Supporting Automation in Spacecraft Activity Planning with Simulation and Visualization

Basak Alper Ramaswamy*, Jagriti Agrawal†, Wayne Chi‡, So Young Kim§, Scott Davidoff¶, and Steve Chien‖
Jet Propulsion Laboratory, California Institute of Technology

Automation is gaining momentum in spacecraft operations, however, at a much slower pace than comparable application domains. The reasons behind slow adoption is (1) the need for high reliability and (2) the limited interaction between human operators and the automated systems. For automated systems to be adopted and trusted by humans, humans need to gain intuition about the decision making process of the automated system and trust in its execution [1]. In this paper, we present how simulation and visualization can enhance adoption of an automated on-board activity scheduling system, specifically in the context of Mars2020 rover mission [2]. The visualization aims to communicate to the users degree of variance and uncertainty in possible schedule execution. Our preliminary validation results suggest that the proposed visualization increases operators' confidence in—and likelihood of adopting—the automated scheduling system.

I. Introduction

Activity planning for spacecrafts is a complex task that involves both a negotiation of priorities of different stakeholders, as well as an optimization of available resources. While past missions have in some cases employed automation to some degree in the activity planning process on the ground, the upcoming Mars 2020 mission will feature an on-board automated scheduling software module which will generate dynamic and responsive schedules.

The benefits of automation is two fold. First of all, the on-board automated scheduler can increase overall science return by optimally utilizing available resources [2]. When scheduling is performed on the ground, engineers typically make choices based on worst case scenarios of resource usage. Estimating resource usage on the ground is a challenge due to variations both in environmental conditions and in efficiency and success of activity executions. For instance, an ample margin is typically added to activity durations to accommodate for longer than expected execution durations or execution failures. Frequently, activities run shorter than their conservative estimates, resulting in idle time and underutilized resources [3]. When scheduling is performed on-board, execution results up to the current point in time are known, allowing the scheduler to take advantage of idle resources. Hence, automation can reduce idle time spent by the rover and increase the overall science return of the mission [3].

The second benefit of autonomous scheduling is the addition of explicitly defined constraints which can reduce time spent by operators on the ground when negotiating a schedule for various activities. Note that activity planning is a highly collaborative task where many science and engineering users are involved at different stages of the process. Traditionally, activity planning involves pinning activities on a timeline, where many constraints, such as temporal dependencies, for activities are implicitly encoded. As planning progresses, frequently a need to edit the timing or ordering of activities arises to optimize for various resources. In this case, operators can be uncertain about whether editing the timing or ordering of an activity will contradict the original intent, leading to time consuming negotiations. When stakeholders explicitly describe their constraints, the allowable edits for an activity's schedule are clearly communicated.

Despite these advantages, automated scheduling poses significant challenges for the operators. First of all, predicting activity scheduling and execution based on a list of constraints is a very challenging cognitive task. Figure 1 shows a plan consisting of a list of activities and each activity's user defined constraints. Deducing how activities might get scheduled from this data is not straightforward. Second, the operators need to validate on the ground that the set of activities and constraints in the plan are "schedulable" under nominal conditions. Lastly, operators must ensure that

*Data Visualization Researcher, Human-Centered Design
†Member of Technical Staff, Artificial Intelligence Group
‡Member of Technical Staff, Artificial Intelligence Group
§Human-System Integration Researcher, Human-Centered Design
¶Group Manager, Human-Centered Design
‖Senior Research Scientist, Artificial Intelligence Group

all necessary or desired constraints are properly and explicitly defined in the plan. The automated scheduler may not respect any implicit intended timings and ordering constraints for activities that are not explicitly defined. Execution outcomes that do not respect intent can undermine the operators' trust in automation significantly.

To address these challenges, we simulate how the automated scheduler of the Mars 2020 rover might behave using Monte Carlo simulations on the ground. We generate many possible execution traces for the given set of activities and constraints, such that operators on the ground can review possible execution variations. However, consuming and comprehending hundreds of execution traces raises a different challenge for the operators. To be able to communicate the variance and uncertainty in a constraint based plan effectively and efficiently, we developed a series of visualizations. These visualizations present an aggregated summary of both (1) the start time and end time variance for each individual activity and (2) parallelism and ordering variations between the activities.

We performed a qualitative assessment to validate the perceived value of these visualizations by operators who have Mars surface mission activity planning experience. The self reported confidence and trust scores improve significantly when simulation results are presented with the visualizations that aggregate execution variations. Moreover, despite finding the aggregate summary tool less familiar, operators indicated strong preference for utilizing our proposed aggregate visualizations if they needed to work with an automated scheduler.

II. Related Work

This work builds upon research in automation in spacecraft activity planning and execution, investigations in operator trust in automated systems, and finally the role visualizations play at easing adoption of automated systems and improving operators trust. In this sections we discuss relevant research in each of these three distinct research areas.

A. Automation in Spacecraft Activity Planning and Execution

Automation in activity scheduling and execution has been a research area of high interest for decades [4–6], where responsiveness, performance and stability have been the major concerns for improvement. CASPER is an automated planning and execution system [6, 7] that controlled and flew on-board the Earth Observing One spacecraft for over a dozen years [8, 9] and the IPEX mission for over one year [10].

Ongoing research explores how to make future rovers more autonomous and productive [11]. Chi et al. discusses a theoretical framework to embed a scheduler in execution, including when to reschedule and how to schedule activities to best respond to uncertainty in execution [2]. Rabideau and Benowitz describe a greedy algorithm to handle activity scheduling to increase responsiveness given the limited flight processing power [12]. The prototype they describe is the automated scheduling system that we refer throughout this paper.

B. Automation and Trust

In the context of human-automation interaction, trust can be defined as the perception that an agent will successfully help an individual achieve their goals in a situation that includes both uncertainty and vulnerability [1]. Many studies have explored the ways in which trust mediates relationships between people and technology [13], and point to a large number of factors, such as dependability, reliability, level of automation, failure rate [14], and transparency[15] that modulate people's trust in automated systems. The construct of trust influences how human operators use and rely on automation; hence, the interaction between human operators and automation impacts performance of the joint human-automation system [14, 16–19]. Automated systems that are designed without proper consideration of human aspects (automation abuse) can result in underutilization (disuse) and over-reliance on automation (misuse) by the human operators [20]. Therefore, building appropriate trust [1] becomes a crucial aspect to achieve successful joint system operation.

