Analyzing the Shifts in Users Data Focus in Exploratory Visual Analysis

ANONYMOUS AUTHOR(S)*

Exploratory visual analysis (EVA) is an essential stage of the data science pipeline, where users often lack clear analysis goals at the start and iteratively refine them as they learn more about their data. Accurate models of users' exploration behavior are becoming increasingly vital to developing responsive and personalized tools for EVA. Yet we observe a discrepancy between the *static* view of human exploration behavior adopted by many computational models versus the *dynamic* nature of EVA. This paper explores potential parallels between the evolution of users' interactions with visualization tools during data exploration and assumptions made in popular online learning techniques. Through a series of empirical analyses using existing experiment data, we seek to answer the question: What are the best learning methods for modeling shifts in users' data focus during EVA? We present our findings and discuss their implications for the future of user modeling for visualization system design.

CCS Concepts: • **Human-centered computing** → *User models.*

Additional Key Words and Phrases: Human-centered Computing, Visual Analytics, User Modeling, Reinforcement Learning

ACM Reference Format:
Anonymous Author(s). 2018. Analyzing the Shifts in Users Data Focus in Exploratory Visual Analysis. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Data analysts interactively query **visual exploration systems (VESs)** such as Tableau [2] and PowerBI [1] to explore large datasets and discover insights [9]. This iterative and complex process is known as **exploratory visual analysis (EVA)** [34]. EVA is particularly challenging as analysts often explore new datasets with unknown structures and content. Complexity increases as analysts may start EVA with vague analysis goals [8], like finding interesting insights [34, 62].

Initially, a user may not know what information is interesting or where to find them. In each interaction, the user focuses on a specific data area to find insights or learn something that might help in future interactions. During EVA, the user shifts to different parts of the data to generate hypotheses/goals and devise potential exploration paths to discover insights. In this work, we use the term ***data focus*** *to refer to the specific data area the user is focusing on at any given time.* However, the scope of data focus can vary. For example, in tabular data, data focus can be a specific data point (in row x, column y) or an entire data column y. In this work, *data focus is defined based on the exploration task, dataset, and VES.* For instance, if a VES recommends visualizations based on data columns (attributes) chosen by the user, it is more effective to analyze the user's data focus in terms of attributes. Likewise if the VES emphasizes specific row filters or data points [29], then data focus should be modeled accordingly. Consider the following scenario *where a user's data focus (attributes) changes as EVA progresses.*

Example 1. *Alice is exploring the Birdstrikes dataset [58] containing records of wildlife strike incidents with aircraft. Her task is to find patterns to improve aviation safety and airline revenue (Figure 1). However, Alice is unfamiliar with this*

Manuscript submitted to ACM

dataset. So she starts by randomly examining visualizations to learn more. After some interactions, she notices that the `number_of_incidents` *with wildlife has been decreasing since the year 2000. Motivated to find other trends using attribute* `flight_date`*, she explores visualization showing* `repair_costs` *over the years. Without finding conclusive trends, she decides to shift to other data columns. She observes a pattern involving* `wildlife_size`*: small birds cause more collisions. Consequently, Alice explores visualizations involving* `wildlife_size` *and discovers that repair costs are significantly higher for collisions with large birds due to the substantial damage they cause. Noting this insight, she continues her interaction to complete the task.*

Alice's strategy for shifting her data focus changes based on what she learns from the data, her information needs, etc. Initially, Alice explores the dataset randomly. However, over time, her shifts become more selective, especially when she identifies an interesting attribute. Even after an unsuccessful exploration of `flight_date`, her data focus shifts are less random because she now understands the dataset better. Eventually, she decides to shift data focus to `wildlife_size`, an attribute she encountered earlier.

By recommending useful visualizations, we can improve EVA experience for users like Alice. For that we need to understand users exploration strategies and data interests [10, 60]. Visualization researchers have studied modeling users' data focus shifts to predict their future actions and data point interests [7, 41, 44]. Besides recommending visualizations, these models can be used to prefetch data areas to speed up interactivity [7] and suggest relevant data regions for insights [24]. These approaches learn data focus shifts from offline interactions using machine learning [7], rule-based pattern matching [24], or online learning algorithms that update parameters after each interaction [41, 44].

However, as users learn more about the data, their strategies for shifting data focus may evolve to complete EVA tasks more efficiently and effectively [8, 34]. With that in mind, knowing Alice will interact with `wildlife_size` in the next step may not be enough. If Alice is following a random exploration strategy, VES should recommend a diverse set of visualizations, providing an overview of the dataset. But in later stages of interaction, when Alice is more selective, it might be better to recommend visualizations showing `wildlife_size`'s relationship with other attributes. Therefore, analyzing and modeling the dynamic shifts in users' data focus complements current models for predicting future actions or data points [29, 41, 44]. Additionally, there is a lack of empirical analysis on how well current algorithms adapt to evolving data focus shifting strategies.

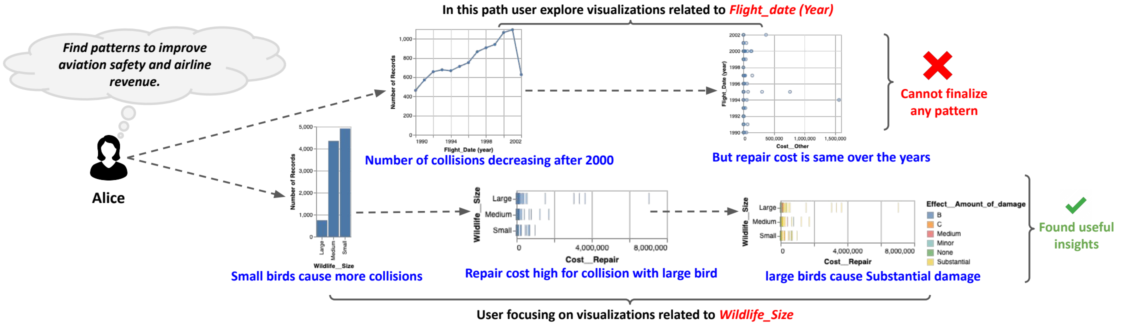

In this path user explore visualizations related to Flight_date (Year)
Find patterns to improve aviation safety and airline revenue.
Alice
Number of Records
1,000
800
600
400
200
0
1990
1994
1998
2002
Flight_Date (year)
Number of collisions decreasing after 2000
Flight_Date (year)
2002
2000
1998
1996
1994
1992
1990
0
500,000
1,000,000
1,500,000
Cost__Other
But repair cost is same over the years
Cannot finalize any pattern
5,000
4,000
3,000
2,000
1,000
0
Large
Medium
Small
Wildlife__Size
Small birds cause more collisions
Wildlife__Size
Large
Medium
Small
0
4,000,000
8,000,000
Cost__Repair
Repair cost high for collision with large bird
Effect__Amount_of_damage
B
C
Medium
Minor
None
Substantial
large birds cause Substantial damage
Found useful insights
User focusing on visualizations related to Wildlife_Size

Fig. 1. A user exploring Birdstrikes dataset [58].

Moreover, existing approaches often overlook many online learning algorithms, such as reinforcement learning (RL). Many online RL algorithms have shown promising results in modeling decision-making strategies in cognitive psychology, neuroscience, etc. [14, 42, 43] In EVA, users may change their data focus strategies in response to what they learn from the data. They may choose suboptimal strategies, like exploring areas that do not contain any insight. However, they help to understand the dataset better and maximize long-term rewards. *Such similarities with online RL algorithms raise an important question: are RL algorithms better suited for modeling these dynamic shifts in data focus?*

The main goal of our paper is to present a benchmark study using a variety of offline and online learning algorithms, including RL which answers: What are the best methods for modeling shifts in users' data focus during EVA? We mirror Feng et al.'s [21], Gathani et al.'s [22], and He et al.'s [29] approach of leveraging prior user studies to address our research questions. Rather than designing a single study covering a single tool, our analysis spans three influential studies [7, 8, 34] from the visualization community that applies distinct approaches to capture exploration behavior. We perform statistical tests to determine *whether users' strategies for shifting data focus evolve* in our selected datasets. This serves as a sanity check for our selection of EVA studies for comprehensive benchmark analysis.

Our analysis is breadth rather than depth-focused, and exploratory in nature. Through this work, we make the following contributions:

- To evaluate how well different algorithms model users' dynamic shifts in data focus, we analyze their ability to predict users' future data focus shifting strategies.
- Our benchmark study includes a comprehensive list of algorithms and user modeling techniques from existing visualization papers, such as Hidden Markov Models and Bayesian Learning [29, 41, 44]. Additionally, we incorporate simple heuristics as baselines that users may adopt to shift their data focus. Our analysis shows *RL algorithms significantly outperform existing methods*, suggesting that current EVA models can be improved .
- We present common findings and challenges from our empirical investigations and suggest future research directions for developing more accurate adaptive data focus models for VESs.

2 Related Work

Understanding the user during Exploratory Visual Analysis (EVA):. We see analyses of users' acquisition of knowledge, specifically in the form of insights [8, 27, 30, 34]. However, these approaches focus on proposing VESs that increase accuracy insights or how fast users can generate insights using their tools [8] rather than analyzing the evolution in users' exploration behavior. Some researchers have proposed visual analysis taxonomies that map users' activities in EVA with users' high-level reasoning process [8, 22]. Several works present users' cognitive frameworks or mental models for analyzing datasets through VESs [26, 36, 45]. However, they do not address how users mental models may evolve as they learn about the data.

