Evaluation Methodology for Comparing Memory and Communication of Analytic Processes in Visual Analytics

Eric D. Ragan
Oak Ridge National Laboratory
1 Bethel Valley Road, Oak Ridge, TN 37831, USA
raganed@ornl.gov

John R. Goodall
Oak Ridge National Laboratory
1 Bethel Valley Road, Oak Ridge, TN 37831, USA
jgoodall@ornl.gov

ABSTRACT
Provenance tools can help capture and represent the history of analytic processes. In addition to supporting analytic performance, provenance tools can be used to support memory of the process and communication of the steps to others. Objective evaluation methods are needed to evaluate how well provenance tools support analysts' memory and communication of analytic processes. In this paper, we present several methods for the evaluation of process memory, and we discuss the advantages and limitations of each. We discuss methods for determining a baseline process for comparison, and we describe various methods that can be used to elicit memory of an analysis for evaluation. Additionally, we discuss methods for conducting quantitative and qualitative analyses of process memory. We discuss the methodology in the context of a case study in using the evaluation methods for a user study. By organizing possible memory evaluation methods and providing a meta-analysis of the potential benefits and drawbacks of different approaches, this paper can inform study design and encourage objective evaluation of process memory and communication.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces and Presentation (e.g., HCI)**]: User Interfaces – *evaluation/methodology*.

General Terms
Experimentation, Human Factors.

Keywords
Analytic provenance, process memory, evaluation, visual history

1. INTRODUCTION
Visual analytics applications provide an interactive means of exploring data and making sense of information, with tools giving analysts the flexibility to approach problems in any number of ways. Many analytic investigations are inherently complex processes due to the size and complexity of available data sets and the nature of hypothesis testing. The resulting complexity and potential variability in human analytic processing can make it difficult to remember the steps and rationale that led to the formation of hypotheses, the generation of specific data views, and the realization of conclusions. Process uncertainty can lead to problems, such as when an analyst needs to recall their steps weeks or months after an investigation to review rationale or explain the process to management or other analysts.

It is not surprising, then, that researchers have designed provenance tools to help capture and visually represent the history of analytic processes [11, 15]. For example, *VisTrails* is a provenance tool that helps track the progression of exploratory visual analytics of scientific data over time [4, 6]. The tool logs the computational steps taken to create different visualizations and generates visualizations of workflow history. Providing analogous functionality, the *GraphTrail* system records and presents analysis pathways taken during exploration of network data [10]. Another visual analytics tool, *CzSaw*, supports provenance of text document analysis with dependency graphs that show entity relationships [19]. *CzSaw* also supports visual history by showing data views that were open at different times throughout an analysis.

Such analytic provenance tools can serve multiple purposes [16]. While conducting an analysis, workflow logs allow analysts to reference previous stages of an analysis to help keep track of data manipulations or previously explored hypotheses. In complex analyses that consist of multiple analysis sessions or extend over long periods of time, reviewing earlier steps can help an analyst clarify memories of past actions and current goals. In addition to supporting the analytic process itself, provenance tools can be used to help communicate the steps of the process to others. It would be expected that having better memory of a process would make it easier to communicate that process, and visual representations of the process might be especially well-suited for communication purposes.

Though many provenance aids exist to support analysis and memory, how do analysts, researchers, and developers know if the tools are effective? From an intuitive standpoint, we could say that tools can be considered effective if analysts find them helpful and continue to use the tools over time. While such a standard makes some practical sense for determining the usefulness of single tool, it is less helpful for improving design and understanding which of the tools' features are most helpful. Objective methods are needed to evaluate how well provenance tools support analysis performance, memory, and communication. Many researchers have conducted studies to evaluate analytic performance, with traditional metrics including task time, instances of insight, and correctness of analytic findings as compared to known solutions [e.g., 1, 2, 13, 22]. However, few attempts have been made to objectively evaluate memory and communication of analytic processes. In this paper, we present several possible methods and discuss the advantages and limitations of each. In addition, we discuss the methodology in the context of a case study testing different evaluation methods for a user study of the effectiveness of visual history tools.

BELIV '14, November 10 2014, Paris, France
http://dx.doi.org/10.1145/2669557.2669563

2. BACKGROUND

Controlled studies are commonly used for evaluating how particular applications or features affect performance on analysis tasks. When paired with quantitative evaluations, controlled studies are well suited for formal statistical analyses to help present results. The drawback is that ecological validity is often sacrificed for increased control, and analysis tasks are often simplified to ease evaluation. In addition to controlled studies with focused tasks, the use of qualitative methods and case studies can be invaluable for understanding how visualizations are used in realistic and meaningful contexts [21]. Qualitative methods, such as those that apply grounded theory, can provide useful and holistic analyses of visual analytics applications [18].

A number of researchers have included evaluations of process history and provenance tools in their work. For example, Dunne et al. [10] conducted a three-month field study with archaeologists to understand the practical effectiveness of their *GraphTrail* visualization and to gain insights about how users build visual history maps. The researchers also conducted a qualitative lab study to better understand how analysts might use the tool's history tracking functionality. Taking a different approach, Heer et al. [15] analyzed interaction logs for the *Tableau* visualization software to better understand how users used the undo/redo functionality when working with visual history interfaces.