People also differentially trust systems with variations in physical proximity [21], personality and anthropomorphism [22–24], and communication styles [25, 26]. Such factors manifest themselves in varying levels of trust depending on the type and complexity of the collaborative tasks [27], and the contexts in which they occur [28].

Measuring trust in automation is a challenge on its own, which is often assessed with surveys administered after exposure to the automated system [19, 29, 30]. In this paper we followed a similar methodology, borrowing keywords suggested in [29] in our custom survey.

C. Visualizations to Support Automation

Data Visualization and Visual Analytics approaches have been taken as ways to augment automation projects across a wide number of domains, including machine translation [31], reinforcement learning [32], image classification [33], image captioning and visual question answering [34, 35].

Interactive visualization are often utilized to enhance the interpretability or explainability of machine learning models [36–38]. Frequently, such tools are designed to help the developers of the autonomy to debug their models, with the hope of expediting the iterative experimentation process, and ultimately to improve performance of the autonomous system [39–41]. Other applications of interactive visualizations include making sense of a model recommendations [42], or to choose a single model among an ensemble of well-performing models [43–45]. In visual analytics domain, several works presented guidelines for the design of uncertainty aware visual analytics tools to improve trust and interpretation, and consequently aid the user in better decision making [46, 47].

Nevertheless, the potential of visualizations in enhancing operators' trust in automated systems in general is an under explored research area. Helldin et al. explored the use of visualizations to enhance trust during an automated car drive scenario [18]. The results from this study show that the drivers who were provided with the uncertainty visualization performed better in take-over situations, and better calibrated their trust in the automatic driving system. The control group, on the other hand, indicated higher trust while performing worse. Other work in human factors domain also suggest the use of visualizations to communicate the state of the automated systems [1].

III. Simulation and Visualization Approach to Improve Trust in On-Board Automated Activity Scheduling

When using the on-board automated scheduler, the operators provide a set of priority-ordered activities with specific resource requirements as well as ordering and timing constraints for these activities. Additionally, the scheduler takes into account plan-level constraints such as allowed total duration, minimum state of charge, and handover state of charge. The scheduler generates an initial schedule adhering to all the constraints while maximizing the number of activities scheduled [12]. However, when activities run significantly longer or shorter available resources change significantly, and the on-board scheduler is reinvoked to generate an updated schedule. Note that activity durations vary significantly during actual execution due to many factors, including, but not limited to, time of execution and environmental factors such as temperature. Hence, predicting a likely activity execution timeline on the ground is a non-trivial problem.

In order to characterize the behavior of the on-board scheduler in the face of uncertain execution, we feed a given plan to a Monte Carlo simulation to generate possible executions. In each simulation execution we vary the activity durations. Therefore, resource consumption (time, energy, data volume) varies in each run, which is the key determinant in execution variation. By analyzing a large number of possible executions, we can identify situations where activities may not execute or other undesired behaviors occur.

Currently we generate one hundred possible execution traces using Monte Carlo simulations. This output is reviewed by the operators, who conventionally review only *a schedule* during activity planning. In order to mitigate the increase in cognitive load and analysis time required to review simulation outputs, we adopted a visualization approach. Visualizations map patterns in the data to visual patterns which can be effectively recognized by humans [48], without resorting to computational analysis approaches which pose higher development costs. Interactive visualizations, in our case, can be highly effective in communicating likely execution possibilities while highlighting outliers in execution. However, the data set is rich, and there are many possible approaches to visualize the data. In the next section, we describe our user-centered design process of narrowing down visualization alternatives and iterating over solutions with feedback from potential users of the visualization tool.

IV. Formative Interviews and Focus Groups with the Operators

We adopted a user-centric, task-driven approach to guide the design of our visualization tool. We started our investigation with a series of interviews and two focus group studies conducted with six representatives of the Mars 2020 Engineering and Science operations team at Jet Propulsion Laboratory. We aimed at understanding how operators may review and validate a constraint-based plan, and identify key questions that they try to answer during this review and validation process.

Note that many of the operators, who participated in the formative research, previously worked on other Mars surface missions such as Mars Science Laboratory (MSL) and Mars Explorer Mission (MER). In these previous missions, a deterministic activity schedule was generated on the ground, and reviewing a plan meant reviewing a Gantt chart of

activities laid out on a timeline. Consistent with this experience, operators initially indicated a desire to review what they called a single 'grounded plan', where the planning tool schedules all activities on the ground with simplified assumptions.

While showing a sample schedule to operators is a simple solution, this approach poses significant issues of lack of communication of how as-run execution might deviate from what operators reviewed on the ground beforehand. We believe that large discrepancies between what operators validated on ground and what actually happened on-board can significantly undermine operators' trust and confidence in the automated scheduling tool. While noting this limitation, operators indicated that reviewing one schedule is a mentally taxing task, and reviewing hundreds of schedules within the five hours tactical planning window of the Mars 2020 mission is not feasible.

To address these challenges, we focused our attention on highly simplified visualizations dedicated to answer specific questions that operators' try to answer when they review and validate a constraint-based plan. We performed two focus group studies with the same group of operators, where we showed them earlier iterations of the visualizations. Among the visualizations we reviewed, aggregated views were found the most informative by the operators. We also asked the operators to comment on the utility of each display, and verbalize their reasonings.

We used feedback from the users of these early versions of the visualization to distill a set of key questions that an appropriate visualization would need to address:

Q1 - What is the overall temporal structure of the plan?
Q2 - When an activity will get executed most likely?
Q3 - Which activities show high variability in terms of execution start times?
Q4 - What is the likelihood of successfully scheduling an activity in all runs?
Q5 - Which activities are likely to run in parallel?
Q6 - Are there any unintended overlaps between activities?
Q7 - In what ways ordering of activities might change?
Q8 - Are there any unintended orderings between activities?
Q9 - How much power profile can vary at what point in the plan?
Q10 - How much data accumulation can vary at what point in the plan?