More recent studies acknowledge the dynamic nature of users' exploration behavior. For instance, Ottley et al. [44] proposed a hidden Markov model approach to maintain a belief over users' evolving *hidden* attention and actions in a *point-based* visualization setting. Utilizing user clicks, they update the model and use particle filtering to infer a set of *top-k* data points for the next click. Monadjemi et al. [41] proposed a Bayesian learning approach to also predict *top-k* data points. Instead of predicting exact data points, we use user interactions to learn how they shift their data focus and predict the next shifting strategy. Users adjust their data focus based on what they learn to find insights effectively and efficiently. Similarly, our RL models attempt a shift in data focus, receive feedback, and adapt their parameters to maximize insights. Typically, visualization papers propose a system focusing on unique EVA characteristics particular exploration task scenarios and user models tailored for such instances [61]. In this benchmark

Characteristics	Tableau user study[8]		imMens user study[34]	Voyager user study[61]	
	Task [T1, T2, T3]	Task [T4]		Task [T1, T2]	Task [T3, T4]
Open-endedness	Focused	Open-ended	Open-ended	Focused	Open-ended
Task Complexity	Analyzing visualizations generated by Tableau based on users' query on dataset		Interactive querying using imMens actions on summarized plots	Specifying data attributes and interacting with visualization recommendations	
Prior experience	15 minutes with a different dataset		15 minutes with datasets	10 minutes demo with a different dataset	

Table 1. Selected user studies with different EVA task characteristics

study, we adopt breath-focused analysis similar to [29] with the goal of analyzing and modeling data focus shifts in various EVA scenarios.

Researchers model users' low-level interactions into high-level exploration phases or analysis patterns to improve how systems support EVA [7, 24, 25]. These models trained on offline user interactions aim to infer users' exploration goals, prefetch corresponding data regions [7], or recommend tailored visualizations [25]. We argue that such static models may not cater to the users' evolving information needs and produce suboptimal recommendations.

Online Learning Algorithms to Model User's Evolving Exploration Strategies: Recent research using real-world query workloads indicates that users learn and modify their keyword queries to express specific and focused intents while interacting with a system [38]. This evolving behavior can be modeled using online learning algorithms. Cen et al. modeled users' evolving information-search strategies from scholarly databases using Reinforcement Learning (RL) [15]. While Luo et al. have modeled users' exploration-exploitation policies in formulating keyword queries for document retrieval [37]. Unlike data querying, users often lack a predefined and concrete intent during EVA. Consequently, EVA presents a significantly larger action space requiring users to make more complex decisions. We posit that the evolutionary nature of EVA provides a more natural setting for the users to learn and shift their data focus.

Use of Reinforcement Learning (RL) and Statistical Metrics in EVA:. Some VESs statistically identify and suggest the underlying data patterns to users. These systems use statistical measures like *diversity, interestingness, and coherency* [4, 5, 18] in their models to determine *how interesting a visualization is from users' perspectives.* However, when utilizing them, they assume these factors have the same contribution, wherein the demand for visualizations based on these elements evolves. Besides, prior studies underscore the need for a framework that utilizes user interaction history for personalized recommendations [20, 47]. Additionally, we see RL models trained by expert user demonstrations to aid future analysts in similar exploration tasks by automatically generating exploration sessions and relevant recommendations [5, 49]. However, relying solely on static statistical metrics, e.g., KL-divergence, and features from expert demonstrations, to identify and suggest visualizations can be problematic. These approaches may not accurately capture/adapt to users' dynamic shifts in their data focus strategies or information needs to gain more insights. DashBot [18] and ATENA [5] have popularized the use of Deep-RL to aid users in finding interesting aspects of the dataset. However, these approaches do not directly incorporate users' preferences or learning as reinforcement to guide their future decisions.

Exploration as a Markov Decision Process (MDP): Because of the popularity of reinforcement learning algorithms, it has become common to model problems using the Markov Decision Process (MDP) framework. MDP-based RL algorithms have been used to explain the workings of the human brain [23] and decision-making processes [16, 42, 43]. MDP has also become popular in modeling data exploration [37] and visual analytics problems [5, 18, 39, 50]. In these works, the RL algorithms try to find the optimal policy that replicates users' analytical processes and use that policy for visualization recommendations.

3 Considerations for EVA-Focused User Studies

Users' shifts in data focus depend significantly on the characteristics of EVA tasks and VESs. In subsection 3.1, we introduce some characteristics that may influence the shifts in users' data focus.

To analyze and model the shifts in users' data focus, it is not optimal to design a single-user study that covers all possible EVA variations. Therefore, we utilize well-known studies by Liu et al. on the **imMens** system [34], Battle et al. on **Tableau** [8], and Zeng et al. on a simulation of **Voyager** [61]. Our benchmark study encompasses a wide range of EVA tasks and characteristics recognized by the visualization community, shown in Table 1.

3.1 Characteristics of EVA User Study Tasks

Task Open-endedness: Researchers categorize exploration tasks into two groups based on how clear users' objectives are [8, 41, 61]. **Open-ended tasks** are exploration tasks without a clear intent or hypothesis. For example, in the imMens user study (Table 1), participants are asked to find *interesting* information from the given datasets. Since "interesting information" is a vague term, users are uncertain about what and where to search. As a result, they may explore different data areas to understand the data and report insights that seem interesting to them. In **focused tasks**, analysts have a precisely defined goal and exploration path that often contributes to a broader objective. For instance, in the Voyager user study task T1 (Table 7), users are asked to find the maximum number of movies for the genre "creative" and the source "book/small story.". Users know what information to retrieve and which data area to explore as the question includes corresponding column names. Nevertheless, with the large data size, users may need additional interactions to find the exact information.

Task Complexity: How users operate a VES to generate insights and the amount of information displayed in the interface increases task complexity, thereby impacting users' exploration strategies in EVA [3, 28, 32]. Voyager aims to recommend the best possible visualizations, whereas Tableau [2] is a more complex system. Tableau has a large action space, allowing users to select, filter, and visualize data in a highly customizable manner. In contrast, imMens summarizes the data into pre-defined four or five visualizations, requiring users to perform only four actions to query these visualizations and identify patterns.

Prior Experience: Users' prior knowledge about the dataset and familiarity with the exploration interface may influence their shifts in data focus and thereby their exploration strategy. While users have a limited time to get familiar with the dataset in imMens, Voyager and Tableau analysis tasks require users to explore an unseen dataset [46].

3.2 Analysis Methodology

Given our selected EVA tasks and characteristics are diverse (Table 1), and the user interaction information these studies have made available or stored, it is challenging to adopt a generalized analysis approach. However, by adopting different techniques and narrowing the scope of data focus, we try to essentially answer the same research questions.

3.2.1 Scope of Data Focus: Defining the scope of data focus for each user study is crucial because data manipulation in VES can range from selecting a data column to filtering parts of it. Existing research has emphasized analyzing attributes (data columns) selection for characterizing user behavior in EVA and recommending visualizations. Thus, in the Tableau and Voyager user studies, we track user data area shifts by analyzing the attributes selected during each interaction.

However, the EVA process differs for imMens, designed to explore large-scale datasets with millions of data points. imMens displays five visualizations showing relationships among the Brightkite traveler check-in attribute, which includes User, Date, Time, Latitude, and Longitude. These visualizations provide interactive querying, allowing users to aggregate or filter data to understand groups of data points. Therefore, imMens operations—such as pan, brush, range-select, and zoom—are indicators of shifts in users' data focus.

3.2.2 Formalizing the Exploration Problem. For each EVA study, we define the scope of data focus and formalize how users shift their data focus to generate insights. Specifically, though not limited to, the following EVA steps: 1. What data area is the user currently focusing on? 2. Based on her current knowledge, what action will the user choose? 3. Consequently, the VES suggests a new visualization or changes the interface. What does the user learn? 4. How does the user update her current knowledge to gather insights?

These steps guide the development of the learning algorithms discussed in the next section for empirical analysis. For instance, in Step 1, we define how to encode the information. In Step 3, we establish a process for quantifying the information learned.

3.2.3 Statistically Analyzing Shifts in Data Focus: We believe users' data focus shifting strategies may change, but we must first validate this claim for the selected studies to benchmark the models accurately.

We use mixed effects models, which are well-regarded in the visualization community for analyzing users' exploration strategies [8, 34]. We ensure that each user's exploration segments have sufficient interactions to capture the strategies effectively. Based on our preliminary evaluation of the interaction logs, we divide each user's exploration session into two exploration halves, ensuring consistency across users performing varying tasks on different datasets, and allowing for a uniform approach for significance testing. Additionally, while users may shift their data focus strategies at different stages of an EVA session, this division also represents the most fundamental level at which we would expect to observe changes in strategy. We investigate: *Do shifts in users' exploration strategies significantly change during EVA (initial half vs. later half)?* In instances where sufficient data is available, a more granular sampling of the exploration session yields consistent results.

We analyze three main factors: (a) *exploration half* (initial vs. later); (b) *task type* (open-ended vs. focused); and (c) *analysis scenario* (participants and datasets). In our mixed effects model design, fixed effects are factors consistent across all groups, such as the exploration half and open-endedness. Random effects vary across different groups and help us account for data variability that fixed effects alone cannot explain.

To determine the importance of fixed effects, specifically the exploration half, we use likelihood-ratio tests [57]. We build two models: a full model that includes all fixed effects, and a null model that excludes the exploration half. By comparing these models using likelihood-ratio tests (lr-est), *we see if the exploration half significantly effects users' data focus shifts.* We assess this by examining the p-values obtained and *find that users' exploration strategies to shift their data focus change between the two exploration halves.* We use the lme4 R library [6] for our analysis.

3.2.4 Modeling Shifts in Data Focus: We aim to identify the best methods for modeling shifts in users' data focus. However, visualization researchers may learn **individual** models for each user, facilitating adaptation to personal preferences. These models are tuned online, using past interactions of that specific user for predicting future interactions, detecting exploration bias, and recommending visualizations [41, 44]. While this approach offers tailored adaptation, it assumes that we have sufficient if not any data for a specific user.

Alternatively, another approach is to learn a single model from multiple users' interaction sessions with the assumption that training on a diverse **population** helps generalize to a new user. This is specifically beneficial for machine learning-based models that require large amounts of training data [7, 18, 50]. In our benchmark study, we test both modeling assumptions. *We aim to investigate how well the online learning algorithms trained on multiple users (population) adapt to individual preferences compared to individual models.*

The algorithms are empirically evaluated based on their accuracy in predicting actions, highlighting how users will change their data focus. We provide technical details on the evaluation procedure in subsubsection 3.2.4.

4 Learning Algorithms for User Model

In this section, we present and justify the selection of learning algorithms used in the benchmark study.

