Constructing User Models from Eye Gaze Data in the Domain of Information Visualization

by

Matthew Gingerich

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Computer Science)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

April 2015

Abstract

A user-adaptive information visualization system capable of learning models of users and the visualization tasks they perform could provide interventions optimized for helping specific users in specific task contexts. This thesis investigates the accuracy of predicting visualization tasks, user performance on tasks, and user traits from gaze data. It is shown that predictions made with a logistic regression model are significantly better than a baseline classifier, with particularly strong results for predicting task type and user performance. Furthermore, classifiers built with interface-independent are compared with classifiers built with interface-dependent features. Interface-independent features are shown to be comparable or superior to interface-dependent ones. Adding highlighting interventions to trials is shown to have an effect on the accuracy of predictive models trained on the data from those trials and these effects are discussed. The applicability of all results to real-time classification is tested using datasets that limit the amount of observations that are processed into classification features. Finally, trends in features selected by classifiers and classifier accuracies over time are explored as a means to interpret the performance of the tested classification models.

Preface

This thesis is an analysis of data collected during the Intervention Study, described in Section 2, and is a part of the larger ATUAV research project referenced in Section 1.1. I was not involved in the design of the Intervention Study or in its collection of raw data, but the applied machine learning experiments presented in this thesis are my original work.

Portions of the text of this thesis were published as “Constructing Models of User and Task Characteristics from Eye Gaze Data for User-Adaptive Information Highlighting” of which I am the lead author [1]. The data processing and machine learning evaluation experimental procedure and scripts I created for this thesis were also used to generate data for Toker et al. [2]. I am also an author of [2], but the statistical analysis of the results in that paper were performed by D. Toker and the details of that work are not discussed in this thesis.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I owe enormous thanks to Dr. Cristina Conati for her guidance on both research topics and on the means to properly document that research. That gratitude extends to the many members of the Intelligent User Interfaces group at the University of British Columbia for continuing to inspire my work in this field.

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

1.1 Background

A user-adaptive interface is an interface that actively changes itself to be personalized to the needs of individual users. In order to provide personalized features for users, user-adaptive systems must first obtain information about users that reveals information about their needs and requirements for the interface; i.e., they need to create a model of their users. A user-adaptive information visualization system capable of learning models of users and the visualization tasks they perform could provide interventions optimized for helping specific users in specific task contexts. This thesis is part of the ATUAV (Advanced Tools for User-Adaptive Visualizations) project, an ongoing multi-year project aiming to uncover knowledge and techniques for designing user-adaptive visualizations [3].

Specifically, this thesis investigates the accuracy of predicting characteristics of visualization tasks, user performance on these tasks, and relevant users' traits from eye tracking data collected during visualization usage. The user and task traits to be predicted include measures of users' performance on tasks, their cognitive abilities such as perceptual speed, and other properties that could influence a user's experience with an interface. Further details about the characteristics being modeled, user modeling, adaptive interfaces, and the context of this research are provided in Section 1.4.

1.2 Research Questions

The work presented in this thesis is governed by the following set of research questions. At a high level, these research questions are asking how accurately user models can represent users. Questions about the predictive accuracy of the model are used to provide a quantifiable measure of the quality of the model, particularly as it relates to knowing what information could be available to a real-time adaptive system.

1. **What** user and task characteristics can we predict as users work with a visualization?
2. **How** can we maximize our classification accuracy?
    a. How does the type of features used affect accuracy?

b. How does the type of intervention presented affect accuracy?

3. **When** can predictions be made?
    a. How do prediction accuracies relate to the amount of observed data?
    b. In practice, what accuracy is achievable for a real-time system?
4. **Why** do predictions succeed or fail?
    a. Can trends in the results be interpreted to give insights into the challenges of making predictions?

1.3 Contributions

To summarize, the goal of this thesis is to explore what value gaze data can have for an adaptive visualization that actively learns the traits of its users and presents customized highlighting of information to meet their individual needs. Gaze patterns are of interest both because they are tightly linked to information processing [4], and because gaze data can be obtained for non-interactive visualizations. Earlier work showed that three different cognitive abilities (perceptual speed, visual working memory and verbal working memory), task performance, and task type can be estimated (albeit with varying degrees of accuracy) from gaze patterns during visualization tasks with bar graph and radar graph visualizations [5]. This thesis complements these previous findings by adding the following contributions.

First, it verifies whether similar results are obtained when predicting the same user and task properties while users interact with bar graphs visualizing more complex datasets. The inclusion of task performance as a target for prediction is intended to address the question of when adaptive interventions should be displayed: users with low predicted performance stand to benefit more from additional support. The prediction of user cognitive abilities and task type can inform the decision of which interventions should be displayed, since prior work has correlated these characteristics with effects on specific aspects of visualization processing [6].

Second, it adds the prediction of a user's locus of control, a personality trait that has been shown to impact visualization performance (e.g. [7]).

Third, it compares classifiers built with interface-independent and interface-dependent features, in order to assess the extent to which these predictive models require having detailed information on the presented visualization.

Finally, it investigates how model accuracy is affected if models are trained with data from tasks that had different types of highlighting interventions added to the visualization. These interventions were designed to highlight graph bars that were relevant to perform the given task. Interventions could eventually be used to provide adaptive support by dynamically redirecting the user's attention to different subsets of the visualized data as needed (e.g. when the visualization is used together with a verbal description that discusses different aspects of a dataset [6]).

1.4 Related Work

This section describes prior work that is related to the topics of user-adaptive visualizations. These topics are first approached from the perspective of disciplines focusing on human-computer interactions through a review of the literature on visualization effectiveness (Section 1.4.1). We then examine the topic of user modeling to infer knowledge about the user of a visualization system and the context of use. We also review work on using eye-tracking data in user modeling (Section 1.4.2)

1.4.1 Visualization Effectiveness

The long-term goal of our research is to provide adaptive *interventions* on visualizations that tailor aspects of the visualization to the individual needs of users. In this research project, these interventions take the form of visual prompts that highlight aspects of graphs by adding elements as overlays on the graphs. This goal is motivated by a number of prior studies that have investigated the relationship between user characteristics and visualization effectiveness.

In 1987, Simkin and Hastie [8] found empirical evidence of interactions between data encoding (representing data values by position, length, or angle) and the type of visualization being performed (comparative or proportional judgements) on the speed and accuracy of judgements. In simple terms, they found that task characteristics influenced users' performance on tasks.

In her doctoral dissertation on the topic of graphic encoding for information visualization, Nowell [9] concluded that color was generally the most effective feature (in comparison to shape and size) for representing data in a way that facilitated visualization tasks. However, she also stressed that the nature of the visualization tasks and the metric by which user performance is assessed can alter the rankings of encoding performance and concludes that task type and performance metrics are the most important features for interface designers to consider.

The influence of user traits on the effectiveness of visualizations has been studied for cognitive abilities. Perceptual speed, visual working memory, and verbal working memory cognitive abilities were found to influence both performance with and preferences for visualizations [10], [11] [12].

In addition to task characteristics and user cognitive abilities, researchers have also studied the relationship between personality traits and visualization effectiveness. Green and Fisher performed tests to assess how user performance with interactive visualizations were influenced by three psychometric factors: Locus of Control, Big Five Extraversion, and Big Five Neuroticism [13]. Green and Fisher reported that all three measures were predictive of completion times on tasks.

Mittal [14] approached the topic of graphical design motivated by the thesis that many design decisions required in the creation of visualizations selectively facilitate inferences in ways that are not apparent to users. The main focus of Mittal's paper is its presentation of a tool that allowed users to see and modify parameters associated with visualizations, with the hope of providing users with an understanding of "visual rhetorical strategies" that modify perceptions of data. As a by-product of creating this tool, Mittal proposed a taxonomy to categorize techniques for focusing attention in informational graphics. Under the category of visual prompts, he categorized strategies into ones that were planned as part of the original design of the visualization versus post-hoc strategies in which information is delivered via an overlay on the original graphic.

1.4.2 User Modeling

The work discussed in the preceding section provides evidence that visualization effectiveness is influenced by a number of factors relating to individual users. User

modeling is the process of learning and storing knowledge about users. The construction of user models is an important component of user-adaptive systems because a user model represents all the information about a user that can inform personalized support.

User modeling has been applied in a variety of contexts and draws on a number of approaches in different domains. To review the work done on this topic, it is helpful to characterize user models along the three dimensions suggested by [15] and [16]:

1. The nature of the knowledge that is being modeled.
2. The structure used to represent knowledge.
3. The approaches used to infer knowledge and maintain the model.

Taken together these three dimensions encompass what the user model represents, the data structures used to represent information, and the machine learning techniques used to build or fit the model. These dimensions are not independent of each other and can be thought of as layers with the method of inference influenced by the structure of data which is in turn shaped by the nature of the represented knowledge.

Beginning with the highest level layer – the nature of the knowledge modeled – there is already a lot of diversity in the user modeling literature. Models can be constructed to track users knowledge [17], their preferences and interests [18], their cognitive abilities [19], their search history [16], or their intentions to act [20]. The choice of what information about users is worth modeling is typically guided by the domain of the application and the potential for adaptations. For example, knowing a user's search history is useful for applications that seek to improve search results as in [16], but this information has little bearing on a user's ability to understand visualizations of arbitrary data. Modeling a user by the level of their cognitive abilities has roots in educational tutoring applications [21] but this approach is also promising for adaptive visualizations given the work on visualization effectiveness discussed in the preceding section.

The structures used to organize and store information are shaped by the nature of that information, but there are still a number of design choices that can be made while choosing which structures will represent data. Numeric values can be used to represent scores on cognitive tests or these values can be discretized into labels. Discretization

steps can involve splitting groups of users based on their median scores or using clustering methods to identify groups that are naturally present in the data as was done in [22] to separate groups of students.

Inferences can be made from a variety of data sources including interface actions, traces of users' gaze patterns, pupil dilation measurements, and other physiological sensors. User modeling based on actions or direct user input is common, but within certain domains with limited interactivity (notably including visualization) it is necessary to use sensors that can provide data when users are passively consuming content.

In user-adaptive visualizations, several researchers have approached the task of user modeling by tracking user interface actions. For instance, Grawemeyer [23] and Mouine and Lapalme [24] recorded user selections among alternative visualizations to recommend visualizations in subsequent tasks. Gotz and Wen [25] tracked suboptimal user interaction patterns to recommend alternative visualizations for the current task. Ahn and Brusilovsky [26] tracked a history of user search terms to customize the display of exploratory search results.

Gaze data has been shown to be a valuable source of information for user modeling in various domains. Eivazi and Bednarik [27] used gaze data to predict user strategies when solving a puzzle game. Kardan and Conati [28] and Bondareva et al. [29] use gaze to predict student learning with educational software, while Jaques et al. [19] leverages it for affect prediction. Liu et al. [30] predict skill level differences between users in collaborative tasks.

In information visualization, user modeling with gaze data was explored by Steichen et al. [5], and Toker et al. [2]. Steichen et al. found that task type and complexity, user performance, and three user cognitive abilities (perceptual speed, visual working memory and verbal working memory) could be predicted with accuracies significantly above a majority class baseline. Their work used simple bar graph visualizations with at most three data points per series. This thesis uses data from a study that involved more complex bar graphs (doubling the maximum data points per series) and added highlighting interventions to the graphs. It also adds the classification of a user's locus of control, and test classification accuracy with gaze features built upon interface-independent areas of interest (AOIs). Toker et al. [2] used the same dataset

leveraged by the results presented in this thesis to model users' skill acquisition. That paper also looked at the performance of interface-independent AOIs and found that they did not perform as well as AOIs based on specific interface features for predicting skill acquisition. This thesis extends the work from [2] on interface-independent AOIs to the classification of task type, user performance, and user cognitive traits.

2 User Study

As discussed in Section 1.4, the analysis in this thesis builds on empirical data collected in a prior user study. The user study, hereafter called the *Intervention Study*, recorded participants' gaze patterns with a Tobii T120 eye-tracker embedded in a computer monitor, as they completed simple bar graph interpretation tasks under several experimental conditions. The experimental software was fully automated and ran in a web-browser. Participants in the study also completed pre- and post-test assessments that measured personal traits and visualization preferences. The study had a total of 62 participants ranging in age from 18 to 42. Most of the participants were students at the University of British Columbia recruited from a variety of departments including psychology, computer science, forestry, finance, and fine arts. In addition to these student participants, seven non-student participants were recruited from the local community [6]. The remainder of this section provides details about the experimental design of the study.

