

A Unified Comparison of User Modeling Techniques for Predicting Data Interaction and Detecting Exploration Bias

Sunwoo Ha, Shayan Monadjemi, Roman Garnett, and Alvitta Ottley

Abstract—The visual analytics community has proposed several user modeling algorithms to capture and analyze users' interaction behavior in order to assist users in data exploration and insight generation. For example, some can detect exploration biases while others can predict data points that the user will interact with before that interaction occurs. Researchers believe this collection of algorithms can help create more intelligent visual analytics tools. However, the community lacks a rigorous evaluation and comparison of these existing techniques. As a result, there is limited guidance on which method to use and when. Our paper seeks to fill in this missing gap by comparing and ranking eight user modeling algorithms based on their performance on a diverse set of four user study datasets. We analyze exploration bias detection, data interaction prediction, and algorithmic complexity, among other measures. Based on our findings, we highlight open challenges and new directions for analyzing user interactions and visualization provenance.

Index Terms—Visual Analytics, Analytic Provenance, User Interaction Modeling, Machine Learning, Benchmark Study

1 INTRODUCTION

Researchers in the visualization community have long viewed interaction as an *analytic discourse* between the analyst and the visualization system [40]. Thus, capturing and analyzing the user's passive interactions has been integral to the visual analytics research agenda [7, 8, 39, 41, 43]. Some believe that this data can provide a *transcript* of the reasoning process, informing more effective visualization encodings, and producing better intelligent algorithms to assist with data exploration, model refinement, etc. [42, 48]. Further, many have seized the opportunity to leverage machine learning techniques to decode the information embedded in the user's interaction log data. This paper adopts the term used by Xu et al.'s [48] recent survey on the analysis of user interactions and refers to the general goal of understanding the user and their sensemaking process as *user modeling*.

Although there has been significant progress in developing algorithms that can reveal valuable information about the user and their analytic process, a unified comparison of the proposed techniques on different datasets, tasks, and analysis scenarios is lacking. Some suggested methods are presented theoretically without in-depth empirical evaluation, while others are validated using controlled user studies, though often with a single dataset. Furthermore, the algorithms are sometimes proprietary, and the community lacks benchmark datasets for easy comparison. These issues point to fundamental problems that can hinder research progress and question the practicality of user modeling techniques outside of academia.

To address this, we present a computational benchmark study, comparing previously proposed user modeling techniques from the visual analytics community. We first narrow the scope to real-time algorithms that learn from low-level interactions with data points that researchers can potentially use for providing real-time support during data exploration. Such user modeling techniques typically fall under two broad categories: (1) *data interaction prediction*, i.e., inferring data points that the analyst is likely to interact with in the near future and (2) *exploration bias detection*, i.e., detecting data features that are disproportionately explored by the user. These two categories, however, are not mutually exclusive as the successful detection of exploration bias might could

also infer potential data points the user may interact with and vice versa. After surveying the body of work, we selected seven proposed techniques and standardized their input and output specifications to account for a variety of datasets. In addition to the selected models, we developed an ensemble model to predict data interaction and detect exploration bias. We then compared all eight modeling techniques' performance with four different publicly available user study datasets that vary in the number of visual attributes shown and the types of tasks the user study participants performed.

We evaluate the user modeling algorithms based on the success rate of their data interaction prediction, their exploration bias detection, and computational complexity. In addition to the comparison, we highlight the open challenges we discovered. In evaluating these algorithms, we aim to provide a better understanding to researchers and practitioners interested in integrating existing user modeling techniques into their system and to encourage further advances in analyzing user interactions and visualization provenance. The implementations, pre-processed data files, and links to the manuscripts and datasets used in this study will be freely available on GitHub¹ to support reproducibility.

We summarize our contributions as follows:

- Using a unified notation, we provide an overview of existing modeling techniques in the visual analytics community for predicting data interaction and detecting exploration bias.
- We compare and evaluate the performance of eight techniques on two aspects of user modeling with four unique user study interaction log datasets.
- Based on our evaluation, we provide recommendations and new research directions for analyzing user interactions and visualization provenance.