4.1 Reinforcement Learning (RL)

Researchers in cognitive science and neuroscience have identified parallels between how RL algorithms learn to make decisions through interacting with an environment and how humans learn to do complex tasks and make decisions [17, 33, 42, 43]. RL algorithms have achieved near-human-level performance in various tasks [40, 51], which justifies their growing popularity in modeling user behavior during exploratory data analysis [5, 18, 39, 50]. To employ RL for modeling users in EVA, it is essential first to formulate *users' interactions with the VES to discover insights* as a Markov Decision Problem (MDP).

Markov Decision Process (MDP): An MDP offers a mathematical framework for sequential decision-making problems by defining interactions through states, actions, and rewards. RL algorithms utilize MDP to learn optimal policies within the environment, i.e., deciding which action to take in a given state and updating the policy based on the rewards received. The structure of the MDP will vary depending on the EVA characteristics of our selected user studies. However, we aim to maintain consistency in MDP design across studies, following the general framework outlined below:

In EVA, user (RL algorithm) learns to find insights through repeated interactions with the VES (environment). The user (RL algorithm) has an exploration policy (π) that guides her shifts in data focus (action) based on the current data area (state). *VES (environment) provides visualizations that contain insights (rewards).* The user uses these rewards to update her current policy (π). Eventually, the user learns an optimal policy (π^*), i.e., better decision-making strategies for shifting data focus to find the desired information. Like an RL algorithm, user's main goal is to maximize the amount of insight (reward).

Selection of RL Algorithms: Our benchmark study includes RL algorithms, broadly classified into value-based, policy-gradient-based, and Actor-Critic methods, which combine elements of the first two.

Value-based RL algorithms learn a *value function (vf)*, which outputs the expected discounted reward of a state or a state-action pair. The algorithm's policy, which determines which action to take in a particular state, depends on this *vf* and is updated through trial and error. Examples of such algorithms include Q-learning and SARSA.

Alternatively, other researchers [12] advocate for simpler algorithms that directly learn a policy without a *vf*. A *vf* may still be used to learn the parameters defining a policy but is not required for action selection [52]. These algorithms are called *value-free or policy-gradient* algorithms, e.g., Reinforce. Bennett et al. [11] demonstrated that combining value-based methods with policy gradients, known as Actor-Critic, also yields promising results in explaining human behavior. Now, let us briefly discuss our selected RL algorithms:

Q-learning (Qlearn): iteratively updates a vf called the Q-function. Qlearn learns the optimal policy through trial and error with a ϵ-greedy policy. That is, choosing either a random action with a small probability, ϵ, or the action with the highest estimated reward with probability, 1 - ϵ. QLearn aims to learn a policy that maximizes the expected reward in an environment [55] based on the Q-function update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \tag{1}$$

Where $Q(s_t, a_t)$ is the value of taking action a in state s at time t. α is the learning rate and controls for the degree of Q-value (Q) update, r_t is the reward received at time t, γ is the discount factor to give more weight to r_t than future rewards, and s_{t+1} is the next state. The last hyper-parameter in this algorithm is ϵ for ϵ-greedy.

SARSA: is value-based like QLearn [48]. But unlike Qlearn, which updates its Q using the action that yields **maximum** Q-value in the next state, SARSA updates Q by following the action based on the ϵ-greedy policy:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{2}$$

SARSA has the same hyperparameters as QLearn : γ, α, ϵ

Reinforce is the simplest policy-gradient method [52]. It directly improves the policy based on the observed rewards without vf [56]. Neural-network-represented parameters define the policy that is improved by following the gradient of the expected reward. Reinforce uses γ and α as hyperparameters.

Actor-Critic extends Reinforce by improving the policy through learning a value function in parallel. It combines value-based (critic) methods with a policy-gradient side (actor) [31]. The actor decides which action to perform based on a given state. The critic uses the vf to tell the actor how good the performed action was and how to update the policy parameters. This algorithm has the same hyperparameters as Reinforce.

4.2 Current baselines for user modeling in visualization

User models in visualization have been used for a broad range of applications, such as studying various forms of biases [41, 44], predicting user's personality attributes [60], and even prefetching next data tiles for responsive VES [7, 19]. For our analysis, we try to include as many models as possible as long as their assumptions are compatible with our benchmark. For example, a key assumption for the Competing Models approach [41] is that users interact with a set of *point-based visual data*, takes as input a *complete underlying dataset* with d attributes and compute 2^d Bayesian models. While it was not prohibitive for the authors to compute exact probabilities (largest model space: 2^7), as the authors note, doing the same for larger datasets can be infeasible and may require additional assumptions or sampling methods [41]. However, we include a more general algorithm that is independent of the underlying dataset but shares the same Bayesian update framework. Specifically, our evaluation includes the following established approaches for comparison.

Momentum: This model assumes that the user's next action will be the same as her previous action in that state.

Hidden Markov Model (HMM): HMM assumes that the user's behavior evolves according to a Markov process—that is, the current state depends solely on the previous state. While a Markov Decision Process (MDP) is defined by states and actions, a Hidden Markov Model (HMM) is characterized by hidden and observable states. To align our user modeling problem and evaluation via predicting users' shifts in data focus (actions), we consider actions as hidden states, and the observed state at time t comprises the data area the user is currently visualizing.

Bayesian Learning: This model assumes a uniform prior on action probabilities, observes new user interaction data (state-action) data and updates the state-action probabilities in light of new observations. At each step, the model picks the next action by sampling from its updated probability distribution.

Support Vector Machine (SVM) [54]: SVMs have been employed in modeling user behavior in Exploratory Visual Analysis (EVA), leveraging offline data to pre-fetch data tiles [7], predict future mouse movements, and infer cognitive traits[13]. In this work, we use information about users' current data areas (states), e.g., the attributes they are interested in, as input features. We then employ SVM to predict the corresponding action for shifting data focus. To ensure a fair comparison with online RL algorithms and emphasize the use of online learning algorithms to model users' shifts in data focus we include an online version of SVM. We implement **OnlineSVM** with `sklearn's` partial-fit function on top of OfflineSVM.

4.3 Simple Decision Making Heuristics

We use the Random strategy as our baseline and some heuristics that users often adopt in decision-making scenarios.
Random Strategy: The agent always picks an action *uniformly at random* from the available choices. Action choice is made irrespective of the rewards received or consideration for potential outcomes.
Greedy Strategies: User picks an action for immediate success based on her previous experience. She chooses the action that has yielded her the highest reward thus far [52].
Win-Stay Lose-Shift (WSLS) Heuristic: Repeats a successful action until it no longer yields rewards, then switches to other actions with equal probabilities. WSLS is a popular heuristic to model human learning in games [53].

5 imMens user study

Unlike the other two interfaces in this paper (Tableau and Voyager), imMens does not require users to select attributes to generate visualizations. Therefore, we first focus on the imMens user study by Liu and Heer [34], which allows us to analyze users' shifts in data focus in a *restrictive setting*.

5.1 Overview of Exploration Task

5.1.1 Analysis Task. The 16 participants in this user study report *interesting findings*, defined as surprising events, data abnormalities, or confirmations of common knowledge [34]. They explore (a) travelers' check-in data from *Brightkite*, a location-based service, and (b) U.S. flight performance data. The Brightkite dataset includes travelers' check-in dates and locations (latitude and longitude), while the U.S. flight data covers airline carriers, flight dates, and arrival and departure delay

5.1.2 Characteristics. The analysis task in this user study is open-ended, as the definition of *interesting findings* is not precise, leading to uncertainty about which data areas to explore. Although the authors provide some examples of interesting findings, the guidance is vague, leaving participants unsure about what specifically to search for. Participants get 15 minutes of **prior experience** with the datasets and interface, which may impact what they learn to a certain degree compared to scenarios where they proceed EVA without familiarity.

5.1.3 imMens Interface. The imMens interface (Figure 2) presents users with four fixed visualizations for exploring the flight performance dataset: (1) Scatter plot: showing the relationship between arrival and departure delays. (2) Carriers: A bar chart for US airline carriers. (3) Year: A bar chart showing flights across years. (4) Month: A histogram with flight over timeframes.

Users interact via *brush & link, pan, zoom, and select* operations [35]. with changes in one visualization updating the others. For instance, if a user *selects 'Year = 2003' in the visualization 'Year'* (Figure 2), the data is filtered, and visualizations update to show information specifically for the year 2003.

Details on the visualizations generated for the travelers' check-in dataset can be found in the imMens study paper by Liu et al. [34].

5.1.4 imMens Interaction Log contains users' imMens operations and the visualizations they interact with at each time step. Additionally, it has users' verbal feedback, where they explain their actions, findings, and reasonings, e.g., what type of information they want to find, if they have discovered anything new, etc.

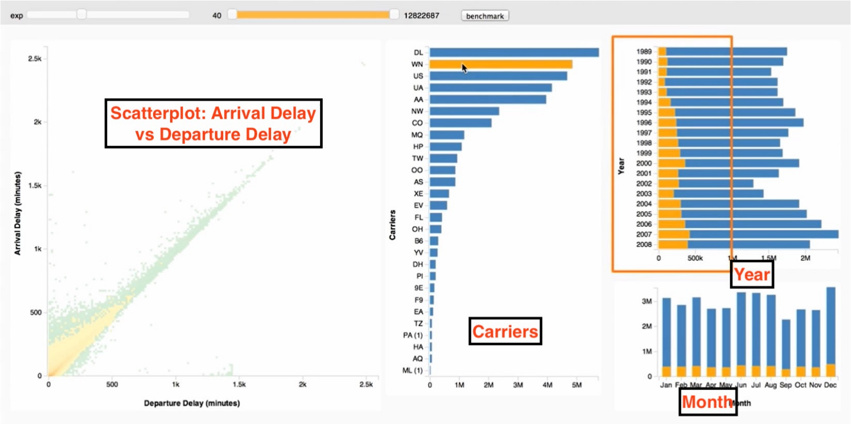


Fig. 2. imMens user exploring flight performance data

5.2 Formalizing User Exploration Problem

5.2.1 User Activities. At each step, users select a visualization, i.e., a data area, and apply imMens operations. The user then analyzes the visualizations in the imMens interface for insights. The information obtained from the interface acts as a reward, guiding the user to either continue focusing on the current data area or shift their data focus elsewhere. If any *interesting findings* are discovered, users report them before continuing their exploration.