These studies demonstrate how case studies, qualitative evaluations, and log analyses can be valuable for understanding how analysts use process-tracking tools. However, other methods are needed in order to objectively quantify tool effectiveness or to formally compare specific design options. The challenge is that it is difficult to evaluate the degree to which process history is beneficial. It is possible to evaluate the effects of analytic history tools on analysis performance by considering analysis outcomes. For instance, Del Rio and da Silva [8] conducted an evaluation of *Probe-It!*, a provenance visualization tool that shows how maps were created via tree representations showing workflow and contributing information sources. Their study found that the majority of participating scientists successfully completed map analysis tasks with the help of the provenance tools, and far fewer successes were observed without the provenance aid.

While evaluation of analysis outcomes can be useful for determining the effectiveness of provenance visualizations during analysis, real-time support for analysis is just one of the potential benefits of provenance tools. Quality of analysis performance is not indicative of the quality of *memory* of the analytic process at a later time, which is often necessary for repeating the analysis or communicating the steps to others. The analytic process used by a particular analyst to achieve a given goal will often be a unique approach. Further, many analyses and investigations are exploratory in nature, leading to nonlinear processes involving backtracking and multiple lines of logic. Gotz and Zhou [12] explain that the concept of *insight provenance* involves both the history of steps and their rationale during an analytic process. Analysts should be able to reproduce the logic and approach taken to achieve insights and reach a conclusion. Memory of the analytic process is important for accurate communication, such as during collaboration or for presentation. In our work, we are studying methods for evaluating how visual design influences the quality of memory of processes and rationale.

Communication of the analysis process is not necessarily the same as presentation of the results or conclusions of an analysis. Results presentations generally involve the final logic and rationale to justify how that data were interpreted, whereas process communication is more concerned with explaining the steps taken to analyze the data and arrive at the conclusions. Process communication is important for collaborative work or meta-analysis of an analytic approach. It can be useful to know which data and hypotheses were considered and which were not.

While evaluation-based research of memory of process history is limited in visual analytics, research in workflow support and personal information management is relevant to evaluation methods for memory of events and processes (e.g., [7, 17, 20]). For example, Czerwinski and Horvitz [7] conducted research with visual reminder systems to aid workflow memory. In a small user study, the researchers recorded participants for an hour of regular computer work. Later, participants were asked to write down the events that happened during that hour. Participants provided their written memories after 24 hours and then again a month later, and the researchers used the written accounts to assess the number of correctly recalled events and the accuracy of the given times of those events.

In a study with a similar type of reminder tool, Park and Furuta [20] evaluated a tool that saved continuous screenshots of computer work and allowed users to browse the history of images. Because the researchers were focusing on supporting task continuity after workflow interruptions, their evaluation consisted of an activity (making travel plans) that was divided over two work sessions separated by one or two days. To evaluate the tool, the researchers measured how long it took participants to resume the task and start new work at the beginning of the second session. This method evaluates memory based on the time it takes to review and refresh workflow memory and to be able to use that memory to make further progress. The approach demonstrates practical usefulness, though the evaluation criterion is based on a participant's somewhat subjective decision of how much review is necessary. In the following sections, we discuss other methods suitable for evaluating the accuracy of process recall.

3. CASE STUDY SCENARIO

To ground our discussion of evaluation methodology in the context of a real scenario, we discuss our experiences conducting a user study of the effectiveness of a visual analytics tool for process memory. The study utilized an intelligence analysis task based on Mini Challenge #1 from the IEEE VAST 2010 Challenge, which involves a collection of text records about illegal arms dealing. To simplify the analysis task, the data set was condensed to a set of 100 records, and participants were asked to investigate whether there was a connection between illegal arms dealing and the spread of disease in a specified time frame. Participants used a visual text exploration tool that allowed users to spatially organize text records, search for keywords, highlight text, add notes, and link records with connection lines. Figure 1 shows a screen shot from the tool.

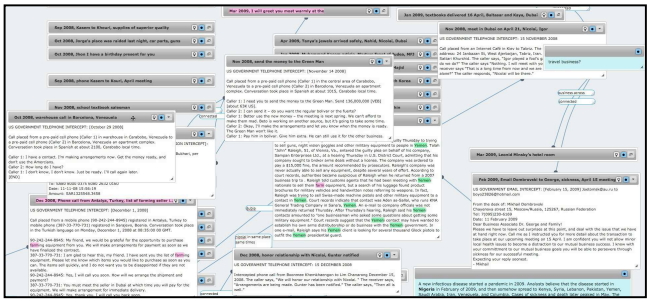

Figure 1. Partial screen capture of the spatial text exploration tool used in the case study.