2.1 Visualization Tasks and Highlighting Interventions

The participants in the study were given bar graph visualizations alongside textual questions about the displayed data. Each graph consisted of eight series elements with six data points each. The question text was always displayed beneath the graph as shown in Figure 1. The graph data is drawn from four domains: student grades per courses, vitamin/mineral levels in pet food brands, movie revenue per city, and company growth/cost per department. Domain knowledge is irrelevant for answering the given questions, but the variety of domains was added to prevent questions from becoming excessively repetitive.

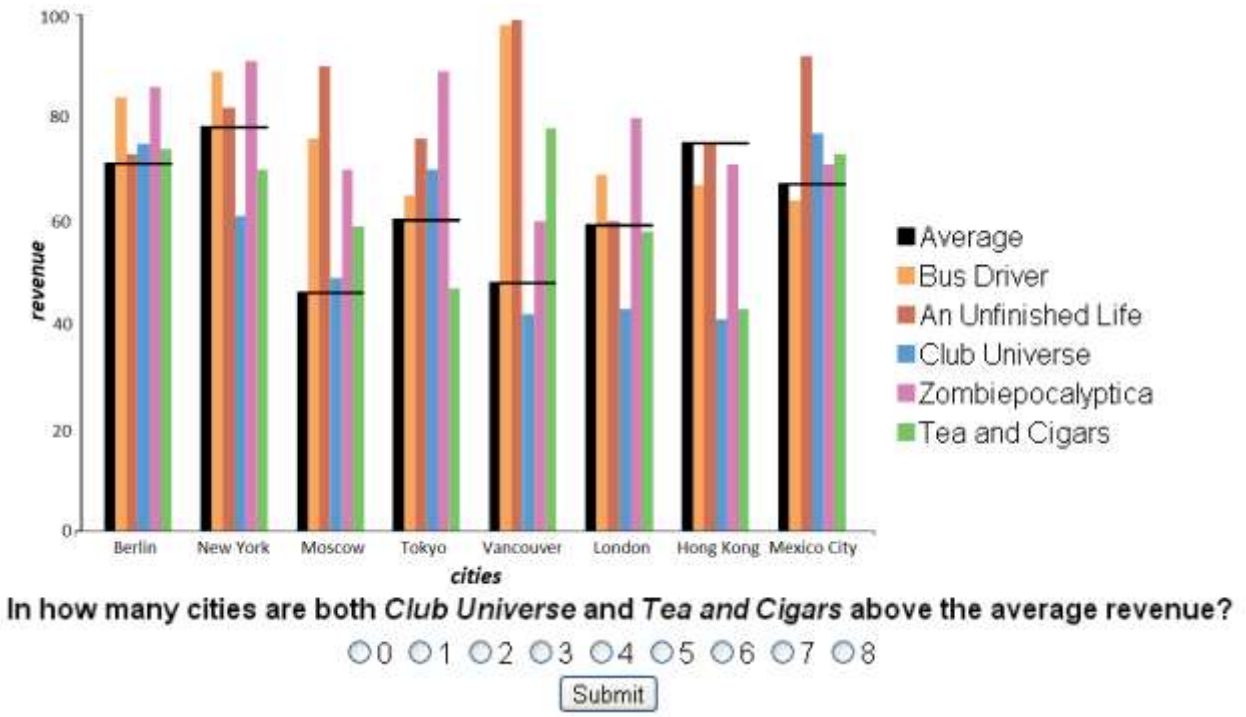

Figure 1. Screenshot of the study interface (text size adjusted for readability).

Task complexity was varied by having subjects perform 2 different types of tasks, chosen from a standard set of primitive data analysis tasks [31]. The first task type was Retrieve Value (RV), one of the simplest task types in [31], which in the study consisted of retrieving the value for a specific individual in the dataset and comparing it against the group average (e.g., “Is John's grade in Philosophy above the class average?”).

The second, more complex task type, was Compute Derived Value (CDV). The CDV task in the study required users to first perform a set of comparisons, and then compute an aggregate of the comparison outcomes (e.g., “In how many cities are both *Club Universe* and *Tea and Cigars* above the average revenue?”, graph shown in Figure 1).

The Intervention Study was designed to test the effectiveness of four different highlighting interventions that draw attention to information within a visualization. The four interventions, shown in Figure 2, were named De-emphasis, Bolding, Connected Arrows, and Average Reference Line.

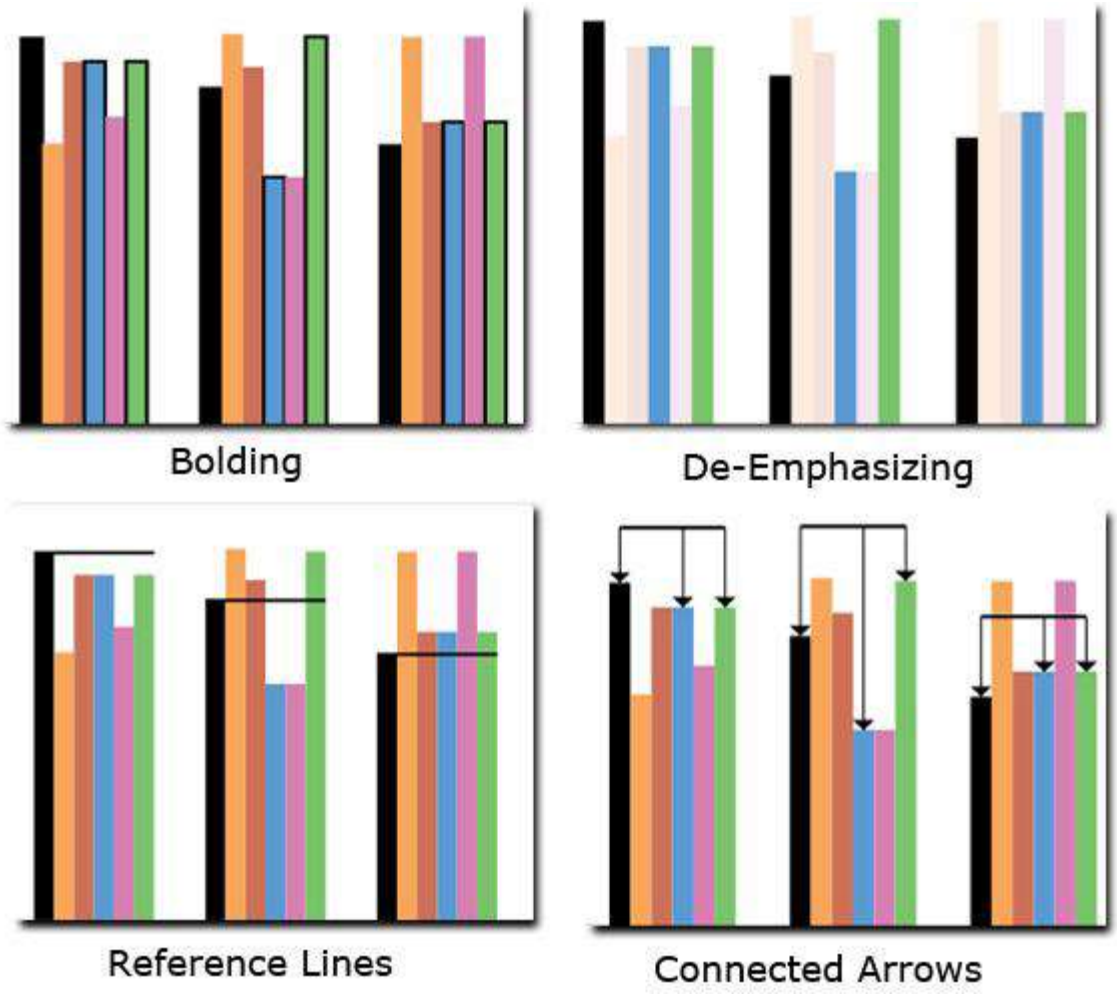

Figure 2. The set of highlighting interventions (figure reproduced from [6]).

The De-emphasis intervention draws attention to a subset of bars in the graph by reducing the opacity of all the other bars in the graph. Bolding highlights a subset of bars by placing a thick border on each selected bar in the subset. The Connected Arrows intervention highlights selected bars with arrows placed above the bars and connects arrows that point to bars within the same group with a line. Finally, the Average Reference Line intervention overlays a horizontal line to selected groups of bars to mark the average value in the group. In the Mittal taxonomy of visual prompts and cues all of these interventions are categorized as task-related visual overlays [14]. They are further classified as either information-preserving (Bolding, De-emphasis, Connected Arrows) or information-adding (Average Reference Line). All of the chosen

interventions highlight information that is relevant for participants as they perform tasks; no deliberately misleading or unhelpful interventions were tested.

Two different *delivery conditions* were used to present participants with interventions. The first condition, named "T0" for "time zero", added the intervention overlay to the graph from the start of the task (i.e. the graph and intervention were shown at the same time). The second condition, named "TX" for "time x" first displayed the question text alone, followed by the visualization without the question or any intervention, and then displayed the visualization, question, and intervention together.

The TX condition was intended to emulate dynamically added interventions, but at the time of the study it was not known when or how to present interventions dynamically (after all, answering these questions is the goal of the present thesis). In lieu of a full adaptive system, the study delayed the transition between each phase of the TX condition until a threshold of gaze fixations was crossed or until a maximum time delay elapsed. Further information about this delivery condition can be found in [6], but for the current analysis, only T0 trials are considered. The reason for discarding the TX trials is that the dynamic delivery is an unnecessary confound for the input features given to the user modeling classifiers. The ultimate goal of this work is to replace the dynamic delivery mechanism altogether with an intervention delivery mechanism informed by user modeling.

The study followed a repeated measures design involving three factors: the task type, the type of intervention shown, and the delivery condition. With 2 task types (RV and CDV), 5 intervention types (the four shown in Figure 2 plus a No Intervention condition), and 2 delivery conditions (T0 and TX) the product of these factors yields 20 experimental conditions. Each condition was repeated four times, for a total of 80 tasks per participant.

2.2 Participant Information

Every participant in the study completed a series of tests for cognitive abilities, a test for the locus of control personality trait, and responded to pre- and post-test questionnaires. This section describes the traits assessed by these tests and questionnaires. It also reports the main conclusions from Carenini et al.'s analysis of the effect of these traits on task performance and user-specific intervention preferences in the Intervention Study [6].

2.2.1 Verbal Working Memory

Verbal working memory is a cognitive trait that represents a user's capacity to store, retrieve, and manipulate linguistic information. This characteristic was assessed in the Intervention Study using a computerized implementation of Turner and Engle's OSPAN test [32]. The OSPAN test measures the span of words that users can remember while completing a secondary, arithmetic task. The Carenini et al. analysis of the Intervention Study data [6] found that verbal working memory had a significant effect overall effect on user task performance. Examining this effect in more depth, their analysis found that there was no significant effect of verbal working memory on user performance on RV tasks; the significant effect was thus attributed to the effect of verbal working memory on CDV tasks. Users with high scores on the verbal working memory assessment were found to have significantly better performance on the complex CDV tasks relative to users with lower scores.

2.2.2 Visual Working Memory

Similarly to verbal working memory, visual working memory is a cognitive trait that represents a user's ability to store, retrieve, and manipulate information. Naturally, visual working memory measures the capacity to retain visual information in contrast to verbal information. The visual working memory trait was assessed with the standardized method described in [33] and [34]. The testing method consists of a visual change detection task. Participants were shown a set of colored rectangles for 0.2 seconds followed by a blank screen for 0.9 seconds. After this retention period, participants were shown a single colored rectangle and were required to identify whether this rectangle was part of the originally displayed set or not.

Participants were shown sets consisting of 4, 6, and 8 colored rectangles during the working memory test. The visual working memory capacity metric was calculated from their results using the formula

$$K = S(H - F)$$

where K is the visual working memory capacity, S is the size of the displayed set of rectangles, H is the true positive hit rate, and F is the false alarm rate. The most variation in participant scores was found when 6 rectangles were shown ($S = 6$), thus this condition formed the basis of the measure used for analysis as it had the most utility in differentiating users.

In the Carenini et al. analysis of the Intervention Study [6], visual working memory was not found to have a significant main effect on task performance; however, in combination with task, there was found to be a significant interaction effect. Specifically, as with verbal working memory, higher working memory capacity was only found to have a significant effect on task performance for complex (CDV) tasks. The Carenini et al. analysis also noted that users with low or average visual working memory rated the Average Reference Line significantly higher than users with high visual working memory. Users with average visual working memory scores rated the Bolding intervention significantly higher than users with either low or high visual working memory scores.

2.2.3 Perceptual Speed

The perceptual speed of participants was measured with Ekstrom et al.'s "Identical Picture Test" [35]. The test consisted of 48 rows of simple geometric figures. On each row, participants were required to mark which of 5 figures were identical to a given target figure. The test was timed to be 1.5 minutes long to exert time pressure during the task.