5.2.2 Extracting Features from Verbal Feedback. Users' verbal feedback logs provide valuable information on their reported insights and the rationale behind imMens operations, i.e., data focus shifting strategies. Liu et al. annotate these verbal feedbacks into seven categories [34]. We leverage user feedback logs and these annotations for reward engineering and features in the user model, thereby enriching our modeling process.

Out of the seven categories, we select four as features for our model, as they encompass 95% of the feedback. These categories reveal user exploration goals, thus offering insights into their data focus-shifting strategies.

Observation: Users discover a piece of information about the data originating from a single visualization (data area). Users may stay in the same data area until they believe no further insights can be gleaned without switching. **Generalization:** Users aggregate information from multiple visualizations and report. **Hypothesis:** Conjecture about the data, made to steer exploration or explain observation/generalization. Thus, indicating possible shifts in data focus. **Question:** Indicates users' desire to explore different data areas. Users' data focus shifting strategies depend on whether the desired information can be found by revisiting previously explored areas or requires venturing into new ones.

More information on this categorization process and examples are available in [34].

5.2.3 Modeling Exploration Using MDP: Users use imMens operations to explore different data areas and discover insights. For example, they might select airlines one by one in the "Carriers" visualization and observe changes in the "Year" and "Scatterplot" views, helping them identify patterns like when an airline started or its delay trends.

We use MDP to model the user's decision-making process in selecting imMens operations to shift their data focus, aiming to maximize insight discovery. In MDP terms, this is referred to as the *exploration policy*, which the user learns to determine which operation will be most beneficial based on their current data area.

However, adhering to a single policy is not optimal. Users may effectively change their exploration policy once they feel they have gathered sufficient information. For example, they may pan the scatterplot to understand the relationships between arrival and departure delays over time or identify which airlines experience the most delays. In this way, users adapt their exploration strategies to discover 'interesting findings'. Next, we define the components of this MDP to model such dynamic shifts.

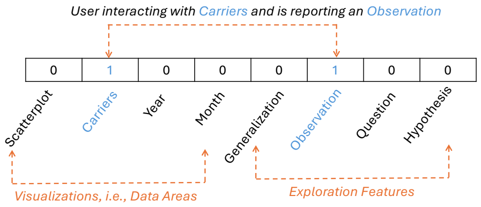


Fig. 3. One-hot encoding of attributes in interactions as State

States: We define the state s_t at time step t to capture the user's current data area of interest and exploration phase introduced in subsubsection 5.2.2. The data area of interest refers to the visualization the user is currently interacting with. We encode this information into a one-hot vector.

Actions: imMens operations *pan, zoom, brush, and range select*

Rewards: In each interaction t, users receive rewards r_t from the information obtained through the imMens visualizations after applying action a_t. This reward $r_t \in [0, 1]$ depends on how much the action contributes to finding an insight. When users report new information, such as an observation or generalization, they receive a reward $r_t = 1$. Similarly, when users generate insights that lead to hypotheses or questions, they also receive a reward $r_t = 1$ for propelling exploration for insights. For the online learning algorithms to better adapt to users' data focus during intermediate interactions, we assign a small reward $r_t = 0.2$.

5.3 Statistically Analyzing Shifts in Data Focus

Following the outline in subsubsection 3.2.3, we analyze the imMens actions users use to shift their data focus. We calculate the probability of each action for the two exploration halves and check, "Does the exploration half significantly effect how users shift their data focus?". The mixed models treat the exploration half as a *fixed effect* and the users and datasets as *random effects*.

The results of the *likelihood-ratio test* between the *full model* and *null model* are shown in Table 2. We find that the exploration half—whether a user is in the initial or later stages of their exploration—significantly influences the probability of selecting an action, indicating shifts in users' focus between data areas (visualizations). For example, initially, users may repeatedly use brush to explore the bar chart 'Carriers' and 'Year' in Figure 2 to find out when the airlines start/end their business. Later, they may repeatedly use pan and zoom to identify patterns between arrival and departure delays.

Actions	χ^2	P-value	Significance
Brush	6.4033	0.01139	*
Pan	8.1423	0.004324	**
Zoom	13.312	0.0002624	***
Select	6.9957	0.00817	**

Table 2. Significance test results for each visualization. Significance: *** p <0.001; ** p <0.01; * p <0.05;

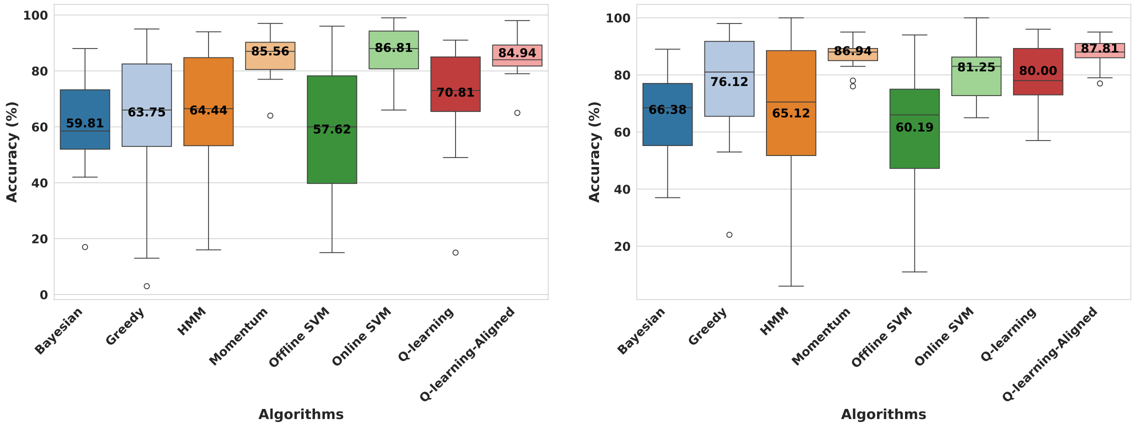
5.4 Performance Evaluation

5.4.1 Evaluation Procedure: In this study, we assess the performance of different learning algorithms in predicting a user's next action for shifting data focus. As outlined in the analysis methodology (subsubsection 3.2.4) we evaluate these algorithms in two different settings:

Individual Setting: For each user, we use the first 80% of their interactions for training and tuning the model. The remaining 20% of interactions are reserved for testing. This helps us evaluate how well the model predicts actions based on strategies from the *same user alone.*

Population Setting: In this case, we use the leave-one-out cross-validation method. To predict a user's strategies for shifting data focus, the model is trained on the strategies of all other users, leaving out the test user. This helps us evaluate how well the model predicts actions given additional data from *other users.*

To ensure consistent comparison between the two settings: (a) we report the test accuracy for the population model over the last 20% of the test user's interactions (b) before making predictions, the population model processes the same initial 80% of the test user's interactions in an online fashion, aligning with the approach used in the individual setting.



(a) Individual Experiments

(b) Population Experiments

Fig. 4. Performance results for two experimental settings in the imMens user study.

5.4.2 Results: To facilitate easier understanding, we aggregate the prediction accuracy of each method *across all EVA tasks and datasets*. Figure 4 presents the experimental results for both settings. To make the plots clearer, we only show the most effective RL model, a simple heuristic, and the current baselines from visualization research. Full results are available in the supplementary materials.

For this study, we find that RL models are slower to adapt to users' shifts in data focus. In the individual experiments, the accuracies are Q-learning (70.81%), SARSA (67.25%), Actor-Critic (67.88%), and Reinforce (64.06%). For the population experiments, the accuracies are Q-learning (80%), SARSA (78%), Actor-Critic (74.94%), and Reinforce (71.88%). We present the best-performing RL algorithm, Q-learning, in Figure 4. However, Momentum and the OnlineSVM outperform all RL algorithms in both settings.

Upon investigation, we attribute this issue to the type of feedback provided to the RL models. Typically, RL models receive feedback on the predicted action and gradually learn the optimal action for each state. Therefore, the RL models predict sequences of interactions one by one, receiving rewards depending on whether their predictions are correct. However, models such as OnlineSVM and Momentum directly use the ground truth action that the user took to adapt their models. RL models lack this direct information, putting them at a disadvantage.

To address this, we modified the best-performing RL model (Q-learning) by assigning a positive reward to the ground truth action, in addition to providing feedback on the predicted action. This extension, called **Q-learning-Aligned** outperforms all other methods in both experimental settings.

However, the performance of the Momentum model does provide insights into users' exploration strategy using the *restrictive* imMens interface. When users focus on a particular visualization, they tend to repeatedly use the same operation until their information needs are satisfied.

Finally, these results highlight the importance of online adaptation in modeling users' dynamic data focus shifts. Notably, OnlineSVM shows a 29.19% accuracy improvement over OfflineSVM in the individual model and 21.06% in the population model.

6 Tableau user study

In this study by Battle et al. [8], participants perform a series of EVA tasks (Table 3). These tasks represent a common exploration progression by which analysts maximize the effectiveness of their EVA sessions.

6.1 Overview of Exploration Task

6.1.1 Analysis Tasks. 27 participants use Tableau [2] to complete a series of analysis tasks (Table 3). Their experience with data analysis and Tableau expertise varied widely. The analysis tasks are from the following dataset: (a) Weather station reports on weather metrics (35 columns, 56.2M rows), (b) U.S. domestic flight performance data (31 columns, 34.5M rows), and (c) Aircraft striking wildlife reports, (94 columns, 173K rows) [8].

6.1.2 Task Characteristics. Let's examine how Battle et al.'s (Table 3) subtask design captures the natural exploration of EVA at a fine granularity. Task T1 captures users' initial exploration to learn the data attributes, facilitating a general understanding of the dataset. Tasks T2 and T3 investigate the statistical relationships or existing patterns between data variables. Finally, task T4 encapsulates more sophisticated explorations, such as prediction and causality analysis. Tasks T1 to T3 are focused, while T4 is open-ended. Users in this study lack prior knowledge of the datasets. To explore datasets using Tableau, users must select data columns (attributes) and choose the best visualization to showcase relationships, adding overhead to the insight-searching process compared to imMens.

