An Analysis of Automated Visual Analysis Classification: Interactive Visualization Task Inference of Cancer Genomics Domain Experts

Connor C. Gramazio, *Student Member, IEEE,* Jeff Huang, David H. Laidlaw, *Fellow, IEEE*

Abstract—We show how mouse interaction log classification can help visualization toolsmiths understand how their tools are used "in the wild" through an evaluation of MAGI – a cancer genomics visualization tool. Our primary contribution is an evaluation of twelve visual analysis task classifiers, which compares predictions to task inferences made by pairs of genomics and visualization experts. Our evaluation uses common classifiers that are accessible to most visualization evaluators: k-nearest neighbors, linear support vector machines, and random forests. By comparing classifier predictions to visual analysis task inferences made by experts, we show that simple automated task classification can have up to 73% accuracy and can separate meaningful logs from "junk" logs with up to 91% accuracy. Our second contribution is an exploration of common MAGI interaction trends using classification predictions, which expands current knowledge about ecological cancer genomics visualization tasks. Our third contribution is a discussion of how automated task classification can inform iterative tool design. These contributions suggest that mouse interaction log analysis is a viable method for (1) evaluating task requirements of client-side-focused tools, (2) allowing researchers to study experts on larger scales than is typically possible with in-lab observation, and (3) highlighting potential tool evaluation bias.

Index Terms—classification, task analysis, visual analysis, biology visualization, visualization, cancer genomics

1 Introduction

In this work we study whether interaction log classification can serve as a new, effective visualization tool design evaluation methodology, and focus on how it can augment traditional qualitative approaches by providing additional context for previously determined tasks. We also explore how predictive task inferences may improve the iterative design process of interactive visualization tools for domain experts. To accomplish this, we ground our exploration in an analysis of MAGI [1] — a cancer genomics visualization tool.

1.1 Contributions

Our first contribution is a discussion that compares the accuracies of twelve automated visual analysis task classification models to hand-coded task inferences made by pairs of genomics and visualization experts. Rather than focusing on sophisticated classification models, our evaluation focuses on classifiers that most visualization researchers could implement themselves: k-nearest neighbors, linear support vector machines (SVMs), and random forests. This way, our findings are more applicable to visualization researchers and practitioners at-large. We discuss the potential benefits that might come from evaluating more complex models in Section 7. Our second contribution is an exploration of common MAGI interaction trends using the predictions from task classification, which expands our present understanding of how visualization is used "in the wild" by cancer genomics domain experts. As part of this investigation, we make our third contribution by exploring how mouse interaction modeling can be used to inform iterative tool design. We also provide design principle hypotheses that can be used to guide future design studies.

These contributions extend current tool evaluation methodologies, which typically focus on field studies and other similar, typically qualitative, types of observation [2]. Although working side-by-side with domain experts in field research yields high levels of detail about analysis workflows, as Carpendale notes, these types of studies are typically smaller in scale and lack precision [3]. Our contributions could provide an important addition to current evaluation methodologies because interaction logs can be passively collected as part of domain experts' natural workflows and also contain precise, quantitative descriptions of visual analysis. Because of this, interaction log analysis can circumvent several common limitations present in more focused and contextual-rich methodologies (e.g., ethnographies). For example, through interaction log analysis, it is easier to study larger populations of domain experts while retaining ecological validity and without potential interference caused from direct observation. Likewise, analyzing large collections of interaction logs may help thwart bias caused from observing small in-lab populations.

Another motivation of our present work was to understand the degree to which anonymized interaction logs could be used to understand analytic intent given the complete omission of context. Our evaluations of visual analysis task inference by humans and computers rely on interaction logs that contain the size and location of each visualization in MAGI and the sequence of mouse events caused by user interaction (i.e., clicks, movements, and scrolls).

- *Connor C. Gramazio, Jeff Huang, and David H. Laidlaw are with the Department of Computer Science at Brown University. E-mail: {connor, dhl}@cs.brown.edu, tvcg@jeffhuang.com.*

Manuscript received April 19, 2005; revised August 26, 2015.

1.2 Outline

We begin with interaction log mining background and related work. Then, we provide a short description of MAGI, including a summary of the application domain. We also explain what types of information we collected in the MAGI interaction logs. Next, we discuss results from a preliminary task inference study in which we worked with two MAGI developers to identify eight common MAGI analysis tasks. We then discuss the results from a task-labeling experiment that provided training data for task classification evaluation. Following our in-lab experiments, we then move on to our classifier evaluation and explore the potential effect that interaction log mining might have on domain expert tool iterative design. Last, we present open research questions and consider the potential broader impact of our contributions.

2 Background and Related Work

2.1 Understanding users: contribution differences

While our present work is related to previous "clickstream" interaction research, our contributions differ: we aim to model less deterministic visual analysis behavior of experts instead of modeling typical navigation behavior of the general population through a sequence of URLs (e.g., to optimize search ranking [4] or commerce [5]). These historically studied clickstream tasks are more deterministic because a user's goal is to find the most relevant search result and will end with a success (search result click) or a failure (search termination or another query). In contrast, visual analysis is typically driven by deriving "insight," which is subjective and variable across applications [6]. Because of these potential empirical differences, we test whether clickstream features from the information retrieval community can accurately model visualization interaction. Hence, another contribution of this work is to assess whether features that were advantageous for classifying these simpler, more deterministic interactions in web search apply as well to more open-ended visual analysis scenarios. However, further evaluating how visual analysis interaction procedures may differ from better-studied and modeled areas of human-computer interaction remains an important area for future research.

2.2 Understanding analytic intent via interaction logs

Our present research complements and expands on automated analytical task inference techniques within visualization and across the broader human-computer interaction community. Although manual interaction analysis has proven useful in smaller case studies such as studying visual analysis in investigative journalism [7] and in understanding collaborative analysis [8], Guo et al. note that hand-coded interaction analyses face myriad scalability issues [9]. As such, many researchers have investigated the automation of visual analysis interaction log evaluation. These techniques often seek to identify design requirements by leveraging interactions as a record of "analytical provenance," which can be loosely defined as a collection of analytical steps undertaken during a visualization's use. Given the scope of provenance research, we recommend Ragan et al.'s survey for a comprehensive overview [10].

Much of this research has focused on action log analysis, which relies on basic software interaction sequences (e.g., `filter` → `sort` → `select`). For example, Zgraggen et al. showed how extracting interaction patterns using regular-expression-like queries from large action datasets helped usability researchers at a large technology company identify key issues in their products [11]. Other visual analysis task reconstruction methods draw on techniques such as multiple sequence alignment [12], [13], [14], [15], graphical modeling [16], and human-in-the-loop qualitative exploration [17]. Etemadpour et al.'s investigation into genomics analysis workflows is more similar to our inquiry into domain expert analysis, but also uses an action representation akin to other previous work [18]. Our present contributions differ from these efforts because we focus on lower-level mouse event analysis (e.g., mouse dwell time) to infer analytic intent, rather than focusing on higher-level interaction representations (e.g., "undo" in a graph-like structure representing workflows [10]).

One benefit to analyzing lower-level mouse events opposed to higher-level representations is the close relationship between mouse movement and gaze, which is a well-studied physiological indicator of intent [19]. Huang et al. as well as Rodden and Fu explore how the relation between gaze and mouse movement can be used to improve web search [20], [21], and Gomez et al. show that the relation also holds for visualization [22]. We utilize this similarity later in our classification evaluation by creating a new feature set inspired by these similarities (Sec. 6.1.3).

Martín-Albo et al. build on the association between intent and mouse interactions to show that intent can be inferred from mouse movement alone without the aid of eyetracking by testing the geometric and temporal similarity between mouse traces [23]. Others like Edmonds et al. and Matejka et al. developed tools to qualitatively analyze mouse traces and intent through heatmaps of frequently interacted-with interface regions [24], [25]. Blascheck et al. pursued a hybridized in-lab approach and tested how event-level interaction logs can be combined with talk-aloud transcripts and eye-tracking to understand interaction [12]. Noting the potential benefits of using higher-resolution interaction logs, Atterer et al. performed a case study to show how interaction strategies and intent can be reconstructed from low-level event logs [26]. Our present work extends knowledge of user analytic intent by analyzing how interaction log classification can lead to insights about domain experts' ecological visual analysis behavior.

2.3 Relation to past biology visualization task analyses

Our present contributions extend previous research that also used biology visualization as a test bed for new evaluation methodology and task modeling. For example, Saraiya et al. developed an evaluation methodology to measure visualization effectiveness based on how many analytical insights it may support [27] and then explored how insights could be used to longitudinally understand visual analysis tasks [28]. O'Brien et al. then extended insight-based methodology to improve its precision while also evaluating another biology-

visualization-motivated application [29]. Instead of just tallying the total number of insights, they suggested that insights – and the tasks that produced them – could be better understood by also measuring a variety of other information such as hypothesis-driven insights and insight complexity. Unlike these past methodological contributions, which rely on hand-coding data, our present line of inquiry investigates how automated modeling can empower initial human classification. Not only does this continue O'Brien's line of research toward quantifying task analysis, but it also allows task analysis to scale to much larger collections of data thanks to automated task inference.