In the next section, we introduce the visualization designed to answer each of these questions and map specific affordances of the visualizations to the questions listed above.

V. Overview of the Visualization

The focus of our simulation and visualization tool is to provide a quick summary of a plan's temporal behavior to a large group of operators who are scientists and engineers and need to adopt the automated scheduling paradigm, while not understanding the inner workings of the algorithm. To this end, the visualization aims to summarize the simulated execution traces with respect to operator-defined constraints. Visualizations aimed at expert users to review the decision-making process of the automated scheduling tool is beyond the scope of this paper.

Accordingly, the visualization communicates three major aspects of the plan. First, it summarizes input constraints defined by the operators. Second, it overlays any specific simulation output on the input constraints. The real value of the visualization is in the third aspect, where users can have an aggregated summary of all simulation outputs. In the following sections, we describe each of these three components.

A. Input Plan Summary

The Figure 1 displays a visualization that summarizes input timing and ordering constraints defined per activity. The activities are ordered according to their average start time in the simulation data. The arcs on the left hand side encode ordering constraints, and the blue end is the activity that has to start or complete before the activity on the orange end. Activities that have to perform at a specific time are shown with pins on the timeline. The gray bars on the timeline indicate allowed start times for an activity, while the pink dot indicate the defined cutoff time. These are called execution windows for an activity. Activities can have up to three disjoint execution windows. The numbers preceding the activity names are their priority, where lower numbers map to higher priority. Priorities and timing constraints are defined independently.

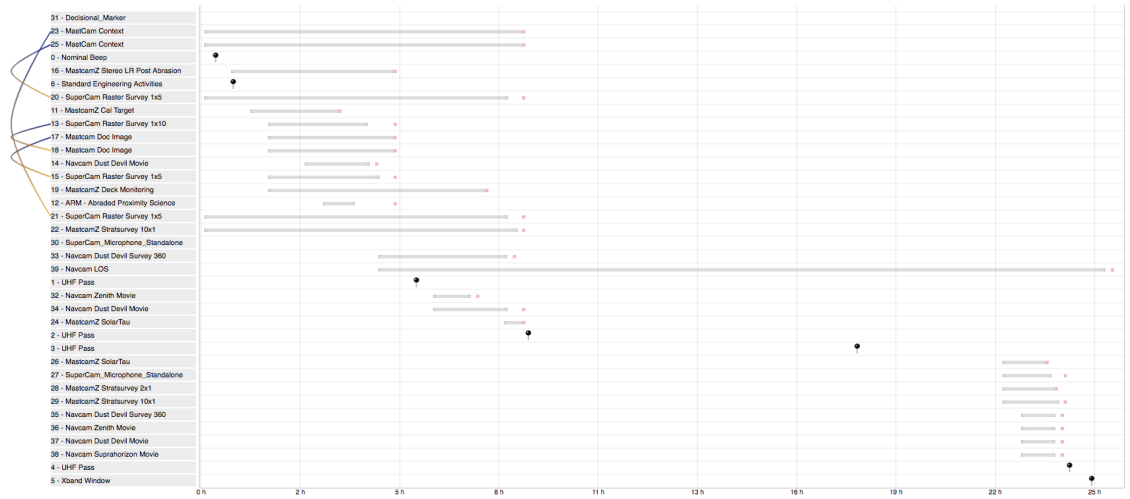


Fig. 1 First, we visualize the temporal and dependency constraints of all activities in the plan.

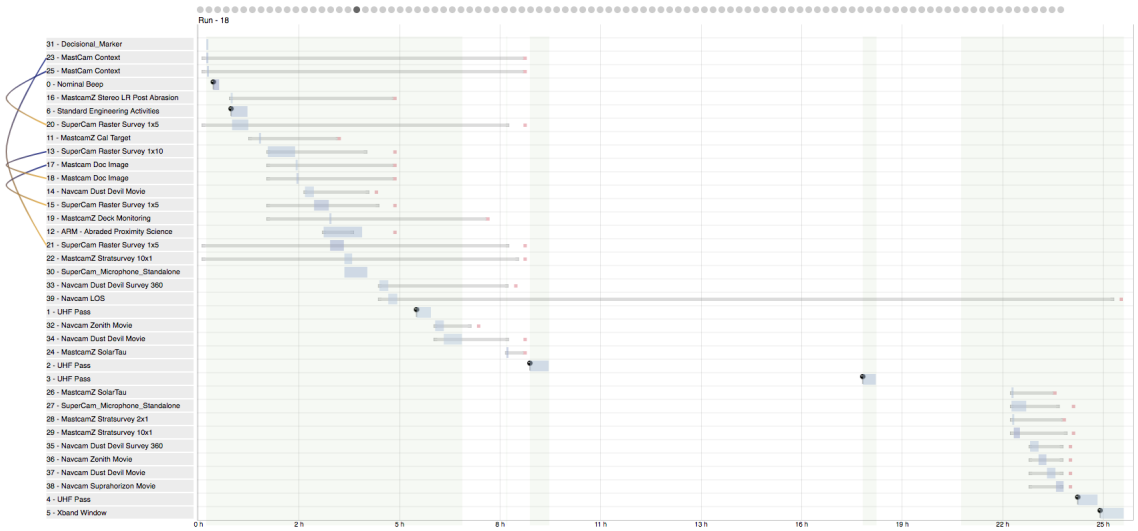


Fig. 2 Users can view any of the simulation outputs against the input constraints.