The study design included two participant sessions separated by one week. In the first session, the experimenter explained the software to participants and allowed them to practice using the tool's features with another data set. Participants were then given 40 minutes to conduct their analyses. To help collect process information, participants were asked to use the think-aloud protocol to describe their thoughts, actions, goals, and intentions. In addition, throughout the analysis, audio and video recordings were used to capture comments and tool usage, and the tool automatically logged user actions and periodically saved screenshots of the workspace. Immediately after the analysis period, participants were asked to explain their findings and the steps taken for the analysis. The first study session took approximately 75 minutes.

Participants then returned one week later for the second session, in which they were again questioned about their analysis processes. The second session took approximately 20 minutes.

The user study was designed to evaluate the extent to which the final state of the visual workspace (i.e., a screenshot taken at the end of the analysis) affects the ability to remember and explain the steps of the analysis. To this end, some participants saw screenshots of the workspace during questioning. The study's 52 participants included student interns (high school, undergraduate, and graduate students) and research scientists. Participant ages ranged from 16 to 60.

With this user study, we were able to test the feasibility of a variety of methods for evaluating process memory. As the focus here is on the evaluation methodology, the results of the experiment itself are not included in this paper. In the following sections, we describe a methodology for evaluating process memory by establishing a baseline for comparison, eliciting process memory, and analyzing collected data to assess memory accuracy. Along with our discussions of possible approaches, we use the study scenario to provide examples of how to use the different methods.

4. EVALUATING PROCESS MEMORY

Memory and communication of a process can be assessed by comparing a recalled process to a baseline process. The goal is to account for specific actions and the rationale for using those actions to move towards an objective. The primary outcomes include accuracy of remembered steps, step ordering, and time estimates for the duration of steps or the entire process. In this section, we discuss methods for establishing a baseline process for comparison with a remembered process, methods for eliciting process memory, and approaches for results analysis.

4.1 Establishing a Baseline

A baseline is needed to accurately evaluate memory of analytic processes through experimentation. Ideally, this baseline process will serve as a record of the steps involved in the analysis in the order that they were executed. Additionally, an accurate event timeline of events would also account for the duration of each step, making it possible to compare times with the perceived times spent on different stages of analysis. Depending on the desired level of granularity for process memory, the baseline process could consist of a detailed list of all actions in the analysis, a thematic list of higher-level stages of the process, or anywhere in between. In addition to the actions or steps taken, a baseline that accounts for the participant's rationale makes it possible to also evaluate memory of logic and intent throughout the process.

4.1.1 Prescribed Baseline

The challenge with creating an accurate baseline is that each participant's process will be different for any real analysis. Creating a baseline consistent across participants would require an approach that required all participants to conduct the analysis in the same way. This could be achieved by asking participants to follow a given procedure or to observe a pre-recorded analysis. As an example using our case study scenario, a video could be created to show pre-recorded screen captures of an analyst using the text exploration tool to investigate the set of documents. Such an approach would greatly simplify the evaluation and increase experimental control, as all participants would have the same analysis experience. On the other hand, forcing participants to follow and remember a predetermined process is not realistic, and the results of such an evaluation may not be as meaningful as one that allows true analytic autonomy.

Further, using a prescribed process eliminates the need for analysts to have any rationale for choosing each step, so evaluation of memory of the participants' intent or rationale may not be meaningful with this approach. It would be possible to provide a summary of contrived rationale at different steps of a given process. For example, using a pre-recorded video of the case study analysis, the video could include voice over of think-aloud comments that explain what the analyst is looking for, why they are reading specific text, or how they are connecting different players in the intelligence analysis scenario. The drawback to providing contrived rationale is reduced realism because the study participants would be remembering arbitrary explanations rather than their own thought processes from an actual critical-thinking activity.

4.1.2 Individualized Baseline

An alternative to a prescribed, consistent baseline is to create a customized baseline for each individual based on his or her actual process. Formalizing the steps of an analytic process is not unknown territory in the field of visual analytics—researchers have shown a number of methods for recognizing steps of analyses. For example, think-aloud protocols require analysts or study participants to verbally explain their thoughts and actions throughout the analysis, providing a record of the both the steps and rationale of the process [3, 5]. In their research of provenance tools, Dou et al. [9] demonstrated the effectiveness of video, think-aloud protocols, and system logs to detail analytic processes. The researchers found that system logs were able to account for approximately 79% of the findings identified by the traditional video and think-aloud methods.

The advantage of using individualized baselines over a prescribed baseline is that it supports more realistic analysis and can account for process variability among different analyses. On the other hand, this natural variability of individualized baselines may result in reduced experimental control for the evaluation. Another drawback is that the individualized approach can take extra time to establish the baseline after reviewing an analysis session. It is important to establish a protocol for determining individualized baselines from the collected process data, and this alone can be challenging. System logs and think-aloud methods can reduce the amount of time needed to identify key events or actions in the analysis, but most types of analysis will still need additional review and coding to determine the baseline.

