# A Case Study of Using Analytic Provenance to Reconstruct User Trust in a Guided Visual Analytics System

Nadia Boukhelifa*
UMR MIA-Paris, AgroParisTech, INRAE
Univ. Paris-Saclay

Evelyne Lutton†
UMR MIA-Paris, AgroParisTech, INRAE
Univ. Paris-Saclay

Anastasia Bezerianos‡
Univ. Paris-Saclay, CNRS,
INRIA

## Abstract

In this paper, we demonstrate how analytic provenance can be exploited to re-construct user trust in a guided Visual Analytics (VA) system, and suggest that interaction log data analysis can be a valuable tool for on-line trust monitoring. Our approach explores objective trust measures that can be continuously tracked and updated during the exploration, and reflect both the confidence of the user in system suggestions, and the uncertainty of the system with regards to user goals. We argue that this approach is more suitable for guided VA systems such as ours, where user strategies, goals and even trust can evolve over time, in reaction to new system feedback and insights from the exploration. Through the analysis of log data from a past user study with twelve participants performing a guided visual analysis task, we found that the stability of user exploration strategies is a promising factor to study trust. However, indirect metrics based on provenance, such as user evaluation counts and disagreement rates, are alone not sufficient to study trust reliably in guided VA. We conclude with open challenges and opportunities for exploiting analytic provenance to support trust monitoring in guided VA systems.

## 1 Introduction

Guidance in visual analytics (VA) is a family of techniques that seeks to gradually and actively close the knowledge gap encountered by users during an interactive visual analytics session [8]. Ceneda et al. characterise this knowledge gap into unknown targets or desired results, unknown paths to reach the desired results, and unknown targets and paths, with each gap necessitating different types and degrees of guidance.

In a comprehensive survey of guidance in visual data analysis, Ceneda et al. [9] describe three levels of guidance which vary in terms of how much instruction is given to the user: prescribing, directing and orienting. In the prescribing category, Ip et al. [22], for instance, developed a guided VA system which provides users with step-by-step instructions to explore the most interesting views in a large collection of images. Directing systems provide less detailed recommendations to the user. For example, Voyager [44] provides suggestions of different visualization alternatives that are ranked based on some statistical and perceptual measures. Similar to directing guidance, orienting systems also provide general suggestions to the user but without a clear order or priority [9]. For example, Vizster [19] guides user attention using color to indicate the presence of communities in social networks.

In guided VA systems, the human and the machine work collaboratively to achieve a task [9]. However, when guidance is discussed it is often implied that the system is guiding the user. For example, Collins et al. [13] describe six broad goals of such guidance: to inform, mitigate bias, reduce cognitive load, for training, for engagement, and to verify conclusions. Nevertheless, there are also guided VA systems (e.g., [2, 15, 16]) where an underlying algorithm also adapts to user feedback, and may therefore be considered as "guided" in-turn by the user input. In EVE [2] for instance, the system is steered to pertinent views of a multi-dimensional search space based on user feedback. Ceneda et al. also distinguish between the two directions of guidance in VA, guidance provided by the system to support the user, and guidance provided by the user to support the system [9].

Combining human and machine intelligence is challenging [21], necessitating a clear identification of roles within the human-machine partnership. And it requires a careful consideration of the various trade-offs that are related to the objectives of these systems, such as optimising for model accuracy and interpretability, or user confidence and trust [3]. Furthermore, because of the intertwined roles of the human and the machine, and the learning component present in many guided VA systems, we do not always know precisely at any moment in time who is guiding who, the system or the user, which may hinder user trust and their confidence in the output of the system.

Various methods to measure user trust are described in the literature, from objective metrics to subjective evaluations. A common method to assess trust is to use questionnaires after an interactive session with the VA system [27]. Whereas questionnaires can provide valuable feedback on users' overall impression of trust or their confidence, they do not provide detailed feedback about the specific aspects of the system that the user may have trusted or distrusted, or at what point in time this happened. Besides, VA users may be biased when asked to reflect a-posteriori on their trust of the VA system (e.g., participants over-emphasising attention to errors, and thus reporting unjustifiably reduced trust, as described in [20]).

We are interested in exploiting analytic provenance [30] data to study trust in guided VA systems. Many interactive systems keep track of user interactions and system states in log files. These are largely more available than subjective user evaluations of trust, particularly when considering trust during VA system usage. In this paper, we describe a case study of a directing guided VA system where we attempt to reconstruct user trust based only on existing interaction log data. We contribute a discussion of the main challenges we encountered and new opportunities for exploiting provenance information to monitor trust in guided VA.

## 2 Various Notions of Trust

Various definitions of trust exist in the literature, with many noting its complex, dynamic and multidimensional nature [27, 38]. Trust is often linked to confidence. For example, Madsen and Gregor [27] define human-computer trust as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid". Shaw [37], however, highlights the affective components of trust, also noted by Madsen and Gregor, which can be in part based on faith. Confidence is seen as more factual and arises as a result of specific knowledge. In a similar vein, and inspired by [28], Han et al. [18] describe trust in the context of VA as "the truster (user)'s belief

*e-mail: nadia.boukhelifa@inrae.fr

†e-mail: evelyne.lutton@inrae.fr

‡e-mail: anab@lri.fr

that the trustee (VA system) will help them correctly identify and visually distill the most valuable and relevant information content". Lee and See [25] also define trust using affective traits as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [11].