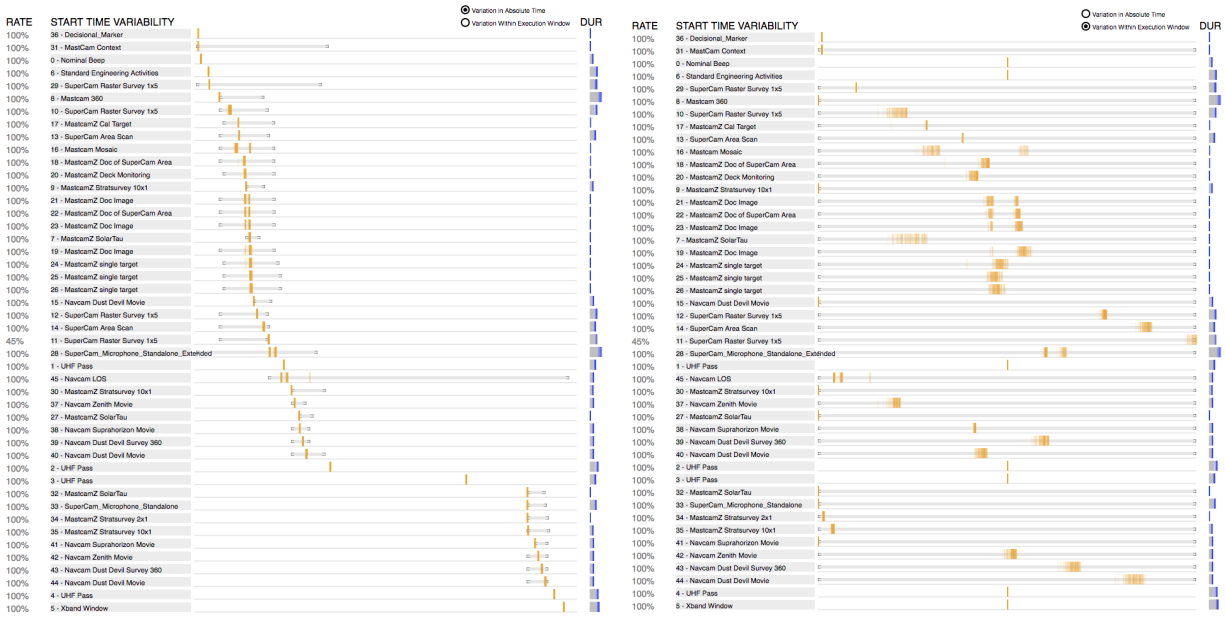
B. Simulated Execution Timelines Visualization

The user can overlay simulated schedules on the input constraint viewer as shown in Figure 2. All hundred simulation outcomes are listed above with a dot. The user can toggle between any simulation output by clicking on the dots. The users can play all simulation results with a continuous animation. When an activity is not scheduled in a simulation run, a red dot at the beginning of the timeline row appears, indicating the failure. The data profiles and power resource profiles associated with each simulation can be displayed below the schedule.

C. Aggregated Summary Visualizations

These visualizations present several key aspects of the simulation data, specifically designed to answer the key questions that operators may need to answer when reviewing and validating a constraint based autonomous plan, which are listed in the previous section.

The first visualization on the right in Figure 3 aggregates start and end times per activity for all hundred simulations. The visualization clearly shows that, despite having wide execution windows, many activities cluster around specific times in the timeline (Q1, Q2). The visualization is also effective in highlighting outliers, indicating high variation in their start times (Q3). The users have the option to toggle on and off end times. Viewing end times along with



(a) Activity start time variation in absolute time

(b) Activity start time variation within execution window

Fig. 3 Aggregated summary visualization of scheduling success rate, start time variability and duration variation.

start times improves understanding of parallelism relationships among activities in this view. The users also have the toggle option to view start and end time variability relative within the execution window or in absolute time as shown in Figure 3 left and right. The gray bars to the right of the display shows duration per activity. The blue strokes at the end of the gray bars show variation in duration for the activity. When a specific activity is not executed in all simulation runs, a success percentage is displayed for the activity (Q4). Note that failure in execution is strictly related to failure in scheduling in our simulations. While not being able to execute opportunistic activities in all runs is an expected outcome, operators expect certain activities to be scheduled and executed in all simulation runs. Hence this information plays a key role in ensuring plans validity.

The next visualization illustrate which activities overlap or run in parallel with which activity, using the connecting arcs. The thickness of arcs encodes frequency of overlap happening between any given activities (Q5, Q6) as shown in Figure 4 left. Similar arc visualization to the right shows order swaps between activities (Figure 4 right). Again arc thickness encodes frequency of the specific order swap happening in simulation runs (Q7, Q8). As noted earlier, activities are sorted by their average start time. If no overlap or order swap relationship is indicated between two activities by the arcs, the viewer should interpret activity above finishes before activity below (Q6, Q8). Note that an earlier iteration of these visualizations utilized an adjacency matrix to encode ordering and parallelism relationships. However, adjacency matrix representation was found harder to interpret and less familiar by the operators during our focus group studies. We then simplified the visualization as arc diagrams.

Finally aggregated power profile and data profile visualizations show variation in these key resources across simulation runs as shown in Figure 5. The power profile visualization for instance clearly shows how frequently minimum state of charge was reached and at what point in the plan (Q9, Q10).

In addition to these questions, the visualization can help users to look beyond well-formed queries, and help glean insights about the behavior of the automated system. Over time, the visualization should help operators build a mental model of the algorithm, so that they can more reliably predict its behavior, which, in turn, will help build trust between the human operator and the automated system.

Fig. 4 Arc visualizations that encode frequency of overlaps and order swaps across activities during all simulation runs. Note that activities are sorted by average start time. If no overlap or order swap relationship is indicated, activity above should be assumed to complete before the activity below.

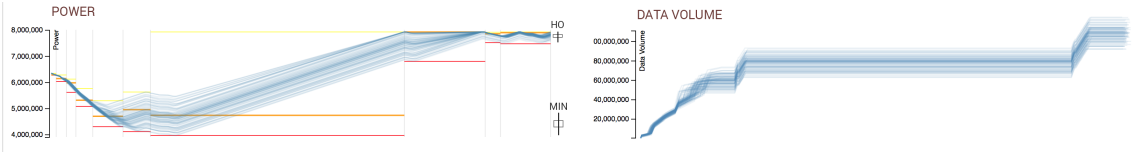


Fig. 5 Aggregated power and data volume resource profiles for all simulation runs.

VI. Qualitative Validation Study

The purpose of the study is to assess whether operators' perception and trust in the automated scheduling would improve when they review the ground simulations of the plan holistically, using the aggregated summary visualizations. Since confidence and trust are subjective constructs, we chose to conduct a qualitative assessment to gather self-reported confidence and trust scores of operators, and looked for different responses modulated by the presence or absence of the aggregated summary visualization.