As part of our case study, we created an individualized baseline for participants conducting the intelligence analysis task about arms dealing and disease. For each participant's analysis, we

reviewed think-aloud comments, video, screen captures, and system logs. We then used thematic coding to summarize the primary steps taken, topics investigated, and strategies used in the intelligence analysis task. For example, if system logs showed that a participant conducted multiple searches for terms related to sickness (e.g., *sick*, *disease*, *ill*, or *infectious*), this helped to establish an investigation theme about sickness. As another example, if a participant provided a think-aloud update about being interested in a certain country, and then the corresponding video and screenshots at that time showed that the participant next reviewed several records involving that country, then this would indicate a theme about investigation of that country.

Our experiences with the individualized baseline approach revealed a major limitation: the baseline will be based on researchers' or coders' interpretations of the process, rather than on the analyst or participant's thoughts or intentions. As such, this approach is more powerful for establishing a baseline about observable actions than for rationale and intentions. While think-aloud comments helped greatly for understanding rationale, the frequency of updates and the amount of given detail varied for participants.

4.1.3 Self-Reported Baseline

If we are most interested in evaluating process memory or communication at a later time, it is possible to ask the analysts or participants to self-report their processes immediately after the analysis. The self-reported summary could then be used as a baseline for later comparisons. In our case study, immediately after the analysis task, we asked participants to explain what they did. We asked them to explain the actions that they took, the topics they considered in the investigation, and the rationale for each step they took.

The advantage of self-reports over coding individualized baselines is that self-reports allow for a baseline based on what participants think they did, rather than what coders think they did.

Unfortunately, the problem with self-reporting approaches is that memory degrades quickly, and even immediate post-analysis reports might not be accurate descriptions of the actual processes. Further, accuracy and reliability of self-reports can be limited by participant communication abilities. From our own experiences and tests thus far, we have found these to be common problems; thus, we do not recommend the use of self-reporting for a process baseline. However, self-reports can still be useful for explaining rationale or highlighting key steps that can be considered along with video coding, think-aloud updates, or event logs to help determine individualized baselines.

4.1.4 Summary of Baseline Trade-offs

Each of the described methods for establishing a baseline of the analytic process has its advantages and disadvantages. Table 1 shows a simplified organization of the trade-offs among the prescribed, individualized, and self-reported baselines. A prescribed baseline is ideal for a consistent process and supports the highest level of experimental control, but a prescribed analysis severely restricts the realism of an analytic process, which may be a serious concern for ecologically valid studies of provenance tools. Individualized baselines are appealing for their high analytic realism, but this comes at the cost of effort and some control. The choice of an appropriate approach for a particular study depends on the goals and priorities of that study, and either an individualized or a prescribed approach could work well in an evaluation of process memory. For the experiment we conducted as part of our case study, we decided to use individualized baselines instead of a prescribed baseline because we prioritized realism and ecological validity over experimental control.

We also tested the use of self-reported baselines in our case study. We found that this method has a major disadvantage of low reliability, which presents a high risk of invalidating an evaluation. Consequently, we recommend against the use of self-reported baseline for evaluation of process memory.

Table 1. Trade-off summary for three methods of establishing a baseline for process memory. High strengths for factors are denoted with '+', moderate strengths are denoted with 'o', and weaknesses are denoted by '–'.

	Type of Baseline		
Factors	**Prescribed**	**Individualized**	**Self-reported**
Cost & Time	+	–	o
Control & Reliability	+	o	–
Analysis Realism	–	+	+

4.2 Eliciting Process Memory

Once a baseline has been established, it is possible to compare a memory or communication of the analytic process with the baseline. Process memory can be assessed based on accuracy of the remembered analysis steps, the order of those steps, and the time taken to complete each step. In this section, we discuss methods for eliciting process memory.

4.2.1 Process Reproduction

A thorough way to assess memory of a process is to require an analyst to repeat the analysis process for the same analytic task using the same tools. For the evaluation of provenance tools, participants would be able to use the tools to help them to recreate the steps of the analysis. This *process reproduction* method has the advantage of allowing highly accurate process replications, which can then be compared to original tool usage and executed actions in the baseline analysis. Because the method for process recall would be the same as the original analysis method, the same coding method could be used for both the baseline and reproduction analyses (though provenance-referencing steps would have to be filtered out from the reproduction phase if provenance tools were available in the memory elicitation analysis).

For our case study example, the reproduction approach would require participants to start over and repeat the investigation with the text exploration tool starting from its original state. It would be important to explain to participants that the goal is to repeat the same approach that they took previously, rather than to gain additional insights about the data.

The process reproduction method can account for both accuracy and order of steps. On the other hand, reproduction may not be appropriate for assessing process times because it would be expected that steps could be completed faster during the second time through. Additionally, if it is possible to rely on provenance aids to recreate actions, the reproduction might not be useful for evaluating memory of rationale for the actions. To mitigate these weaknesses, the reproduction approach could be augmented by asking participants to explain the rationale for each step and to

estimate the time taken for steps. The given rationale and times could then be compared to those of the original analysis as determined by think-aloud protocols or post-analysis interviews.

A major disadvantage of the reproduction approach is that reproducing an entire analysis can be time consuming, which can significantly increase the cost of the evaluation. Also, because process reproduction relies primarily on actions, this approach does not necessarily account for communication.