6.1.3 Interface and Interaction Log. Users choose which attributes to explore and add them to the Tableau worksheet (Figure 5). Tableau then suggests visualizations with different visual encodings. The interaction log records users' interactions with Tableau, such as their choice of attributes and visualization type.

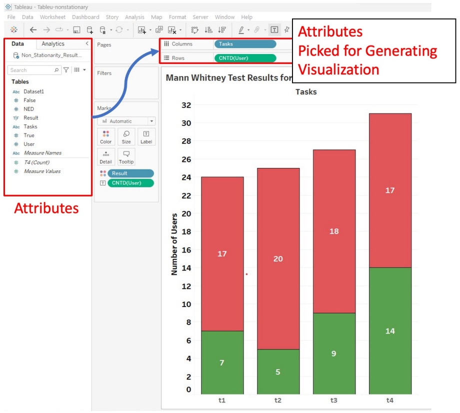


Fig. 5. Analyzing attributes using Tableau interface

6.2 Formalizing Exploration Problem

6.2.1 User Activities. To analyze how users shift their data focus in Tableau, we examine how they select and modify attributes during analysis tasks. Users choose attributes to generate visualizations, then either (a) spend more time exploring the current data area or (c) modify the attribute set to shift focus. This process continues until they reach their desired insights. As users explore, they gain deeper understanding of the data, helping them make informed decisions about shifting their data focus.

6.2.2 Modeling Exploration Using MDP: Let's say a user's exploration session has T interactions. Now, we will define MDP components for Tableau exploration tasks.

States: State s_t represents a user's data focus in interaction $t \in [1, T]$, *specifically the attributes she has selected to generate a visualization.*

Actions: A user's action a_t showcases her strategy for shifting data focus. Our proposed actions $a_t \in A$ provide fine granularity to analyze these shifts. The action **Keep** is used when a user finds the current attributes useful and wants to investigate further. Alternatively, the user can modify the attribute set to explore new information. Actions **Modify-1**, **Modify-2**, **Modify-3**, and **Modify-4+** are used to modify one, two, three, and more than three attributes in a single

Task	Task Description
T1	Consider the following weather measurements: Heavy Fog[Heavy Fog], Mist [Mist], Drizzle [Drizzle], and Ground Fog [Ground Fog]. Which measurements have more data?
T2	How have maximum temperatures [T Max] and minimum temperatures [T Min] changed over the duration of the dataset (i.e., over the [Date] column)?
T3	How do wind measurements [High Winds] compare for the northeast and southwest regions of the US?
T4	What weather predictions would you make for February 14th 2018 in Seattle and why?

Table 3. Analysis tasks for Weather dataset

interaction, respectively. These actions capture how exploratory the user wants to be in their shifts in data focus, Our analysis of the interaction logs reveals that over 99.5% of the time, users modify fewer than four attributes in a single step. Therefore, we do not need to go beyond Modify-4+.

Reward: We want our MDP reward to reflect how useful a particular data area is for finding insights. To do this, we quantify the usefulness of a set of attributes so that users will find insights when investigating that data area. This reward r_a, calculated using Equation 3, quantifies the importance of attribute a from the dataset D.

$$r_a = \frac{\text{number of users that used } a}{\text{total number of users who explored } D} \tag{3}$$

Consequently, if α is the selected attributes in t, users receive reward:

$$R_t = \sum_{a \in \alpha} r_a \tag{4}$$

6.2.3 Exploration Problem in the Context of MDP: Based on the visualization generated from a user's selected attributes (s_t) at interaction $t \in T$, the user decides how to modify these attributes using an action a_t. The strategy by which the user picks a_t on s_t is called the user's exploration policy π. After taking the action, the user reaches state s'_t and receives reward r_t. The user updates π based on new insights r_t. The goal is to maximize the chances of discovering attributes with desired insights: $\max_\pi \mathbb{E}\left[\sum_{interaction=t}^{T} \text{reward}(\pi, interaction)\right]$. To achieve this, the user may need to adapt her policy to $\pi^* : State \longrightarrow Action$.

As a simple example, in task T4, from Table 3, a user might continue exploring `Temperature` and `Fog` (state) in detail or shift focus (action) to `Snow` and `Precipitation` to achieve a more accurate weather prediction. Insights gained (reward) from the explored visualizations influence how the user shifts her data focus (policy update).

6.3 Statistically Analyzing Shifts in Data Focus

Using the actions introduced in subsubsection 6.2.2, we determine the probability of using a particular data focus shifting strategy in each exploration half. We treat the exploration half as a fixed effect, with individual users and datasets as random effects. Table 4 presents the likelihood ratio test results.

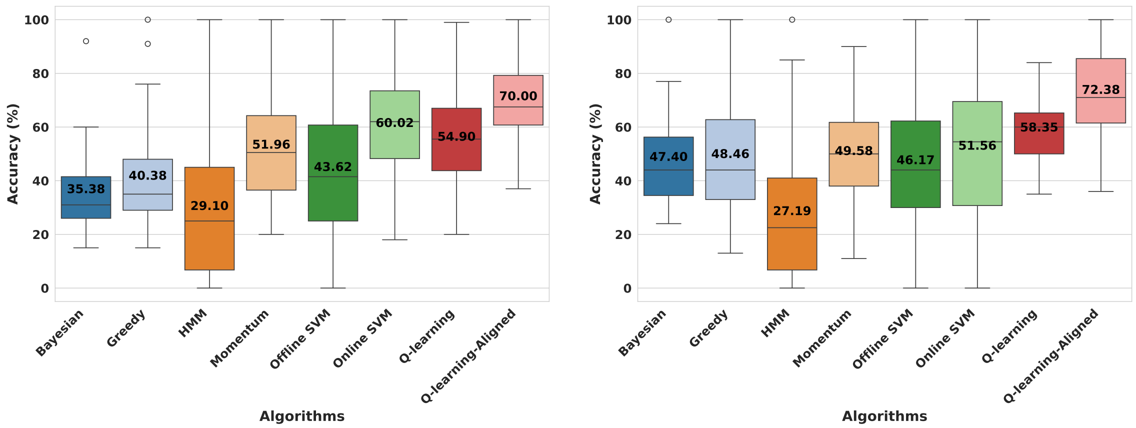
The results reveal a significant shift in users' data focus strategies between the two exploration halves. For instance, the action *Keep*, which allows users to further investigate the current data area, is notably influenced by the *exploration half.* This is because, users initially explore multiple attributes to get a generic idea about the data before performing a drill-down analysis by repeatedly using *Keep* on selected attributes.

Actions	χ^2	P-value	Significance
Keep	5.1005	0.02392	*
Modify-1	5.3292	0.02097	*
Modify-2	6.748	0.009385	**
Modify-3	6.3406	0.0118	*
Modify-4+	6.152	0.01313	*

Table 4. Significance test results for each visualization. Significance: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$;

6.4 Performance Evaluation

Following the same evaluation setup and summarization procedure described in subsection 5.4, the results of the Tableau experiment are presented in Figure 6.



(a) Individual Experiments

(b) Population Experiments

Fig. 6. Performance results for two experimental settings in the Tableau user study.

Results: Similar to subsubsection 5.4.2, in this study, Q-learning-Aligned outperforms all other methods. The accuracy of the RL models in the individual experiments are Q-learning-Aligned (70%), Q-learning (54.90%), SARSA (53.38%), Actor-Critic (58.79%), and Reinforce (54.65%). For the population experiments, the accuracies are Q-learning-Aligned (72.38%), Q-learning (58.35%), SARSA (55.27%), Actor-Critic (55.79%), and Reinforce (55.79%). We present Q-learning and Q-learning-Aligned in Figure 4.

Compared to current baselines in the visualization community, Q-learning-Aligned explicitly integrates users' attribute interests with reward signals after each interaction and uses hyperparameters to anticipate shifts in their data focus, leading to more accurate predictions. This approach is beneficial because, while predicting a high-reward (insight-yielding) attribute may bring initial success, repeatedly using the same strategy (e.g., with the *Keep* action) becomes less effective over time, as the user may have already gathered the needed insights and shifted focus to other attributes.

Additionally, in this experiment, we observe that with more data in the population setting, OfflineSVM (+2.55%), Q-learning (+3.45%), SARSA (+1.89%), and Q-learning-Aligned (+2.38%) see performance improvements. However, OnlineSVM (-8.46%), and Actor-Critic (-3.83%) do not benefit from training over large populations. This finding intrigued us, prompting further investigation into what these models are actually predicting.

First, we examine the distribution of actions in the test data: Keep (55.15%), Modify-1 (32.48%), Modify-2 (5.39%), Modify-3 (1.23%), and Modify-4+ (5.76%). Each row of Table 5 and Table 6 represents the accuracy of predicting a specific action in individual and population settings, respectively. In the population setting, Actor-Critic and Q-learning are overfitting to the action Keep (Table 5). However, this is somewhat lessened in the individual setting. On the other hand, in both settings, Q-learning-Aligned performance improves over generic Q-learning on the actions that occur less frequently than Keep. Most algorithms struggle to predict Modify-2, Modify-3, and Modify-4+ actions, which appear in less than 15% of interactions. However, leveraging this imbalance and always predicting the most frequent action, Keep won't lead to high accuracy. For better performance, models must at least accurately predict smaller shifts like Modify-1. Although Momentum did well in the imMens user study, it lacks learning capability and depends on users to shift data focus, making it less suitable for modeling but still informative about the data.