A. Participants

We recruited eight participants with Mars surface mission operations experience. We excluded the operators who reviewed the visualizations beforehand. Our gender distribution was three female and five male. Five participants were in age range of 20 to 30, two participants were in age range of 30 to 40, and one participant reported above 50. Their operational experience ranged from two to seven years with an average of four years. Five out of eight participants had background in aerospace engineering, and the remaining operators had background in software engineering. None of the participants reported other engineering or science fields as their primary background. All participants had performed activity planning related roles.

B. Data Sets and Visual Material

We used three planning data sets for the study. One data set was used for training, and the other two were used for the experimental conditions. Training data consisted of 18 activities, each with a temporal constraint. There were 12 dependency constraints defined across the 18 activities. The experimental data sets included 40 activities (+/-2), each with temporal constraints. Each experiment data had about six dependency constraints across activities. Simulations of both plans showed comparable order swap and parallelism relationships among activities.

In order to provide better experimental comparison between conditions, we decided to simplify some of the presented simulation data and interaction features in the experimental condition. For example, we removed continuous animation of simulation results. We assumed that while users would be entertained with an animation initially, it would be time consuming during the study and less useful compared to a user directed switch between different simulations. Second, we removed the data profile and power plots in both conditions in order to focus participants' attention to the temporal structure of the plan. During the training, we conveyed to the users that they should assume that plan is safe to execute in terms of power and data resources. Finally, we did not show users percentage of scheduling success and used data sets where all activities are scheduled in all simulation runs.

C. Study Protocol

We designed an evaluation to assess whether the aggregated summary visualizations improved operators' perception and trust in the automated scheduling tool. Because we sought to test with space flight operations engineers experienced with a particular spacecraft, which represent a very limited population, we elected to use a within-subjects design. Figure 2 shows the control condition (Condition A), which uses a simple rendering of simulation results on a conventional timeline visualization. Figures 3 (either left or right) and 4 show Condition B, which supplement the control condition with the aggregated summary visualizations, displaying them above the condition A visualization. Hence, the difference between conditions is the absence or presence of aggregated summary visualizations. Since condition A was the control, we kept the order of conditions fixed across users. However, we counterbalanced the order of datasets across users. Therefore, half the participants viewed the first data set with Condition A, the other half viewed the second data set with Condition A and vice versa for Condition B.

At the beginning of the study we gave a brief explanation of the automated planning tool, and explained the purpose of the study to the participants as evaluating visualizations to review ground simulations of automated scheduling. We then gave a five minute training to the participants for condition A, explaining to them visual encodings, and explained to them how to interpret the data. During this training, we explained to the users that they will be reviewing simulation data, and actual execution of plan may not be exactly like any of the simulated schedules that they are going to review. We also told them to assume that plan is safe to execute with respect to power and data resources.

The users then viewed one data set in Condition A. We let them explore different simulations freely. We prompted several questions during their investigations, such as 'what can you say about ordering between activity x and y?', or 'how likely you think activity x overlaps with activity y?'. The questions were focused on activities that showed high variation in their timing and ordering, meant to situate the participants in interpreting data using the visualizations. These questions motivated them to seek further variations in the data. We asked open-ended questions about their general observations of the plan's temporal structure. After about 10 minutes, we asked the participants whether they feel confident in their understanding of how the plan may behave. Once they confirmed that they are sufficiently understanding the plan's temporal structure, we moved onto the second part.

Before the second part of the study, we gave the participants a brief training explaining visual encodings in the aggregated visualizations. We explained to them how to interpret the data going through few examples. Afterwards, the participants viewed condition B with the other data set. We repeated the exact same procedure, letting them to investigate freely, and then prompted them with similar questions related to specific activities. Towards the end, we again asked open-ended questions about their general observations of the activity plan.

After participants interacted with both visualizations, we gave them a demographics survey, followed by a qualitative evaluation survey with a 7-point Likert scale, ranging from *"not at all"* (1) to *"extremely"* (7). Questions 1 - 14 were paired, repeating same question for each condition. The survey questions were modified from a general-purpose confidence / trust assessment survey proposed in [29]. We did not use the general purpose trust survey questions in this study, because we found that the participant in the pilot study evaluated the underlying system (the automated planning tool itself), instead of whether the visualizations improve the trust in the automated planning tool. However, we incorporated the dimensions suggested in [29].

The question pairs asked per condition can be summarized as follows:

- Q1-2: **Comprehension**: I was able to understand temporal structure of the plan with...
- Q3-4: **Unpredictability**: I detected unpredictable schedules with ...
- Q5-6: **Variance**: I felt confident in my understanding of the variance in schedule with ...
- Q7-8: **Uncertainty**: I felt confident in my understanding of uncertainty in execution outcomes with...
- Q9-10: **Familiarity**: I felt familiar with the planning tool with ...
- Q11-11: **Reliability**:I found the planning tool reliable with ...
- Q13-14: **Trust**: I can trust this planning tool to create schedules in the future with ...

Questions 15-19 were specific to Condition B. These questions assessed operators' perceived performance using the aggregated summary visualization as well as their preference to utilize it. These questions can be summarized as follows:

- Q15: **Helpfulness**: Aggregated summary visualization was helpful.
- Q16: **Perceived speed**: I was able to detect variance in schedule faster with the aggregated summary visualization.
- Q17: **Perceived accuracy**: I was able to detect variance in schedule more accurately with the aggregated summary visualization.
- Q18: **Increase in confidence**: Aggregated summary visualization increased my confidence in safety of plan execution.
- Q19: **Preference**: I would need the aggregated summary visualization to make informed decisions about constraint-based automated planning in the future.

Finally, we audio-recorded all eight sessions, which lasted about 30 minutes on average. The results are summarized below.