Others, like Streit et al., used biology visualization to study visual analysis in areas where there are diverse types and formats of data [30]. Whereas Streit et al. focused on constructing a model for heterogeneous biological data analysis, Murray et al. synthesized common analysis tasks in biological network analysis [31]. Although both sought to explain cancer genomics visual analysis, the aims of our present work are distinct. Differences between our present contributions and these past two models might be best understood through Brehmer and Munzner's task typology [32]: Streit et al. primarily focused on "what" each task was operating on, Murray et al. primarily focused on "why" each task was being performed, and our present research primarily focuses on "how" each task was being performed.

3 MAGI AND LOG COLLECTION

Our investigation into visual analysis task classification is anchored by studying MAGI mouse interaction logs. MAGI is an online visualization tool that allows cancer genomics researchers to explore a variety of genetic mutation data across many cancers in five visualizations [1]. A screenshot of a query in MAGI is shown in Figure 1. Given cancer genomics specialization variety, MAGI was designed to support a diversity of expertise through its multiple views (e.g., basic science vs. pharmaceutical research; wet lab biologist vs. bioinformatician).

The top-most visualization in MAGI is an aberration matrix, which uses color to show mutations (cells) in user-queried genes (rows) across different sequencing samples (columns; i.e., patients). Below the aberration matrix, the linked heatmap can show related continuous data (e.g., gene expression) for the same combinations of genes and samples. The third visualization row in MAGI shows a network view and a transcript annotation chart. The network view shows how the proteins that each queried gene encodes can interact with one another, whereas the transcript annotation chart shows the physical location where mutations occur. The last visualization shows the physical location of copy number aberrations which affect large swaths of the genome.

While researchers might use only one visualization for analysis, visualizations may also be used together. For example, a researcher might use the aberration matrix to identify stair casing patterns of "mutual exclusivity," which are an indicator of biological significance. Or, they might continue that line of inquiry after detecting mutual exclusivity for a subset of mutations and examine where they physically occur in the transcript annotation chart.

Type of information	Attributes
Mouse events	{click, move, scroll}, time, x, y
Tooltip events	x, y, width, height
MAGI components ($\times 6$)	x, y, width, height
Window state	width, height
Query	number of genes and datasets

TABLE 1
Data contained in each MAGI mouse trace interaction log. MAGI components refer to the five visualizations and control panel.

Like with many other visual analysis tools for domain experts, one difficulty in evaluating MAGI is that cancer genomics researchers are geographically distant and are often hard to schedule for observation. This poses a hurdle for user-centered design because these limitations often result in studies that consider only small numbers of tool users. Although small case studies can provide useful information about tool-use, they can be susceptible to sample bias without careful recruitment consideration. This is particularly true in cancer genomics, which has many distinct foci that use the same data (e.g., applied pharmaceutical vs. basic science research). As such, it is possible that relying on small population observations could cause iterative design decisions to overfit a tool to the requirements of a small number of users at the expense of a large, unstudied subpopulation. If successful, interaction log classification would provide a way for understanding task requirements of entire populations in ecological settings, and would provide a way to help counter sample bias using the smaller scale, in-lab methodologies that tool evaluators already utilize.

3.1 Mouse interaction log schema

Our interaction classification evaluation classification focuses on analyzing mouse interaction logs collected on MAGI's gene set query results page. We provide an example query about the Notch pathway, which is implicated in a variety of cancers [33], in Figure 1. Each collected log contains information about all mouse events, each visualization's size and location, the window size, and anonymized queries. In addition to the five visualizations, we also collected the size and location of MAGI's control panel and tracked when tooltips were activated in each of MAGI's visualizations. Given that users can toggle visualization visibility, we also tracked how size and location of the visualizations might have differed over time. The full collection of log attributes is listed in Table 1.

3.2 Log culling

We applied a two-step culling process to remove interaction logs that were unlikely to contain important information about visual analysis tasks. The first step in log culling involved the removal of logs without mouse interactions, which were created by web crawlers. This removal resulted in 1,616 logs with mouse event data. Afterwards, we then removed 63 logs that were deemed to have too few events to describe visual analysis tasks. For example, a user might realize that they mistyped their query and immediately navigate backward. While this scenario might provide important usability information about tool-use, it does not express information about the analytic intent of what the

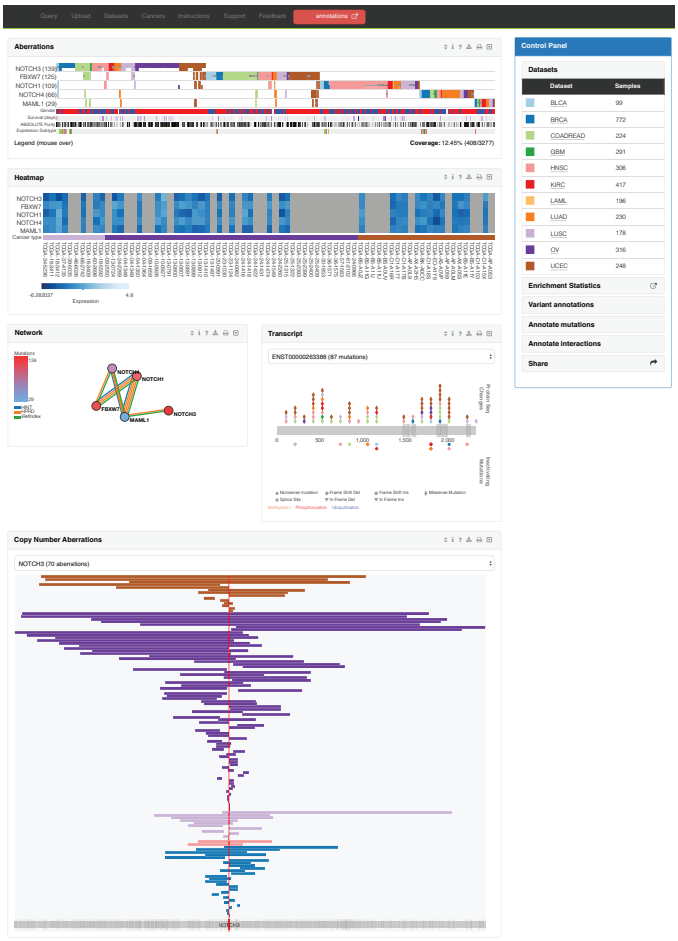

Fig. 1. A screenshot of MAGI showing the aberration matrix (top), heatmap (second top), network view (middle-left), transcript chart (middle-right), copy-number aberration browser (bottom), and control panel (right).

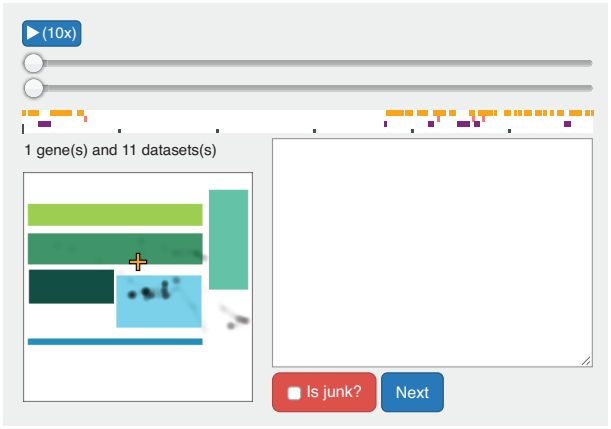

Fig. 2. Example free-text label trial where participants were asked to provide a 1 to 2 sentence description of what type of task was performed in the visualized interaction log. Interaction logs were summarized in a visualization in each trial, which showed the location for each of MAGI's five visualizations in differently colored rectangles, and mouse activity with a black heatmap overlay. Users could watch the mouse and tooltips appear/disappear by using the playback button and two sliders to change time. The timeline below the sliders showed mouse movement (orange), click (red), and scroll (purple) events. Users could play the log by clicking on a $10\times$ playback button or manually control playback with two sliders (top: whole-log, bottom: small adjustments to top).

user hoped to accomplish. We defined "too few events" as any log with a mouse event count under the central 95% interval's lower bound. To compute the central 95% interval, we used an estimated lognormal distribution after visually analyzing the data's distribution with a quantile-quantile plot ($\mu = -71.99, \sigma = 773.38$, threshold=38.5 events).

4 Task Identification with MAGI Creators

Our first analysis of the MAGI interaction log data involved a free-text labeling task with two of the developers of MAGI, where our overall approach resembles thematic analysis. The purpose of this was twofold: (1) to pilot the feasibility of labeling analysis tasks from interactions alone, and (2) to derive a shortlist of categories, which could be used as classifier labels and as multiple choice options in our planned follow-up user study.

Here, we use "task" to refer to Gotz and Zhou's interaction characterization for visual analysis tools [34], which defines tasks and sub-tasks as "high-level, logical structures of a user's analytic process, such as the user's cognitive goals and sub-goals." For convenience, and due to their similarity, we refer to both as "task" for the remainder of the manuscript as their distinction is not critical for our present contributions.