	Actor-Critic	**Q-learning**	**Q-learning-Aligned**	**Online SVM**	**Offline SVM**	**Greedy**
Keep	76.33	87.67	88.00	61.67	50.67	51.67
Modify-1	21.67	3.33	37.67	43.67	31.00	18.33
Modify-2	3.33	16.00	23.00	24.33	21.33	24.67
Modify-3	0.33	0.67	2.00	2.00	0.00	3.33
Modify-4+	1.33	3.33	15.67	13.00	4.67	8.33

Table 5. Individual Setting: Each row shows the accuracy (%) of predicting a specific data focus shifting action

	Actor-Critic	**Q-learning**	**Q-learning-Aligned**	**Online SVM**	**Offline SVM**	**Greedy**
Keep	98.67	93.67	88.67	69.33	54.67	67.67
Modify-1	2.67	13.67	53.00	47.00	30.33	31.67
Modify-2	10.67	21.67	40.00	52.33	48.33	24.33
Modify-3	0.00	9.67	9.67	9.67	11.00	4.67
Modify-4+	0.00	4.00	32.00	12.67	3.67	23.00

Table 6. Population Setting: Each row shows the accuracy (%) of predicting a specific data focus shifting action

7 Zeng et al. Voyager User Study

The user study by Zeng et al. [61] comprises analysis tasks (e.g., T1–T3 in Table 7) similar to those in the Tableau study subsubsection 6.1.1, and open-ended analysis tasks (e.g., T4 in Table 7) as in subsubsection 5.1.1, where users explore datasets and self-report any insights.

Task ID	Task	Task Objective
T1	Focused	Which creative type has the max number of movies based on Book/Short Story (Source)?
T2	Focused	Among Disney (Source) movies, what's the running time of the highest-grossing?
T3	Open-Ended	What kinds of movies will be the most successful based on data?
T4	Open-Ended	Explore data for [15 mins]. Use bookmarks to save patterns, trends, or insights.

Table 7. Focused and Open-ended EVA tasks for Movies Dataset

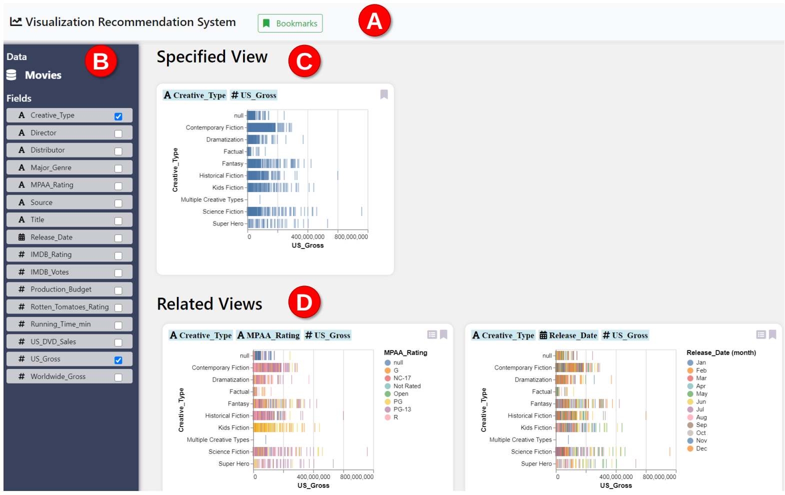
7.1 Overview of Exploration Task

7.1.1 Analysis Task. In this study by Zeng et al., 72 participants complete four data exploration **tasks** using a visualization recommendation system (Figure 7) [61]. The study includes two **datasets**: (a) *Movies* dataset containing 3,101 records and 16 attributes, (b) *Birdstrikes* dataset (Example 1) containing 10,000 records and 14 attributes [58].

7.1.2 Task Characteristics: During the user study participants complete two focused tasks and two open-ended tasks (see Table 7). The **focused** tasks (T1 and T2) provide clear hints on which data area to explore. Whereas, for the **open-ended** tasks (T3 and T4) participants are expected to explore the dataset for interesting insights. Participants received a study overview and a 10-minute demo with a dataset distinct from the ones used in the study. Hence, they have no **prior experience** with the dataset.

7.1.3 Interface and Interaction Log. The **interface** (Figure 7) for this user study is inspired by the Voyager systems [58, 59]. Similar to Tableau, users can select attributes to explore from panel (B), and the system recommends visualizations based on these selected attributes in panel (C). Additionally, the system suggests visualizations from related data areas in panel (D). Users can analyze these visualizations and bookmark them for later review, accessible through the bookmark gallery button in panel (A).

The **interaction log** contains information on the user's actions, including selecting attributes, bookmarking charts, mouse hover, and scroll-over charts. The logs also contain metadata about the user's chart bookmarks for open-ended tasks. To account for unintentional noise, we follow the same approach as Zeng et al. [61] and exclude mouse movement logs that lasted less than half a second.



A Creative_Type ☑
A Director ☐
A Distributor ☐
A Major_Genre ☐
A MPAA_Rating ☐
A Source ☐
A Title ☐
Release_Date ☐
IMDB_Rating ☐
IMDB_Votes ☐
Production_Budget ☐
Rotten_Tomatoes_Rating ☐
Running_Time_min ☐
US_DVD_Sales ☐
US_Gross ☑
Worldwide_Gross ☐

Fig. 7. Interface for the Voyager user study.

Actions	χ^2	P-value	Significance
Keep	5.5818	0.01815	*
Modify-1	35.449	7.594×10^{-6}	***
Modify-2	2.9127	0.08789	.
Modify-3	0.8095	0.3683	-

Table 8. Significance test results for each action. Significance: *** p <0.001; ** p <0.01; * p <0.05; . p <0.10; - p >0.10

7.2 Formalizing the Exploration Problem

7.2.1 User Activities. Users' exploration activities in this user study have close similarity to the Tableau study (subsubsection 6.2.1). In both cases, the exploration is driven by the data attributes users select in each interaction. The Voyager system generates visualizations relevant to the selected attributes. For the next exploration steps, (1) the user may analyze the specified chart, (2) shift to exploring other attributes in the system-generated related views, or (3) modify her current attribute selections in the data panel.

As discussed in subsubsection 6.2.1, the visualizations contain rewards that may come in the form of relevant answers to the task questions for focused tasks (T1, T2) or insights/data characteristics that the users may bookmark for open-ended tasks (T3, T4). Rewards help the user decide how to shift her data focus.

7.2.2 Modeling Exploration Using MDP. To maintain consistency in our MDP formulation, we aligned it with the Tableau user study described in subsubsection 6.2.2. While the state representation (s_t) remains the same, there are minor modifications to the action space (A) due to the Voyager interface.

Actions: In contrast to the Tableau interface, the visualization recommendation system used in Zeng et al. [61] limits users to select a maximum of three attributes in a single interaction. Therefore, the maximum degree of shift is limited to changing three attributes in a single interaction; thus, we remove the *Modify-4+* action from the Tableau user study action space described in subsubsection 6.2.2. The final set of actions are: *Keep*, *Modify-1*, *Modify-2*, and *Modify-3*.

Reward: In this user study, during open-ended tasks, users bookmark *interesting visualizations*. We leverage this bookmark metadata to extend the reward function defined in Equation 4. Specifically, for each user, we augment R_t by adding a positive scalar value n_a, corresponding to the number of times an attribute (a) appears in the user's bookmarked charts:

$$R_t = \sum_{a \in \alpha} r_a + n_a \tag{5}$$

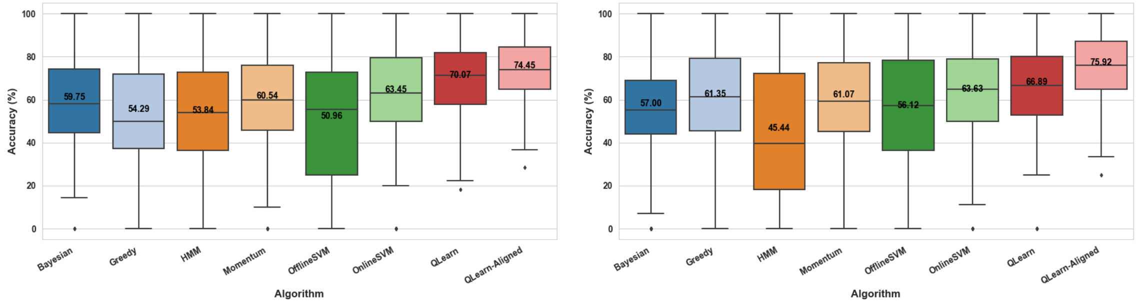
7.3 Statistically Analyzing Shifts in Data Focus

For this study, the *full model* includes the *exploration phase* and *open-endedness* as fixed effects and the *user* and *dataset* as random effects. We opted to include open-endedness as a fixed effect because each participant encountered both levels of open-endedness. To sanity check, we also investigate treating open-endedness as a random effect yielding the same significance test results.

Table 8 shows the likelihood test results, revealing significant effects of the exploration halves on the actions *Keep*, *Modify-1*, and *Modify-2*. Overall, the findings align with those from the Tableau user study (subsection 6.3): users tend to favor drill-down analysis of the same attributes (*Keep*) or gradually modify them (*Modify-1,2*), with these behaviors significantly changing across exploration phases. In contrast, drastic shifts using *Modify-3* account for only 1.36% of all interactions and are not influenced by the *exploration half*.

7.4 Performance Evaluation

Following the same evaluation setup and summarization procedure described in subsection 5.4, the results of the Voayger user study experiments are presented in Figure 6.



(a) Individual Experiments (b) Population Experiments

Fig. 8. Performance for two experiment settings for Voyager Study

Results: RL-based algorithms outperform all other methods for both individual (Figure 8a) and population (Figure 8b) experiments. Similar to subsubsection 5.4 and 6.4, RL algorithms (Tables 9, 10) perform better than Bayesian and HMM because they explicitly incorporate the user's *interesting insights* and *relevant attributes* with the reward signal. For instance, user 31 (Figure 9) explores `Title`, `IMDB Rating` and `Gross Profit` in early interaction, next she shifts one attribute to `Rotten Tomatoes Rating`, then for the bulk of following interaction user returns to `IMDB Rating` for a drill-down analysis (i.e keep action). However HMM and Greedy are still stuck in a sub-optimum and associate the shift of one attribute (modify-1) as the best action for future interactions on `Title`, `IMDB Rating` and `Gross Profit`.