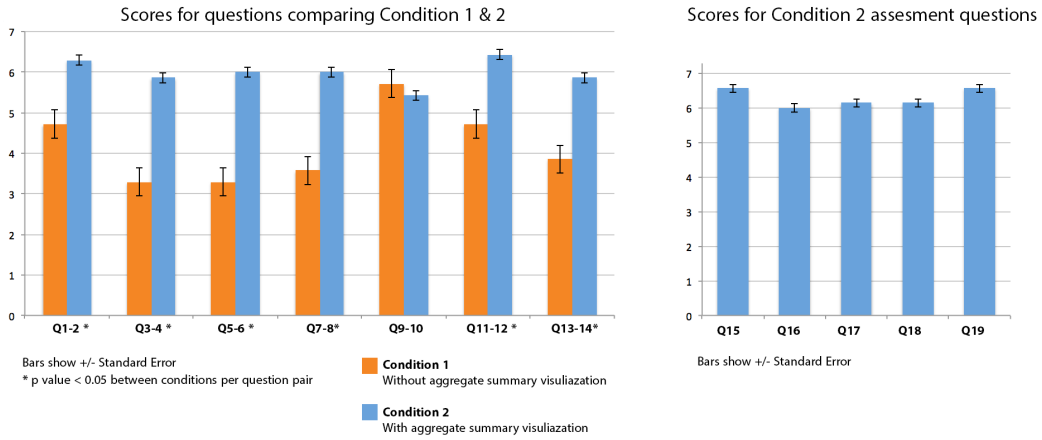


Fig. 6 Summary of survey study results for both questions comparing both visualization conditions (from 1 to 14) and for questions assessing benefits and preference for aggregated summary visualization (from 15 to 19).

D. Summarized Survey Questions and Results

Despite having a small participant pool, the survey results showed consistent pattern, indicating a clear increase in users' perceived performance as well as improved trust and confidence in the automated scheduling when simulations were presented with the aggregated summary visualizations (Condition B). The bar chart shown on the left of Figure 6 shows that scores were higher (more positive) for Condition B for six out of seven dimensions of trust and confidence. The participants reported that they could understand the temporal structure of the plan (Q1 vs Q2), and detected outliers more effectively (Q3 vs Q4) in Condition B. They also reported higher confidence in their perception of variance and uncertainty in the plan in Condition B (Q5 vs Q6, and Q7 vs Q8 respectively). With respect to familiarity-related questions (Q9 vs Q10), we believe that some participants evaluated their familiarity with the visualization rather than with the planning tool for that question. Hence, they gave higher scores to the more conventional timeline visualization used in Condition A. In fact, participants who gave a low familiarity score to Condition B, verbally indicated that they had to spend time to learn how to read that visualization. Finally, participants reported increase in reliability to and confidence in the underlying planning tool for Condition B.

The bar chart in 6 right shows that all participants reported that the aggregated summary visualization was helpful (Q15) with a 6.57 mean score. Participants reported that they were faster and more accurate to detect variance in possible schedule outcomes using the aggregated summary visualizations (Q16, Q17). They also reported higher confidence in plan's execution in Condition B (Q18) as well as a strong preference to utilize the aggregated summary visualization to accompany the automated scheduling (Q19).

E. Qualitative Feedback

One of the major concerns we had about the aggregated summary visualization was it's novelty as opposed to the familiarity of the visualization used in Condition A. However, after a brief training, all participants were able to interpret the visual encodings in the aggregated summary visualizations. In fact, in Condition B, where both visualizations were present, all participants spent more time studying the aggregated summary visualization. Only 3 out of 8 participants viewed the timeline visualization shown in Figure 6.

6 out of 8 participants reported that clicking through individual execution timelines was cumbersome. While some participants clicked through many simulations before arriving at conclusions, others felt confident after viewing randomly-selected subset of simulations. One participant, P3, indicated that "...I think there is excess of information...you are not going to click through 100 plans, I think I clicked about 15 to 20". Another participant, P5 commented "I can see it [the activity] moving around, but you have to know the right one to click to see different behavior." In fact, four out of eight participants explicitly indicated a need to have a statistical summary visualization. For instance P5 commented "...but if I clicked on an activity, it would be nice to see a distribution of where that activity did execute in all the runs."

When using the timeline visualization (Condition A), we observed that participants' perception of variation in the temporal structure of the plan showed a diverging trend. While half the participants quickly noted that the plan follows a similar trend in all simulations (P3, P4, P5), others used a less certain language when commenting on possible activity execution outcomes. For instance, P2 commented that activities can execute any time in their execution window, and added "...ordering can be pretty much random." P7 commented "...they [activities] seem to happen at pretty wide range of times, ... this [one activity] can happen pretty much anytime." P7 further commented that "I am observing some trend but yet again I see some simulations where things don't follow that trend." We believe that not being able to see a holistic picture led these participants to assume a higher variation.

On the other hand, when using the aggregated summary visualization participants verbalized insightful comments about the behavior of the plan. For instance, P4 commented "even though the activity has a large execution window, it always gets executed before this UHF pass [communication window], so I can expect to have the data from that observation on the ground." P8 commented on the value of arc visualizations indicating that they are more effective than a timeline display to show whether activities do actually overlap. P6 indicated "that's kind of hard to see ... [on a timeline] whether activities do overlap, or barely one is starting when one is ending, here [on the arc visualization] it is clear that they do sometimes overlap... in fact things overlap more than I would assume looking at this plot [timeline visualization]".

At the end of the study, we asked participants to provide general comments about the visualizations. The participant with the most expertise in activity planning, P2, indicated that "...Would I trust this [on-board automated scheduling] without human in the loop? ... I think there is a threshold to overcome to say yes ... with a lot of validation like this [aggregated summary visualization] so the behavior is understood, the visualization helps me get more insight to see what is happening under the hood. " Similarly, P5 noted "Looking at the top view [aggregated summary visualization], it's a lot easier for me to identify questions I might have as an approver... [such as] these two activities run in parallel very infrequently, and maybe nobody thought about that ... 'hey Mastcam are you OK with your doc imaging and z-deck monitoring running in parallel? ' ... 'oh no, that's bad, we missed a constraint' ". On the same topic P8 noted that "...checking for specific science intent is a lot easier with this [aggregated summary] visualization... if I see things running in parallel, I can quickly pick out 'why are those running in parallel, that shouldn't be allowed.'"