2 RELATED WORK

User modeling techniques can fall into broad categories based on shared objective and can exhibit extreme diversity, as presented by the extensive survey from Xu et al. [48]. For example, researchers have proposed techniques to infer data interactions [23, 32, 38], various forms of bias [21, 32, 45], and even user attributes [3, 29]. Additionally, other techniques aim to assist the user by data prefetching [1, 24], recommending visualizations [22], providing interface guidance [4, 5], and improving data selection [15, 18]. In our work, we aim to evaluate existing models within the categories of data interaction prediction and exploration bias detection. We selected these two categories because of their broad applicability to varying datasets and system designs.

2.1 Exploration Bias Detection

While exploring datasets and making decisions, humans are susceptible to cognitive limitations and biases that arise naturally from perception

¹<https://github.com/washuvis/vis2022usermodels>

- Sunwoo Ha is with Washington University. E-mail: sha@wustl.edu.
- Shayan Monadjemi is with Washington University. E-mail: monadjemi@wustl.edu.
- Roman Garnett is with Washington University. E-mail: garnett@wustl.edu.
- Alvitta Ottley is with Washington University. E-mail: alvitta@wustl.edu.

Manuscript received 31 March 2022; revised 1 July 2022; accepted 8 August 2022.
Date of publication 26 September 2022; date of current version 2 December 2022.
This article has supplementary downloadable material available at <https://doi.org/10.1109/TVCG.2022.3209476>, provided by the authors.
Digital Object Identifier no. 10.1109/TVCG.2022.3209476

and intuition [10, 14]. There are many different types of biases. For example, one form of bias, *confirmation bias*, was described by Nickerson et al. [37] as interpreting evidence in ways that are partial to existing beliefs already in mind. Gigerenzer et al. [19] defined *cognitive bias* as a deviation between human judgment and a rational norm. Lee et al. [27] define *drill-down fallacies* as wrongfully attributing a deviation in trend to a local change, while it is in fact a more general phenomena. Cho et al. [6] investigate the impact of *anchoring bias* (i.e. focusing too much on one piece of information) while making decisions.

This paper focuses on models for *exploration bias*. Exploration bias is akin to *sample selection bias*, a generic phenomenon in social sciences when the investigator fails to capture a random sample of the target population [47]. Much like sample selection bias can lead to biased inferences about a study population, exploration bias can result in false deductions from the data. Using this as motivation, Wall et al. [45] introduced *Attribute Distribution*, which is a metric for quantifying exploration bias by observing user interactions while inspecting individual data points. Their work aimed to draw the connection between cognitive biases and exploration bias (i.e., cognitive biases can cause exploration bias), with proposed subsequent solutions for increasing bias awareness during data analysis [36].

While exploration bias in the context of visual analytics often has a negative connotation, it may at times be intentional and be interpreted as a user's data interest [45]. Other works have investigated several techniques to detect and quantify exploration bias a user may manifest while interacting with datasets. For example, Gotz et al. [21] presented *Adaptive Contextualization*, a statistical approach to detecting exploration bias by observing user interactions while filtering high-dimensional data. In their recent work, Monadjemi et al. [32] detected exploration bias via Bayesian model selection. While the motivation behind each of these modeling techniques may be different, they all share a critical characteristic: in essence, they all compare the distribution of the underlying data set with the distribution of data points with which the user has interacted.