4.1 Methods

4.1.1 Participants

Two participants remotely completed the free text log labeling task through screen sharing software. Each participant was involved with the development of MAGI and was familiar with MAGI's interface and the full range of ways MAGI could be interacted with.

4.1.2 Design and Displays

Instead of predefining a set number of interaction logs for participants to label, the experimental environment created trials on-demand by randomly sampling as many interaction logs as a participant could label within 45 minutes.

In each trial, an interaction log summary visualization was rendered alongside playback controls (Fig. 2). In the visualization, each of MAGI's charts were shown as a differently colored rectangle. A heatmap was overlaid on top of the visualization rectangles, which showed regions that users commonly interacted with. Participants could also watch the mouse move (orange crosshair) and tooltips appear (red rectangles) throughout the log's duration by either clicking a $10\times$-speed play button, or by dragging one of two sliders that controlled the playback time. The top slider was used to make large changes, and the bottom slider was used to fine-tune time navigation, which was useful for longer logs. Below the sliders, we included a small timeline showing click, movement, and scroll events. Additionally, the number of genes and datasets in each MAGI query was shown above the interaction log visualization.

4.1.3 Procedure

Each participant was instructed to work with the experimenter to infer the predominant analytical task for as many interaction logs as possible within 45 minutes. For each log, the participant would brainstorm with the experimenter

about what type of task the trial's interaction log depicted. Afterwards, the experimenter would write a 1-2 sentence description of the task and verify with the participant that the description summarized the brainstormed task. If there was no recognizable task, or if the task wasn't considered useful, the log would be labeled as "junk." After entering the log description, participants continued to the next trial.

4.2 Results and Discussion

We collected 50 labels in total (25/participant). Because we were interested in identifying a shortlist of commonly performed analytical tasks we then performed two rounds of manually grouping similar labels. To accomplish this, we printed out cards for each label response that contained the written description and accompanying interaction log visualization, along with a unique ID. Then, referencing the text summary for each card, we grouped similar cards in a manner similar to hierarchal clustering. After, we performed a second round of grouping to consolidate thematically similar groups. The resultant categories were as follows:

Aberration matrix and transcript chart cross-referencing: Frequent back-and-forth analysis between the transcript chart and aberration matrix. For conciseness, we will refer to this task as "cross-referencing" unless otherwise noted.

All-encompassing or undirected browsing: Interactions with MAGI that appear undirected, that are typically diffuse, and that use many or all of MAGI's visualizations.

Co-occurrence or exclusivity analysis: Interactions that concern the aberration matrix, typically characterized by mousing over columns (co-occurrence) or exclusivity (staircases from column-exclusivity; Fig 1).

Copy-Number-Focused Analysis: Analysis characterized by heavy use of the copy-number aberration browser.

Junk: Logs that have no discernible analysis behavior (e.g., immediate page refresh after < 1 second or short, temporally distant bursts of movement).

Targeted gene, mutation, or annotation lookup: Targeted search behavior when a user has a specific piece of information they want to find (e.g., a particular patient-column in the aberration matrix).

Transcript mutation distribution analysis: If users interact with the transcript chart, they typically focus on certain distributional characteristics such as towers of mutations at a single point in the transcript ("hotspots") or at mutations that fall along coding regions.

Other: Behavior that falls outside of what was labeled in this experiment (e.g., use of the network view).

This procedure was guided by previous analyses that were part of MAGI's formative iterative design, which identified hypothesis formation and testing tasks targeted on biological significance as two of MAGI's largest use cases.

One question that arises from these results is how consistently these tasks can be inferred using only low-level interaction logging data, which is critical for reliable classification. We test this in the next study.

5 User Study: Log Task Labeling

The primary goal of this experiment was to collect labels to train, validate, and test interaction log classifiers. We also wanted to test whether humans could reliably infer analytical tasks from mouse interaction logs alone. Our prediction was that interaction-task inference would be reliable between interaction log observers. To these ends, we asked five pairs of visualization and genomics experts (1 of each/pair) to label tasks in a series of MAGI logs using the eight labels from our prior evaluation (Sec. 4).

5.1 Methods

5.1.1 Participants

10 participants (5 pairs) completed the study. Five participants were recruited through university mailing lists for graduate students and had formal knowledge of genomics. The remaining five participants were recruited from human-computer interaction research groups in our institution. Each participant had at least one year of academic or professional experience in either genomics or visualization. The median number of years each participant had spent in their degree program was 2 years (range: 0-5). Figure 4 shows participant expertise. Each was compensated $10/hour. The experimental protocol was approved by our university's IRB.

5.1.2 Design and Displays

The user study was held in pairs such that each session had one genomics expert and one visualization expert. The study was designed for pairs of participants because we believed pair coding would help control labeling variance and because the experiment required expert knowledge of visualization and genomics, which presented single-person recruitment limitations. Another motivation was that fatigue was too prohibitive in a pilot with single participants.

Each pair of participants saw 96 random-order trials, which consisted of 2 replications of a 48-trial design. One replication contained a unique set of interaction logs while the second replication contained logs that were identical between subjects to analyze inter-rater reliability (IRR). We settled on a 48-trial design after performing a power analysis for Fleiss' kappa [35] ($\kappa_0 = 0.6, \kappa_1 = 0.4, \alpha = 0.05, \beta = 0.2$ with 5 raters), which suggested including at least 41 trials.

The 48-trial design consisted of 24 randomly sampled logs and another 24 logs that were sampled based on three feature sets we had planned to use in our eventual classification evaluation (Sec. 6.1). To sample the 24 feature-based trials, we first had a MAGI expert create example ground truths for each of the eight previously defined task labels, where we knew the full context of each query (e.g., "the expert was interested in exploring a particular biological pathway"). Then, using each of the three feature sets and eight ground truths, we sampled 24 nearby neighbors.

To create the six unique sets of logs (5 pairs + 1 IRR), we generated all feature-set-based trials at the same time by picking the 6-closest logs for each of the 24 $\{\text{feature set}\} \times \{\text{label}\}$ combinations. Next, we semi-randomly shuffled the samples so that each pair of participants would be given an unordered, complete collection of the 24 combinations. For example, the first participant would be given one of

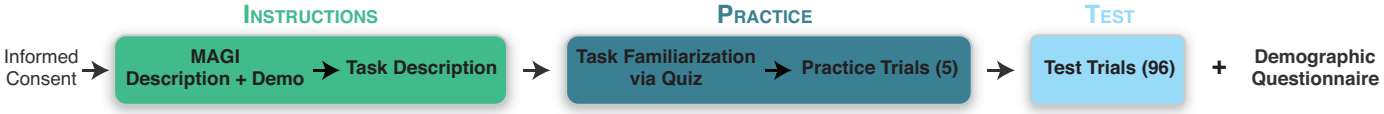

Fig. 3. The procedure for our pair-participant task labeling study.

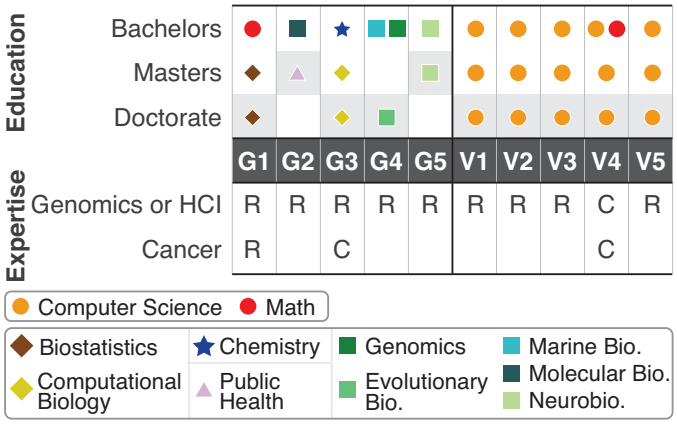

Fig. 4. User study participant demographics. Non-circle degree glyphs relate to genomics expertise. Shaded cells mark currently-pursued degrees. "G" columns refer to genomics experts, whereas "V" refers to human-computer interaction (HCI) and/or visualization experts. "R" expertise entries refer to hands-on research experience, whereas "C" refers to coursework exposure.

each 24 combinations, but these 24 logs would not always be the first-closest-neighbors. This procedure was designed to control for potential bias stemming from nearest-neighbor ordering while still including all 24 conditions.

The remaining 144 random-sample logs were then sampled without replacement from the set of remaining logs.

5.1.3 Procedure

Following informed consent, the study took place over three stages: instructions, practice, and test (Fig. 3). All participants took between 1.5 and 2 hours to complete the study.

Instructions: In the MAGI overview, each pair read through a description of each MAGI chart and watched a short video of MAGI being used by an expert. In the experimental task overview, participants were provided text descriptions for each of the MAGI task labels and were shown example stimuli.

Practice: Participants were presented a grid of 8 anonymized ground-truth logs along with task label descriptions, and were asked to discuss with their partner which label they believed should be assigned to each log. After guessing, participants could reveal the answer by clicking on a "show" button. Following the quiz, participants then completed five practice trials per the test procedure below.