4.2.2 Written or Verbal Walkthrough

Rather than having analysts or participants reproduce the analysis, another option is to ask the participants to walk through the steps of the analysis. For many types of analysis, this approach can be faster than full reproduction. Similar to the reproduction approach, *written or verbal walkthroughs* accounts for free recall of steps as well as order. Additionally, participants' reports can include time estimations.

In our case study, we asked participants to provide verbal walkthroughs of their analyses of the text records about illegal arms dealing. The study revealed that the most challenging aspects of the walkthrough were explaining the level of detail that participants needed to provide and then encouraging them to continue explaining the entire analysis process. We suspect that it was tedious for participants to describe the entire process from the 40-minute activity. Many participants tried to describe vague high-level summaries (e.g., *"I kept searching for things, and then I would read anything that looked interesting, and then I would keep doing that."*). The experimenter often needed to provide continual prompting and ask clarifying questions to elicit a complete account of the analysis process.

Walkthroughs are dependent on communication ability, which can be viewed as an advantage if process communication is a major element of interest. But the dependency on communication ability undoubtedly adds complexity to the walkthrough method. A consistent interview protocol is needed for eliciting details, and responses will need to be coded for comparison to the baseline process. Some people are likely to be more inclined to give more details than others, so consistency can be a challenge. While an interviewer can help encourage a participant to provide additional details in a verbal walkthrough, collecting sufficient details may be more problematic with a written report without a moderator. Further, clearly accounting for content and order can be difficult for both written and verbal methods. Descriptions are often not chronological—people will add more details as they remember them, so the coding effort will need to construct a coherent chronology from the given events.

4.2.3 Step Ordering

The reproduction and walkthrough methods both require participants to recall the steps of their processes, which makes the evaluation of order dependent on recall. The *step ordering* evaluation method avoids this problem by eliminating free recall of steps and focusing only on order. In step ordering, the baseline process must first be coded and broken down into key steps for each participant. Then, to evaluate memory of order, the participant or analyst is asked to organize the steps of the process into the correct ordering. This could be done, for example, by labeling index cards or using a simple software application (such as PowerPoint) that allows ordering labeled items. The given order can then be easily compared to the true order of the baseline process (as will be discussed in 4.3.2).

We asked participants to complete a step ordering activity for the second session of our case study (i.e., one week after completing the intelligence analysis activity). To prepare the method, we referred to the steps and themes from the thematic coding that was done for the individualized baseline approach (as previously explained in section 4.1.2). We then summarized 10 themes on PowerPoint slides (one slide for each step or theme) and jumbled the order. Then, when participants returned for the second session, we asked them to put the slides into the order that they completed the steps in the investigation.

The primary advantages of the step ordering method are that it is fast, easy to quantify, and maintains a fairly high level of control for assessing order recall. One disadvantage is that it may likely involve more recognition than recall because it gives participants the correct steps. Furthermore, it may be possible to guess a plausible order based on available steps. If the progression of steps is obvious, the evaluation will not be useful. In addition, the step ordering method is generally not well suited for assessing immediate post-analysis memories because it requires time to code and break down the key steps (unless the analysis relies on a prescribed process, rather than a free analysis).

From our experiences, we also found that it was difficult to select key steps for analyses that included cyclic investigations. For example, in the case study, if participants first searched for terms related to sickness at the beginning of the analysis, and they again searched for sickness later in the analysis, then including a *sickness* theme would introduce complications for linear ordering. To avoid such problems, it was sometimes necessary to either exclude repeated topics from the selected themes or to provide additional details about the steps to remove sequential ambiguity.

4.2.4 Modified Step Ordering

As an alternative to conducting step ordering by providing only the steps from the baseline process, a *modified step ordering* method can introduce extraneous steps into the set of steps to be ordered. For example, for the intelligence analysis activity used in the case study, additional steps could be created for topics that could be plausible for the task but were not included in the data set (e.g., *mustard gas incident*, *sabotaged satellite launch*, or *farm equipment manufactured in Texas*). It would also be possible to include actions that were not performed in the baseline process (e.g., *move all records about car parts to the far right* could be a extra step for an analysis that did not involve moving the indicated records).

Analysts or participants can be told that some of the steps are wrong or extraneous, and they will have to both recognize and order the correct steps. The advantage to this modified step ordering approach is that it could make the correct ordering less obvious. Additionally, this method makes it possible to include a measure of recognition accuracy based on the inclusion or exclusion of erroneous steps in the guessed ordering.

The drawback of the modified step ordering method is that it increases complexity of the evaluation and reduces control. It could be difficult to determine what extraneous steps to add. If the extra steps are created from a set of plausible steps for the analysis task and data set, then all participants might not be able to have the same additional steps. Again, the exception would be if a prescribed analysis process were used as the baseline, in which case the same extraneous steps could be used for all participants.

4.2.5 Summary of Memory Elicitation Trade-Offs

The strengths and weaknesses of the discussed methods for eliciting process memory are summarized in Table 2. Process reproduction has many advantages but suffers from the high cost of reproducing an analysis and the time needed to code the steps.