2.2 Data Interaction Prediction

This paper also examines algorithms that predict users' data interactions. There has been several techniques for inferring which data points the user will likely explore in the future. For example, Healy et al. [23] dynamically identified and tagged data elements in their visualization that were of potential interest based on the user's actions using Bayesian classification. Battle et al. [1] modeled a user's sequence of interactions to pre-fetch data for a map visualization. They maintained a Markov chain model of interactions, and ranked the data for pre-fetching according to the likelihood of the users taking actions corresponding to said data. Similarly, Ottley et al. [38] observed users' mouse interaction and succeeded in modeling the attention of users during visual data exploration with a hidden Markov model. Their algorithm leverages a user's observed clicks with a visualization system to obtain next click predictions. In addition to bias detection, Monadjemi et al. [32] showed that their framework can predict future interactions and summarize analytic sessions. Zhou et al. [49] observed users to develop a model of their analytical focus and use that model to surface relevant medical publications to users during visual analysis of a large corpus of medical records. As evident in this brief overview, there are many variations of techniques that learn data points that may be relevant to a user by observing their data exploration. However, the visual analytics literature lacks a unified comparison among these techniques.

2.3 Evaluation of User Modeling Techniques

There are many ways to evaluate an algorithm or technique, including complexity analysis, implementation analysis, and laboratory user study [34]. The choice of evaluation approach depends on the project goals and contributions. As a result, we observed a diverse set of validation methods ranging from crowdsourced user studies and case studies to interaction simulations (e.g., [33, 38, 49]). For this paper, we utilize user study datasets to assess how the current modeling approaches might perform under actual user data to gain insight into their potential

use in a real-time system. Therefore, we review the prior work on evaluated user modeling techniques with user study datasets.

For example, Ottley et al. [38] validated their model's ability to predict future interactions on one user study. While their study validated the technique successfully, it did not evaluate their technique against any baselines. Zhou et al. [49] validated their technique's ability to identify relevant medical concepts via a user study. They successfully validated their technique by comparing it to the ground truth elicited from the study participants and measured a set of usability metrics via a post-study survey. However, their analysis also did not include any baselines. Additionally, Gotz et al. [21] conducted a formal user study to investigate how their technique (Adaptive Contextualization) can mitigate selection bias. Their user study included a two experimental groups: one with access to bias mitigation features, and one without access to such features (baseline). With their work being one of the earlier work in detecting selection bias, they would not have been able to compare against other techniques that were proposed in later years.

Most similar to our work, Monadjemi et al. [32] evaluated the performance of their technique (Competing Models) with two existing techniques by Ottley et al. [38] and Wall et al. [45] for data interaction prediction and bias detection, respectively. Their evaluation, while including baselines and multiple user studies, was still limited in that it did not include all existing techniques as baselines and did not consider as many user studies. We extend upon their work by standardizing and comparing the performance of four additional user modeling techniques with two additional datasets. Moreover, we introduce an ensemble approach that combines the selected modeling techniques for both exploration bias detection and data interaction prediction. We hope that our analysis of these techniques will shed some light for researchers and practitioners who are interested utilizing them.

3 PURPOSE AND SCOPE

This paper presents a benchmark study to compare the user modeling methods proposed in the visual analytics literature, using publicly available user interaction datasets. As demonstrated in section 2, the body of work on analyzing user interactions is diverse. However, instead of a comprehensive comparison of all existing user models, this paper focuses on techniques for predicting data interaction and detecting exploration bias. These categories of user modeling algorithms are complementary. For example, inferring the next data interaction could aid exploration bias detection, and we can use exploration bias inferences to identify data points relevant to the user's current tasks. Furthermore, we can observe from the literature that researchers have already developed algorithms that can already perform both tasks (e.g., [32, 38]). Additionally, we selected these two user modeling categories because of their potential application in providing real-time support for data exploration. We pose the following research questions to guide our evaluations:

RQ1: Which algorithm is most accurate at predicting next data interactions for dataset with (i) goal-driven tasks and (ii) open-ended tasks?

RQ2: Which algorithm is most accurate at detecting exploration bias for datasets with (i) goal-driven tasks and (ii) open-ended tasks?