The Carenini et al. analysis [6] of perceptual speed in the intervention study found that there was a significant effect of perceptual speed on task performance, with high perceptual speed users completing tasks significantly faster than users with low or average perceptual speeds. As was the case for both working memory traits, the effect of perceptual speed on task performance was only statistically significant for CDV tasks.

2.2.4 Locus of Control

Whereas visual working memory, verbal working memory, and perceptual speed are cognitive traits of users, locus of control is a personality trait. The locus of control trait is a measure of the extent to which a person feels they are in control of the events that surround them [36]. People who believe they have a large degree of control over these events have an *internal* locus of control, while people who attribute surrounding events to outside influences beyond their personal control have an *external* locus of control. Locus of control was measured in the Intervention Study via a questionnaire following the method presented in [36]. The Carenini et al. analysis of the study data found no statistical correlations between users locus of control scores and their task performance or intervention preferences.

2.2.5 User Visualization Expertise and Preferences

The expertise level of participants with bar graphs was assessed prior to their interaction with the study tasks. Participants were asked how frequently they: looked at simple bar graphs, created simple bar graphs, looked at complex bar graphs, and created complex bar graphs. Each of these frequencies was self-reported on a five point Likert scale. The simple and complex graphs shown as examples on the pre-study questionnaire are shown in Figure 3.

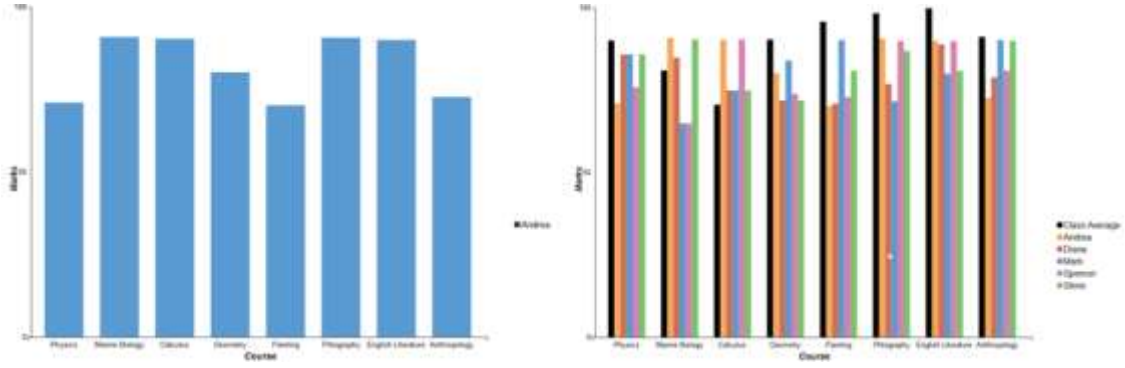

Figure 3. The simple (left) and complex (right) graphs shown to participants in the Intervention Study on questions about visualization expertise.

A post-study questionnaire was given to participants to collect data about their preferences towards the different highlighting interventions. Participants first used a five point scale to rate the usefulness of each intervention (including no intervention) on RV

tasks. They then ranked the interventions for RV tasks, again including no intervention, and were asked to add justifications for their ranking. The questionnaire then collected analogous ratings and rankings for interventions that were applied during CDV tasks. Finally, participants were given space to add any additional comments.

The Carenini et al. analysis of user traits in this study found no significant correlations between self-reported expertise and either task performance or visualization preferences [6]. The analysis did report a strong correlation between expertise with simple and complex graphs. The statistical analysis of intervention preferences and performance found that the subjective ratings of participants' preferences aligned with their objective performance results. De-emphasis was found to be the most preferred intervention overall and also the one that correlated with the fasted task completion times. De-emphasis had a significantly better performance than Connected Arrows and Bolding, which in turn had a significantly better performance than the Average Reference Line intervention. All highlighting interventions performed significantly better than the no intervention condition.

3 Eye-Tracking

The Tobii T120 eye-tracker [37] used in the Intervention Study integrates eye-tracking cameras into a 17” computer monitor with a resolution of 1280 × 1024 pixels. The eye-tracker was used without a chin rest, allowing for free head movement, though participants were instructed to remain still and not avert their gaze away from the screen during trials.

3.1 Defining Summative Gaze Features and AOIs

The Tobii Studio software that accompanies the eye-tracked used in the Intervention Study processes gaze data into a series of fixations. Fixations are clusters of gaze samples around a point on the screen that represent a lingering gaze at that point. Fixations are of interest for user modeling as they represent locations where information was most likely to have been registered by the user [38]. Quick movements between these fixations are called saccades and can be simply represented by the time between fixations, the distance between the fixations, and the inclination of the movement, as measured by the angle of the saccade relative to an absolute coordinate system and the relative angle between saccades. Figure 4 illustrates these basic gaze features.

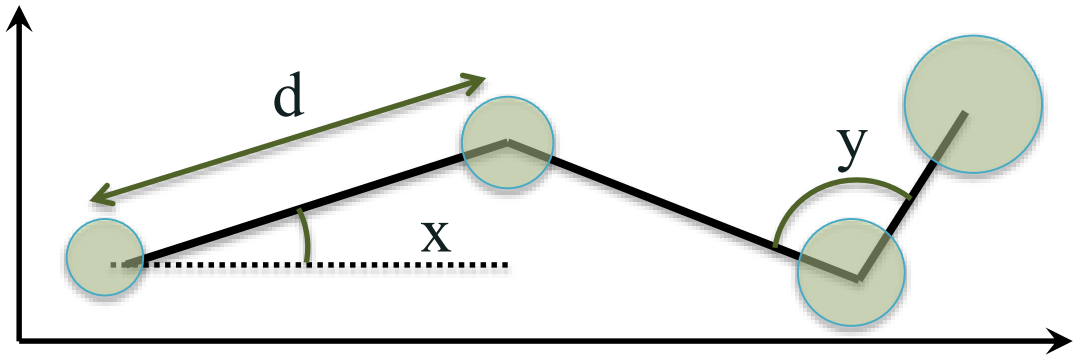

Figure 4. Basic eye tracking measures. Circles represent fixations while lines between circles represent saccades. Saccade length is represented by the measure d, while y represents relative saccade angle and x represents absolute saccade angle.

The fixation data provided by Tobii Studio was converted into a set of basic measures including the measures shown in Figure 4 as well as fixation count, rate, and duration using the open-source Eye Movement Data Analysis Toolkit (EMDAT) [39]. For

classification experiments, these basic measures were further processed using EMDAT into features that summarized the measures observed during a period of time. Sum, mean, and standard deviation measures for fixation durations, saccade lengths, relative saccade angles, and absolute saccade angles were added to a count of the total number of fixations, and the fixation rate over the given window of time to form the set of basic task-level gaze features.

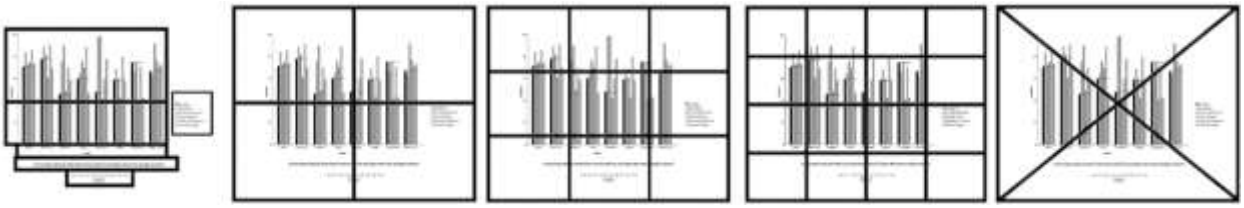

Figure 5. Areas of interest (AOIs). Left to right: Custom AOIs, 2×2 Grid, 3×3 Grid, 4×4 Grid, and X-shaped Grid

In addition to task-level features calculated on fixations occurring anywhere within the study interface, several feature sets were defined to include summative measures of fixations occurring only within *areas of interest* (AOIs). Five different sets of definitions of AOIs were constructed, as shown in Figure 5. The Custom AOI set was manually defined by the authors of [6] to place AOIs around functionally distinct areas of the interface. The areas chosen for these Custom AOIs were the graph legend, the graph labels, the text of the question prompt in the study, the region of high data variability within the graph, the region of low data variability within the graph, and the input field for responses in the study. The high and low data variability regions in the graph were constants in the study, with the data generated from uniform random distributions constrained to the high variability region. The Grid AOIs shown in Figure 5 were first introduced in [2] and represent a visualization-independent approach to defining AOIs and hence gaze features.

Given a set of AOI definitions, the following AOI-level features were calculated for each AOI: total number of fixations in the AOI, sum and mean of fixation durations in the AOI, time to the first fixation within the AOI, time to the last fixation within the AOI, longest fixation in the AOI, the proportion of the total number of fixations that occurred within the AOI, the proportion of total fixation durations that was spent within the AOI, and the proportion of the number of transitions from the AOI to every other AOI. Each

transition from one AOI to another is a separate feature (i.e. the total number of transition features is the number of AOIs squared). One feature set was created from each of the five feature sets shown in Figure 5. These feature sets all included task-level features as well as the AOI-level features. Additionally, one feature set was created that contained only task-level features and no AOI-level features. There was therefore a total of six feature sets based on summative gaze features. A list of all features is given in Table 1.

Table 1. Listing of all features. The count is the total number of features grouped into a row, where *n* is the number of AOIs.

Feature Description	Count
Task-level features	14
Sum, mean, and standard deviation of fixation durations	3
Sum, mean, and standard deviation of saccade distances	3
Sum, mean, and standard deviation of relative saccade angles	3
Sum, mean, and standard deviation of absolute saccade angles	3
Fixation rate	1
Total number of fixations	1
AOI Features	$2n^2 + 8n$
Fixation rate on AOI	n
Longest fixation duration on AOI	n
Time of first and last fixation on AOI	$2n$
Sum and proportion of fixation durations on AOI	$2n$
Sum and proportion of number of fixations on AOI	$2n$
Count and proportion of transitions from AOI to every other AOI and itself	$2n^2$

3.2 Data Validation

A number of factors within the user study could lead to invalid eye tracking data samples. Participants looking away from the screen, moving excessively, or even blinking could result in the eye tracker being unable to confidently locate gaze fixations. The Tobii Studio software that accompanied the eye tracker automatically filtered samples for measurement noise resulting from minor eye movements such as tremors and microsaccades [40]. However, this filtering process cannot guarantee that a segment of data will consist of entirely valid data.

EMDAT provided several options to verify the integrity of segment data in the process of creating summative features [39]. A minimum threshold could be set for the proportion of valid time samples in a segment, a maximum gap threshold could be set for the longest consecutive period without valid samples within a segment, or a combination of these methods could be used. Setting a minimum threshold for the proportion of valid samples ensures overall validity, while setting a maximum threshold on gap lengths limits discontinuities within the segment. For the machine learning tests in this thesis, a minimum validity threshold of 70% was used in addition to a maximum gap threshold of three seconds.

Segments were removed from the dataset if they failed to meet the validity thresholds defined above. In the Intervention Study, one segment consisted of a single visualization task represented by a graph-related question given to a participant. Every participant completed 80 such visualization tasks, thus there were 80 segments of task data per user. Data from tasks were independently classified with no information about the participants involved in the tasks beyond their gaze features. Segments of data were individually filtered for validity, thus participants who had invalid data on many tasks were not removed entirely from the dataset, provided they had at least one task that met the validity criterion.

Figure 6 displays the proportion of data samples that were discarded due to validity filtering with the parameters given above. The number of discarded samples can be seen to decrease with larger windows of time because short segments were more likely to fail to meet the required proportion of valid samples.