A different take on trust is provided by Visser et al. [14], who examine trust as a process that can be updated continually over time rather than the final product or the goal. They propose trust cues, which are information elements and visualizations useful to make a trust assessment about a human being or an agent (e.g., by showing the risk and uncertainty of the provided information). Similarly, Sperrle et al. [38] argue that calibration and re-calibration of trust are necessary over time. Trust calibration has been the focus of much recent work, and has looked at ways to align user trust of a VA system and the system's actual trustworthiness such as by communicating uncertainty or providing visual cues [18].

In general, trust in guided VA follows the traditional definitions of trust [11, 18], where the trustee is the user and trust is the level of confidence and belief that the system is presenting the most relevant information. There are two aspects that seem unique to guided VA. In guided VA work, we often see the object of trust being the learning process itself, i.e., user trust on whether the system adapts and learns appropriately from the user's feedback and input [23, 50]. Moreover, although we are not aware of any studies of trust per-se in guided VA systems, other work on mixed-initiative systems more generally (not necessarily with a VA component) exist. For example, in a system where users collaborate with a robot to perform navigation tasks [12], it is hypothesised that personal characteristics of the user (locus of control) may affect whether users trust system recommendations and how they decide to collaborate (or not) with the system. While this last work does not necessarily include a learning component in the system side, it highlights how trust in human-machine partnerships can be affected by personal traits and can manifest as different levels of interaction and collaboration with the system.

In this paper we also follow the traditional definitions of trust [11, 18] and use the terms trust and confidence interchangeably. However in our case, the system can also be considered a "truster" and the user a trustee, as the VA system guides the user to "interesting" areas of the search space, but also the user steers the system towards most pertinent views through interactive feedback.

## 3 Provenance and Trust

Previous work has investigated the use of provenance information about the data [47] and human analytical paths [45, 46] to make trust assessments. Venters et al. [42] track personalised provenance such as based on a person's role in an organisation, and recommend using the provenance of the data and the analytical process to gauge user trust in the process. Sacha et al. [33] discussed the role of uncertainty, awareness and trust in VA and argued that users' confidence in the VA results depends on their degree of awareness of the different types of uncertainty that are present or generated in the system. They also advocate for capturing analytic provenance and for using human analytic paths to build trust measures that are then contrasted with system uncertainty. Other work looked at combining provenance information with different types of metrics. To estimate trust, Ceolin et al. [10] proposed to combine analytic provenance with user reputation. They built and evaluated a computational pipeline whereby relevant provenance features are extracted, then used to generate 'stereotypes' of user behaviour. They then estimate the reputation of both stereotypes and users, and use this information to determine the trustworthiness of artifacts.

Findings from previous studies who explored analytic provenance of the sensemaking process suggest a link between characteristics of the knowledge discovery process and the level of user trust [35, 42]. Basing their work on the knowledge generation model for VA [35], Sacha et al. suggested that trust can be inferred by examining how long or how often the user stays in each exploration or verification loop [34]. As Xu et al. [46] put it: "tighter exploration loops suggest confidence, whereas scattered exploration suggests distrust of the process". However, formally measuring and ultimately quantifying a trust inference model from user's analytic provenance (such as their knowledge discovery cycles) is still an open research challenge.

Inspired by these different works on provenance and trust, we propose to record not only users' exploration paths, but also the system state at key stages of the exploration. Similar to Sacha et al., our approach consists in confronting user trust metrics, constructed from their analytic provenance, and uncertainty of the system, that we also compute from our system logs.

## 4 Measuring User Trust in the System

Researchers in machine learning and artificial intelligence systems use different methods to ensure or measure user trust and reliability. Frequent interactions with the algorithmic process seems to develop user trust and satisfaction (in terms of goals and expectations) [17] while in the same time cognitive workload [51] and user fatigue [24, 41] remain a limiting factor. Some Interactive Evolutionary Computation (IEC) systems propose a self-trust assessment of user rating [32] to provide information to the machine learning process (i.e. the surrogate model computation). Findings from Yin et al. [48] show that trust is directly impacted by user's own estimations of the system's accuracy. Nourani et al. also found that the degree to which users agree with the system's outputs can be considered a proxy or indirect measure for reliance and trust [31], and therefore can be used to identify a variety of trust situations including distrust, overtrust and calibrated trust [25]. Building on this measure, Yu et al. [49] propose a reliance rate based on the number of times the user agreed with the system answers out of all their decisions. Similar to previous work, we experiment with various metrics to infer user trust from past analytical provenance information, as discussed in the next section.

## 5 A Case Study: Reconstructing Trust from Analytic Provenance

We present a case study in guided VA where we infer user trust in a VA system, based only on interaction and system log data that we collected from a previous study. In what follows, we describe the VA system, the user study and provenance data, and our exploration of four metrics that we use to indirectly measure user confidence and trust in the VA system and its learning component. We also compare the inferred user trust with system uncertainty regarding user goals and exploration strategies.