3.1 Algorithm and Dataset Selection

To form our collection of algorithms and datasets discussed in this paper, we started with the pool of papers that were discussed in Xu et al.'s [48] 2020 survey on the analysis of user interactions and provenance in visual analytics. The authors built a corpus of 105 manuscripts with seed papers collected from provenance-related prior work, supplemented by manually scanning all issues of major journals and all proceedings in the Visualization and HCI communities from 2009 to 2019. We accounted for the papers published around the same time and after the survey by performing a forward citation search on the manuscripts in the survey and the survey itself, resulting in a corpus of 110 manuscripts.

We narrowed the pool to the final selection of seven papers with algorithms within scope [21, 23, 32, 33, 38, 45, 49]. To the best of our

Table 1: A summary of the characteristics of the user models in this study and their overall success on the tested datasets for next interaction prediction.

	Goal	Data	Schema	Complexity		Overall Success
User Modeling Methods	Data Interaction Exploration Bias	Continuous Discrete	Preprocessing Priors Required Off the Shelf	Inference	Prediction	Accuracy
ENSEMBLE	■ ■	■ ■	■ ■ ■	$O(2^d t + 2^d d^3)$	$O(nd^2)$	82%
k -NN: k -Nearest Neighbors [33]	■ ■	■ ■	■ ■ ■	None	$O(nd^2)$	65%
B-BN: Boosted Naive Bayes [23]	■ ■	■ ■	■ ■ ■	$O(nd)$	$O(nd)$	59%
AF: Analytic Focus [49]	■ ■	■ ■	■ ■ ■	$O(cd)$	$O(nd)$	59%
HMM: Hidden Markov Model [38]	■ ■	■ ■	■ ■ ■	$O(p)$	$O(np)$	67%
CM: Competing Models [32]	■ ■	■ ■	■ ■ ■	$O(2^d t + 2^d d^3)$	$O(nd^2)$	68%
AD: Attribute Distribution [45]	■ ■	■ ■	■ ■ ■	None	$O(d)$	
AC: Adaptive Contextualization [21]	■ ■	■ ■	■ ■ ■	None	$O(d)$	

c - # of unique values for an attribute (specific to AF [49]); d - # of data attributes; n - # of points in the underlying dataset;
 p - # of particles (specific to HMM [38]); t - # of observed interactions;

knowledge, this paper includes all real-time algorithms for detecting data interaction prediction and exploration bias that are (i) freely available, (ii) have precise code, or pseudocode descriptions, or (iii) can successfully be implemented following a reasonable amount of troubleshooting and debugging. A few edge cases could theoretically be included but are excluded from this analysis. For example, the *Fore-cache* system proposed by Battle et al. [1] captures movement data as a user navigates a large image, predicts where they will likely explore next, and uses this prediction to drive a prefetching algorithm. We can theoretically reformulate this problem and use the expected next move to infer the user's interaction, assuming the tasks is navigation-related. However, we did not include this or any other algorithm that required such reformulation.

We used a similar process for our dataset selection. We manually examined our corpus of 110 papers to identify manuscripts with user studies and publicly available datasets. Four user study datasets logged participants' interactions with data elements throughout the study session. We included all four datasets in our evaluations.

4 USER MODELING METHODS

In this section, we provide a brief overview on the existing modeling techniques considered in our benchmark study. For most of the selected modeling techniques, their implementations were not readily available. For ones that were readily available, the implementations were hard-coded to work with specific datasets. In order to easily evaluate the modeling techniques across multiple datasets, we implemented every

technique with a standardized syntax and input/output format.

The \mathcal{R} icon indicates that the technique is capable of data interaction prediction. The \mathcal{E} icon indicates that the technique is capable of exploration bias detection. We compare a total of eight user models, based on a variety of theoretical approaches and summarized in Table 1. Of the algorithms we selected, three can only infer next data interaction (k -Nearest Neighbors [33], Boosted Naive Bayes [23], and Analytic Focus [21]), two can only detect exploration bias (Attribute Distribution [45] and Adaptive Contextualization [21]), and two can infer both (Hidden Markov Model [38] and Competing Models [32]). In addition to the selected algorithms, we also introduce and evaluate an ensemble model, developed by averaging the output of all modeling techniques.