Test: Each trial had a single log, and participants were asked to mark which task they thought was most characteristic. Marking "other" required an accompanying short text description. Each trial included task descriptions and examples to the right of the response area and in printed handouts. To encourage faster responses, each trial displayed a timer and a beep would play after 45 seconds; however, participants could take as long as they needed.

5.2 Results and Discussion

5.2.1 Inter-rater reliability and accuracy: similar strategies

Our planned analysis of the 48 inter-rater reliability (IRR) trials for each participant-pair using Fleiss' κ was 0.405. According to Landis and Koch, this maps onto fair-to-good reliability [36]. Fair-to-good reliability suggests that there was a moderate amount of subjectivity between pair responses, but that the individual differences across trials was low enough to be confident in the response reliability. To supplement Fleiss' κ we also measured the *modal accuracy* of each participant, which defines a correct response as any response that matches the most frequently assigned label(s) for a given interaction log. Participant accuracies, in order of study completion date, were: 69%, 73%, 73%, 65%, and 77%. Both Fleiss' κ and accuracies suggest that all participants had similar, consistent labeling strategies.

5.2.2 Task label diversity and frequency shows consistency

To understand participant-pair task labeling strategy similarity we analyzed labeling frequencies and labeling consistency across participant-pairs (Fig. 5).

To measure similarity, we calculated Shannon diversity indexes for each pair-participant using label frequencies. The diversity indexes were 1.90, 1.97, 1.86, 1.76, and 1.91. Values closer to $\ln 8 \approx 2$ refer to more uniform label frequency distributions and values closer to 0 refer to skewed distributions. Diversity indexes are calculated through Shannon entropy: $H' = -\sum_{i=1}^{L} p_i \ln p_i$. L is the number of labels and p_i is the ith label frequency's proportion of the 96 total labels for a given participant-pair. Each diversity index fell within the top 15% of the potential range of diversity ($[0, \ln(8)]$), which suggests that participants applied similarly uniform task labeling strategies. These results also support our initial task selection methodology because our synthesized task labels were used with little favoritism.

We also made several qualitative observations based on labeling frequency to drill down beyond reliability summary statistics. First, participant-pair 4's poor accuracy may stem from slightly-deviant labeling proportions: they never provided a cross-referencing task label IRR response, had only one targeted analysis response, and over half of their responses were either "junk" or undirected labels. This skew is the likely source for their comparatively lower accuracy and Shannon diversity index. Another distinction is that participant-pair 3 never provided an "other" response, though this is not necessarily abnormal given the relatively

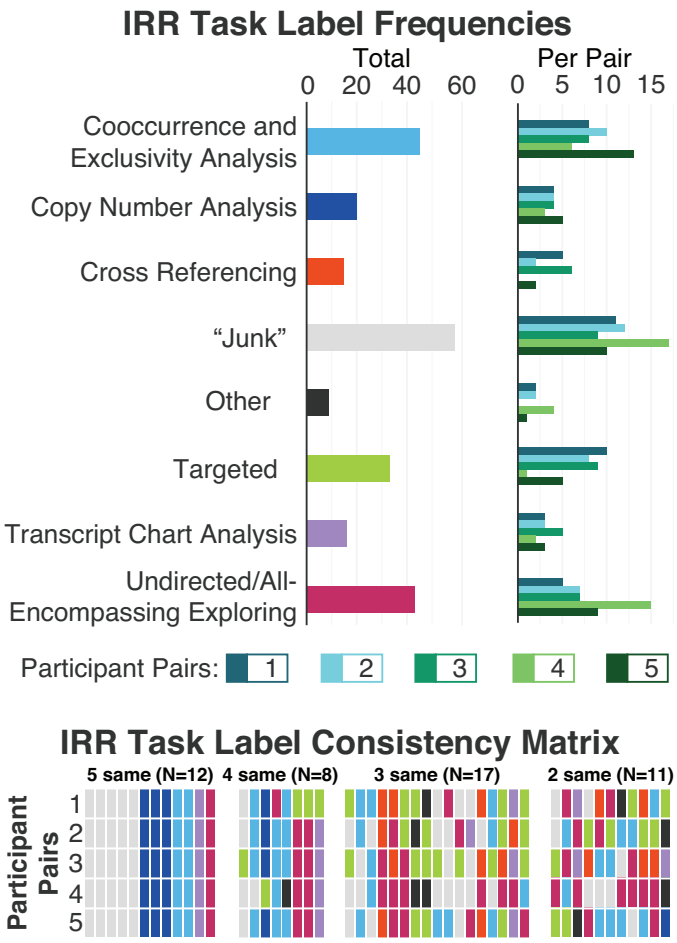

Fig. 5. Task label frequencies (top) and ordered labeling consistency between participants for each interaction log (bottom; rows: participants, columns: interaction logs).

low "other" response rates of the other pairs. Aside from these two deviations, participants' strategies were largely consistent; 20 of the 48 IRR trials had 4 or 5 identical labels out of the 5 labels given by participant-pairs, and 17 IRR trials had 3 identical labels across the participant-pair responses. There were no trials with five different labels.

We also found no significant difference between modally correct labels between random and feature-based sampling methods through a two-sided Fisher's exact test ($p = 0.57$; feature-set: $73\%, 88/120$; random: $69\%, 83/120$).

5.2.3 "Other" label descriptions

There were 23 "other" labels across the 480 total responses ($< 5\%$). The most frequent reason for selecting "other" was to report different types of cross-referencing task behavior (9), given that the provided cross-referencing task label only pertained to interactions between the aberration matrix and transcript chart. Other responses pertained to other MAGI features not covered by the eight labels (e.g., the network visualization and control panel) (11), or to simple page exploration without analytic purpose (2). Only once did participants respond that they were unable to determine what type of task a user was pursuing.

While it is possible that there are other tasks than the eight we identified, they are likely to be rare outliers. Similarly, the comparative scarcity of "other" responses suggests that our eight task categories were effective at describing typical MAGI interactions.

5.2.4 "Junk" assignment strategies

One concern we had while designing the experiment was whether participants would put potentially meaningful logs in "junk." Our intent was for junk to be a catch-all for logs that slipped past our prefiltering, which eliminated empty or near-empty logs. For example, there was one log that we would have considered to be undirected exploration due to its diffuse interactions; however, the pair of participants could not identify a behavior and marked it as junk (opposed to marking it as "other" as one other participant did). Although we saw some instances of undesirable junk labeling while proctoring the study, we found that participants were overall consistent with our junk-labeling expectations.

5.2.5 Takeaway: reliable, consistent human task inference

Overall, these quantitative and qualitative trends both point to similar conceptual understanding of how each task mapped onto mouse interactions and suggest that participant-pairs used similar labeling strategies. This is an important discovery because it shows that tool evaluators can reconstruct meaningful information about tool use from interaction logs alone. The reliability and presumed reproducibility of these findings establishes a foundation for our next evaluation. Using these results from our human-centered evaluation we can establish a baseline from which automated machine classification can be compared against.

6 Log-Task Classification

We evaluated 12 classifiers to test whether automated classification could predict visual analysis tasks with comparable accuracy to domain experts from the previous experiment. Each classifier was built from a selection of three models (k-nearest neighbors, linear support vector machines, random forests) and four feature sets, as described below. Our evaluation predictions focused on identifying a best-performing classifier to use in a follow-up exploratory analysis of the entire MAGI interaction log corpus. To test each model's effectiveness, we used the 48 IRR trials from our previous in-lab experiment and used the non-IRR trials for training and cross-validation.

Our model selection was guided by selecting models that would be accessible to typical visualization researchers and practitioners. We determined accessibility by how widely classification models were used in-practice and how readily they could be used "out of the box" with well-documented machine learning libraries (e.g., Python's `scikit-learn`). Another selection criterion was to select models that would perform well given few training data, which can be a common-place limitation in domain-expert-focused research. It is important to note that there are many potentially promising, but more complex, alternative classification methods that could also be used, which might result in more accurate predictions (Sec. 7). We opted to pursue simpler models for two reasons. First, we wanted to pursue a systematic approach to studying classifiers' given our present knowledge of interaction mining applications

ROI Transition [37]	**Dwell** [4]	**Mouse Tracking** [38]
transition count	total time	stationary H
transitioned-to count	μ dwell time	transition H
	σ dwell time	total time $\forall$ ROI
	# datasets	active time $\forall$ ROI
	# genes	dwell time $\forall$ ROI
		μ active time $\forall$ ROI
		μ dwell time $\forall$ ROI

TABLE 2
An overview of three feature sets used in our classification (not shown: "all," the combination of these sets). ROI transition count is short-hand for the complete adjacency matrix of transition features between each ROI. Transitioned-to count sums one dimension of the complete matrix. μ: mean, σ: deviation, H: entropy.

in visualization, and thought there would be too great a number of unbound decisions to use more complex classification pipelines. Second, we wanted to focus our evaluation on models that would not be too elaborate for much of our target audience to easily use.

6.1 Feature Sets

In our present classification evaluation, we consider three feature sets: dwell, region-of-interest (ROI) transition, and a novel "mouse tracking" approach. A summary of each feature set is listed in Table 2. "Region of interest" (ROI) corresponds to MAGI's five visualizations and control panel (Fig. 1).