VII. Discussion and Future Work

The qualitative assessment results validated the potential role that visualizations can play at easing the adoption of constraint-based activity planning and automated scheduling that will be utilized in the Mars 2020 mission. The aggregated summary visualizations present rich information about the variance in execution of a constrained-based plan neither burdening users with too much information, nor hiding uncertainty from them. Being able to observe in a concise form, how the automation may behave under different circumstances does increase the self-reported trust and confidence rating in the system. However, our results should be interpreted as preliminary both in exploring the design

space of possible other visualizations, as well as the validation of the benefits.

In the future, we aim to enrich the visualization tool to address other challenges related to constraint-based planning. For instance, we plan to use visualizations to diagnose *broken plans* where activities fail to execute during simulations. Note that activities can fail to schedule and execute due to shortage of any resource that they require, ranging from power, heating, time, to a busy instrument. A failure can be due to complex interactions between activities where multiple resource shortages come into play. Providing deterministic diagnostics for the reasons of a scheduling and execution failure is a non-trivial problem to solve. We aim to leverage operators' visual perceptual abilities and help them diagnose possible failures. These visualizations will focus on presenting the availability of resources throughout simulation timelines.

We also aim to validate our approach by gathering more quantitative data. While participants reported increased confidence when using our visualization tool, this result brings forth the question of whether having such increased confidence is justified. The visualization is only useful if users' confidence is increasing in parallel to their comprehension of the possible execution outcomes of the constraint-based plan. If not, a sense of confidence that is only perceptual and subjective can even be harmful during actual activity planning operations. Hence, we plan to measure self-reported confidence against quantitative assessment of comprehension of the temporal structure and variance of a plan.

VIII. Conclusion

This project emanated through challenges faced in adopting an automated on-board scheduling system that is to replace a highly conventional and deterministic mode of activity planning. Even if the safety of execution of activity plan is guaranteed, human operators still need to understand how the system can behave before approving a constraint-based plan. While Monte Carlo simulations provide possible execution outcomes to study the behavior of the on-board scheduler, the amount of data produced at the end of these simulations poses a comprehension challenge for the operators. Our approach to aggregate and visualize such simulation data has been positively embraced by the operators who can potentially work with automated activity scheduling tools. The preliminary qualitative assessment indicates a strong user preference for visualizing simulation data in the proposed manner.

Acknowledgements

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. We would like to thank the Mars 2020 mission operations team, especially Elyse Fosse, James Biehl, Rachael Collins and Kimberly Steadman for their valuable input and feedback during iterations of this work. We also would like to thank Jet Propulsion Laboratory Data Science program office managers Dan Crichton and Richard Doyle for funding this research.

References

[1] Lee, J. D., and See, K. A., "Trust in automation: Designing for appropriate reliance," *Human factors*, Vol. 46, No. 1, 2004, pp. 50–80.

[2] Chi, W., Chien, S., Agrawal, J., Rabideau, G., Benowitz, E., Gaines, D., Fosse, E., Kuhn, S., and Biehl, J., "Embedding a Scheduler in Execution for a Planetary Rover," *International Conference on Automated Planning and Scheduling (ICAPS 2018)*, 2018.

[3] Gaines, D., Anderson, R., Doran, G., Huffman, W., Justice, H., Mackey, R., Rabideau, G., Vasavada, A., Verma, V., Estlin, T., et al., "Productivity challenges for mars rover operations," *Proceedings of 4th Workshop on Planning and Robotics (PlanRob)*, London, UK, 2016, pp. 115–125.

[4] Chien, S., Smith, B., Rabideau, G., Muscettola, N., and Rajan, K., "Automated planning and scheduling for goal-based autonomous spacecraft," *IEEE Intelligent Systems and their applications*, Vol. 13, No. 5, 1998, pp. 50–55.

[5] Rabideau, G., Knight, R., Chien, S., Fukunaga, A., and Govindjee, A., "Iterative repair planning for spacecraft operations using the ASPEN system," *Artificial Intelligence, Robotics and Automation in Space*, Vol. 440, 1999, p. 99.

[6] Chien, S. A., Knight, R., Stechert, A., Sherwood, R., and Rabideau, G., "Using Iterative Repair to Improve the Responsiveness of Planning and Scheduling." *AIPS*, 2000, pp. 300–307.

[7] Knight, S., Rabideau, G., Chien, S., Engelhardt, B., and Sherwood, R., "Casper: Space exploration through continuous planning," *IEEE Intelligent Systems*, Vol. 16, No. 5, 2001, pp. 70–75.

[8] Chien, S. A., Sherwood, R., Tran, D., Cichy, B., Rabideau, G., Castano, R., Davies, A., Mandl, D., Trout, B., Shulman, S., et al., "Using Autonomy Flight Software to Improve Science Return on Earth Observing One." *Journal of Aerospace Computing Information and Communication*, Vol. 2, No. 4, 2005, pp. 196–216.

[9] Tran, D., Chien, S., Rabideau, G., and Cichy, B., "Flight Software Issues in Onboard Automated Planning: Lessons Learned on EO-1," *International Workshop on Planning and Scheduling for Space*, 2004.

[10] Chien, S., Doubleday, J., Thompson, D. R., Wagstaff, K. L., Bellardo, J., Francis, C., Baumgarten, E., Williams, A., Yee, E., Stanton, E., et al., "Onboard Autonomy on the Intelligent Payload EXperiment CubeSat Mission," *Journal of Aerospace Information Systems*, 2016.

[11] Gaines, D., Rabideau, G., Doran, G., Schaffer, S., Wong, V., Vasavada, A., and Anderson, R., "Expressing Campaign Intent to Increase Productivity of Planetary Exploration Rovers," *Proc. Int. Workshop on Planning and Scheduling for Space (IWPSS'17), Pittsburgh, Pennsylvania, USA*, 2017.

[12] Rabideau, G., and Benowitz, E., "Prototyping an Onboard Scheduler for the Mars 2020 Rover," *Proceeding of International Workshop on Planning and Scheduling for Space, Pittsburgh, PA*, 2017.