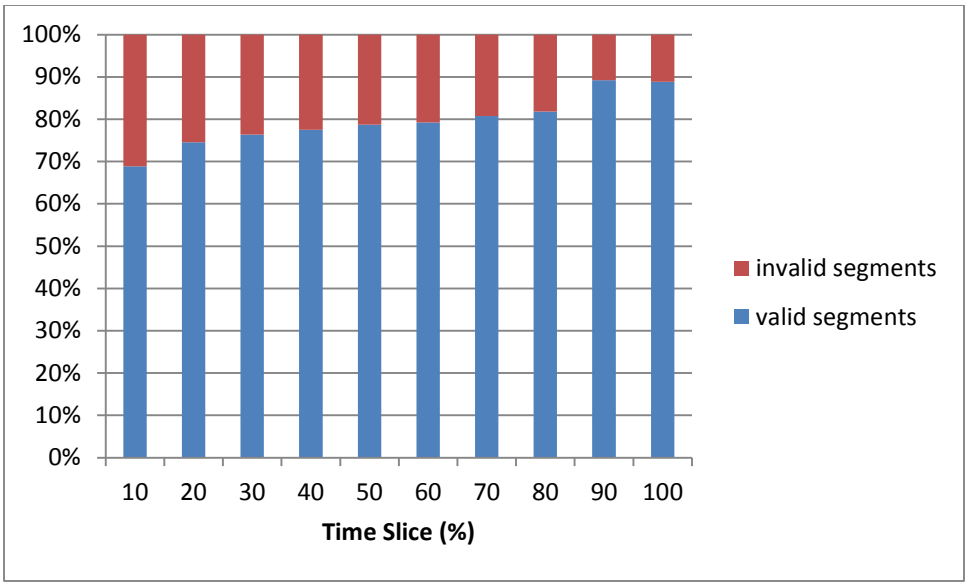

Figure 6. Percentage of valid data by time slice.

4 Classification and Analysis Methodology

The Intervention Study described in Section 0 created a dataset that combines information about users' gaze patterns during visualization tasks with their performance on different types of tasks and their individual characteristics. To investigate the research questions outlined in Section 1.2, classification experiments were run on this dataset. The broad objective in each of these experiments was to evaluate how well user and task characteristics could be predicted from the eye-tracking features gathered during a users' interactions with the visualization. This section describes the experimental procedure used to fit classifiers and analyze the resulting classification accuracies.

4.1 Classification Setup and Evaluation

The classification experiments all assessed the ability of classifiers to accurately predict the following traits: completion time, task type, visual working memory, verbal working memory, locus of control, user expertise with basic graphs, and user expertise with complex graphs. All of these traits were transformed into binary labels in the input data set, prior to the training or testing of classifiers.

Task type is by default a binary label as there were only two tasks tested in the Intervention study: the RV (simple) tasks and CDV (complex) tasks. The cognitive and personality-based user traits were all originally recorded as numerical values and were thus transformed into binary labels via median splits. For all traits, the median split never provided a 50% split, since the discretization of traits in measurement resulted in many users being equal to the median for each trait. As an example, the median score for the visual working memory trait was a word span of five words; this was the score of more than 40% of participants, so labelling this median group as either high or low visual working memory users skews the balance of classes. To partly mitigate this effect, the median split used the size of the high and low classes as a tiebreaker to place all participants with trait values equal to the median into the smaller class.

Finally, the "completion time" labels were created as a measure of users' performance on visualization tasks. Completion time refers to the amount of time required for a participant to complete one visualization task (i.e. submit an answer to

one question about a bar graph in the study). To form the labels for classification, separate median splits were performed for completion times within RV and CDV trials. The rationale for this is that the simpler RV trials took much less time on average than CDV trials, thus performing a median split across all trials would entangle information about the task type into the label. The resulting class indicates whether a participant answered a particular question faster or slower than the median response across all participants for questions of the same type.

As a measure of participants' performance on tasks, completion time has the drawback that it does not address whether participants were successful in accurately answering the given question. Slow performance on tasks is still a valid metric of performance even when users answer questions incorrectly, as it implies a user had difficulty even after deliberating for a long time. The problem is with the quick and inaccurate cases, which may reflect users rushing through tasks inattentively. Trials that were both quick and inaccurate accounted for fewer than 4% of the collected data, so it was deemed unnecessary to alter the completion time metric due to the scarcity of these inaccurate responses.

Each of the user characteristic labels and task characteristic labels discussed above was a classification target of a machine learning test. All machine learning tests consisted of predicting the target label using features computed during a single task. The training and test datasets were not partitioned by user or by time, so different participants and task types were mixed together. The constructed classifiers were not given task type or participant identifiers as features. The dataset was, however, partitioned by the type of intervention shown during tasks. In other words, classifiers were separately trained and tested for each intervention condition. The rationale for this experimental design decision was that in an adaptive visualization, the presence of interventions would be regulated by the predictions made by classifiers, thus the adaptive system would be aware of the intervention condition. Furthermore, by their nature as information highlighting aids, the visualization interventions were designed to alter the way users process information and could be expected to alter gaze patterns. Testing each intervention condition separately enabled the effect of these conditions on classification accuracy to be quantified.

Tests were evaluated using ten-fold cross-validation. This evaluation method divides the classification dataset, consisting of features and labels, into ten equally sized subsets of the data with equal label distributions. In each “fold” of cross-validation, a different subset of the data is selected as a test set and the other subsets are used as training data for classifiers. Additional runs of cross-validation were created by randomly generating different subdivisions of the dataset for folds. These additional cross-validation runs increase the robustness of the evaluation results and provide a measure of their variability. All of the machine learning tests reported in this thesis employed ten runs of ten-fold cross-validation.

4.2 Classifier Selection and Feature Selection

Preliminary tests were run with the following machine learning algorithms: logistic regression, support vector machines (SVM), naïve Bayes, multilayer perceptron, random forest, and decision tree classifiers. The selected algorithms were run using the Weka machine learning toolkit developed by the University of Waikato [41]. Overall, logistic regression showed the highest classification accuracies which is in line with the results found in previous work with similar features and target classes [5]. Due to these preliminary findings, the full set of machine learning experiments were run with logistic regression as the classifier.

A correlation-based feature selection (CFS) filter [42] was applied to reduce the number of features. The CFS algorithm selects features that correlate highly with the target class labels, but that are uncorrelated with other features. The feature selection filter was nested within folds of cross-validation such that features were selected using only information in the training sets. A wrapper-based feature selection method [43] was also considered in preliminary tests, but this method did not yield significantly higher accuracies on the given dataset and has the drawback of being much more computationally intensive than the CFS filter, especially for more complex classification algorithms. However, as there was no emphasis in this work on tuning parameters for classifiers or experimenting with novel classification algorithms, there is still a lot of room for research into more optimal classification schemes in this area.

4.3 Emulating Partial Observations with Time Slice Datasets

The features given as input to classification algorithms (described in Chapter 3) are all summaries of gaze samples taken over a period of time. In preliminary tests, this temporal window for computing features was the length of a single visualization task (i.e. a user's response to one question in the study). However, features that require data from entire tasks cannot be computed until tasks are over. To address the research question of *when* predictions can be made and what accuracy is practically achievable in a real-time system, features must be computed that emulate having only partial observations from the beginning of tasks.

The method used to emulate partial observations was to create "time slice" datasets that restrict the window of time used to compute features. The time windows used to create time slices always began at the start of visualization tasks and had a length that was a percentage of the total task length. They were created in increments of 10%, from 10% to 100% of the task length.

4.4 Experiment Definition and Statistical Analyses

A machine learning experiment was conducted for each of the eight traits of interest: task type, completion time, perceptual speed, visual working memory, verbal working memory, locus of control, expertise with basic graphs, and expertise with complex graphs. The structure of the experiment was identical for each of the traits; in the remainder of this section, the term *classification target* will refer to an arbitrary trait.

In every experiment, datasets were formed by first separating the full dataset into intervention-specific datasets as described in Section 4.1. Further datasets were then created with the time slicing method described in Section 4.3. Within each dataset, classifiers with a variety of feature sets were evaluated with 10 runs of 10-fold cross-validation. The evaluated feature sets differed in their definition and use of AOIs (see Section 3.1 for details).

A 3-way repeated measures analysis of variance (ANOVA) was conducted for each classification target (for a total of eight ANOVAs) with classification accuracy as the dependent measure. The factors in the ANOVA were classifier type (with 2 levels: baseline or logistic), feature set (with 6 levels: using no AOIs, manually defined AOIs,

and 4 levels for the 4 visualization-independent AOIs), and intervention type (with 5 levels: the no intervention condition plus 4 types of intervention). The repeated measure of classification accuracy for each condition in the ANOVA was drawn from the time slice datasets, with one run of cross-validation within one time slice providing one measure of classification accuracy for each combination of factors (to use other terms, each cross-validation run is a “subject” in the ANOVA). Pairwise comparisons within cross-validation runs (subjects) in the ANOVA were used to interpret significant effects. Reported effects are significant at the $p < 0.05$ level after all tests are corrected for multiple comparisons using the Bonferroni correction.

5 Classification Results

This section discusses the results of using eye-gaze features to classify user and task characteristics. The results for each classification target are first presented individually.

Discussion is structured to roughly follow the research questions posed in Section 1.2: first, the overall accuracies of the logistic classifier and the baseline classifier are compared to check whether classification of the trait is at all feasible. Next, comparisons are made between different classification conditions to determine how classification accuracies can be maximized with feature set and intervention selections. Then, to address the question of *when* predictions can be made, the top classification results are further unpacked by examining the accuracy achieved within time slice datasets. After the analysis of results for specific traits, Section 5.7 summarizes trends in the results between traits.

5.1 Task Type

The classification results for task type are the strongest out of all the characteristics investigated in this work. The logistic regression classifier was able to obtain an accuracy of 91.5% overall (with a standard deviation of 6.38%) on predictions of whether a given task was of the computed derived value (CDV) or retrieve value (RV) type based on eye-gaze data. This value is quite high relative to the majority-class baseline of 51.6% (with a standard deviation of 2.00%). Figure 7 summarizes these mean and standard deviation values in a bar graph. The overall accuracy is an average of the results from classifiers constructed with different AOI feature sets, displayed interventions, and amounts of observed data.

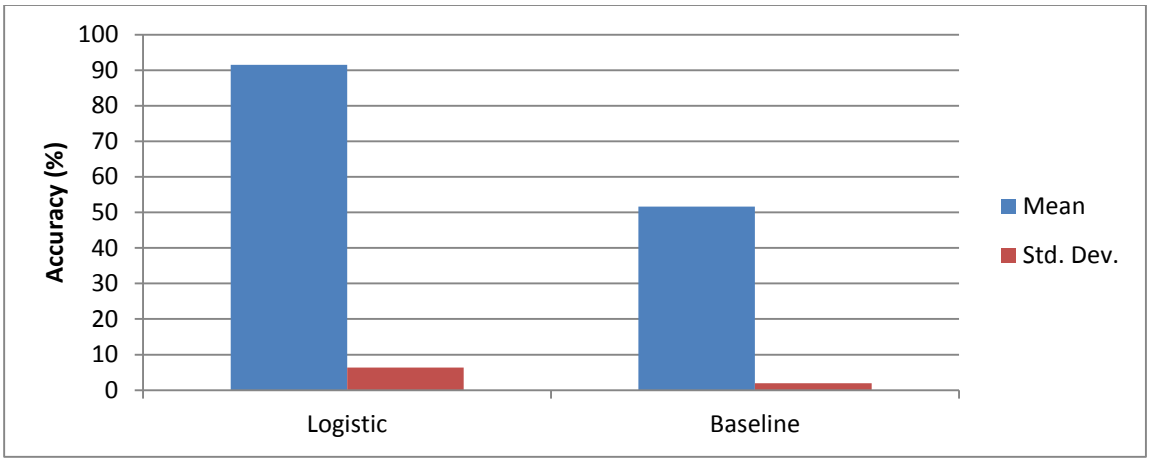

Figure 7. Overall classification results for Task Type.

ANOVA results indicate that the main effect of classifier type is significant, meaning that the mean accuracy of the logistic regression classifier is unlikely to have been so different from the mean accuracy of the majority-class baseline classifier by chance alone. Moreover, significant main effects were found for the two other experimental factors: AOI feature sets and displayed interventions. Significant interaction effects were also found between all combinations of factors.

Pairwise comparisons between results of the six feature sets revealed that classification with no AOIs had significantly lower accuracy than classification with any other feature set. The X-shaped AOIs performed significantly better than no AOIs, but significantly worse than all other AOIs. Manually-defined AOIs had significantly lower performance than the rectangular grid-based AOIs. The 3×3 grid AOIs had significantly lower performance than the 2×2 and 4×4 grid. Finally, results with the 2×2 grid had a higher mean accuracy than results with the 4×4 grid, but this difference was not found to be statistically significant. To summarize, the ordering of feature sets is listed below from the least predictive feature set (No AOIs) on the left to the most predictive feature sets on the right; feature sets with no statistically significant differences are grouped with together with underlining:

No AOI < X Grid < Custom (Manually-Defined) < 3×3 Grid < 4×4 Grid < 2×2 Grid

Figure 8 graphs the mean accuracy and standard deviation of classifiers on task type prediction with each of the AOI feature sets. Standard deviation is shown instead of standard error as standard error measures are too small to be shown on this scale.