### 5.1 A Guided VA System using IEC

The guided VA system we are dealing with [7] is a SPLOM-based exploratory visualization tool (Fig. 1). The user can select a scatterplot from the SPLOM and create various selections in a zoomed in view. Using Interactive Evolutionary Computation (IEC) [40], the tool progressively builds combined dimensions as axis-parallel projections to explore the original dimension space (using principle component analysis) and non-axis parallel projections for a more extended search. The system then proposes new views (i.e. scatterplots) to the user based on a fitness function which takes into account: (i) the amount of visual pattern in the 2D projection. The assumption is that users are interested in finding patterns in their data, such as linear relationships and outliers. These patterns are detected using the scagnostics library [43] (Fig. 2); (ii) the complexity of the proposed combined dimension - the IEC favours simple mathematical functions; and (iii) the user evaluation of the view which is captured interactively via a slider, on a scale of 1 to 5 (from least to most interesting).

In this VA system, the user steers the system by providing an evaluation of the scatterplots they visit, and the system guides the user

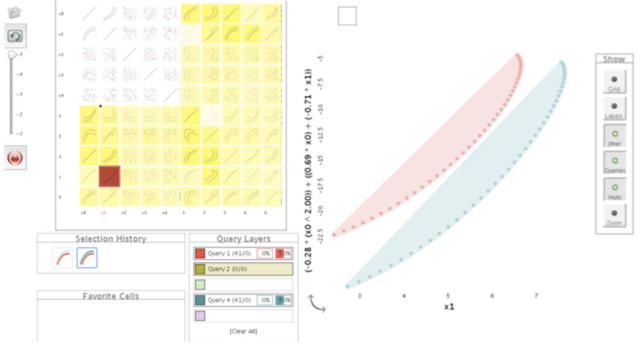

Figure 1: The guided visual analytics (VA) tool we used in our case study. An interactive evolutionary algorithm evolves new scatterplots (in yellow background color) and guides the user towards views that are more "interesting" (cells with darker background color). The user can modify the system evaluation of each view using a slider, then press a button to evolve a new generation of scatterplots.

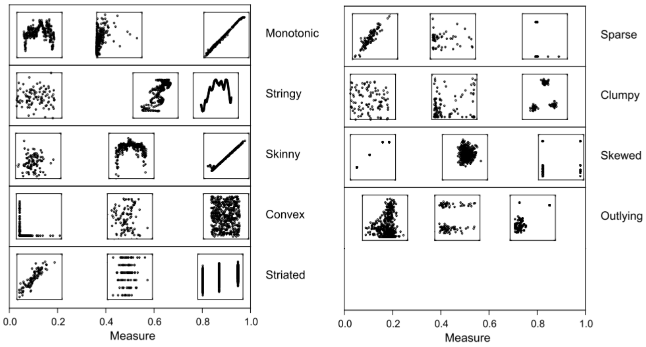

Figure 2: Nine scagnostics measures from [43] to compute the amount of visual patterns in scatterplots.

by providing a ranked list of views, based on the fitness function and scatterplots from past user exploration cycles (called generations). An approximated user model (also called a surrogate function) is learned from past user interactions and is used to filter obvious bad solutions and identify possible interesting views.

To steer user exploration, the IEC provides *directing guidance* [9] through a ranked list of recommended scatterplots. In our guided VA, rank is visualised using background color intensity, the darker the color the higher the rank. Users also influence the system's recommendations through direct actions when they evaluate the scatterplots using the slider, and indirectly based on their past SPLOM visits. In either case, the feedback provided by the user informs the creation of the next generation of recommended scatterplots (i.e. this type of guidance is directed towards the past, as described in [9]).

## 5.2 User Study and Provenance Data

We collected analytic provenance data in the form of XML log files from a training task as part of a user study published in [2]. We run our study with 12 participants who had limited experience with guided VA tools. The goal of the study was to evaluate the VA system in two parts: a training part, run as a game, for which we collected the provenance data; and an open exploration part where participants explored their own dataset looking for insights.

For the game task, we generated a 5D dataset in which we introduced an enclosed curvilinear dependency between two variables ($x_0$ and $x_1$) and random data for the other dimensions. Participants were instructed to use the tool to evolve a scatterplot in which the two curves in Fig. 3 (left) could be separated by a straight line. This task relates to the 'characterise distributions' visualization task of Amar et al. [1] as to solve the task, participants need to decide on which target scagnostic distribution, or combination of distributions, will help them best separate the two curves.

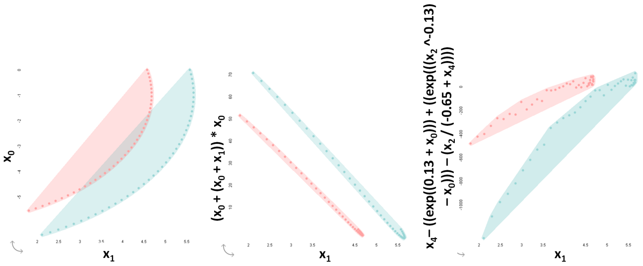

Figure 3: Screenshots of two alternative solutions to the game task (left) that use a basic dimension combination (middle) and a more complex formula (right).

We provided participants with two types of *explanations* about how the system functions: (i) a global explanation by showing participants the different types of scagnostics that the system is favouring when creating new views (Fig. 2); and (ii) a more local explanation within the VA tool that shows the system ranking (guidance) using background color as (the higher the system evaluation the darker the color).

Participants had around 20 minutes to accomplish the game task. Ten participants successfully separated the two curves (Success Sessions SS), while two participants evolved views with partial overlap but within the allocated time (Fail Sessions FS).