6.1.1 Dwell

The features in dwell are: total session time; mean and standard deviation of dwell time; and the number of datasets and genes in a query. Each feature is taken from a subset of Agichtein et al.'s features for modeling web search ranking [4]. We include only a subset due to differences in application areas and in interaction log schemas (multiple-page vs. single-page sessions).

One difficulty raised by the dwell feature set was how to best quantize mouse traces into active and dwell periods. To accomplish this, we chose a dwell threshold (100ms) using the interquartile mean of all contiguous-event time differences across all interaction logs. We operationalized the threshold using the interquartile mean opposed to other methods (e.g., median split) because the distribution of time differences had a long right tail that skewed whole-range averages. A common issue causing the skewed distribution were sessions where a user would leave MAGI open for days, whereas most differences were fractions of a second.

6.1.2 ROI Transition

The ROI transition feature set is comprised of the adjacency matrix describing transition frequencies between ROIs and the total number of transitions to each ROI. The two groups of features are adapted from Brown et al.'s binary classifiers for visual search task completion time and personality factors such as locus of control [37]. Although Brown et al. tested several predictive models, we use only their state-based feature set, which had the highest predictive accuracy for task slow vs. fast completion time (83%).

6.1.3 Mouse Tracking

The mouse tracking feature set includes five types of times for each ROI and two types of entropy that measure how users transitioned between ROIs. The name "mouse tracking" alludes to its adaptation of eye tracking features.

The first three types of time included in mouse tracking are the total cumulative time spent in each ROI, the cumulative active time spent in each ROI, and the cumulative dwell time spent in each ROI. The last two times are the mean active and dwell times for each ROI. These measures are inspired from region-of-interest analysis in scan path clustering analyses [39], [40], and were calculated with the same methods as the dwell feature set.

The other two mouse tracking features describe different kinds of entropy to summarize how users interacted with MAGI at a more global scale. Within the context of MAGI, entropy can be thought of as how deterministic a user's interactions are between ROIs (i.e., targeted vs. diffuse). To calculate entropy, we consider MAGI ROIs ($\Re$), the transition frequency probabilities between each ROI (M), and the stationary distribution of each ROI (π). The stationary distribution (i.e., the limiting probability distribution) represents the probability that the mouse will be over a given ROI at any point in time [41, p. 199]. Both entropies are based on Krejtz et al. scanpath classification methods [38].

The first measurement of entropy uses Shannon entropy to calculate whether the distribution of ROI transitions is equal, where entropy values closer to 1 represent equal distributions and values closer to 0 represent focal distributions. Our use of $\log_{10}$ constrains entropy to a unit scale:

$$H_{Shannon} = -\sum_{i \in \Re} \pi_i \log \pi_i \tag{1}$$

The second measurement of entropy is similar, but also considers the transition frequency probabilities to understand whether interaction was more random (closer to 1) or more deterministic (closer to 0):

$$H_{Transition} = -\sum_{i \in \Re} \pi_i \sum_{j \in \Re} M_{ij} \log M_{ij} \tag{2}$$

6.1.4 All: Dwell + ROI Transition + Mouse Tracking

We also tested a composite "all" feature set, which combined the features from all three aforementioned sets.

6.2 Classification Evaluation Methods

Our final experimental design consisted of twelve classification models (3 classifiers $\times$ 4 feature sets), all of which were implemented in Python's `scikit-learn`. To select parameters for each model, we performed an exhaustive search for all parameter combinations using 3-fold cross validation. Parameter selections for each model are listed in Table 3. Then, to examine predictive variance, we evaluated each model fifty times using the same parameters across runs.

Classifier	Feature Set	Parameters
k-nearest	All	$k = 9$, w=distance
k-nearest	Dwell	$k = 10$, w=uniform
k-nearest	ROI Transition	$k = 5$, w=distance
k-nearest	Mouse Tracking	$k = 7$, w=uniform
Linear SVM	All	$c = 69.519$
Linear SVM	Dwell	$c =< 0.001$
Linear SVM	ROI Transition	$c = 0.001$
Linear SVM	Mouse Tracking	$c = 0.004$
Random Forest	All	estimators=75
Random Forest	Dwell	estimators=40
Random Forest	ROI Transition	estimators=40
Random Forest	Mouse Tracking	estimators=40

TABLE 3
Parameter selection for each tested classifier. w: weight

6.3 Classification Evaluation Predictions

Before conducting the comparative classifier evaluation, we made the following predictions:

- P1 Random forest models would be more accurate compared to k-nearest neighbor and linear SVM accuracies.
- P2 Mouse tracking features would be more accurate compared to dwell and ROI transitions for predicting task labels.

We predicted that random forests would be the most accurate because it was unclear whether our feature sets were linearly separable. Further, random forests provide a way to down-weight less effective features based on how their decision trees are trained, whereas k-nearest neighbors treats all features equally because it uses Euclidean distance. We predicted that mouse tracking would be the most accurate feature because it considered both time and transition, but at multiple levels of detail. In contrast, dwell considers only entire-session times and ignores regions of interest. Similarly, ROI transition focuses only on individual transitions, ignores more global descriptions of behavior, and does not consider interaction times.

6.4 Classification Evaluation Results and Discussion

6.4.1 Analysis of classifier performance

Because our test data has five "correct" labels for each interaction log (1 label/participant) we tested P1 and P2 with two types of accuracies: *match-any* and *modal* accuracy.

Match-any accuracy is calculated based on whether a classifier prediction matches any of the five labels provided by participants and is a lower-bound measure of classifier performance.

Modal accuracy is the same accuracy that was used in our previous user study: predictions are correct only if they match the most frequently assigned label(s) for each interaction log.

We used two accuracies — one loose and one strict — due to the qualitative, under-defined nature of what a "reasonably correct" prediction could be. It is important to note that the difference between the two accuracies is also meaningful: if match-any accuracy is 75% and modal accuracy is 50%, then 2/3 of the match-any-correct labels are also modally-correct responses and 1/3 are modally incorrect responses. For this reason, we planned our analyses to first examine match-any accuracy and use modal-accuracy as a mechanism to break match-any accuracy ties.

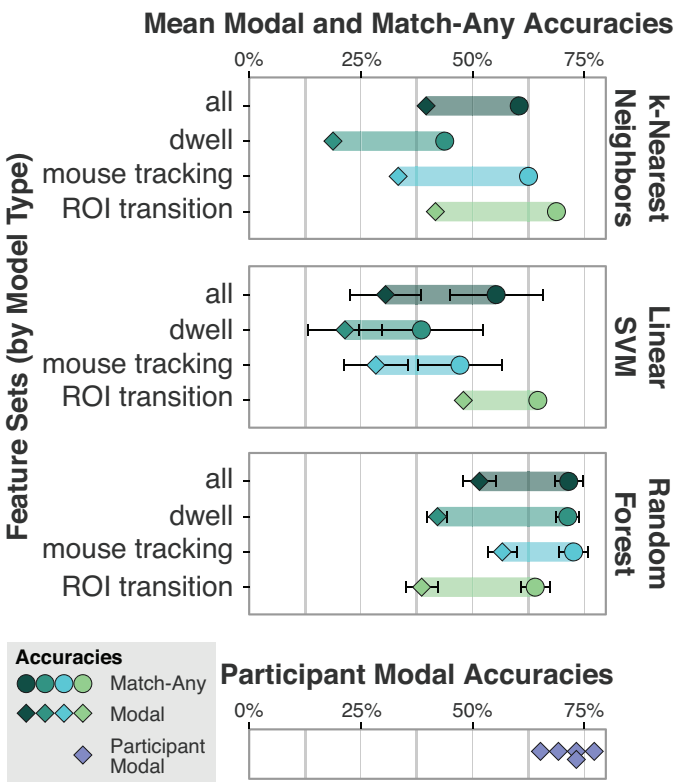

Fig. 6. Means and standard deviations of classifier accuracies after running each model 50 times. Match-any accuracy is calculated based on whether predictions matched any label assigned to an interaction log by participants. Modal accuracy is calculated based on whether predictions match the most frequently assigned label(s) for an interaction log. Higher accuracies with smaller accuracy intervals are better. k-Nearest Neighbors has no standard deviation because successive runs will always select the same k shortest Euclidean-distance points. Below the three models we also include modal accuracies for each of the five participant-pairs for easier comparison (stacked glyphs represent multiple participants with the same accuracy).

The twelve models' match-any accuracies ranged from 38% (linear SVM, dwell) to 73% (random forest, mouse tracking) and the modal accuracies ranged from 18% (k-nearest neighbors, dwell) to 56% (random forest, mouse tracking). The full-range of results are shown in Figure 6.

Previous visual analysis interaction classification has achieved similar accuracies. For example, Brown et al.'s task completion time predictive models [37] had between 62% and 83% accuracy and their personality-attribute models had between 61% to 67% accuracy when testing for traits like locus of control and neuroticism. In comparison, our models were similarly accurate, but modeled a more complex and nuanced characterization of interaction (e.g., binary vs. octenary models).