Each model assumes there is a visualization of a dataset $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$, where each data point in \mathcal{D} corresponds to a visual element. Each data point $x_i \in \mathcal{D}$ has d attributes which may be continuous, categorical, or ordinal. Users may take different actions while interacting with data points using the visualization (i.e. click, hover, etc.). Thus, the models also assume a non-empty set of interactions supported by the system as $\mathcal{C} = \{c_1, c_2, \dots\}$. As users interact the visualized data points, they maintain a set of observation $\mathcal{I} = \{(x_{i,t}, c) \mid t = 1, 2, \dots\}$, where $x_{i,t}$ denotes an interaction of type c with data point x_i at time step t .

4.1 k -Nearest Neighbors Classifier as seen in [33] \mathcal{R}

A straightforward method for learning users' data interest by observing interactions is to train a k -nearest neighbors (k -NN) binary classifier. This approach assumes that proximity drives a user's exploration patterns. Monadjemi et al. [33] built a k -NN classifier that relies on a notion of distance between data points and computing the matrix of k nearest neighbors for every data point given the defined distance formula as a pre-processing step. The complexity for this pre-processing step is $O(n^2 d)$. Refer to Table 1 for more information and complexity variables. This model assumes that every data point in the dataset, $x_i \in \mathcal{D}$, has a binary label $y_i \in \{0, 1\}$. A label of $y_i = 1$ means the data point x_i is relevant to the task at hand, whereas a label of $y_i = 0$ means x_i is irrelevant. By considering the labels of the nearest neighbors, this model is able to provide us with the probability of any given point being relevant in light of past observations: $\Pr(y_i = 1 \mid x_i, \mathcal{I})$. Hence, we use this posterior belief as our ranking function:

$$f_r(x_i) = \Pr(y_i = 1 \mid x_i, \mathcal{I}). \quad (1)$$

4.2 Boosted Naive Bayes Classifier as seen in [23] \mathcal{R}

Healey et al. [23] propose a naive Bayes classifier for maintaining a belief over users' interest in data points. This approach tracks the frequencies of attributes explored and assumes that the user's latent data interest is related to the dataset attributes' occurrence rate. In

Table 2: Notations used in this section.

Set Notation	Description
$\mathcal{D} = \{x_1, x_2, \dots, x_n\}$	The set of data points visualized.
$\mathcal{A} = \{a_1, a_2, \dots, a_d\}$	The set of data attributes.
$\mathcal{E} = \{e_1, e_2, \dots, e_m\}$	The set of interaction types (e.g. hover, click, etc.).
$\mathcal{I} = \{(x_{i,t}, c) \mid t = 1, 2, \dots\}$	The sequence of interactions in a session. $(x_{i,t}, c)$ denotes taking action c on data point x_i at time step t .
$f_r: \mathcal{D} \mapsto [0, 1]$	The data ranking function for next data interaction prediction. \mathcal{D} , is mapped to a numerical value between 0 (low) and 1 (high).
$f_b: \mathcal{A} \mapsto [0, 1]$	The bias function where the set of attributes, \mathcal{A} , is mapped to a numerical value between 0 (low) and 1 (high).

a similar formulation as the k -NN approach, each data point $x_i \in \mathcal{D}$ has a binary label $y_i \in \{0, 1\}$. A label of $y_i = 1$ means the point x_i is relevant to the task at hand and a label of $y_i = 0$ means it is not relevant. The goal of this approach is to reason about the unknown labels in light of frequencies of observations calculated via the Bayes' law:

$$\Pr(y_i = 1 \mid x_i) \propto \Pr(x_i \mid y_i = 1) \Pr(y_i = 1). \quad (2)$$

Extending x_i to each of its individual d dimensions and assuming conditional independence among attributes, we get:

$$\begin{aligned} & \Pr(y_i = 1 \mid x_{i,1}, x_{i,2}, \dots, x_{i,d}) \\ & \propto \Pr(x_{i,1}, x_{i,2}, \dots, x_{i,d} \mid y_i = 1) \Pr(y_i = 1) \\ & = \Pr(y_i = 1) \prod_{j=1}^d \Pr(x_{i,j} \mid y_i = 1), \end{aligned} \quad (3)$$