We chose to analyse logs from the training study part instead of the open exploration, because the game task was well-defined and participants were more likely to explore multiple generations before reaching the solution (mean 21 generations, 9 for open exploration). Our log data contained three main types of information pertaining to: (i) user interactions with the tool including their evaluations via the slider; and (ii) the IEC (genetic engine) status at each generation including details about the individuals in each generation (i.e. the combined dimensions), their fitness components and scagnostics scores; and (c) the overall learned scagnostics weights that govern the relative importance of each of the nine scagnostic measurements. The weights are initialised uniformly to 1/9, then updated via a simple multiple linear regression as soon as enough user interactions are recorded.

## 5.3 Computing Trust Measures

We propose to investigate two types of indirect trust metrics. The first aims to study the stability of user exploration strategies over time [**M1**]. We consider this measure an indirect indication of tighter exploration loops, that have been associated with increased confidence [46]. And second, the user agreement with system evaluations per exploration cycle (or generation) [**M2–4**]. We consider this measure as an indirect way to capture the users' trust and confidence in the learning process. We elaborate more on these measures in the next sections.

We note that our goal here is not to reach conclusions with statistical significance about trust and/or success and failure of experiments, but rather to explore how different metrics discussed in the literature can reveal different aspects of user trust. As such we do not report mean values across experiments, but we illustrate our observations with example sessions. More controlled experiments and user studies are needed to verify these observations and findings.

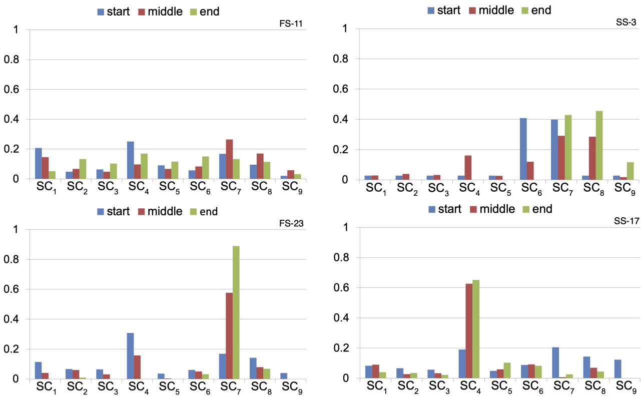

Figure 4: Learned scagnostics weights ($SC_{1..9}$) over three generation bins (start, middle, end) for Fail Sessions FS (left) and Success Sessions SS (right). The weights are scattered for one fail session (FS-11, top left) but more focused for the success sessions (SS-3,17, right). FS-23 (bottom left) appears focused as the participant was very close to finding a solution before time run out.

### 5.3.1 Stability of user exploration strategies

In previous work [2, 6] we found that our VA tool supports a variety of exploration strategies [**M1**] centered around three major scagnostics (skinny, convex and sparse) that appear to be important for the game task. We then wanted to verify whether success sessions tend to have "tighter exploration loops" thus suggesting user confidence, and fail sessions tend to have "scattered exploration" patterns, suggesting in this case user distrust of the process, as discussed in [34, 46].

In our case, the knowledge discovery cycles or loops pertain to the successive generation runs provided by the underlying evolutionary algorithm. By analysing the scagnostics scores and weights over the different generation runs, we found that the stability of the exploration strategy, which we define as the user's persistence in searching for the same visual pattern from one part of the session to the next, to be a key factor in deciding the outcome of the exploration and the speed with which it converges (in terms of number of generations). *The more scattered exploration cycles, as shown in Fig. 4 for a fail session (FS-11, top left) may indeed suggest lower user confidence than for a success session (e.g., SS-17, bottom right).*

Since we do not have access to direct user assessments of trust, we will look next at implicit ways to infer it, such as from user scoring strategies and how often they agreed or not with system feedback. This is inspired by related work [31,48,49] which suggests a direct link between trust and user's own estimations of the system's accuracy. We will use the two fail sessions (FS-11,23), and select a long and a short success session (SS-3,17) to illustrate the different metrics.

### 5.3.2 User agreement with system evaluations

First we start by defining three metrics that we hypothesise are linked to user trust and confidence in our guided VA system. We compute these metrics from the log files per generation for each experiment:

**[M2]** *Evaluation count*: the number of times a user evaluated scatterplots using the slider. High evaluation counts could indicate high disagreements with system suggestions, thus less user trust.
**[M3]** *Evaluation score delta*: evaluation score values range from 1 to 5, that the user or the system assigns to scatterplots (1 for least interesting, 5 for most interesting). The score delta is the (absolute or signed) difference between user and system evaluation scores. The bigger the absolute score delta the higher the disagreement, and thus the lower the user trust.

**[M4]** *Evaluation rank delta*: the rank occupied by each evaluation score in a given generation. In contrast to evaluation scores, ranks are relative evaluations of views within a single generation (rather than absolute scores). We decided to calculate ranks as they are more in coherence with the way the genetic algorithm works. There are up to 5 possible ranks (1 for lowest rank, 5 for highest). The rank delta is the (absolute or signed) difference between user and system evaluation ranks. The bigger the absolute rank delta the higher the disagreement for that generation, and thus the lower the user trust.