Before testing our predictions, we first analyzed the variance of model type and feature set with respect to match-any accuracy, and found a significant main effect for each (model: $F(2, 588) = 483.74, p < 0.001$; feature: $F(3, 588) = 164.39, p < 0.001$) as well as a significant interaction between the two ($F(6, 588) = 95.53, p < 0.001$).

The significant interaction between model type and feature set likely refers to the dissimilarities in accuracy for k-nearest neighbors and linear SVM models compared to random forest models. Match-any accuracy across model types was largely fixed for ROI transition features and varied for the other three such that ROI transition features were most-accurate for k-nearest neighbors and linear SVM models and were least-accurate for random forests. This suggests that dwell and mouse tracking are not linearly separable and, for similar reasons, are not well-suited for simple Euclidean-distance-based classification models. The lack of separability is supported by close-to-zero SVM margin parameter selections, which suggests that across all feature sets the data was too noisy to define a hyperplane that cleanly separated data. It would be interesting to test whether certain subsets of data are more easily separated to achieve better performance; however, such analysis falls outside the present comparative model analysis goals.

To better understand the performance differences between model types and feature sets we systematically tested our planned predictions for match-any accuracy using 2-sample Welch's t-tests. We first tested match-any accuracy by model type (P1) and found that random forests were significantly better than both k-nearest neighbor ($t(291.26) = 15.03, p < 0.001$) and linear SVM models ($t(242.83) = 17.99, p < 0.001$), and also that k-nearest neighbor models were better than linear SVM models ($t(348.73) = 6.37, p < 0.001$). Thus, random forests were best, followed by k-nearest neighbors and then by linear SVM models.

After finding that random forests were the most match-any accurate classifiers, we then tested whether mouse tracking was the most accurate feature set (P2) using only random forest predictions. Our second prediction partially held: mouse tracking was significantly more accurate than dwell ($t(93.01) = 2.17, p = 0.03$) and ROI transition ($t(97.99) = 13.09, p < 0.001$), but was not significantly different compared to all ($t(97.83) = 1.68, p = 0.1$). The non-significant difference between all and mouse tracking may suggest that "all" accuracy primarily stems from mouse tracking and has nearly no benefit from dwell and ROI transition features. Another important result was that the ROI transition feature set performed significantly worse than the three other feature sets (all: $t(97.91) = 11.67, p < 0.001$; dwell: $t(93.43) = 12.35, p = 0$; ROI transition: $t(97.99) = 13.09, p < 0.001$).

While random forest mouse tracking classifiers were significantly more match-all accurate compared to the other random forest classifiers, we also compared modal accuracies due to the small in-practice accuracy differences between all, dwell, and mouse tracking features (Fig. 6). As before, mouse tracking was significantly more modally accurate than dwell ($t(86.93) = 25.97, p < 0.001$) and ROI transition ($t(96.99) = 26.29, p < 0.001$), and was also significantly more modally accurate than "all" ($t(96.88) = 7.45, p < 0.001$). Although "all" includes mouse tracking features, mouse tracking may have performed better because the ROI transition and dwell features could have been maladaptive for predicting modal task labels.

Thus, these analyses indicate that random forest mouse tracking classification models were best.

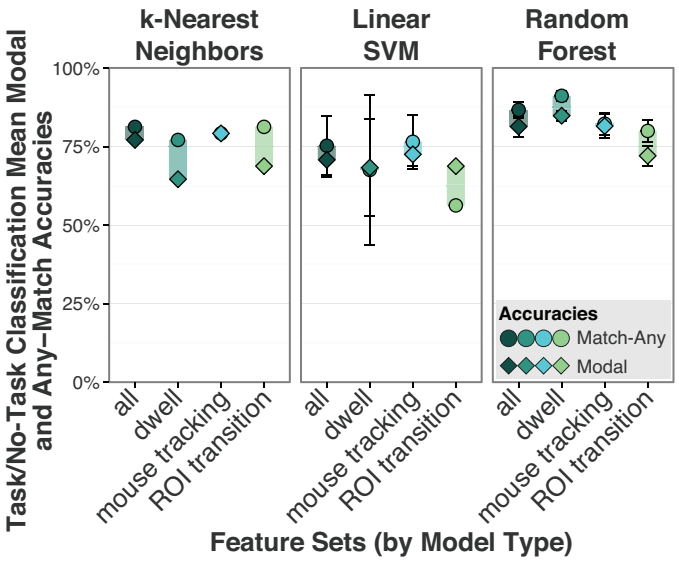

Fig. 7. Means and standard deviations of task/no-task classifier accuracies after running each model 50 times. Predictions were taken by transforming the earlier multi-class predictions into binary junk vs. non-junk categories.

6.4.2 *Binary classification: detecting visual analysis*

One remaining question after comparing classification accuracies was whether certain task labels were more difficult to predict than others. The previous analysis provided overall model accuracies compared to expert-coded "groundtruth," but did not elaborate on why model accuracies differ. Unfortunately, answering "why" is challenging with our present results because of the number of labels. Therefore, we framed our analysis of why accuracies might differ based on how easy it was for the classifiers to detect the presence vs. absence of visual analysis tasks. Rather than consider eight labels, classifiers that use this simplified task/no-task representation need only consider two. We tested task/no-task classification accuracy by retaining "junk"-label predictions as "no-task" labels and by transforming the rest to "task-present" labels. If accuracies across the 12 binary models were to be universally higher, it would signify that it is easier to distinguish whether there was salient visual analysis compared to differentiating what specific visual analysis task a user was undertaking. We predicted that

P3 Binary task-present/no-task classification would result in higher accuracies.

We based P3 on qualitative inferences that "junk"-labeled logs generally have different looking mouse trails compared to the other seven labels. For example, it is easier to differentiate an empty log from one with lengthy interaction sequences, but it may be much harder to identify whether a lengthy interaction sequence depicts undirected exploration or cross-referencing tasks.

We report both match-any and modal binary classification accuracies in Figure 7. As predicted (P3), ranges for match-any and modal accuracies were both higher (match-any: 56%–91%; modal: 65%–85%). Random forest mouse tracking classifiers had the same match-any accuracy as modal accuracy (82%). The best performing task-present/no-task classifier was random forest dwell, which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2017.2734659, IEEE Transactions on Visualization and Computer Graphics

had both the best match-any accuracy (91%) and a modal accuracy (85%).

The smaller task-present/no-task accuracy intervals between match-any and modal accuracies compared to octenary classification suggests that most of octenary modal error was due to error between non-junk labels opposed to confusion between the "junk" label vs. other labels (P3). For example, random forest mouse tracking classification had no difference between accuracies in task-present/no-task classification unlike in octenary classification. This difference in labeling confusion between task-present/no-task and octenary classification is an important distinction because it means that both binary and multi-class classifiers can be used as a method for pruning uninteresting interaction logs that lack visual analysis tasks.

The support for P3 also suggest that it is more difficult to differentiate visual analysis tasks from one another opposed to deciding whether an interaction log contains a visual analysis task. We qualitatively validated this by visually exploring predicted "junk" labels and found that most histories showed short or otherwise sparse interactions compared to more lengthy or short, but consecutive, sequences of mouse events. Most often we found that no-task "junk" logs contained interactions indicative of user error such as "quickbacks:" logs where users immediately navigated backward. In contrast, the other labels were often associated with longer-duration logs with greater numbers of events, which creates a separable boundary between the "junk" and non-junk labels.

6.5 Exploring possible classification benefits to design

In this section, we explore several possible ways that automated visual analysis task classifiers can improve the iterative design process. Our aim is to provide insight about how MAGI is used, to identify how this insight can be incorporated into iterative design, and to enumerate testable hypotheses about cancer genomics visualization interaction, which can be used to inform future design studies. Our discussion is based on exploratory analysis after using random forest mouse tracking classification to predict analysis tasks for the remaining 1,267 logs that were not part of our prior in-lab study (Sec. 5). While interpretation of these results is limited by a lack of ground-truth, our previous analyses show that task/no-task separation, and therefore comparison, is reliable. Additionally, we can be sufficiently confident in comparisons where there are large label-count differences given classification error rates.

Prediction results are shown in Figure 8. Junk labels were the most common (326) followed by cooccurrence and exclusivity analysis (287), undirected or all-encompassing exploration (253), and targeted analysis (226). The other tasks were assigned smaller label amounts: copy number analysis (2), other (12), cross referencing (45), and transcript chart analysis (113).

