST 2014, Paris, France, October 25-31, 2014* (2014), pp. 13–22. doi:10.1109/VAST.2014.7042477. 3

[KGS11] KANG Y., GÖRG C., STASKO J. T.: How can visual analytics assist investigative analysis? design implications from an evaluation. *IEEE Trans. Vis. Comput. Graph. 17*, 5 (2011), 570–583. doi:10.1109/TVCG.2010.84. 2

[LGK*10] LIU Z., GÖRG C., KIHM J., LEE H., CHOO J., PARK H., STASKO J. T.: Data ingestion and evidence marshalling in jigsaw VAST 2010 mini challenge 1 award: Good support for data ingest. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010* (2010), pp. 271–272. doi:10.1109/VAST.2010.5653042. 2

[NXW14] NGUYEN P. H., XU K., WONG B.: A survey of analytic provenance. *Middlesex University* (2014). 1

[PC05] PIROLLI P., CARD S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis* (2005), vol. 5, pp. 2–4. 1

[SGL09] SHRINIVASAN Y. B., GOTZ D., LU J.: Connecting the dots in visual analysis. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2009, Atlantic City, New Jersey, USA, 11-16 October 2009, part of VisWeek 2009* (2009), pp. 123–130. doi:10.1109/VAST.2009.5333023. 1, 2

[SSK*16] SACHA D., SENARATNE H., KWON B. C., ELLIS G. P., KEIM D. A.: The role of uncertainty, awareness, and trust in visual analytics. *IEEE Trans. Vis. Comput. Graph. 22*, 1 (2016), 240–249. doi:10.1109/TVCG.2015.2467591. 1, 3

[SSS*14] SACHA D., STOFFEL A., STOFFEL F., KWON B. C., ELLIS G. P., KEIM D. A.: Knowledge generation model for visual analytics. *IEEE Trans. Vis. Comput. Graph. 20*, 12 (2014), 1604–1613. doi:10.1109/TVCG.2014.2346481. 1, 2

[SVK*07] SCHEIDEGGER C. E., VO H. T., KOOP D., FREIRE J., SILVA C. T.: Querying and creating visualizations by analogy. *IEEE Trans. Vis. Comput. Graph. 13*, 6 (2007), 1560–1567. doi:10.1109/TVCG.2007.70584. 2

[SvW08] SHRINIVASAN Y. B., VAN WIJK J. J.: Supporting the analytical reasoning process in information visualization. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April 5-10, 2008* (2008), pp. 1237–1246. doi:10.1145/1357054.1357247. 2

[UGMT04] UGGIRALA A., GRAMOPADHYE A. K., MELLOY B. J., TOLER J. E.: Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics 34*, 3 (2004), 175–186. 2

[WSP*06] WRIGHT W., SCHROH D., PROULX P., SKABURSKIS A. W., CORT B.: The sandbox for analysis: concepts and evaluation. In *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006* (2006), pp. 801–810. doi:10.1145/1124772.1124890. 1, 2

[XAJ*15] XU K., ATTFIELD S., JANKUN-KELLY T. J., WHEAT A., NGUYEN P. H., SELVARAJ N.: Analytic provenance for sensemaking: A research agenda. *IEEE Computer Graphics and Applications 35*, 3 (2015), 56–64. doi:10.1109/MCG.2015.50. 1