Written or verbal walkthroughs offer most of the advantages of process reproduction at lower cost. Walkthroughs are also greatly dependent on communication ability, which could be considered as either a benefit or a limitation depending on whether a study is focusing on memory only or on communication as well. The step ordering methods are limited to evaluating order only, but they may be appealing for their simplicity, control, and low cost.

While we have separated the types of methods for eliciting process memory in order to organize the discussion of trade-offs, it is important to note that method selection does not have to be limited to any single method. It is certainly possible to combine properties of different methods, such as how verbal explanations could be used during process reproduction to provide additional information about rationale or step duration.

It is also possible to use multiple elicitation methods in sequence. For example, because step ordering methods are fast but do not account for free recall, it could work well to use step ordering after measuring process memory using a technique that does involve free recall (i.e., walkthrough or reproduction). Of course, the order that methods are used is important. It would not make sense, for instance, to ask participants to provide a walkthrough after they complete a step ordering activity because the step ordering method provides the steps of the process. It would be expected that participants might recall the steps listed from the step ordering method rather from their memories of their processes.

In our case study, we used both verbal walkthroughs and step ordering to elicit process memory. We wanted to evaluate process memory both immediately after the analysis and then again one week later. In each participant's first session, we asked the participants to provide a verbal walkthrough of the investigation immediately after the analysis. We decided against using process reproduction because of time constraints. In the second session, we again asked participants to provide verbal walkthroughs, and we then asked them to complete the step ordering task. We decided against the use of the modified version of step ordering for the sake of simplicity.

Table 2. Summary of trade-offs for methods of eliciting process memory. High strengths for factors are denoted with '+', moderate strengths are denoted with 'o', and weaknesses are denoted by '–'.

	Process Memory Elicitation Method			
Factors	**Repro-duction**	**Walk-through**	**Step Ordering**	**Mod. Step Ordering**
Cost & time	–	o	+	o
Control & Simplicity	–	–	+	o
Step recall	+	o	–	o
Rationale recall	+	+	–	–
Process order	+	+	+	+
Step duration	–	+	–	–
Communication independent	+	–	+	+

4.3 Analyzing the Results

After establishing a baseline process and recording a remembered version of the analysis, the next step is to compare the two to assess the accuracy of the process memory. Quantitative measures and methods can be helpful for clear comparisons of results for different analysis trials or participants, though these methods depend on qualitative methods for identifying the component steps for quantification. Qualitative analyses of process memories and explanations are also important for meaningful interpretations of quantitative findings and differences.

4.3.1 Percentages for Process Coverage

One example of a quantitative method for assessing process memory is to approximate the percentage of process coverage from the remembered version. The number of recalled correct steps can be counted and compared to the number of steps in the baseline process. Similarly, counting the number of extraneous or erroneous steps can allow for calculation of error percentages. These measures are relatively simple to apply to free recall methods such as process reproduction, written walkthroughs, or verbal walkthroughs. In addition, because the measures are normalized as percentages, they can accommodate processes of different numbers of steps, as might be found with individualized baselines. The quantitative results can be easily compared for different visualization tools, analyses, or experimental conditions using descriptive or inferential statistics.

Of course, counting steps requires some type of coding method to identify the steps in both the baseline and remembered processes. Percentage of recalled steps is a simple measure, but it can serve as a straightforward means of comparison. For more meaningful explanations, step percentages can be combined with qualitative descriptions of process memories and their differences.

We considered objectively measuring process coverage in our case study, in which we established individualized baselines for participants and had them provide verbal walkthroughs. We found that it was difficult to clearly distinguish between steps in our open-ended, text-based analysis task. Additionally, the steps included in participants' walkthroughs were often reported at varying levels of detail, with some steps glossed over and others discussed in depth. As a result, we were ultimately dissatisfied with the option of using percentages of process coverage as a measure of recall accuracy. We suspect that this approach may work better for types of analyses that involve more concrete steps or less exploratory analysis, but more work is needed to understand the effectiveness and limitations of objectively measuring process coverage.

4.3.2 Rank Correlations for Step Order

In addition to the step coverage, quantitative methods can be used to help assess the accuracy of the order of steps in the remembered process. To compare step orderings to the baseline ordering, rank correlations, such as Spearman's rank correlation, can be used. This worked well in the analysis the step ordering results collected in our case study. Rank correlation analyses for process order cannot account for extraneous steps because the correlations assume a correspondence between the steps in the recalled and baseline processes; however, this is reasonable because it does not make sense to test the order of events that have no correct place in the sequence of steps. Correlation measures are useful because they provide standardized test values regardless of the number of steps, which is important when evaluating against an individualized process of variable length. Rank correlations can be easily applied to results from the *step*

ordering method of eliciting process memory because steps have already been identified, and the given steps match those in the baseline. Rank correlations could also be used with recall methods (i.e., process reproduction, walkthroughs, or step ordering) by coding the key steps of the recalled process and including only the correctly recalled steps for the correlation with the correct ordering of those steps in the baseline.