6.5.1 Understanding behavior via interaction frequency

Figure 8 shows that the aberration matrix was interacted with most frequently compared to the other visualizations. This information provides several testable hypotheses about user behavior that can be used to inform future iterative

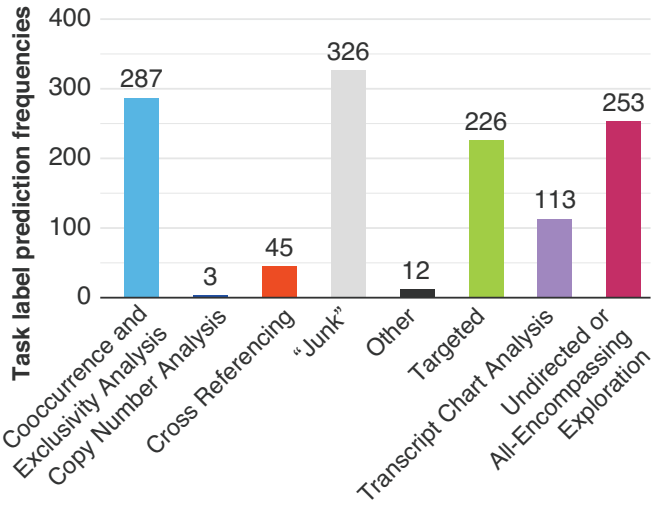

Fig. 8. Distribution of predicted task labels for the 1,267 logs that were not included in our in-lab labeling study using a random forest classifier and "mouse tracking" feature set.

design decisions. One possibility is that most researchers use MAGI to test co-occurrence and exclusivity predictions and therefore use the aberration matrix more than the other features of MAGI. Another possibility is that the aberration matrix is used most frequently because its spatial positioning at the top of MAGI causes an availability or similar spatial cognitive bias since it is the first chart users see on the query. Or, it could be that the aberration matrix is used most often because of a combination of the two other possibilities. These classification-based hypotheses lend themselves naturally to established iterative design evaluation methodologies such as A/B testing, which could help MAGI designers understand whether the spatial positioning of the aberration matrix is a large factor for its frequent use.

This location proximity effect might also be supported by the comparatively low interaction frequencies associated with the copy number analysis task, as the copy number aberration browser is located at the bottom of the page. Hence, just as the aberration matrix was favored because it is first, the copy number aberration browser might not be utilized because it is last. Another possible explanation for its infrequent use is that the copy number analysis task may be partially subsumed by other tasks given that the mouse would need to move over other visualizations en route to the browser. However, anecdotally, we did not find predominantly copy number focused interactions in other labels.

6.5.2 Which exploration strategy is more common: Top-down or bottom-up?

Two common visualization design heuristics are to support either top-down or bottom-up exploration. Top-down strategies refers to Ben Schneiderman's popular tool design mantra: "overview first, zoom and filter, then details-on-demand" [42]. In contrast, bottom-up strategies refer to diving into details first: "search, show context, [then] expand on demand" [43]. This is a critical point for tool design because supporting detail-oriented, bottom-up exploration can often be at odds with supporting top-down exploration. As such,

typical visualization design patterns maintain that it is best to focus on dominant tool-use patterns (e.g., Ziemkiewicz et al.'s evaluation of immunobiology visualization [44]).

Hence, the predictive classification results highlight an open research problem: researchers use MAGI for both top-down analysis tasks (e.g., "undirected exploration") in similar proportion to bottom-up strategies (e.g., "targeted search"). As such, both exploration procedures should be supported in future design iterations. This raises an important design question given Ziemkiewicz et al.'s finding that visual analysis tools that seek to support all analysis behavior may lead to substandard designs [44]. What, then, is the best strategy for supporting tasks that are equally common without creating two separate tools?

To address this open research question in the design of MAGI, we implemented and deployed a new resizable and repositionable layout so that researchers can alter MAGI's components to better match their individual requirements.

6.5.3 Can classification counter incorrect generalization?

The predictive classification results were in many ways a surprise to us given past observations of MAGI, which led us to expect that cross-referencing was a common and important task requirement; however, our modeling suggests that this might not be true. The surprise that our prior observations did not generalize to the larger collection of interaction logs is an example of how bias can affect experimental analysis, which we also discuss in Section 6.5.2 with respect to overfitting search task support. In particular, our revelation about cross-referencing task frequency highlights how human tendency to use a representativeness heuristic when generalizing information [45] can be maladaptive in design evaluation. The difficulty lies in the fact that most design studies typically rely on field studies with small populations [3]. Because humans tend to generalize through representativeness, which does not take sample size into consideration, it means that evaluators are likely to overfit task requirements if they do not take extraordinary care. Based on our present findings we hypothesize that interaction log classification can benefit iterative design and task requirement analysis by helping counter sample biases in tool evaluation by showing the distribution of tasks for larger sample sizes than what it typically attainable through in-person observation. By showing an alternative hypothesis to tool evaluators, it is possible that evaluators could not just avoid jumping to incorrect conclusions about tool use, but they might also make new discoveries about qualitative in-person observations.

Our top-down vs. bottom-up predictive findings also highlight another application of how interaction mining may help offset bias. Given that it is so common to only consider one strategy in tool design, evaluators of smaller in-lab observations may dismiss both requirements' similar importances (or not see them at all due to sample bias).

For these reasons, we suggest supplementing in-lab observation with interaction log analysis of how a tool is used remotely by a larger sample of users. By using both methodologies, designers can make detailed predictions with in-lab observations and better identify potential sources of bias by consulting the interaction logs of larger tool-use samples. Pursuing this mixed-method evaluation design would preserve the realism of field observations while also affording designers greater generalizability confidence.

Thus, classification can critically serve as a tool to test in-lab ecological validity, and with the right data can paint a comprehensive picture of the types of tasks a tool is most used for. This insight into tool use will also likely benefit from future development and adoption of more advanced approaches, which we discuss in Section 7. However, knowledge acquisition from such automated approaches is inherently limited by the lack of context of interaction logs. Environmental factors, true ground truth, and the cognitive state cannot be known, only inferred. In contrast, these shortcomings are what talk-out-loud qualitative methods excel at collecting. Hence, even with the advent of more powerful task data mining techniques, we believe qualitative evaluation will remain an equally valuable, rather than replaceable, aspect of tool design. The strength of classification should therefore not be tested by whether such approaches can serve as a replacement to qualitative inquiry, but rather how they can supplement it.

7 Broader Impact and Potential Open Questions

7.1 Generalizability of contributions

Although our contributions use MAGI as a case study, they also demonstrate how interaction log analysis can serve as a viable evaluation methodology for the broader visualization research community independent of application area. These contributions also show that mouse interaction analysis generalizes from the more deterministic text-focused applications in Section 2 to more open-ended visual analysis environments that incorporate not just text, but also interactive visualizations.

7.2 More accurate modeling may result in different types of generalizability and implications for design

While the classification error is low enough in our evaluation to infer user behavior and possible design implications, the development of more accurate classification could lead to more precise predictions and discussion about the relation between tasks and effective design. One potential way to achieve higher accuracy is to include an explicit feature selection step in future task classification pipelines. An alternative potential approach to increase accuracy is to model tasks as mixtures. For example, mixture models would break away from modeling only the most dominant session interactions, and could provide more robust understanding of likely-heterogeneous tasks such as "all-encompassing exploration."

7.3 Can unsupervised learning achieve comparable accuracy?

Our present evaluation only considers supervised learning approaches, which leaves the potential effectiveness of unsupervised approaches an open problem. This open problem can be tested in the future by evaluating whether clustering based on geometric-temporal distances of interaction segments [23] can accurately predict visual analysis

tasks. However, one barrier to this approach, which must also be examined, is how to best segment interaction logs into discrete components that accurately represent stages of visual analysis. While it is possible that segmentation could be skipped, it is unlikely that clustering would produce accurate results without it because of the large geometric-temporal variability of entire minutes-long mouse movement between users. One benefit to clustering, as opposed to classification, is that the phylogenies produced by hierarchical approaches could be used to test the quality of existing theoretical interaction taxonomies that are either based on literature surveys or qualitative observation.

8 Conclusion

Our findings illustrate the potential utility of mouse interaction log analysis as a new method for analyzing typically hard-to-access domain expert populations.

Using 1,553 interaction logs of MAGI, an online cancer genomics visualization tool, we first showed through in-lab evaluation that low-level interaction data alone is sufficient for reliable task inference. We then discussed how accessible classification methods matched our in-lab study inferences with up to 73% accuracy and could separate interaction logs with visual analysis tasks from those without with up to 91% accuracy. Unlike previous interaction log analysis research, our investigation considered whether interactions could be inferred by humans and machines from mouse event data opposed to higher level representations of interaction that explicitly contain richer semantic information.

We conclude that domain expert tool evaluation can be improved by combining contextually-rich qualitative observation with larger-scale interaction log analysis. By leveraging a mixed-methods approach, tool designers can retain a deep understanding of the environment that their tool is used in and the analytical goals their tool is used to achieve; they can then test specific task-based predictions based on qualitative observation by analyzing interaction logs of larger population samples to assess the ecological validity of their in-lab findings.

Acknowledgments

The authors would like to thank Mark D.M. Leiserson and Benjamin J. Raphael for their helpful discussion and support in the early stages of this work. This material is based upon work supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-1058262.