[13] Sheridan, T. B., and Hennessy, R. T., "Research and modeling of supervisory control behavior. Report of a workshop," Tech. rep., NATIONAL RESEARCH COUNCIL WASHINGTON DC COMMITTEE ON HUMAN FACTORS, 1984.

[14] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R., "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, Vol. 53, No. 5, 2011, pp. 517–527.

[15] Boyce, M. W., Chen, J. Y., Selkowitz, A. R., and Lakhmani, S. G., "Effects of agent transparency on operator trust," *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, ACM, 2015, pp. 179–180.

[16] Sorkin, R. D., and Woods, D. D., "Systems with human monitors: A signal detection analysis," *Human-computer interaction*, Vol. 1, No. 1, 1985, pp. 49–75.

[17] Rovira, E., and Parasuraman, R., "Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance," *Human factors*, Vol. 52, No. 3, 2010, pp. 411–425.

[18] Helldin, T., Falkman, G., Riveiro, M., and Davidsson, S., "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving," *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications*, ACM, 2013, pp. 210–217.

[19] Rezvani, T., Driggs-Campbell, K. R., Sadigh, D., Sastry, S. S., Seshia, S. A., and Bajcsy, R., "Towards trustworthy automation: User interfaces that convey internal and external awareness." *ITSC*, 2016, pp. 682–688.

[20] Parasuraman, R., and Riley, V., "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, Vol. 39, No. 2, 1997, pp. 230–253.

[21] Bainbridge, W. A., Hart, J., Kim, E. S., and Scassellati, B., "The effect of presence on human-robot interaction," *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, IEEE, 2008, pp. 701–706.

[22] Powers, A., Kiesler, S., Fussell, S., Fussell, S., and Torrey, C., "Comparing a computer agent with a humanoid robot," *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, ACM, 2007, pp. 145–152.

[23] Kiesler, S., Powers, A., Fussell, S. R., and Torrey, C., "Anthropomorphic interactions with a robot and robot–like agent," *Social Cognition*, Vol. 26, No. 2, 2008, pp. 169–181.

[24] Li, D., Rau, P. P., and Li, Y., "A cross-cultural study: Effect of robot appearance and task," *International Journal of Social Robotics*, Vol. 2, No. 2, 2010, pp. 175–186.

[25] Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., and Hagita, N., "Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior," *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, ACM, 2009, pp. 69–76.

[26] Rau, P. P., Li, Y., and Li, D., "Effects of communication style and culture on ability to accept recommendations from robots," *Computers in Human Behavior*, Vol. 25, No. 2, 2009, pp. 587–595.

[27] Biros, D. P., Daly, M., and Gunsch, G., "The influence of task load and automation trust on deception detection," *Group Decision and Negotiation*, Vol. 13, No. 2, 2004, pp. 173–189.

[28] Yan, Z., Liu, C., Niemi, V., and Yu, G., "Exploring the impact of trust information visualization on mobile application usage," *Personal and ubiquitous computing*, Vol. 17, No. 6, 2013, pp. 1295–1313.

[29] Jian, J.-Y., Bisantz, A. M., and Drury, C. G., "Foundations for an empirically determined scale of trust in automated systems," *International Journal of Cognitive Ergonomics*, Vol. 4, No. 1, 2000, pp. 53–71.

[30] Rovira, E., McGarry, K., and Parasuraman, R., "Effects of imperfect automation on decision making in a simulated command and control task," *Human Factors*, Vol. 49, No. 1, 2007, pp. 76–87.

[31] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al., "Google's multilingual neural machine translation system: enabling zero-shot translation," *arXiv preprint arXiv:1611.04558*, 2016.

[32] Zahavy, T., Ben-Zrihem, N., and Mannor, S., "Graying the black box: Understanding DQNs," *International Conference on Machine Learning*, 2016, pp. 1899–1908.

[33] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H., "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[34] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D., "Show and tell: A neural image caption generator," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[35] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D., "Vqa: Visual question answering," *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[36] Lipton, Z. C., "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.

[37] Montavon, G., Samek, W., and Müller, K.-R., "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, 2017.

[38] Miller, T., "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.

[39] Wongsuphasawat, K., Smilkov, D., Wexler, J., Wilson, J., Mané, D., Fritz, D., Krishnan, D., Viégas, F. B., and Wattenberg, M., "Visualizing dataflow graphs of deep learning models in TensorFlow," *IEEE transactions on visualization and computer graphics*, Vol. 24, No. 1, 2018, pp. 1–12.

[40] Pezzotti, N., Höllt, T., Van Gemert, J., Lelieveldt, B. P., Eisemann, E., and Vilanova, A., "Deepeyes: Progressive visual analytics for designing deep neural networks," *IEEE transactions on visualization and computer graphics*, Vol. 24, No. 1, 2018, pp. 98–108.

[41] Strobelt, H., Gehrmann, S., Pfister, H., and Rush, A. M., "Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE transactions on visualization and computer graphics*, Vol. 24, No. 1, 2018, pp. 667–676.

[42] Rong, X., and Adar, E., "Visual tools for debugging neural language models," *Proceedings of ICML Workshop on Visualization for Deep Learning*, 2016.

[43] Alexander, E., and Gleicher, M., "Task-driven comparison of topic models," *IEEE transactions on visualization and computer graphics*, Vol. 22, No. 1, 2016, pp. 320–329.

[44] McMahan, H. B., Holt, G., Sculley, D., Young, M., Ebner, D., Grady, J., Nie, L., Phillips, T., Davydov, E., Golovin, D., et al., "Ad click prediction: a view from the trenches," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2013, pp. 1222–1230.

[45] Kahng, M., Fang, D., and Chau, D. H. P., "Visual exploration of machine learning results using data cube analysis," *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, ACM, 2016, p. 1.

[46] Chuang, J., Ramage, D., Manning, C., and Heer, J., "Interpretation and trust: Designing model-driven visualizations for text analysis," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2012, pp. 443–452.

[47] Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., and Keim, D. A., "The role of uncertainty, awareness, and trust in visual analytics," *IEEE transactions on visualization and computer graphics*, Vol. 22, No. 1, 2016, pp. 240–249.

[48] Ware, C., *Information visualization: perception for design*, Elsevier, 2012.