4.3.3 Times for Step Duration

Besides step order and the percentage of recalled steps, another metric to consider for process memory is memory of the amount of time taken to complete individual steps or the entire analysis. Reported time estimations are trivially quantitative for numerical estimations given in a walkthrough. Alternatively, estimations of perceived step times could be assessed using relative times by ranking the duration of steps. Rankings could then be compared to the true ordering using rank correlation tests.

In our case study, we found that analysis processes did not always have phases or steps with clear beginnings or ends. However, it would have been possible to have coders estimate step times and to compare those times with estimations given by the participants. For tools that involve the creation of multiple visualizations or views, the time taken to generate visualization could be measured more accurately than in our case study scenario.

4.3.4 Times for Recall Efficiency

Another possible time measure is the amount of time taken to recall, explain, or reproduce the process. While comparing perceived step times from the remembered process to those of the baseline makes it possible to evaluate the accuracy of perception of step duration, measuring the time needed to recall the process provides a measure that corresponds to the difficulty or efficiency of recalling the process. Collecting this measure involves recording the time taken to reproduce the analysis with the process reproduction approach, explain the steps using the walkthrough approach, or order the process steps using the step ordering approach. For analysis, faster times can indicate easier recall, but recall times must be analyzed in conjunction with accuracy measures (e.g., coverage percentages or order correlations) for meaningful interpretation. For example, a fast explanation that is incorrect is not better than a correct slow explanation. An efficiency ratio of recall accuracy to recall time can aid interpretation of speed and accuracy with a single metric that can be analyzed with traditional quantitative methods.

In our experience using verbal walkthroughs in the case study, the major problem with considering recall time was that participant personality and communication ability greatly affected walkthrough times. Recall efficiency might be better suited for process reproduction or step ordering methods that are less dependent on communication. Alternatively, recall efficiency might be more appropriate for within-subjects comparisons that can account for participants' individual differences.

4.3.5 Subjective Ratings

Depending on the complexity of the analytic task and the communication abilities of study participants, it can be difficult to code steps to assess process coverage or step ordering. As an alternative to establishing an objective protocol to determine quantitative measures of memory quality, a more subjective rating system may be used. Human raters can review both the original analysis process and the remembered version, and then they can assign a score to indicate the quality or accuracy of the memory or explanation.

The rating approach is best suited for free-recall methods, such as process reproduction or walkthroughs. Because ratings are subjective, the primary advantage of a rating approach is that it allows flexibility and can accommodate human judgment for the quality of a process memory or explanation. The scoring protocol can include holistic ratings to account for the combination of process accuracy, process ordering, duration accuracy, and recall time. Additionally, more focused criteria can ask raters to score specific qualities of a remembered process, such as communication clarity, speed, accuracy of process coverage, or accuracy of step order. Quantitative ratings can be analyzed with traditional quantitative methods to compare memory and communication results from different study conditions.

A potential disadvantage of a rating approach is that it can be costly or difficult to recruit and train raters. Ideally, the raters should be blind to the study conditions. Also, rating necessitates consideration for the number of raters and rater reliability [14, 23].

Due to the exploratory nature of the intelligence analysis task of our case study, subjective ratings were the most useful means of quantifying accuracy of process walkthroughs. Two raters separately reviewed the recorded data for each participant's analysis (video, audio, screen shots, and system logs) to establish a baseline, and then the raters separately scored each participant's verbal walkthrough. We found that this method worked well for rating the overall quality of the walkthrough and coverage of the main themes or topics. However, we were unable to obtain ratings for participant intentions or rationale, as these were not always apparent from participant comments and the captured data. Similarly, participants were inconsistent in the amount of detail they provided about intentions and thoughts in their walkthroughs.

4.3.6 Selection of Analysis Methods

Comparisons of memory results from different tools or study conditions are useful for identifying differences in effectiveness and outcomes. The components of an evaluation methodology should complement each other; appropriate analysis methods will depend on the goals of the study and type of data collected. We encourage the use of multiple types of data analyses (including both quantitative and qualitative methods, when appropriate) in order to provide a more complete understanding of visual analytics tools.

5. CONCLUSION

We can evaluate memory and communication of analytics processes by comparing remembered or reproduced processes to a baseline process. When analyzing responses and coding actions, primary metrics can include presence of remembered steps, step ordering, and time estimates. For objective comparisons, it is important to establish a baseline sequence of steps. While relying on a prescribed analytic process is useful for high experimental control and simplifying evaluation, such an approach reduces the realism and meaningfulness of process memory. The creation of individualized baselines for each participant's analysis supports a more realistic evaluation but at the cost of ease and control.