From the figure, it is clear that while the accuracy of all classifiers is well above the baseline, the practical differences between classifiers built with different feature sets are minimal. The largest of these differences is the gap in performance obtained with any AOIs versus no AOIs, but even with no AOIs classifiers obtained a mean accuracy of 86.5% compared to 91.7% obtained with X-shaped grids, 92.2% with custom AOIs, 92.6% with 3×3 grids or 92.9% with 4×4 or 2×2 grids.

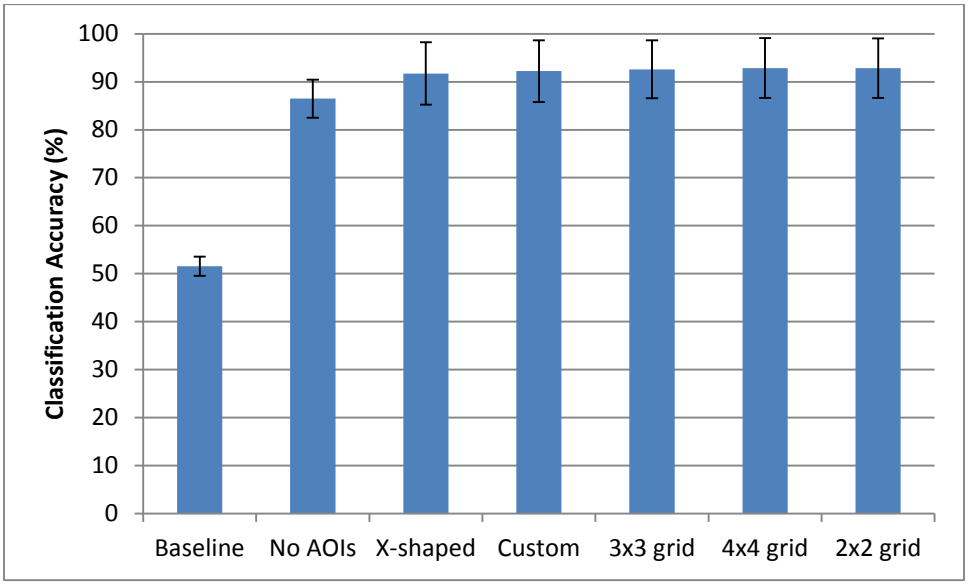

Figure 8. Accuracy of task type classification with eye gaze data given AOI feature sets. Error bars show one standard deviation.

Pairwise comparisons of results from classifiers for different intervention conditions also produced an ordered effect. The No Intervention condition and the Average Reference Line intervention are tied for the lowest performance, followed by Bolding, Connected Arrows, and De-emphasis in that order. With the exception of the No Intervention and Average Reference Line conditions, the differences in performance under each condition are all significant. Using the notation established above of grouping conditions without statistically significant differences together, the effects of interventions on overall task type prediction can be summarized as follows:

No Intervention < Average Ref. Line < Bolding < Connected Arrows < De-emphasis

A caveat to this result is that the main effect of the intervention condition considers neither the type of classifier nor the feature set. As interaction effects for all factors are significant, this leads to some counter-intuitive results. For example, part of

the main effect of intervention is due to variability in the accuracy of the baseline majority class classifier. The baseline is expected to be 50% given that the task type in the study was controlled to be even, but as some trials were discarded for their lack of valid data (see Section 0), the dataset can be skewed slightly towards one class distribution. Classifiers for the Bolding intervention were found to have significantly better performance than those built for the Average Reference Lines or No Intervention conditions, but this is only true because of the higher baseline classifier performance for the Bolding intervention. Examining only results from logistic classifiers, the Bolding intervention yielded the lowest overall mean.

Analyzing the main effects of the ANOVA factors provides some insight into the conditions in which predictions can be made about target classes, but the previous discussion highlights the danger of assuming these general effects will hold within specific conditions. Two conditions that are of particular interest for classification are the No Intervention condition and the intervention condition that yields the highest classification accuracy. The No Intervention condition is the condition that is most relevant to providing adaptive help because the decision to add highlighting interventions to a visualization would typically be made after observing only users' interactions with the default visualization. However, an adaptive system could also continue to refine its user model after presenting interventions or it could even show interventions explicitly for the purpose of building a more precise user model. For task type classification, the intervention that corresponded to the highest classification accuracies was De-emphasis.

In the No Intervention condition, the best performing feature set used the 3×3 grid AOIs to achieve a mean classification accuracy of 92.3% with a standard deviation of 6.28% compared to a baseline accuracy of 51.0% with a standard deviation of 0.403% (these figures are summarized in Figure 9). In the De-emphasis condition, the best performing feature set used the 2×2 grid AOIs to achieve a mean classification accuracy of 95.2% with a standard deviation of 3.67% compared to a baseline accuracy of 52.0% with a standard deviation of 3.33% (these figures are summarized in Figure 10). These results demonstrate the impact of AOI and intervention selection on

classification accuracy, but they still ignore the influence of time because they average together results from time slice datasets of all lengths.

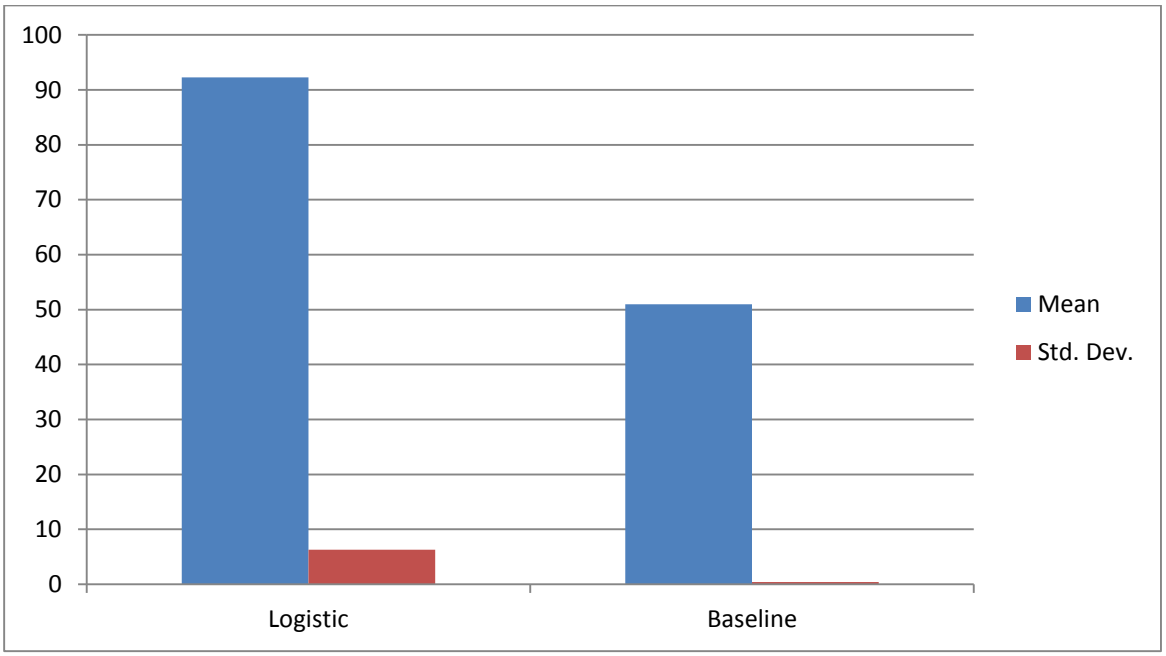

Figure 9. Classification results for Task Type with the best performing feature set in the No Intervention condition.

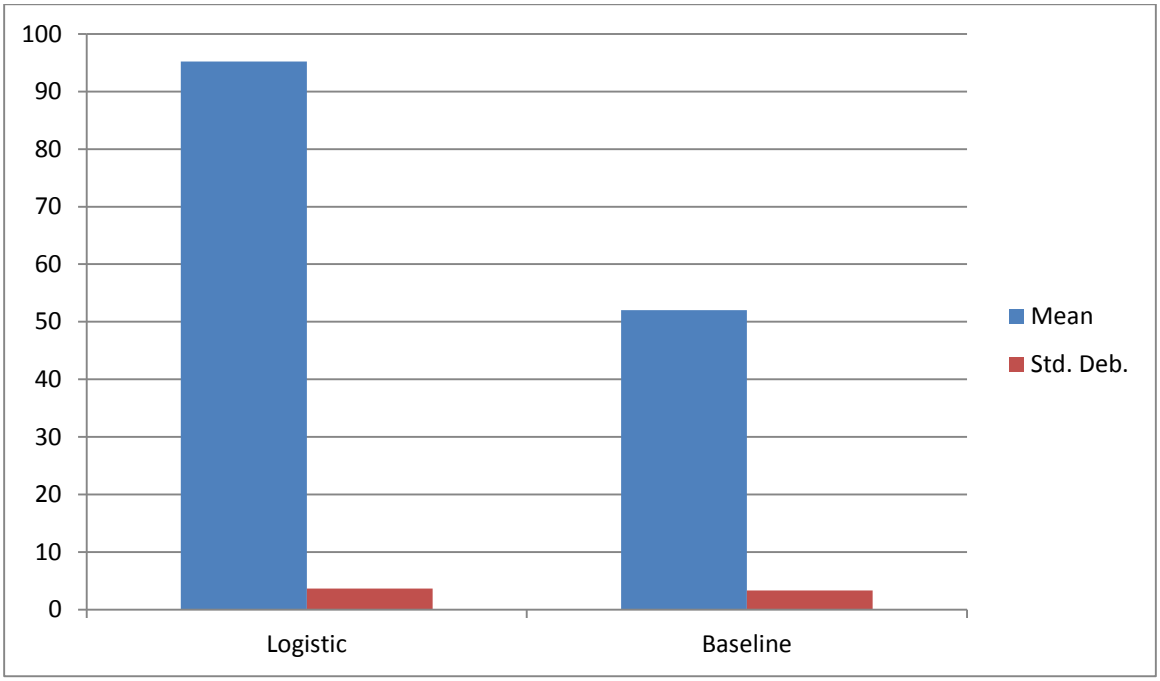

Figure 10. Classification results for Task Type with the best performing feature set over all conditions.

Figure 11 depicts trends in classification accuracy over time in the No Intervention and De-emphasis conditions. The trends are shown for the AOI feature sets with the best mean performance in these conditions which are the 3×3 grid for the No Intervention condition and the 2×2 grid for the De-emphasis condition. In both conditions, classification accuracies increase monotonically with time and reach their peak with 100% of the data observed. However, looking only at the first time slice with 10% of the task data, classification accuracy in the No Intervention condition is 81.7% with a standard deviation of 0.60% versus a baseline of 51.8% with no variability. In the De-emphasis condition, logistic regression achieves an accuracy of 89.3% given 10% of the task data with a standard deviation of 1.18% versus a baseline of 61.5% (with no variance). The trend of the baseline in the De-emphasis intervention shows that the class distribution is slightly skewed when data is only available for the first 10% of tasks, but becomes less skewed over time and is roughly even after 30% of the data is observed. This trend can be attributed to problems obtaining sufficient valid gaze data over small time intervals as the skewed in classes indicates that tasks are being dropped from the dataset for data validity reasons. The De-emphasis condition is correlated with fast completion times overall and may thus be particularly prone to these validity issues as the number of gaze samples is smaller.