Consistent with earlier study results (subsections 5.4 and 6.4), QLearn benefits from alignment. QLearn-Aligned accurately *predicts rare shifts in data focus* (e.g., correctly predicts Modify-3 for user 31) and *quickly adapts to shifts* as seen in Figure 9 for user 23, whose strategy is to explore 5 different data insights across different analysis segments, explicitly telling the model the correct shifts after it makes an incorrect prediction helps better anticipate shifts between data segments.

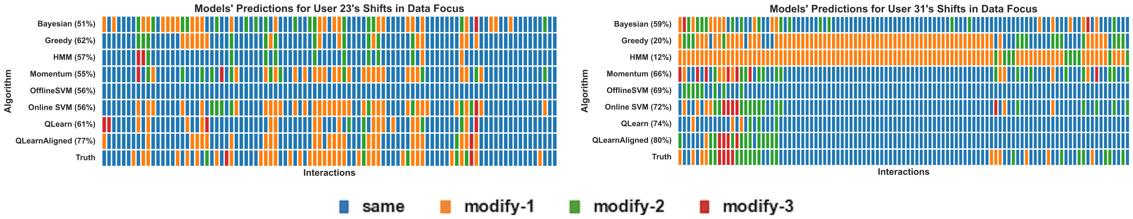


Fig. 9. Comparison between model's predictions and ground truth for two users' shifts in data focus

Open-endedness: Learning algorithms generally model users better (+8.48%) in focused tasks than in open-ended tasks. This is because opened tasks (Table 7) entail a larger search space for models as well as users [8]. The top-performing Q-learning-Aligned (QLearn-Aligned) model averages a 9.44% improvement in focused tasks across both experiment settings. This result suggests that although QLearn-Aligned and other RL models consistently outperform other models in predicting shifts in data focus, they are also not entirely robust to the higher exploration noise during open-ended exploration.

Experiment Setting: We observe a tradeoff between the two modeling designs. Consistent with the results from Tableau study, Greedy (+7%) and OfflineSVM (+5.2%) perform significantly better in population experiments while HMM (−8.2%) fails for individual setting.

To understand this difference, we perform a deeper investigation of the model's predictions similar to Tables 5 and 6. We find that training on the population pulls model predictions for data focus shifts toward the average, thereby making more accurate predictions. For instance, OfflineSVM's prediction accuracies for actions *Keep*, *Modify-1*, *Modify-2*, and *Modify-3* are 64%, 28%, 17%, and 1.3%, respectively, when trained on individual users. However, when trained on the population, the training data accumulates users' preference for *Keep* and its accuracy increases by 8% for this dominant action.

In contrast, HMM drastically underperforms in the population experiment. In focused tasks like T1 (Table 7), HMM performs well, achieving 66% accuracy. However, in open-ended tasks like T4, where users' exploration patterns differ significantly, its accuracy drops to 28%. This highlights the compounding risk of models learning noise as relevant patterns and failing to adapt when introduced to new strategies for shifting data focus.

8 Discussion

This paper presents an in-depth analysis of three prominent EVA studies [8, 34, 61] and compares the performance of popular online learning algorithms with existing EVA modeling techniques. Our initial statistical analyses confirm that analysts' data focus shifts from the first to the second half of their exploration sessions. These results suggest that algorithms that take data focus shifts into account will be well suited to EVA scenarios, which is reinforced by the competitive performance of the RL algorithms over existing methods. In this section, we summarize our key findings, discuss their implications, and suggest new avenues for future research.

8.1 Reinforcement Learning (RL) for user modeling in EVA

A clear takeaway from our analyses is that **RL algorithms, specifically QLearn-Aligned consistently outperform existing techniques for modeling users' shifts in data focus during EVA.** In Tables 9 and 10, we summarize the accuracy gain of all tested algorithms. We believe these performance improvements come from RL algorithms mimicking users' decision-making strategies for analyzing the next data area based on the current one's usefulness.

Model Class	Model	imMens	Tableau	Voyager
Reinforcement Learning	QLearn-Aligned	**84.94%**	**70.00%**	**74.45%**
	QLearn	70.81%	54.90%	70.07%
	SARSA	67.25%	53.38%	69.73%
	Actor-Critic	67.88%	58.79%	65.91%
	Reinforce	64.06%	54.65%	67.36%
Simple Decision Making Heuristics	Greedy	**63.75%**	**40.38%**	**54.29%**
	WSLS	22.81%	24.35%	25.76%
	Random	25.38%	19.92%	25.26%
Current baselines in Visualization	Momentum	85.56%	51.96%	60.54%
	Bayesian	59.81%	35.38%	59.75%
	HMM	64.44%	29.10%	53.84%
	OnlineSVM*	**86.81%**	**60.02%**	**63.45%**
	OfflineSVM	57.62%	43.62%	50.96%

Table 9. A summary of models' performance for Individual Experiment Setting. *OnlineSVM is a new adaptation to current SVM baseline in visualization

Model Class	Model	imMens	Tableau	Voyager
Reinforcement Learning	QLearn-Aligned	**87.81%**	**72.38%**	**75.92%**
	QLearn	80.00%	58.35%	66.89%
	SARSA	78.38%	55.27%	66.65%
	Actor-Critic	74.94%	54.96%	66.57%
	Reinforce	71.88%	55.79%	63.25%
Simple Decision Making Heuristics	Greedy	**76.12%**	**48.46%**	**61.34%**
	WSLS	23.19%	23.81%	25.96%
	Random	27.56%	25.08%	24.23%
Current baselines in Visualization	Momentum	**86.94%**	49.58%	61.07%
	Bayesian	66.38%	47.40%	57.00%
	HMM	65.12%	27.19%	45.44%%
	OnlineSVM*	81.25%	**51.56%**	**63.63%**
	OfflineSVM	60.19%	46.17%	56.12%

Table 10. A summary of models' performance for Population Experiment Setting. *OnlineSVM is a new adaptation to current SVM baseline in visualization

QLearn models users more effectively than other RL algorithms, such as Actor-Critic and Reinforce (section 4). We believe the effectiveness stems from Q-learning's policy updates, where the chosen action is compared to the best-estimated action in the next state $[\arg\max_a Q(s_{t+1}, a)]$. This resembles how humans evaluate actions by considering the potential benefits of the best option identified so far. Additionally, more complex algorithms generally need larger datasets for effective training. Being a simpler algorithm, QLearn-Aligned is a strong candidate for modeling users's dynamic data focus within VESs; where a model may need to learn from limited streaming interaction data and adapt to the user's shifts in data focus in real-time.

Overall our results suggest that **popular algorithms within the visualization community, specifically Bayesian and HMM, have limited functionality to capture the nuances of users' dynamic shifts in data focus.** RL algorithms leverage hyperparameters (section 4) to dynamically adjust to the user's exploration rate and integrate users' data interests online. In contrast, HMMs are more popular in unsupervised scenarios with abundant relevant unlabeled data and do not perform as well in noisy and constrained data environments.

8.2 Influence of Exploration Task Characteristics:

We find that EVA tasks are difficult to model due to their open-endedness and complexity, which directly impacts algorithm performance. In the imMens study, users have a limited decision space (subsubsection 5.1.1), exploring a preselected set of attributes with fixed visualizations, making it easier for models, including RL, to track users' data focus shifts. Whereas, the Tableau and Voyager studies users have a broader decision space, users navigate a number of attributes, adjusting their data focus strategies as they explore. They may choose suboptimal strategies, like exploring areas that do not contain any insight. However, such strategies help to better understand the dataset and maximize long-term rewards. This highlights the importance of using RL, which excels at learning decision-making policies in complex scenarios.

Task open-endedness further complicates the modeling process, as users are often exploring unknowns and need to investigate more attributes. This is supported by the algorithm's performance in the more focused tasks of the Voyager user study, where models achieved an 8.48% higher average accuracy compared to open-ended tasks (subsection 7.4). These findings allude to a broader hypothesis that current learning models can be further improved to better capture users' exploration behavior in open-ended tasks.

We also hypothesize that users with prior experience may require less effort to complete tasks of similar complexity. It stems from the idea that, if we want to observe how users shift their strategies, we must place them in environments that challenge them to shift. We believe we encountered this issue when analyzing the Tableau interaction logs (subsubsection 6.2.1), where in task T4, users often reused attributes analyzed in previous tasks rather than exploring new ones. Changing the task objective seems to reduce the effect of prior experience, as we observe in the open-ended task in the Voyager user study (subsection 7.3).

8.3 Takeaways for Future Experimenters/Model Designers

Considerations for experiment design: Given that the performance of many models are not significantly different across individual (Table 9) and population settings (Table 10), we take care in reporting our observations and hesitate to make broad statements. However, we note some interesting trends from our empirical evaluation.

There is a clear difference between offline and online models when comparing the individual and population experiment settings. For instance, OfflineSVM performs better in the population setting, which can be attributed to the availability of more user experiences. This increased data allows the model to learn from broader data-focus shifting strategies across users as we discuss in Voyager study (subsection 7.4). In contrast, OnlineSVM shows poorer performance in population experiments, likely because the real-time feedback it depends on is drowned out by patterns the model has already internalized. Additionally, this effect may compound when models learn from noisy data shifting strategies during open-ended tasks, as encountered with HMM in the Voyager study (subsection 7.4).

RL model, specifically QLearn-Aligned, performs slightly better in the population setting. However, it already shows strong results in the individual experiments. This reliability makes QLearn-Aligned a robust baseline for future systems.

Reusing rather than creating study data: The diversity of our selected studies (Table 1) provides an opportunity to observe, formalize, and model users' shifts in data focus in already established and diverse EVA scenarios. Additionally, designing a new user study poses significant challenges. First, recruiting a diverse user base is time-consuming. Second, devising tasks that span a wide range of open-endedness requires extensive research and domain expertise. Third, determining the appropriate visualization tools and level of detail to capture from user interactions adds to the layer of complexity [22]. Finally, as Zgraggen et al. [62] observe, users lack the awareness or desire to explain everything they learn during EVA. Thus, conducting a new user study might not be more beneficial to analyze users' dynamic data interests than pursuing established studies.