**Fig. 5-1** shows the evaluation counts **[M2]** for failed sessions FS-11,23 (left) and success sessions SS-3,17 (right). The evaluation count varies across experiments, with some participants (e.g., SS-3) only scoring a few views per generation (on average, max of 5) and other participants actively scoring a large number of views at each run (on average up to 25 evaluations per generation). The frequency of evaluations also varies across generations within the same experiment. Both SS-17 and SF-23 have a bell shaped graph. In SS-17, the exploration starts with a relatively low number of evaluations per generation (4) and increases steadily till the middle of the exploration, then decreases as the participant gets closer to the final solution, presumably also as the exploration strategy stabilises (see the prominent clumpy scagnostic $SC_7$ in Fig. 4 bottom right). In contrast, SS-3 and FS-11 have almost an inverse-bell shaped pattern, but the shape of the pattern does not seem to determine the outcome of the exploration. We may consider that high evaluation counts imply lower user confidence as participants appear to make frequent adjustments to system evaluations. This is difficult to verify based solely on log data. Moreover, although the views that are not evaluated could imply consensus between the user and the system, this may well be due to user over-reliance on the system, or user fatigue. *In the context of our VA system, we do not think that the evaluation count alone is a reliable metric to measure user trust.*

We next look more deeply into the details of user disagreements with the system **[M3]**. There are cases where the IEC scores higher than the user (i.e. negative evaluation score delta in **Fig. 5-2**), such as at the start of the exploration for SS-17 (2nd generation) while the system is still learning about user exploration and evaluation/ranking strategies. This type of disagreement happens at various stages for the fail sessions FS-11,23, presumably because the exploration focus is scattered, as described earlier (Fig. 4). However, even S3 has a predominant negative delta for the majority of the exploration session. These findings might indicate that participants adopt different evaluation strategies, such as by focusing on views that are not pertinent to their task and giving them a penalty score to make sure they are removed from subsequent generations (SS-3, FS-23), or reinforcing already good solutions by providing higher evaluation scores to ensure these views (or close ones) are taken into account by the system (FS-11, SS-17). **Fig. 5-3** shows similar patterns with error bars for the absolute evaluation rank delta **[M4]**. What is interesting here is that the rank delta tends to drop towards the end of the exploration for most experiments, indicating perhaps more user trust as the system is approaching convergence. *User frequent disagreements with system scores does not always imply lack of trust. Similar to findings from previous work [2], user adjustments of system scores, favoring penalising or encouraging feedback strategies, could be a personal approach adopted by users to give stronger signals to the system with regards to what patterns they like or dislike.*

The final row of plots in **Fig. 5-4** contrasts *user trust* of system evaluations, modelled as the normalised evaluation score delta (as described in this section), and the *system uncertainty* about the user goal inferred from the type of visual patterns they are interested in. The system uncertainty is calculated as the $r^2$ error from a multiple linear regression on the nine scagnostic values for each cell evaluated by the user. We use the fitting error $r^2$ to estimate this