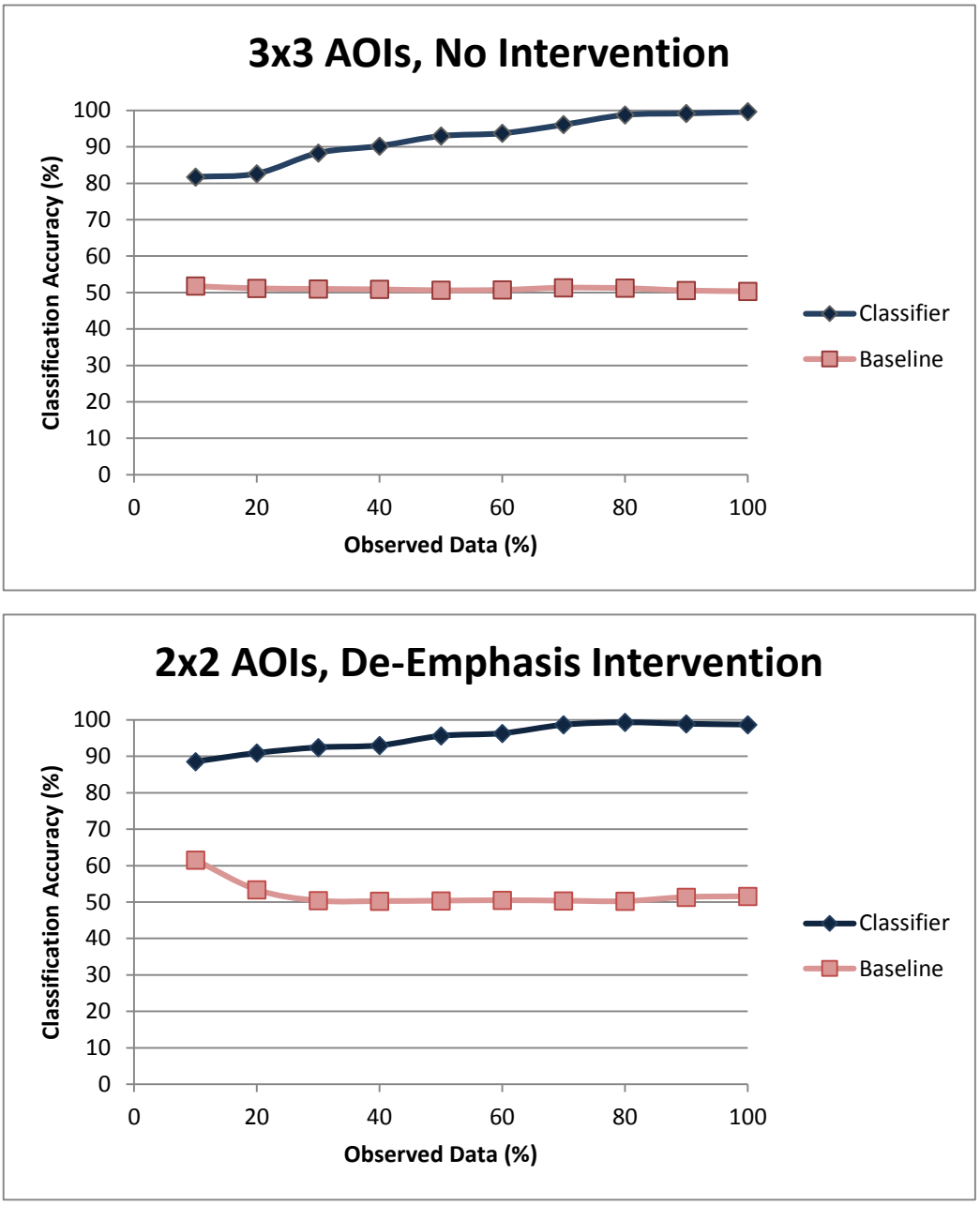

Figure 11. Task type classification accuracy with the best-performing AOI features for the No Intervention condition (top) and the De-emphasis intervention (bottom) plotted over time.

5.2 Completion Time

Participants' completion time on tasks was used as the primary measure of their performance on visualization tasks. Recall from Section 4.1 that participants were labelled as Fast or Slow on a per-task basis and that these labels are relative to the median times obtained by all participants within each task type. The overall accuracy obtained on this classification target was 88.2% (standard deviation of 3.50%) with a baseline of 73.6% (standard deviation of 6.55%).

Pairwise comparisons between feature sets of different AOI types show that the No AOI feature set has significantly lower performance than all others. The 2×2 grid and 3×3 grid AOIs yielded the highest accuracies, with no significant difference between them. The results from X-shaped AOIs, manually-defined AOIs, and 4×4 grid AOIs are not significantly different from each other, but are significantly better than the results with no AOIs and significantly worse than the results with the two smaller rectangular grids. These effects are summarized below using the notation described in Section 5.1 (recall that the least predictive feature sets are on the left):

No AOI < Custom (Manually-Defined) < X Grid < 4×4 Grid < 2×2 Grid < 3×3 Grid

Pairwise comparisons on the effects of interventions on classification accuracies show exactly the same trends for completion time classification as for task type classification. Once again, the interventions that yielded the lowest to highest accuracies are as follows:

No Intervention < Average Ref. Line < Bolding < Connected Arrows < De-emphasis

Within the No Intervention condition, the AOI feature set with the best performance used the 3×3 grid AOIs to achieve a mean accuracy of 86.2% (standard deviation of 2.45%) relative to a baseline of 65.1% (standard deviation of 0.77%). The No AOI feature set had the best performance in the De-emphasis condition with a mean accuracy of 93.2% but this reflects an extremely high baseline of 84.6%.

As seen in Figure 12, the accuracy of completion time prediction in the No Intervention condition is 85.0% (standard deviation 0.66%) after only 10% of the data from a task is observed, relative to a baseline of 65.9%. Despite the skewed baseline, completion time is predicted with accuracies significantly and substantially above the baseline from the beginning of tasks.

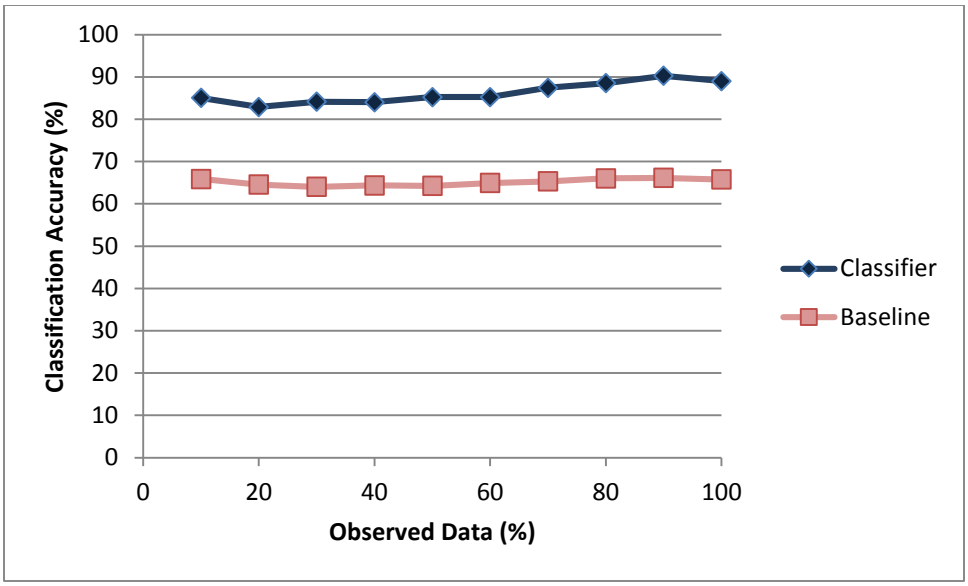

Figure 12. Accuracy of completion time classification from gaze features with 3 x 3 grid AOIs in the No Intervention condition plotted with respect to time.

5.3 Perceptual Speed

Logistic regression significantly outperformed the majority class baseline on the classification of binary perceptual speed labels. The overall mean accuracy obtained with logistic classifiers was 59.9% (standard deviation of 3.90%) while the overall mean accuracy of the baseline was 51.2% (standard deviation of 0.92%).

The type of AOIs used in feature sets and the displayed intervention were found to have significant main effects on classification accuracy as well as significant interaction effects. Pairwise comparisons between feature sets found a strictly ordered effect on classification accuracy with the mean accuracy of each feature set being significantly different from every other feature set. These effects are reported below using the notation described in Section 5.1 (recall that the least predictive feature sets are on the left):

No AOI < Custom (Manually-Defined) < X Grid < 4 × 4 Grid < 3 × 3 Grid < 2 × 2 Grid

Pairwise comparisons between the intervention conditions found that significantly lower accuracies were obtained for this trait with the Bolding intervention than with any other intervention condition. The highest classification accuracies were obtained with the Connected Arrows and Average Reference Line interventions, which were not significantly different from each other, but were significantly above the other

interventions. The No Intervention and De-emphasis conditions were between these two extremes and were not significantly different from each other. These effects are summarized as follows:

Bolding < No Intervention < De-emphasis < Connected Arrows < Average Ref. Line

Although the difference between the Connected Arrows and Average Reference Line interventions was not significant, Average Reference Lines did have a marginally higher mean accuracy of 61.9% compared to 60.1% for Connected Arrows. The Average Reference Line intervention also had a significantly lower baseline than the Connected Arrow intervention (50.7% compared to 52.1%). For these reasons, the Average Reference Line intervention has a better claim to being the “best” intervention condition for classifying perceptual speed than the Connected Arrow intervention.

The feature set that resulted in the highest classification accuracy for perceptual speed in the Average Reference Line intervention condition used the 2×2 grid AOIs. The mean accuracy with these AOIs was 63.0% (standard deviation of 3.41%) with a baseline accuracy of 50.8% (standard deviation of 0.61%). The 2×2 grid AOI feature set also had the highest performance of any feature set in the No Intervention condition, obtaining an accuracy of 61.4% (standard deviation of 3.32%) with a baseline of 50.9% (standard deviation of 0.19%).

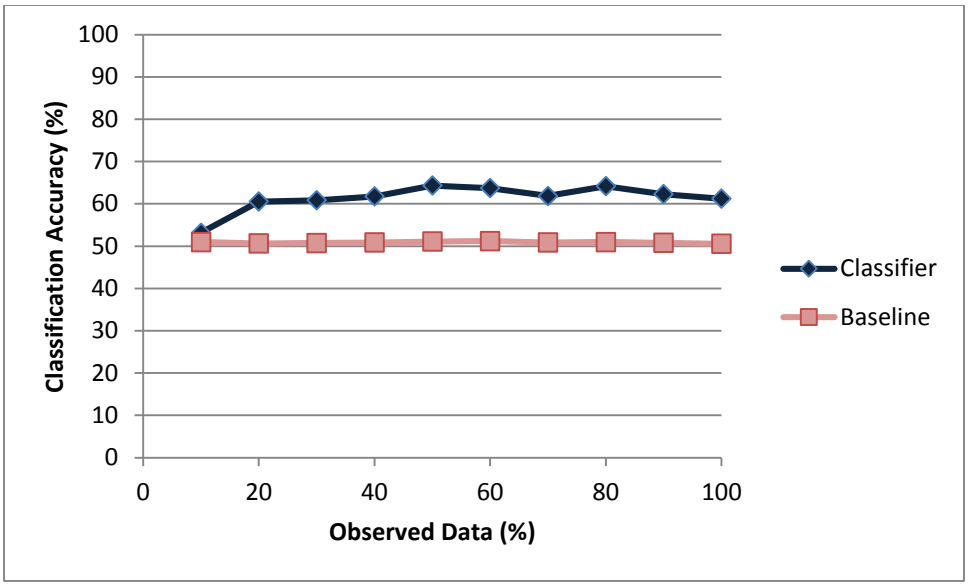

Figure 13. Accuracy over time of perceptual speed classification using gaze data with 2×2 grid AOIs.

The trend in classification accuracy over time using the 2×2 grid AOIs in the No Intervention condition is shown by Figure 13. The accuracy of the logistic regression classifier is significantly above the baseline during the first time slice with 10% of the data, but there is a considerable jump in accuracy in the second time slice with 20% of the task data. Quantitatively, the accuracy in the first time slice was 53.1% (std. dev. 1.27%) with a baseline of 50.9%, while the accuracy at the second time slice was 60.5% (std. dev. 1.44%) with a baseline of 50.6%. The classifier has a peak accuracy of 64.3% (std. dev. 1.08%; baseline 51.1%) that occurs with 50% of the task data observed.

5.4 Visual Working Memory

Visual working memory was classified by the logistic regression classifiers with an overall mean accuracy of 56.0% (std. dev. 3.52%). ANOVA results indicate that the difference between this accuracy and the baseline accuracy of 54.4% (std. dev. 0.84%) is statistically significant. The ANOVA results also indicated that the main effects of AOI feature sets and interventions on classification accuracy were significant, as were the interaction effects of all factors.

The pairwise comparisons used to interpret these statistical effects found that the feature set with no AOIs had significantly lower classification accuracies than any other

feature sets. 2×2 grids and 4×4 grids had significantly higher classification accuracies than any other feature sets but did not have significantly different effects relative to each other. The remaining feature sets – 3×3 grids, manually-defined AOIs, and X-shaped AOIs – did not have significant differences in their classification accuracies. These effects are summarized as follows:

No AOI < 3×3 Grid < Custom (Manually-Defined) < X Grid < 2×2 Grid < 4×4 Grid

The analysis of the effects of interventions on classification accuracy found that the Bolding intervention corresponded to the most accurate predictions of visual working memory. Accuracies in the Bolding condition were significantly higher than accuracies in the No Intervention condition which were in turn significantly higher than the accuracies in the other three intervention conditions. The Connected Arrows intervention had significantly better accuracies than the De-emphasis intervention; however, the Average Reference Lines intervention had results between these two interventions and was not found to be significantly different from either one. These effects are summarized as follows (note that Average Reference Line is underlined twice as it is not significantly different from De-emphasis and Connected Arrows but there are significant differences between each those two when compared directly):

De-emphasis < Average Ref. Line < Connected Arrows < No Intervention < Bolding

In the No Intervention condition, classification accuracy was highest with the 4×4 grid AOIs. The 4×4 grid AOIs were also the feature set with the best performance in the Bolding intervention, which was the intervention condition with the highest mean accuracy for classifying visual working memory. However, despite the Bolding intervention having a higher mean accuracy when average over all feature sets, the feature set with 4×4 grid AOIs had its highest accuracy in the No Intervention condition. The maximum accuracy of visual working memory classification was therefore achieved when the 4×4 grid AOIs were used in the No Intervention condition. The mean accuracy, averaged over time slices, of the 4×4 grid AOIs in the No Intervention condition was 61.5% (std. dev. 1.91%) which is relative to a baseline of 54.3% (std. dev. 0.43%).