Comparing against all relevant baselines: Testing current approaches for user modeling in visualization, including Momentum, Bayesian, HMM, SVMSs, and three additional simple decision-making heuristics, gave us valuable information about how RL algorithms perform. Even though Momentum and Greedy are simple, they outperform other algorithms in a few scenarios (Table 10), when users exactly repeat past actions regardless of their outcomes before reaching a decision. This finding underscores the importance of having relevant baselines to test our modeling assumptions. While these simple heuristics are common in RL and machine learning literature, we observe that *similar natural baselines are often overlooked in the evaluation of sophisticated user models* in visualization [29, 41, 44]. This methodology is even more important for visualization recommendation scenarios, where we know there are already many algorithms available to compare against [61].

References

[1] [n. d.]. Microsoft PowerBI. https://powerbi.microsoft.com.

[2] [n. d.]. Tableau Software. http://www.tableausoftware.com.

[3] Jonathan Back and Charles Oppenheim. 2001. A model of cognitive load for IR: implications for user relevance feedback interaction. *Information Research* 6, 2 (2001), 6–2.

[4] Calvin S Bao, Siyao Li, Sarah G Flores, Michael Correll, and Leilani Battle. 2022. Recommendations for visualization recommendations: Exploring preferences and priorities in public health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[5] Ori Bar El, Tova Milo, and Amit Somech. 2020. Automatically generating data exploration sessions using deep reinforcement learning. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1527–1537.

[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (Oct. 2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[7] Leilani Battle, Remco Chang, and Michael Stonebraker. 2016. Dynamic prefetching of data tiles for interactive visualization. In *Proceedings of the 2016 International Conference on Management of Data*. ACM, 1363–1375.

[8] Leilani Battle and Jeffrey Heer. 2019. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 145–159.

[9] Leilani Battle and Alvitta Ottley. 2022. A programmatic definition of visualization insights, objectives, and tasks. *arXiv preprint arXiv:2206.04767* (2022).

[10] Leilani Battle and Carlos Scheidegger. 2020. A Structured Review of Data Management Technology for Interactive Visualization and Analysis. *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. https://doi.org/10.1109/TVCG.2020.3028891

[11] D. Bennett, G. Davidson, and Y. Niv. 2022. A model of mood as integrated advantage. *Psychological Review* 129, 3 (April 2022), 513–541. https://doi.org/10.1037/rev0000294 Epub 2021 Sep 13.

[12] Daniel Bennett, Yael Niv, and Angela J. Langdon. 2021. Value-free reinforcement learning: policy optimization as a minimal model of operant behavior. *Current opinion in behavioral sciences* 41 (2021).

[13] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. 2014. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics* 20, 12 (2014), 1663–1672.

[14] Robert R Bush and Frederick Mosteller. 1953. A stochastic model with applications to learning. *The Annals of Mathematical Statistics* (1953), 559–585.

[15] Yuxiang Cen, Lu Gan, and Chuanhe Bai. 2013. Reinforcement learning in information searching. *Information Research* 18, 1 (2013), paper 569. http://InformationR.net/ir/18-1/paper569.html

[16] Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. 2011. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron* 69, 6 (March 2011), 1204–1215.

[17] Peter Dayan and Yael Niv. 2008. Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology* 18, 2 (2008), 185–196.

[18] Dazhen Deng, Aoyu Wu, Huamin Qu, and Yingcai Wu. 2022. DashBot: Insight-Driven Dashboard Generation Based on Deep Reinforcement Learning. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–11. https://doi.org/10.1109/tvcg.2022.3209468

[19] Punit Doshi, Elke Rundensteiner, and Matthew Ward. 2003. Prefetching for Visual Data Exploratio. 195– 202. https://doi.org/10.1109/DASFAA.2003.1192383

[20] Will Epperson, Doris Jung-Lin Lee, Leijie Wang, Kunal Agarwal, Aditya G Parameswaran, Dominik Moritz, and Adam Perer. 2022. Leveraging analysis history for improved in situ visualization recommendation. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 145–155.

[21] Mi Feng, Evan Peck, and Lane Harrison. 2018. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 501–511.

[22] Sneha Gathani, Shayan Monadjemi, Alvitta Ottley, and Leilani Battle. 2022. A Grammar-Based Approach for Applying Visualization Taxonomies to Interaction Logs. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 489–500.

[23] Paul W. Glimcher. 2011. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences* 108, supplement_3 (2011), 15647–15654. https://doi.org/10.1073/pnas.1014269108

[24] David Gotz and Zhen Wen. 2009. Behavior-driven visualization recommendation. In *Intelligent User Interfaces*.

[25] David Gotz and Michelle X Zhou. 2009. Characterizing users' visual analytic activity for insight provenance. *Information Visualization* 8, 1 (2009), 42–55.

[26] Tera Marie Green, William Ribarsky, and Brian Fisher. 2009. Building and Applying a Human Cognition Model for Visual Analytics. *Information Visualization* 8, 1 (2009), 1–13. https://doi.org/10.1057/ivs.2008.28

[27] Hua Guo, Steven R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. 2016. A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 51–60. https://doi.org/10.1109/TVCG.2015.2467613

[28] Jacek Gwizdka. 2010. Distribution of cognitive load in web search. *Journal of the American Society for Information Science and Technology* 61, 11 (2010), 2167–2187.

[29] Sunwoo Ha, Shayan Monadjemi, Roman Garnett, and Alvitta Ottley. 2022. A unified comparison of user modeling techniques for predicting data interaction and detecting exploration bias. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 483–492.

[30] Chen He, Luana Micallef, Liye He, Gopal Peddinti, Tero Aittokallio, and Giulio Jacucci. 2021. Characterizing the Quality of Insight by Interactions: A Case Study. *IEEE Transactions on Visualization and Computer Graphics* 27, 8 (2021), 3410–3424. https://doi.org/10.1109/TVCG.2020.2977634

[31] Vijay R Konda and John N Tsitsiklis. 2003. Actor-critic algorithms. *SIAM journal on control and optimization* 42, 4 (2003), 1143–1166.

[32] Heidi Lam. 2008. A framework of interaction costs in information visualization. *IEEE transactions on visualization and computer graphics* 14, 6 (2008), 1149–1156.

[33] Yuan Chang Leong, Angela Radulescu, Reka Daniel, Vivian DeWoskin, and Yael Niv. 2017. Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron* 93, 2 (2017), 451–463.

[34] Zhicheng Liu and Jeffrey Heer. 2014. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2122–2131.

[35] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. 2013. imMens: Real-time visual querying of big data. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 421–430.

[36] Zhicheng Liu and John Stasko. 2010. Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 999–1008. https://doi.org/10.1109/TVCG.2010.177

[37] Jiyun Luo, Sicong Zhang, and Hui Yang. 2014. Win-win search: Dual-agent stochastic game in session search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 587–596.

[38] Ben McCamish, Vahid Ghadakchi, Arash Termehchy, Behrouz Touri, and Liang Huang. 2018. The data interaction game. In *Proceedings of the 2018 International Conference on Management of Data*. 83–98.

[39] Tova Milo and Amit Somech. 2018. Next-step suggestions for modern interactive data analysis platforms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 576–585.

[40] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[41] Shayan Monadjemi, Roman Garnett, and Alvitta Ottley. 2020. Competing models: Inferring exploration patterns and information relevance via Bayesian model selection. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 412–421.

[42] Yael Niv. 2009. Reinforcement learning in the brain. *Journal of Mathematical Psychology* 53, 3 (2009), 139–154.

[43] Yael Niv, Jeffrey A. Edlund, Peter Dayan, and John P. O'Doherty. 2012. Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. 32, 2 (2012), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

[44] Alvitta Ottley, Roman Garnett, and Ran Wan. 2019. Follow the clicks: Learning and anticipating mouse interactions during exploratory data analysis. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 41–52.

[45] Robert E Patterson, Leslie M Blaha, Georges G Grinstein, Kristen K Liggett, David E Kaveney, Kathleen C Sheldon, Paul R Havig, and Jason A Moore. 2014. A human cognition framework for information visualization. *Computers & Graphics* 42 (2014), 42–58.

[46] Adam Perer and Ben Shneiderman. 2008. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 109–118.

[47] Xin Qian, Ryan A. Rossi, Fan Du, Sungchul Kim, Eunyee Koh, Sana Malik, Tak Yeon Lee, and Nesreen K. Ahmed. 2022. Personalized Visualization Recommendation. *ACM Trans. Web* 16, 3, Article 11 (sep 2022), 47 pages. https://doi.org/10.1145/3538703

[48] Gavin A Rummery and Mahesan Niranjan. 1994. On-line q-learning using connectionist systems. In *Proceedings of the 1994 connectionist models summer school*. Citeseer, 1–11.

[49] Mariia Seleznova, Behrooz Omidvar-Tehrani, Sihem Amer-Yahia, and Eric Simon. 2020. Guided exploration of user groups. *Proceedings of the VLDB Endowment (PVLDB)* 13, 9 (2020), 1469–1482.

[50] Danqing Shi, Yang Shi, Xinyue Xu, Nan Chen, Siwei Fu, Hongjin Wu, and Nan Cao. 2019. Task-oriented optimal sequencing of visualization charts. In *2019 IEEE Visualization in Data Science (VDS)*. IEEE, 58–66.

[51] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484–489.

[52] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement learning: An introduction.* MIT press.

[53] Kohei Tamura and Naoki Masuda. 2015. Win-stay lose-shift strategy in formation changes in football. *EPJ Data Science* 4 (2015), 1–19.

[54] Vladimir Vapnik, Steven Golowich, and Alex Smola. 1996. Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems* 9 (1996).

[55] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.

[56] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[57] Bodo Winter. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499* (2013).

[58] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658.

[59] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2017. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2648–2659. https://doi.org/10.1145/3025453.3025768

[60] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. 2020. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 757–783.

[61] Zehua Zeng, Phoebe Moh, Fan Du, Jane Hoffswell, Tak Yeon Lee, Sana Malik, Eunyee Koh, and Leilani Battle. 2021. An evaluation-focused framework for visualization recommendation algorithms. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 346–356.

[62] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–12.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009