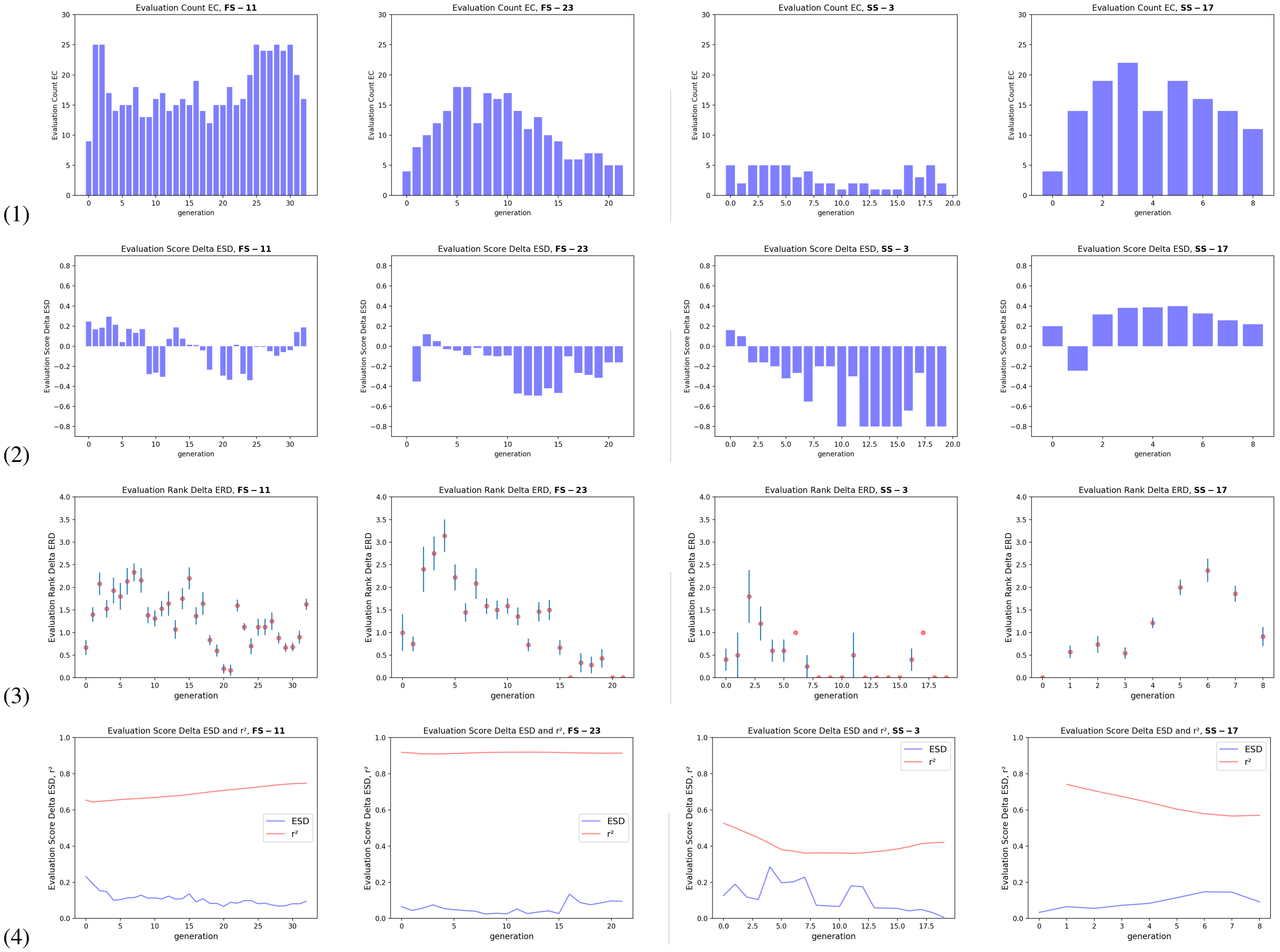

Figure 5: Rows 1–3 show the three user agreement metrics for four experiments: (left) two Fail Sessions (FS-11,23), and (right) two Success Sessions (FS-3,17). We report the mean per generation for: (1) User evaluation counts, (2) Signed evaluation score deltas, and (3) Absolute evaluation rank deltas. Row (4) shows system error or uncertainty against absolute evaluation score deltas (all values normalised).

uncertainty, where high $r^2$ values indicate better system fitting and therefore less uncertainty about the goal of the user. We can observe an inverse relationship between user trust and system uncertainty. Overall, the more user trust (i.e. small user evaluation delta) the smaller the system uncertainty (i.e. bigger $r^2$). This is expected and shows a correct functioning of the guided VA system, regardless of the outcome of the experiment. *We observed that when the user adds new information to the system such as by disagreeing with its evaluation score, the system uncertainty increases. Contrasting user trust and system uncertainty at various stages of the exploration may reveal how the VA system and the user "co-evolve" their mutual trust over time* [39].

## 6 Discussion and Perspectives

We discuss next our observations and identify perspectives and challenges as we move towards capturing trust in guided VA systems.

**Indirect & direct measures of trust.** Our exploration of metrics based on provenance alone as a way to indirectly measure confidence and trust, do not seem enough to study trust reliably in guided VA. It is likely, that in order to more reliably interpret these measures we will need the occasional direct user assessment of trust. A combination of the two approaches will likely reduce user fatigue and increase the accuracy of the indirect measures. How to best combine direct and indirect trust measures remains future work.

**Individual differences in trust.** In our results we observed a wide variability between participants. For example, we show participants that adopted a positive and others a negative re-reinforcement to train the system. Moving forward, it may be of interest to take into account more systematically the way users approach exploration and their ranking/evaluation profiles when attempting to gauge trust indirectly. More generally, as it has been suggested in different mixed-initiative systems [4, 12, 29], we need to consider how users personalities, preferences and expertise may affect how they interact and whether they trust the guided VA system.

**User & system agreement as a measure trust.** Even with seemingly different levels of disagreement between the ranking of the system and that of the user, the system can still converge, as indicated by 10/12 participants that succeeded in the task. This disagreement could actually be due to the adopted user evaluation strategy which may consist in exaggerating the evaluation scores to nudge the system, or strongly penalising certain patterns to give a clear signal to the IEC. Thus the agreement rate between the user and the system may not be best metric to judge user trust after all, unlike what has been suggested in the literature [49]. Nevertheless, it does provide visualization designers with the opportunity to aid users form and implement their strategies, by providing feedback on how their actions impact the system evaluation and the learning process. These can include explanations to make the user more aware of the sensitivity of their input and their scoring strategies (e.g., encourage/confirm/sensor; or adopt fine/coarse strategies in terms of range of scores used) and how this input affects the system.

**Trust in stages of analysis.** Currently, most of the work on trust focuses on whether it is appropriate to trust the system and what aspects of it to trust. Nevertheless, in guided VA systems it is possible that the notion of trust is fluid and appropriate trust may change in the course of the analysis. Past work in guided VA systems has identified two separate phases in the collaboration of users and system [5]. It has been observed that analysts may spend some time actively training the system and exploring their data (exploration phase) and other times use the trained system to find evidence of specific hypotheses and drill down in particular parts of their data (exploitation). While we did not study this in our current paper, we expect that trust needs and expectations are different in these stages. For example maybe lower trust or user reliability on the system are acceptable or even desired during exploration, but high trust is needed during exploitation. Indeed, an excess of users trust may inhibit their creativity. This point is particularly crucial in artistic systems [26] where sometimes the question of who is the artist, the human or the machine, is raised. We thus need to identify what are the appropriate levels of trust at each stage of the exploration, and possibly situations where trust is irrelevant or even inappropriate (for example when abusing the system by assigning on purpose extreme or contradicting weights to understand how the underlying model reacts [3]). This highlights future research directions in studying appropriate trust, as well as calibration in VA systems taking into account different cycles or phases in the analysis.