References

[1] M. D. M. Leiserson, C. C. Gramazio, J. Hu, H.-T. Wu, D. H. Laidlaw, and B. J. Raphael, "Magi: visualization and collaborative annotation of genomic aberrations," *Nature Methods*, vol. 12, no. 6, pp. 483–484, 06 2015. [Online]. Available: http://dx.doi.org/10.1038/nmeth.3412

[2] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *Trans. Vis. Comput. Graphics*, vol. 18, no. 9, pp. 1520–1536, Sept 2012. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2011.279

[3] S. Carpendale, *Evaluating Information Visualizations*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 19–45. [Online]. Available: https://dx.doi.org/10.1007/978-3-540-70956-5

[4] E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by incorporating user behavior information," in *Proc. of Research and development in information retrieval (SIGIR)*, 2006, pp. 19–26. [Online]. Available: http://doi.acm.org/10.1145/1148170.1148177

[5] Q. Guo and E. Agichtein, "Ready to buy or just browsing?: Detecting web searcher goals from interaction data," in *Proc. of Research and development in information retrieval (SIGIR)*, 2010, pp. 130–137. [Online]. Available: http://doi.acm.org/10.1145/1835449.1835473

[6] S. R. Gomez, H. Guo, C. Ziemkiewicz, and D. H. Laidlaw, "An insight- and task-based methodology for evaluating spatiotemporal visual analytics," in *Proc. of Visual analytics, science, and technology (VAST)*, Oct 2014, pp. 63–72. [Online]. Available: http://dx.doi.org/10.1109/VAST.2014.7042482

[7] M. Brehmer, S. Ingram, J. Stray, and T. Munzner, "Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists," *Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2271–2280, Dec 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346431

[8] P. Isenberg, A. Tang, and S. Carpendale, "An exploratory study of visual information analysis," in *Proc. of Human Factors in Computing Systems (CHI)*, 2008, pp. 1217–1226. [Online]. Available: http://doi.acm.org/10.1145/1357054.1357245

[9] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw, "A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights," *Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 51–60, Jan 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2467613

[10] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 31–40, Jan 2016.

[11] E. Zgraggen, S. M. Drucker, D. Fisher, and R. DeLine, "(s|qu)eries: Visual regular expressions for querying and exploring event sequences," in *Proc. of Human Factors in Computing Systems (CHI)*, 2015, pp. 2683–2692. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702262

[12] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl, "VA^2: A visual analytics approach for evaluating visual analytics applications," *Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 61–70, Jan 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2467871

[13] J. A. Fails, A. Karlson, L. Shahamat, and B. Shneiderman, "A visual interface for multivariate temporal data: Finding patterns of events across multiple histories," in *Proc. of Visual analytics, science, and technology (VAST)*, Oct 2006, pp. 167–174. [Online]. Available: http://dx.doi.org/10.1109/VAST.2006.261421

[14] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2659–2668, Dec 2012. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2012.225

[15] K. Wongsuphasawat and J. Lin, "Using visualizations to monitor changes and harvest insights from a global-scale logging infrastructure at twitter," in *Proc. of Visual analytics, science, and technology (VAST)*, Oct 2014, pp. 113–122. [Online]. Available: http://dx.doi.org/10.1109/VAST.2014.7042487

[16] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse, "The lumière project: Bayesian user modeling for inferring the goals and needs of software users," in *Proc. of Uncertainty in Artificial Intelligence*, 1998, pp. 256–265. [Online]. Available: http://dl.acm.org/citation.cfm?id=2074094.2074124

[17] P. H. Nguyen, K. Xu, A. Wheat, B. W. Wong, S. Attfield, and B. Fields, "Sensepath: Understanding the sensemaking process through analytic provenance," *Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 41–50, Jan 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2467611

[18] R. Etemadpour, M. Bomhoff, E. Lyons, P. Murray, and A. Forbes, "Designing and evaluating scientific workflows for big data interactions," in *2015 Big Data Visual Analytics (BDVA)*, Sept 2015.

[19] C.-M. Huang, S. Andrist, A. Saupp, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers in Psychology*, vol. 6, p. 1049, 2015. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01049

[20] J. Huang, R. W. White, and S. Dumais, "No clicks, no problem: Using cursor movements to understand and improve search," in *Proc. of Human Factors in Computing Systems (CHI)*, 2011, pp. 1225–1234. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979125

[21] K. Rodden and X. Fu, "Exploring how mouse movements relate to eye movements on web search results pages," *SIGIR Workshop on Web Information Seeking and Interaction*, pp. 29–32.

[22] S. Gomez, R. Jianu, R. Cabeen, H. Guo, and D. Laidlaw, "Fauxvea: Crowdsourcing gaze location estimates for visualization analysis tasks," *Trans. Vis. Comput. Graphics*, vol. PP, no. 99, pp. 1–1, 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2016.2532331

[23] D. Martín-Albo, L. A. Leiva, J. Huang, and R. Plamondon, "Strokes of insight: User intent detection and kinematic compression of mouse cursor trails," *Information Processing & Management*, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2016.04.005

[24] A. Edmonds, R. W. White, D. Morris, and S. M. Drucker, "Instrumenting the dynamic web," *Journal of Web Engineering*, vol. 6, no. 3, p. 243, 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=2011250.2011254

[25] J. Matejka, T. Grossman, and G. Fitzmaurice, "Patina: Dynamic heatmaps for visualizing application usage," in *Proc. of Human Factors in Computing Systems (CHI)*, 2013, pp. 3227–3236. [Online]. Available: http://doi.acm.org/10.1145/2470654.2466442

[26] R. Atterer, M. Wnuk, and A. Schmidt, "Knowing the user's every move: User activity tracking for website usability evaluation and implicit interaction," in *Proc. of World Wide Web (WWW)*, 2006, pp. 203–212. [Online]. Available: http://doi.acm.org/10.1145/1135777.1135811

[27] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *Trans. Vis. Comput. Graphics*, vol. 11, no. 4, pp. 443–456, July 2005. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2005.53

[28] P. Saraiya, C. North, V. Lam, and K. A. Duca, "An insight-based longitudinal study of visual analytics," *Trans. Vis. Comput. Graphics*, vol. 12, no. 6, pp. 1511–1522, Nov 2006. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2006.85

[29] T. O'Brien, A. Ritz, B. Raphael, and D. Laidlaw, "Gremlin: An interactive visualization model for analyzing genomic rearrangements," *Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 918–926, Nov 2010. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2010.163

[30] M. Streit, H. J. Schulz, A. Lex, D. Schmalstieg, and H. Schumann, "Model-driven design for the visual analysis of heterogeneous data," *Trans. Vis. Comput. Graphics*, vol. 18, no. 6, pp. 998–1010, June 2012. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2011.108

[31] P. Murray, F. McGee, and A. Forbes, "A taxonomy of visualization tasks for the analysis of biological pathway data," in *Proceedings of BioVis*, 2016.

[32] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2376–2385, Dec 2013. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2013.124

[33] J. Sjölund, C. Manetopoulos, M.-T. Stockhausen, and H. Axelson, "The notch pathway in cancer: Differentiation gone awry," *European Journal of Cancer*, vol. 41, no. 17, pp. 2620 – 2629, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959804905007276

[34] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," in *Proc. of Visual analytics, science, and technology (VAST)*, Oct 2008, pp. 123–130. [Online]. Available: http://dx.doi.org/10.1109/VAST.2008.4677365

[35] M. A. Rotondi, *kappaSize: Sample Size Estimation Functions for Studies of Interobserver Agreement*, 2013. [Online]. Available: http://CRAN.R-project.org/package=kappaSize

[36] G. G. K. J. Richard Landis, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[37] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang, "Finding waldo: Learning about users from their interactions," *Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 1663–1672, Dec 2014. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2014.2346575

[38] K. Krejtz, T. Szmidt, A. T. Duchowski, and I. Krejtz, "Entropy-based statistical analysis of eye movement transitions," in *Proc. of Eye Tracking Research and Applications*, 2014, pp. 159–166. [Online]. Available: http://doi.acm.org/10.1145/2578153.2578176

[39] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof, "A comparison of scanpath comparison methods," *Behavior Research Methods*, vol. 47, no. 4, pp. 1377–1392, 2015. [Online]. Available: http://dx.doi.org/10.3758/s13428-014-0550-3

[40] J. H. Goldberg and J. I. Helfman, "Scanpath clustering and aggregation," in *Proc. of Eye-Tracking Research & Applications*, 2010, pp. 227–234. [Online]. Available: http://doi.acm.org/10.1145/1743666.1743721

[41] H. M. Taylor and S. Karlin, *An introduction to stochastic modeling*. Academic press, 2014.

[42] B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations," in *Proc. of Symposium on Visual Languages*, Sep 1996, pp. 336–343. [Online]. Available: http://dx.doi.org/10.1109/VL.1996.545307

[43] F. van Ham and A. Perer, ""search, show context, expand on demand": Supporting large graph exploration with degree-of-interest," *Trans. Vis. Comput. Graphics*, vol. 15, no. 6, pp. 953–960, Nov 2009. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2009.108

[44] C. Ziemkiewicz, S. Gomez, and D. Laidlaw, "Analysis within and between graphs: Observed user strategies in immunobiology visualization," in *Proc. of Human Factors in Computing Systems (CHI)*, 2012, pp. 1655–1658. [Online]. Available: http://doi.acm.org/10.1145/2207676.2208291

[45] D. Kahneman and A. Tversky, "Subjective probability: A judgment of representativeness," *Cognitive psychology*, vol. 3, no. 3, pp. 430–454, 1972. [Online]. Available: http://dx.doi.org/10.1016/0010-0285(72)90016-3