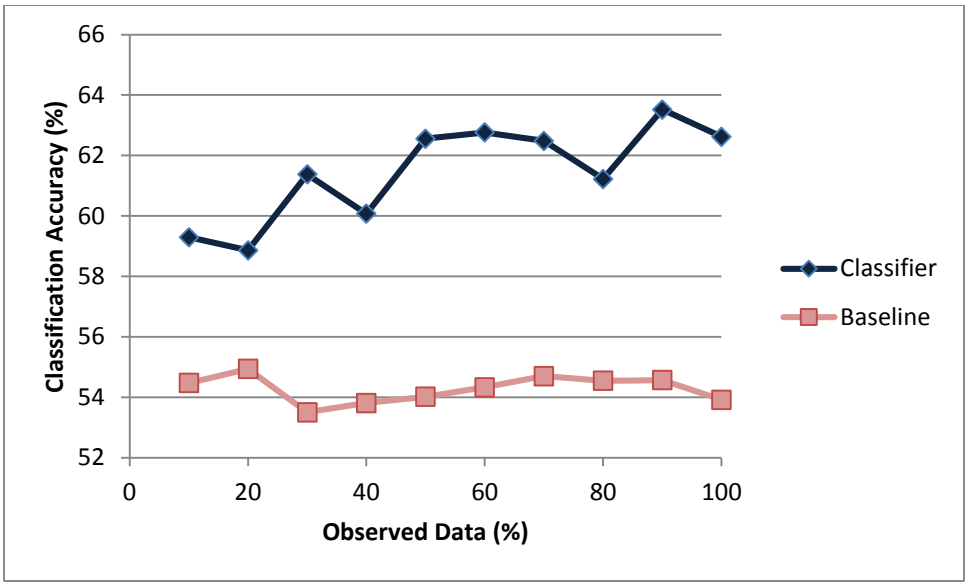

Figure 14. Accuracy of visual working memory classification from gaze features with 4 x 4 grid AOIs in the No Intervention condition plotted with respect to time.

The accuracy obtainable for visual working memory classification in the No Intervention condition with 4×4 grid AOIs varies over time as shown in Figure 14. The trend in accuracy generally increases with more observed data, but there are several local peaks in accuracy during the task. The absolute peak classification accuracy of 63.5% (std. dev. 1.12%; baseline 54.6%) was obtained with 90% of the data observed. However, an accuracy of 62.6% (std. dev, 1.60%; baseline 54.0%) was obtained only 50% of the observed data and this accuracy is both within 1% of the absolute peak accuracy and relative to a lower baseline. The first major peak in accuracy occurred with 30% of the data observed. At this peak, which also corresponds to the lowest baseline accuracy and hence the most even distribution of classes in the test set, an accuracy of 61.4% (std. dev. 1.94%; baseline 53.0%) was obtained.

The peaks in accuracy observed in the trends of classification over time may correspond to occasions within tasks in which visual working memory was commonly engaged. The decrease in accuracy from observing 90% of task data to observing 100% of task data suggests that in the final 10% of tasks users with both high and low visual working memory have more similar patterns than they do in the preceding 90% of the task. Since features are aggregates of gaze behavior over time, periods that are

relatively non-discriminatory between the predicted classes would have the effect of diluting the differences in features from previous times.

5.5 Verbal Working Memory

Similarly to visual working memory, logistic regression classification of verbal working memory was able significantly exceed the baseline results with a moderate difference in mean accuracy. The overall mean of logistic classifiers on this classification target was 61.6% (std. dev. 2.55%) versus a baseline of 59.8% (0.81%).

The type of AOI feature set was found to have a significant effect on classification accuracy. Pairwise comparisons found that the feature set with no AOIs had significantly lower classification accuracies than any other feature sets. 2×2 grids and manually-defined custom AOIs had significantly higher classification accuracies than any other feature sets but did not have significantly different effects relative to each other. The 4×4 grids, 3×3 grids, and X-shaped AOIs did not have significant differences in their classification accuracies. These effects are summarized as:

No AOI < 4×4 Grid < 3×3 Grid < X Grid < 2×2 Grid < Custom (Manually-Defined)

Interventions were also found to have a significant effect on classification accuracy. Classification results with the Bolding intervention were significantly higher than with any other intervention. Classification results with the connect arrows or Average Reference Line intervention were significantly above those with either the De-emphasis intervention or no intervention. There was no significant difference in classifier performance between the Connected Arrows and the Average Reference Line intervention, nor was there a significant difference between the De-emphasis intervention and no intervention. These effects are summarized as:

De-emphasis < No Intervention < Connected Arrows < Average Ref. Line < Bolding

In the No Intervention condition, the 4×4 grid AOI feature set had the highest classification accuracies, with a mean accuracy of 61.7% (std. dev. 1.67%) relative to a baseline of 59.5% (std. dev. 0.43%). In the intervention condition with the highest classification accuracies overall, Bolding, X-shaped AOIs obtained the highest mean accuracy of any feature set of 65.0% (std. dev. 1.33%) relative to a baseline of 59.5% (std. dev. 0.60%).

The trend in classification accuracy over time for the prediction of verbal working memory from gaze data in the No Intervention condition using the best performing feature set is shown in Figure 15. The accuracy on this classification task drops off noticeably towards the end of tasks and, as seen for visual working memory classification, there are multiple local peaks in classification accuracy during the task. The first peak, with 20% of the data observed, has an accuracy of 63.3% (std. dev. 0.99%) relative to a baseline of 59.5% The second peak, at 50% of the data, represents an accuracy of 63.5% (std. dev. 1.01%) relative to a baseline of 59.6%.

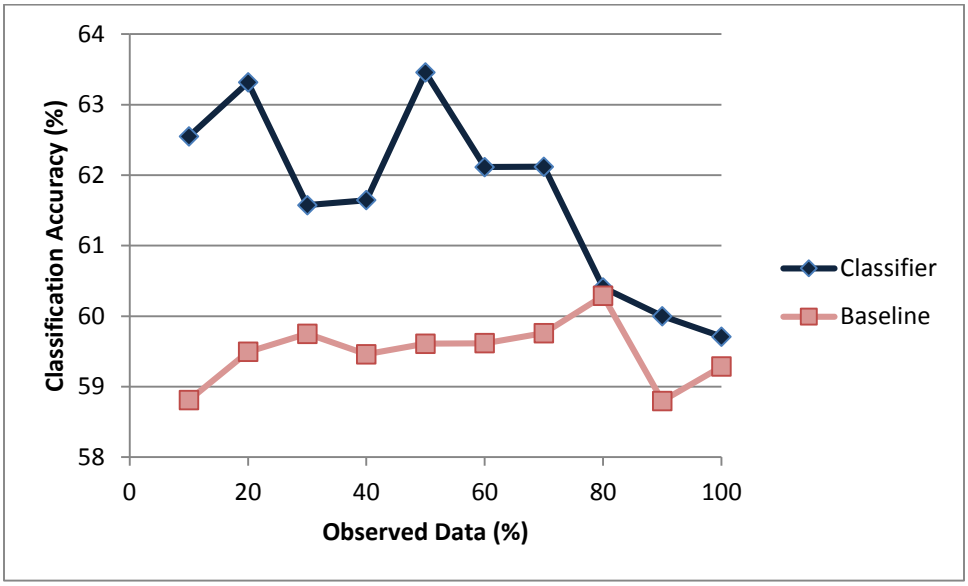

Figure 15. Accuracy of verbal working memory classification from gaze features with 4 x 4 grid AOIs in the No Intervention condition plotted with respect to time.

5.6 Expertise and Locus of Control

Classification of users' self-reported expertise did not achieve significantly better results than the baseline classifier for either expertise with simple bar graphs or expertise with complex bar graphs. Classification of the locus of control personality trait also failed to exceed the baseline. These traits were previously found not to correlate with task performance or intervention preference in the Carenini et al. statistical analysis of the intervention study [6].

5.7 Comparison across Classification Targets

This section summarizes Sections 5.1 to 5.6 by discussing results from all classification targets alongside each other. Trends that are common between results from multiple classification targets are also explored.

5.7.1 Classifier Effects

For all our classification targets except visualization expertise and locus of control, main effects of classifier in the ANOVAs indicate that logistic regression performs significantly better than the baseline. Classification targets on which significant regression results relative to the baseline were not found will not be discussed further in this section. As Table 2 shows, the mean accuracy of the logistic classifier is above 90% for task type and 85% for completion time. Accuracies for user characteristics are lower, consistent with the findings reported by Steichen et al. [44]. It should be noted, however, that the means reported in Table 2 are obtained by averaging over the results from all the tested AOI feature sets, interventions, and varying amounts of observed data. Thus, these means are conservative estimates of the accuracies that could be achieved in practice since less informative feature sets would not be used.

Table 2. Main effects of classifier type on classification.

Classification Target	Logistic Accuracy (%)		Baseline Accuracy (%)	
	Mean	Std. Dev.	Mean	Std. Dev.
Task Type	91.5	6.38	51.6	2.00
Completion Time	88.2	3.50	73.6	6.55
Verbal Working Memory	61.6	2.55	59.8	0.81
Visual Working Memory	56.0	3.52	54.4	0.84
Perceptual Speed	59.9	3.90	51.2	0.92

Looking only at tasks in which users received no interventions, Table 3 reports, for each classification target, the mean accuracy over time achieved by the best performing AOI set.

Table 3. Best obtainable classifiers with no intervention.

Classification Target	Logistic Accuracy (%)			Baseline Accuracy (%)	
	Best AOIs	Mean	Std. Dev.	Mean	Std. Dev.
Task Type	3 × 3	92.3	6.28	51.0	0.40
Completion Time	3 × 3	86.2	2.45	65.1	0.77
Verbal Working Memory	4 × 4	61.7	1.67	59.5	0.43
Visual Working Memory	4 × 4	61.5	1.91	54.3	0.43
Perceptual Speed	2 × 2	61.4	3.32	50.9	0.19

5.7.2 Effects of AOI Type

Table 4 summarizes all pairwise comparisons between AOI types for all traits with significant classification results. In this table, underlining groups factor levels for which there are no statistically significant differences in classification accuracy. For example, the table shows that Custom AOIs and X-shaped grid AOIs do not have significantly different accuracy in the classification of perceptual speed.

Table 4. Effects of AOI type on classification. See text for details.

Classification Target	AOI Type (Lowest Classification Accuracy to Highest)
Perceptual Speed	No AOIs < Custom < X < 4×4 < 3×3 < 2×2
Visual WM	No AOIs < 3×3 < Custom < X < 2×2 < 4×4
Verbal WM	No AOIs < 4×4 < 3×3 < X < 2×2 < Custom
Task Type	No AOIs < X < Custom < 3×3 < 4×4 < 2×2
Completion Time	No AOIs < Custom < X < 4×4 < 2×2 < 3×3

Several trends are visible from the pairwise comparisons in Table 4. The No AOI feature set consistently performs worse than the AOI-based sets, thus showing the value of the finer-grained information provided by the AOI-based measure. On the other hand, the Custom AOIs are generally not better than generic feature sets. In fact, for each classification target except verbal working memory, there is a generic AOI feature set with significantly higher accuracy than the Custom AOI. These findings suggest that manually isolating functional regions in the interface may not always be

necessary for creating informative AOIs, and thus provide encouraging, although preliminary, evidence that gaze-based classifiers can be built without a priori information on the target visualizations. This result is considered preliminary as its generality has yet to be tested on a wide variety of visualizations or data distributions.

Verbal WM and visual WM share several trends with respect to their accuracy with each type of AOI (e.g. the 2×2 grids were better than both the 3×3 grids and the X grids, which were in turn better than having no AOIs). The similarity in these trends could be a reflection of the similarity of the role of these two traits in task processing.