**Guidance and trust.** In our study, we considered a directing guided VA system, where the system proposes a ranked list of potentially interesting views to the user. It would be interesting to study and compare the impact of other types of guidance on user exploration strategies and trust. For example compare this directing guidance with more orienting guidance (e.g., no ranking of scatterplots provided), or a more prescriptive guidance by showing a step-by-step strategy to evolve a winning solution. We suspect that the trust stemming (and expected) in each of these guidance types will differ, with more prescriptive strategies fostering higher trust and expectations of trust, and if in turn the system fails to deliver this would impact user exploration the most.

Furthermore, in our work the direction of user guidance was backward facing, since the system was learning from past user interactions. It will be interesting to study trust in the context of feedforward guidance where the user proactively drives the analysis by directly specifying the visual patterns they are interested in in future exploration stages, such as using sketching [36]. It is likely that such guidance will be perceived by users as more uncertain and exploratory, and thus influence users strategies, for example make them less stable, and impact trust.

We also found that less stable exploration strategies led to task failure and likely reduced user trust. It is possible that this may be due to the nature of the game task we used, that had a clear goal for the outcome of the exploration and was known a-prior. It would be interesting to see if this stability is also inherent in open-ended visual analysis tasks that are deemed objectively successful. And more generally, more work is needed to study the impact of different types of guidance on the choice and stability of user exploration strategies, and whether and how this can in turn impact trust and user confidence. For example, can guidance be used to stabilise the exploration, and does a more stable strategy increase user trust?

## 7 Conclusion

We discuss the use of provenance data from automatic logs, in an attempt to implicitly calculate user trust in guided VA systems. We contribute a case study, based on a game experiment that utilises interactive evolutionary computation to explore a multi-dimensional dataset. In this context, we also contribute several possible methods to calculate trust. Our results show that alone, these implicit calculations do not always align with what we expect from a user-system collaboration. We discuss additional perspectives and future directions, including aspects that are unique to guided VA systems: the caution that user strategies to train the system may vary and a disagreement between system and user may be an artifact of this training process rather than an indication of low trust; the need to understand what levels of trust are required at different stages of the analysis; and the challenges of evaluating the impact of different types and degrees of guidance on user trust in human-machine collaboration.

## References

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

[2] N. Boukhelifa, A. Bezerianos, W. Cancino, and E. Lutton. Evolutionary visual exploration: evaluation of an IEC framework for guided visual search. *Evolutionary computation*, 25(1):55–86, 2017.

[3] N. Boukhelifa, A. Bezerianos, R. Chang, C. Collins, S. Drucker, A. Endert, J. Hullman, C. North, and M. Sedlmair. Challenges in evaluating interactive visual machine learning systems. *IEEE Computer Graphics and Applications*, 40(6):88–96, 2020.

[4] N. Boukhelifa, A. Bezerianos, I. C. Trelea, N. M. Perrot, and E. Lutton. An exploratory study on visual exploration of model simulations by multiple types of experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[5] N. Boukhelifa, W. Cancino, A. Bezerianos, and E. Lutton. Evolutionary visual exploration: evaluation with expert users. In *Computer Graphics Forum*, vol. 32, pp. 31–40. Wiley Online Library, 2013.

[6] W. Cancino, N. Boukhelifa, A. Bezerianos, and E. Lutton. Evolutionary visual exploration: experimental analysis of algorithm behaviour. In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*, pp. 1373–1380, 2013.

[7] W. Cancino, N. Boukhelifa, and E. Lutton. Evographdice: Interactive evolution for visual analytics. In *2012 IEEE Congress on Evolutionary Computation*, pp. 1–8. IEEE, 2012.

[8] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE transactions on visualization and computer graphics*, 23(1):111–120, 2016.

[9] D. Ceneda, T. Gschwandtner, and S. Miksch. A review of guidance approaches in visual data analysis: A multifocal perspective. In *Computer Graphics Forum*, vol. 38, pp. 861–879. Wiley Online Library, 2019.

[10] D. Ceolin, P. Groth, V. Maccatrozzo, W. Fokkink, W. R. V. Hage, and A. Nottamkandath. Combining user reputation and provenance analysis for trust assessment. *Journal of Data and Information Quality (JDIQ)*, 7(1-2):1–28, 2016.

[11] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, vol. 39, pp. 713–756. Wiley Online Library, 2020.

[12] M. Chiou, F. McCabe, M. Grigoriou, and R. Stolkin. Trust, shared understanding and locus of control in mixed-initiative robotic systems. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 684–691. IEEE, 2021.

[13] C. Collins, N. Andrienko, T. Schreck, J. Yang, J. Choo, U. Engelke, A. Jena, and T. Dwyer. Guidance in the human–machine analytics process. *Visual Informatics*, 2(3):166–180, 2018.

[14] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality*, pp. 251–262. Springer, 2014.

[15] S. M. Drucker, D. Fisher, and S. Basu. Helping users sort faster with adaptive machine learning recommendations. In *IFIP Conference on Human-Computer Interaction*, pp. 187–203. Springer, 2011.

[16] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 473–482, 2012.

[17] R. S. Gutzwiller and J. Reeder. Dancing With Algorithms: Interaction Creates Greater Preference and Trust in Machine-Learned Behavior. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 63(5):854–867, Aug. 2021. doi: 10.1177/0018720820903893

[18] W. Han and H.-J. Schulz. Beyond trust building—calibrating trust in visual analytics. In *2020 IEEE workshop on trust and expertise in visual analytics (TREX)*, pp. 9–15. IEEE, 2020.