We discussed several methods for measuring process memory and communication. The process reproduction method has the advantage of supporting realistic and accurate replications of analyses that can be directly compared to a baseline analysis. However, process reproduction can be time consuming and require additional modification to elicit memory of step rationale. Free recall through written or verbal process walkthroughs can be faster than full reproduction and can more naturally incorporate

rationale reports, but dependency on communication ability adds complexity to the approach. Finally, the step ordering method is fast and provides a convenient means of evaluating process order, but the approach involves recognition rather than recall, and it would not work well for processes composed of obvious progressions of steps. These limitations can be partially mitigated by modifying the step ordering method to include extraneous analysis steps, but the trade-off is added complexity for results analysis.

Thus far, we found that existing evaluation efforts of process memory is limited in the realm of visual analytics. In ongoing research, we are working on a controlled experiment to evaluate the effectiveness of process memory aids, and we are testing the presented evaluation methods to study the effectiveness of visual history tools. As we progress with our research, we will further our knowledge of which methods are effective and useful for evaluating memory and communication of analytic processes. By organizing potential methods for evaluating analytic process memory, we hope that this paper can inform study design and encourage objective evaluation of process memory and communication. Every method has its own set of advantages and disadvantages, but a meta-analysis of methods is useful for helping researchers to select the best methods for their purposes.

6. ACKNOWLEDGMENTS

This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

7. REFERENCES

[1] Amar, R., Eagan, J. and Stasko, J. Low-level components of analytic activity in information visualization. IEEE Symposium on Information Visualization (2005), 111-117.

[2] Andrews, C., Endert, A. and North, C. Space to think: large high-resolution displays for sensemaking. Proceedings of the 28th international conference on Human factors in computing systems (2010), 55-64.

[3] Andrews, K. Evaluating information visualisations. Proceedings of the 2006 AVI BELIV workshop (2006), 1-5.

[4] Bavoil, L., Callahan, S. P., Crossno, P. J., Freire, J., Scheidegger, C. E., Silva, C. T. and Vo, H. T. Vistrails: Enabling interactive multiple-view visualizations. IEEE Visualization (2005), 135-142.

[5] Boren, T. and Ramey, J. Thinking aloud: Reconciling theory and practice. IEEE Transactions on Professional Communication, 43, 3 (2000), 261-278.

[6] Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T. and Vo, H. T. VisTrails: visualization meets data management. Proceedings of the 2006 ACM SIGMOD international conference on management of data (2006), 745-747.

[7] Czerwinski, M. and Horvitz, E. An investigation of memory for daily computing events. People and Computers XVI-Memorable Yet Invisible (2002), 229-245.

[8] Del Rio, N. and da Silva, P. P. Identifying and explaining map imperfections through knowledge provenance visualization. University of Texas at El Paso, 2007.

[9] Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R. and Chang, R. Recovering reasoning processes from user interactions. IEEE Computer Graphics and Applications, 29, 3 (2009), 52-61.

[10] Dunne, C., Henry Riche, N., Lee, B., Metoyer, R. and Robertson, G. GraphTrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2012), 1663-1672.

[11] Freire, J., Koop, D., Santos, E. and Silva, C. T. Provenance for computational tasks: A survey. Computing in Science & Engineering, 10, 3 (2008), 11-21.

[12] Gotz, D. and Zhou, M. X. Characterizing users' visual analytic activity for insight provenance. Information Visualization, 8, 1 (2009), 42-55.

[13] Griffin, A. L., MacEachren, A. M., Hardisty, F., Steiner, E. and Li, B. A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. Annals of the Association of American Geographers, 96, 4 (2006), 740-753.

[14] Gwet, K. Handbook of inter-rater reliability. Gaithersburg, MD: STATAXIS Publishing Company (2001), 223-246.

[15] Heer, J., Mackinlay, J., Stolte, C. and Agrawala, M. Graphical histories for visualization: Supporting analysis, communication, and evaluation. Visualization and Computer Graphics, IEEE Transactions on, 14, 6 (2008), 1189-1196.

[16] Heer, J. and Shneiderman, B. Interactive dynamics for visual analysis. Queue, 10, 2 (2012), 30.

[17] Horvitz, E., Dumais, S. and Koch, P. Learning predictive models of memory landmarks. Proceedings of the CogSci (2004.

[18] Isenberg, P., Zuk, T., Collins, C. and Carpendale, S. Grounded evaluation of information visualizations. Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (2008), 6.

[19] Kadivar, N., Chen, V., Dunsmuir, D., Lee, E., Qian, C., Dill, J., Shaw, C. and Woodbury, R. Capturing and supporting the analysis process. IEEE Symposium on Visual Analytics Science and Technology (2009), 131-138.

[20] Park, Y. and Furuta, R. Keeping narratives of a desktop to enhance continuity of on-going tasks. Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (2008), 393-396.

[21] Plaisant, C. The challenge of information visualization evaluation. Proceedings of the working conference on Advanced Visual Interfaces (AVI) (2004), 109-116.

[22] Saraiya, P., North, C. and Duca, K. An insight-based methodology for evaluating bioinformatics visualizations. Visualization and Computer Graphics, IEEE Transactions on, 11, 4 (2005), 443-456.

[23] Shrout, P. E. and Fleiss, J. L. Intraclass correlations: uses in assessing rater reliability. Psychological bulletin, 86, 2 (1979), 420.