where $x_{i,j}$ denotes the j th attribute of the i th data point. Therefore, we get the ranking function as:

$$f_r(x_i) = \Pr(y_i = 1) \prod_{j=1}^d \Pr(x_{i,j} \mid y_i = 1). \quad (4)$$

While Naive Bayes often performs as well as more sophisticated models in practice [26], its performance can further be improved by *boosting*. As recommended by Healey et al. [23], we used AdaBoost to train a stronger classifier.

4.3 Analytic Focus Modeling as seen in [49]

Zhou et al. [49] propose a technique for modeling *analytic focus* (AF) during a session. Their technique is based on an abstract notion of *concepts* defined as “*meaningful data attribute[s] in the problem domain.*” This technique tracks user focus on each of the concepts by observing their interactions and maintain an importance score for each concept. **The model assumes that the user's latent data interest is related to the occurrence rate of discrete concepts observed in their data interactions.** With concepts being left open for practitioners to define in their particular domain, here, we let each unique value appearing in the discrete attributes of the dataset to represent a concept, performing a pre-processing step on continuous attributes to convert them into discrete bins.

Each interaction α is defined by two values: an initial importance score $I_\alpha(0)$ and a persistence score P_α . In a sense, $I_\alpha(0)$ refers to how intentional an action is and how informative it is in revealing analytic focus while P_α is a measure of how long the visual changes resulting from action α remain. Using these parameters and the Ebbinghaus forgetting curve [12, 13, 35], the authors define the per-action importance score function t time steps after an action is taken as:

$$I_\alpha(t) = I_\alpha(0) \times e^{-\frac{t}{P_\alpha}}. \quad (5)$$

Using the equation above, the authors propose an additive model for calculating the importance score for each concept c at time τ :

$$I^c(\tau) = \sum_{(x_i, t, \alpha) \in \mathcal{I}_{\tau, c}} I_\alpha(\tau - t), \quad (6)$$

where $\mathcal{I}_{\tau, c}$ denotes the set of interactions with points involving concept c up until time step τ . Considering that each data point $x_i \in \mathcal{D}$ may subscribe to multiple concepts (one concept per dimension in our case), we extend this method so that the importance of a data point is the product of the importance of its individual concepts. In other words, given a data point $x_i \in \mathcal{D}$ and the current time step τ ,

$$f_r(x_i) = \prod_{c \in C(x_i)} I^c(\tau), \quad (7)$$

where $C(x_i)$ is the list of concepts with which x_i is associated.

4.4 Hidden Markov Model as seen in [38]

Ottley et al. [38] propose a *Hidden Markov Model* (HMM) approach for modeling user attention during visual exploratory analysis. **Their approach assumes that the visual features of the interface will drive the user's attention (it considers all possible combinations)** and their interactions such that attention at time $t + 1$ will be similar to the observation at time t . They constructed a hidden Markov model, presuming the user's behavior evolves under a Markov process (that is, the behavior at a particular time only depends on their behavior at the previous time step), and interaction events are generated conditionally independently given this sequence of attention shifts. The set of visible visual features (e.g., color, position, size, etc.) creates the *mark space* and the latent states. Further, their algorithm uses particle filtering [11, 20] to sample from the model's posterior distribution and predict the user's attention for the next time step. Integral to this technique is a *bias vector* that captures the relative importance of the mark space components. We use the particle filter output to predict next data interaction and the bias vector for bias detection.