[19] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Symposium on Information Visualization, 2005. InfoVis 2005.*, pp. 32–39. IEEE, 2005.

[20] D. Honeycutt, M. Nourani, and E. Ragan. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 8, pp. 63–72, 2020.

[21] E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.

[22] C. Y. Ip and A. Varshney. Saliency-assisted navigation of very large landscape images. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1737–1746, 2011.

[23] M. Kahng and D. H. P. Chau. How does visualization help people learn deep learning? evaluation of GAN lab. In *2019 IEEE workshop on EValuation of Interactive VisuAl Machine Learning systems (Eviva-ML)*. IEEE, 2019.

[24] R. Kamalian, E. Yeh, Y. Zhang, A. Agogino, and H. Takagi. Reducing human fatigue in interactive evolutionary computation through fuzzy systems and machine learning systems. pp. 678 – 684, 01 2006. doi: 10.1109/FUZZY.2006.1681784

[25] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[26] E. Lutton. Evolution of fractal shapes for artists and designers. *IJAIT, International Journal of Artificial Intelligence Tools*, 15(4):651–672, 2006. Special Issue on AI in Music and Art.

[27] M. Madsen and S. Gregor. Measuring human-computer trust. In *11th australasian conference on information systems*, vol. 53, pp. 6–8. Australasian Association for Information System, 2000.

[28] S. Marsh and M. R. Dibben. Trust, untrust, distrust and mistrust–an exploration of the dark (er) side. In *International conference on trust management*, pp. 17–33. Springer, 2005.

[29] B. M. Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

[30] C. North, R. Chang, A. Endert, W. Dou, R. May, B. Pike, and G. Fink. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pp. 33–36. 2011.

[31] M. Nourani, C. Roy, T. Rahman, E. D. Ragan, N. Ruozzi, and V. Gogate. Don't explain without verifying veracity: An evaluation of explainable ai with video activity recognition. *arXiv preprint arXiv:2005.02335*, 2020.

[32] A. D. Piemonti, M. Babbar-Sebens, S. Mukhopadhyay, and A. Kleinberg. Interactive genetic algorithm for user-centered design of distributed conservation practices in a watershed: An examination of user preferences in objective space and user behavior. *Water Resources Research*, 53(5):4303–4326, 2017. doi: 10.1002/2016WR019987

[33] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE transactions on visualization and computer graphics*, 22(1):240–249, 2015.

[34] D. Sacha, H. Senaratne, B. C. Kwon, and D. A. Keim. Uncertainty propagation and trust building in visual analytics. In *IEEE VIS 2014*, 2014.

[35] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014.

[36] L. Shao, M. Behrisch, T. Schreck, T. von Landesberger, M. Scherer, S. Bremm, D. A. Keim, et al. Guided sketching for visual search and exploration in large scatter plot spaces. In *EuroVA@ EuroVis*, 2014.

[37] R. B. Shaw. *Trust in the balance: Building successful organizations on results, integrity, and concern*. Jossey-Bass San Francisco, 1997.

[38] F. Sperrle, M. El-Assady, G. Guo, D. H. Chau, A. Endert, and D. Keim. Should we trust (x) ai? design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433*, 2020.

[39] F. Sperrle, A. V. Jeitler, J. Bernard, D. A. Keim, and M. El-Assady. Learning and teaching in co-adaptive guidance for mixed-initiative visual analytics. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pp. 61–65, 2020.

[40] H. Takagi. Interactive evolutionary computation: System optimization based on human subjective evaluation. In *IEEE International Conference on Intelligent Engineering Systems*, vol. 1998, pp. 17–19, 1998.

[41] H. Takagi. Interactive evolutionary computation: fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275–1296, 2001. doi: 10.1109/5.949485

[42] C. C. Venters, J. Austin, C. E. Dibsdale, V. Dimitrova, K. Djemame, M. Fletcher, S. Fores, S. Hobson, L. Lau, J. McAvoy, et al. To trust or not to trust? developing trusted digital spaces through timely reliable and personalized provenance. In *Proc. IEEE VIS Int'l Workshop Analytic Provenance for Sensemaking*, 2014.

[43] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Information Visualization, IEEE Symposium on*, pp. 21–21. IEEE Computer Society, 2005.

[44] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658, 2015.

[45] K. Xu, S. Attfield, T. Jankun-Kelly, A. Wheat, P. H. Nguyen, and N. Selvaraj. Analytic provenance for sensemaking: A research agenda. *IEEE computer graphics and applications*, 35(3):56–64, 2015.

[46] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovitch. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, vol. 39, pp. 757–783. Wiley Online Library, 2020.

[47] G. Yeo. Trust and context in cyberspace. *Archives and Records*, 34(2):214–234, 2013.

[48] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12, 2019.

[49] K. Yu, S. Berkovsky, R. Taib, J. Zhou, and F. Chen. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 460–468, 2019.

[50] Y. Zhang, B. Coecke, and M. Chen. On the cost of interactions in interactive visual machine learning. In *2019 IEEE workshop on EValuation of Interactive VisuAl Machine Learning systems (Eviva-ML)*. IEEE, 2019.

[51] J. Zhou, S. Arshad, S. Luo, and F. Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. pp. 23–39, 09 2017. doi: 10.1007/978-3-319-68059-0_2