Aside from the trends noted above, the relative accuracies obtained with each set of AOIs vary among classification targets. This finding can be attributed to each target having distinct influences on the way users interact with the interface, which in turn influences which interface elements and related gaze patterns are most predictive. For example, task type influences the number and position of bars that participants must examine, while verbal WM influences how long users can retain information given in the question text, the graph labels, and the legend. The varied performance of different AOI across targets may reflect their different abilities in capturing the relevant interface elements and related attention patterns. For example, in the 4×4 Grid AOIs, three of the grid cells lie near the top of bars in the bar graph and may collectively capture the majority of visual comparisons of bar length. Features from these three AOIs were often selected in the classification of visual WM.

5.7.3 Effects of Intervention Type

Table 5 shows the significant main effects of intervention type on classification accuracy for all classification targets. As with Table 4, underlining groups conditions which did not have significantly different mean classification accuracies from each other.

Table 5. Effects of interventions on classification. The Average Reference Line intervention is abbreviated as 'Line' and the Connected Arrow intervention as 'Arrow'.

Classification Target	**Intervention (Lowest Classification Accuracy to Highest)**
Perceptual Speed	Bold < None < De-emphasis < Arrow < Line
Visual WM	De-emphasis < Line < Arrow < None < Bold
Verbal WM	De-emphasis < None < Arrow < Line < Bold
Task Type	None < Line < Bold < Arrow < De-emphasis
Completion Time	None < Line < Bold < Arrow < De-emphasis

The effects in Table 5 show that classification accuracy with the *None* condition is often worse than accuracy with an intervention. For every classification target there is at least one intervention that correlates with statistically significant improvements in predictions. For instance, for visual working memory classification, significantly better predictions can be obtained with the Bolding intervention (59% mean accuracy) than with no intervention (57% mean accuracy), but presenting other interventions reduced the accuracy of predictions. This variation in prediction accuracies across interventions may be due to the fact that, in some cases, interventions may make classification more difficult by reducing the differences in user gaze behaviors between the two groups to be predicted (e.g. helpful interventions may make the gaze behavior of low perceptual speed users closer to that of their high perceptual speed counterparts).

In addition to the main effects reported above, there were interaction effects between the AOI and intervention factors, as well as classifier type. One general implication of these results is that the effect of using a particular feature set is dependent on the intervention displayed during trials, but the specific interactions are difficult to interpret owing to the large number of conditions that could be individually considered.

5.7.4 Trends for Accuracies over Time

The trends in classification accuracy over time for user cognitive traits consistently show peaks occurring before the end of trials. One explanation of these peaks is that they may correlate with times during which the cognitive trait is most necessary during

tasks. The decrease in accuracy towards the end of tasks would thus be due to the more predictive patterns of gaze from earlier in tasks being diluted in the process of aggregating gaze data into summative features. Another possible interpretation of this trend is that features that are correlated with a trait at the beginning a task may be inversely correlated with the trait later in the same task. For example, users with high visual working memory may fixate on the graph legend early in the task while users with low visual working memory may need to fixate more on the legend later in the task.

6 Conclusions

This thesis investigated the accuracy of predicting user tasks, performance and traits, while users are performing visualization tasks with bar graphs. This research is a step toward user-adaptive visualization systems that can model their users' needs during interaction and provide real-time intervention personalized to satisfy these needs. We showed that user performance, task type, and four of the user characteristics investigated can be predicted from eye gaze data with accuracies significantly above a majority class baseline, with particularly strong results for task type and performance. These findings mirror results from previous work in which users used bar graphs for solving similar tasks with simpler datasets, thus supporting the robustness of the results to changes in visualization complexity. Furthermore, it was shown that using gaze features not customized to the specific interface used in the study delivered comparable accuracies as interface-dependent feature sets. This finding is an encouraging sign that the classification methods discussed in the thesis could be generalized across interfaces without requiring the definition of custom features for each one. Finally, it was found that classification accuracy is influenced by the highlighting interventions added to bar graphs to support visualization processing. This influence can be either negative or positive, depending on the intervention and the classification target. Interventions that facilitate continued user and task modeling could be preferred in practice over interventions that are otherwise comparably effective in improving user performance.

The relevance of this work should also be considered within the context of the broader ATUAV project [3]. As a precursor to further experimentation with different machine learning methods or more interactive visualizations, one practical benefit of this work is that it serves as a trial for testing methodology that could be reused. A couple of the lessons learned in this process include: software verification tests are needed to establish the validity of tools used in the testing process, and the statistical testing methods used in this thesis were found to require strong corrections that could induce type II statistical errors which could be mitigated with a holdout test set. Both of these methodological issues have now been addressed to some extent in ongoing work in the ATUAV project.

Further work on this project could investigate the features selected in the learned models, add more sources of data, use different feature generation methods, or use new machine learning algorithms. An interesting open question surrounding this work is how a learned user model would be used in practice. Before moving to a full adaptive system, a study could be run that used participants' cognitive abilities as measured in the pre-study tests to adapt. The construction of a functional adaptive interface and its evaluation with users is also left to future work.

References

[1] M. J. Gingerich and C. Conati, "Constructing Models of User and Task Characteristics from Eye Gaze Data for User-Adaptive Information Highlighting," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[2] D. Toker, B. Steichen, M. Gingerich, C. Conati, and G. Carenini, "Towards Facilitating User Skill Acquisition - Identifying Untrained Visualization Users through Eye Tracking," in *Proceedings of the 2014 international conference on Intelligent user interfaces*, 2014.

[3] C. Conati, G. Carenini, D. Toker, and S. Lallé, "Towards User-Adaptive Information Visualization," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[4] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, Oct. 1976.

[5] B. Steichen, C. Conati, and G. Carenini, "Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data," *TIIS*, 2014.

[6] G. Carenini, C. Conati, E. Hoque, B. Steichen, D. Toker, and J. T. Enns, "Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[7] C. Ziemkiewicz, R. J. Crouser, A. R. Yauilla, S. L. Su, W. Ribarsky, and R. Chang, "How locus of control influences compatibility with visualization style," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011, pp. 81–90.

[8] D. Simkin and R. Hastie, "An Information-Processing Analysis of Graph Perception," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 454–465, Jun. 1987.

[9] L. Nowell, "Graphical Encoding for Information Visualization: Using Icon Color, Shape, and Size to Convey Nominal and Quantitative Data," PhD Dissertation, Virginia Polytechnic Institute and State University, p. 75, 1997.

[10] C. Conati and H. Maclaren, "Exploring the role of individual differences in information visualization," in *Proceedings of the working conference on Advanced visual interfaces*, New York, NY, USA, 2008, pp. 199–206.

[11] M. C. Velez, D. Silver, and M. Tremaine, "Understanding visualization through spatial ability differences," in *IEEE Visualization, 2005. VIS 05*, 2005, pp. 511–518.

[12] D. Toker, C. Conati, G. Carenini, and M. Haraty, "Towards adaptive information visualization: on the influence of user characteristics," in *Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization*, Berlin, Heidelberg, 2012, pp. 274–285.

[13] T. M. Green and B. Fisher, “Impact of personality factors on interface interaction and the development of user profiles: Next steps in the personal equation of interaction,” *Information Visualization*, vol. 11, no. 3, pp. 205–221, Jul. 2012.

[14] V. O. Mittal, “Visual Prompts and Graphical Design: A Framework for Exploring the Design Space of 2-D Charts and Graphs,” in *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, Providence, Rhode Island, 1997, pp. 57–63.

[15] D. Sleeman, “UMFE: A User Modelling Front-end Subsystem,” *Int. J. Man-Mach. Stud.*, vol. 23, no. 1, pp. 71–88, Jul. 1985.

[16] P. Brusilovsky and E. Millán, “User Models for Adaptive Hypermedia and Adaptive Educational Systems,” in *The Adaptive Web*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer Berlin Heidelberg, 2007, pp. 3–53.

[17] J. C. L. Jonathan P. Rowe, “Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments.,” 2010.

[18] J. Fink and A. Kobsa, “User Modeling for Personalized City Tours,” *Artificial Intelligence Review*, vol. 18, no. 1, pp. 33–74, Sep. 2002.

[19] N. Jaques, C. Conati, J. M. Harley, and R. Azevedo, “Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring System,” in *Intelligent Tutoring Systems*, S. Trausan-Matu, K. E. Boyer, M. Crosby, and K. Panourgia, Eds. Springer International Publishing, 2014, pp. 29–38.

[20] Y.-M. Jang, R. Mallipeddi, and M. Lee, “Identification of human implicit visual search intention based on eye movement and pupillary analysis,” *User Model User-Adap Inter*, vol. 24, no. 4, pp. 315–344, Oct. 2014.

[21] B. Woolf, E. Aimeur, R. Nkambou, and S. Lajoie, Eds., *Intelligent Tutoring Systems*. Montreal, Canada, 2008.

[22] C. Conati and S. Kardan, “Student Modeling: Supporting Personalized Instruction, from Problem Solving to Exploratory Open Ended Activities,” *AI Magazine*, vol. 34, no. 3, pp. 13–26, Sep. 2013.

[23] B. Grawemeyer, “Evaluation of ERST – An External Representation Selection Tutor,” in *Diagrammatic Representation and Inference*, vol. 4045, 2006, pp. 154–167.

[24] M. Mouine and G. Lapalme, “Using Clustering to Personalize Visualization,” in *2012 16th International Conference on Information Visualisation (IV)*, 2012, pp. 258–263.

[25] D. Gotz and Z. Wen, “Behavior-driven visualization recommendation,” in *Proceedings of the 14th international conference on Intelligent user interfaces*, New York, NY, USA, 2009, pp. 315–324.

[26] J. Ahn and P. Brusilovsky, “Adaptive visualization for exploratory information retrieval,” *Information Processing & Management*, vol. 49, no. 5, pp. 1139–1164, Sep. 2013.

[27] S. Eivazi and R. Bednarik, “Predicting Problem-Solving Behavior and Performance Levels from Visual Attention Data,” presented at the 2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI 2011, 2011, pp. 9–16.

[28] S. Kardan and C. Conati, “Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation,” in *In: Proc. of UMAP, 21st Int. Conf. on User Modeling, Adaptation and Personalization*, 2013.

[29] D. Bondareva, C. Conati, R. Feyzi-Behnagh, J. M. Harley, R. Azevedo, and F. Bouchet, “Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning,” in *Artificial Intelligence in Education*, vol. 7926, H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 229–238.

[30] Y. Liu, P.-Y. Hsueh, J. Lai, M. Sangin, M.-A. Nussli, and P. Dillenbourg, “Who is the expert? Analyzing gaze data to predict expertise level in collaborative applications,” in *IEEE International Conference on Multimedia and Expo, 2009. ICME 2009*, 2009, pp. 898–901.

[31] R. Amar, J. Eagan, and J. Stasko, “Low-Level Components of Analytic Activity in Information Visualization,” in *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, Washington, DC, USA, 2005, pp. 15–21.

[32] M. L. Turner and R. W. Engle, “Is working memory capacity task dependent?,” *Journal of Memory and Language*, vol. 28, no. 2, pp. 127–154, Apr. 1989.

[33] K. Fukuda and E. K. Vogel, “Human Variation in Overriding Attentional Capture,” *J. Neurosci.*, vol. 29, no. 27, pp. 8726–8733, Jul. 2009.

[34] E. K. Vogel, G. F. Woodman, and S. J. Luck, “Storage of features, conjunctions, and objects in visual working memory,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 1, pp. 92–114, Feb. 2001.

[35] R. B. Ekstrom and U. S. O. of N. Research, *Manual for Kit of Factor Referenced Cognitive Tests*. Educational Testing Service, 1996.

[36] J. B. Rotter, “Generalized expectancies for internal versus external control of reinforcement,” *Psychological Monographs: General and Applied*, vol. 80, no. 1, pp. 1–28, 1966.

[37] Tobii Technology, “An introduction to eye tracking and Tobii Eye Trackers.” 27-Jan-2010.

[38] K. Rayner, “Eye movements in reading and information processing: 20 years of research,” *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998.

[39] S. Kardan, “Eye Movement Data Analysis Toolkit (EMDAT) User Manual.” 05-Sep-2012.

[40] A. Olsen, “Tobii I-VT Fixation Filter - Algorithm Description.” 20-Mar-2012.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[42] M. Hall, “Correlation-based Feature Selection for Machine Learning,” PhD, University of Waikato, 1999.

[43] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997.

[44] B. Steichen, G. Carenini, and C. Conati, “User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities,” in *Proceedings of the 2013 international conference on Intelligent user interfaces*, New York, NY, USA, 2013, pp. 317–328.