4.5 Competing Models as seen in [32]

Monadjemi et al. [32] propose a Bayesian model selection approach for detecting exploration bias and inferring next data interaction during visual exploratory analysis. This approach, called *Competing Models* (CM), enumerates a set of probabilistic models each of which represent exploration based on a subset of data attributes. Given a dataset with d attributes, they enumerate the model space $\mathbf{M} = \{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{2^d}\}$ where each model represents exploration based on one subset of attributes. **Thus, it assumes and one or more data attributes (it considers all possible combinations) are sufficient to represent a users data interest.** As interactions arrive, they maintain a belief $\Pr(\mathcal{M}_i \mid \mathcal{I})$ which is interpreted as the viability of model \mathcal{M}_i in explaining user interactions. To rank the data points to the task at hand in light of past interactions, they use Bayesian model averaging to get:

$$f_r(x_i) = \Pr(x_i \mid \mathcal{I}) = \sum_{j=0}^{2^d} \Pr(x_i \mid \mathcal{I}, \mathcal{M}_j). \quad (8)$$

To quantify bias towards each attribute, they use the law of total probability as follows:

$$f_b(a_i) = \sum_{\mathcal{M}_j \in \mathbf{M}_{a_i}} \Pr(\mathcal{M}_j \mid \mathcal{I}), \quad (9)$$

where $\mathbf{M}_{a_i} \subset \mathbf{M}$ denotes the set of models involving attribute a_i .

4.6 Attribute Distribution as seen in [45]

Wall et al. [45] propose a set of metrics for quantitatively detecting various types of cognitive biases. The most relevant to our work is the metric they call *attribute distribution*. **Their approach assumes that in an unbiased exploratory session, the distribution of values explored by the user will resemble the distribution of the underlying data.** It relies on well-known statistical tests, Chi-Square [31] and Kolmogorov-Smirnov [30], which are used for discrete and continuous attributes respectively. In each case, the test takes the underlying data as well as the set of points with which the user has interacted as inputs and returns a p -value indicating whether or not the two samples could be reasonably drawn from the same distribution. Then, they define their attribute distribution metric for a given attribute $a_i \in \mathcal{A}$ to be:

$$f_b(a_i) = 1 - p\text{-value}_{a_i}. \quad (10)$$

4.7 Adaptive Contextualization as seen in [21]

Gotz et al. [21] propose *Adaptive Contextualization*, a metric to measure selection bias in real-time. In order to detect such biases, they propose measuring the Hellinger distance between the datasets before and after filtering. The Hellinger distance is a statistical measure designed to quantify the similarity between two distributions. Thus, **this**

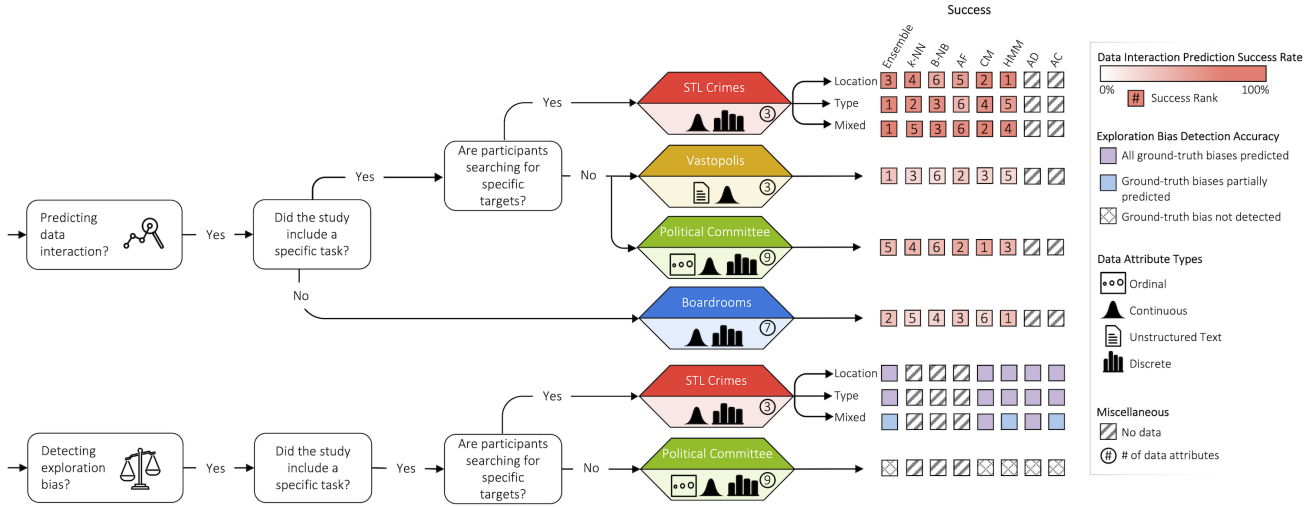


Fig. 1: A summary of the characteristics of the selected user interaction datasets and the performance of the user modeling techniques with said datasets. We represent success using each technique’s performance at top-100 points for next data interaction prediction.

approach assumes that we can detect bias by comparing the observed distribution to the expected data distribution. The authors propose using the Hellinger distance defined for discrete probability distributions with a pre-processing step to discretize the continuous attribute into discrete bins.

Consider a particular dimension of interest, a_i , represented as a set of m discrete values $\{v_1, v_2, \dots, v_m\}$. The distribution of these values in the underlying data and interaction data are $\{p_1, p_2, \dots, p_m\}$ and $\{q_1, q_2, \dots, q_m\}$ respectively. We compute the Hellinger distance as:

$$f_b(a_i) = H(\mathcal{D}_{a_i}, \mathcal{I}_{a_i}) = \sqrt{\frac{1}{2} \sum_{j=1}^m (\sqrt{p_j} + \sqrt{q_j})^2}, \quad (11)$$

where \mathcal{D}_{a_i} and \mathcal{I}_{a_i} denote only the a_i attribute of the underlying and interaction datasets respectively. While this technique was intended to detect selection bias (i.e. before vs. after filtering), we adopt it to quantify exploration bias.

4.8 Ensemble Approach

We anticipate that no model will be able to correctly recognize relevant points in every system, task, or dataset. This phenomena was also observed by the machine learning community which led to the development of *ensemble methods* [9, 44]. Ensembles are a collection of classification models combined to construct a model with higher predictive power. Using this approach, we consider a final technique which combines predictions from all five techniques that predict data interactions and all four techniques that detect exploration bias reviewed in this paper. To predict next data interaction, for a given data point, $x_i \in \mathcal{D}$, we compute the ranking function, $f_r(x_i)$, using each of the modeling techniques and average the ranking values to get the ensemble prediction. Then, we proceed with ranking the data points, a process identical to the one in the preceding modeling techniques. Similarly, we average the bias values towards each data attribute from each modeling technique to get the ensemble prediction for exploration bias.

5 USER STUDY INTERACTION LOGS

We used four publicly available interaction logs from user study experiments with visualization systems. These studies contained varying interfaces as well as task designs. We aimed to evaluate and compare the performance of the user modeling techniques with a variety of interaction logs, ranging from studies with very directed and specific tasks to the polar opposite, open data exploration. Refer to Figure 1 for a detailed summary of the characteristics of each dataset.

5.1 STL Crimes as seen in [38]

The *STL Crimes* dataset involved a highly specified task thus capturing less exploration noise. We selected this dataset as it represents the best-case observational scenario for user modeling.

Curated by Ottley et al. [38], this dataset consists of 1,951 crime instances reported in the month of March 2017 with a total of eighteen attributes. Each instance of crime on the map was displayed as a dot with a position and color, indicating the location and type of the crime respectively. Eight types of crime exist in the database: homicide, theft-related, assault, arson, fraud, vandalism, weapons, and vagrancy. 30 participants were recruited from Mechanical Turk and they completed three categories of directed search tasks that were designed to encourage them to either click on: (1) *Type-Based* data points with similar crime categories (e.g. Homicide, Assault, etc.), (2) *Location-Based* data points that are within in the same vicinity, or (3) *Mixed* data points that are of the same crime category and in the same vicinity. Each participant completed a total of six tasks, two from each category. Since the authors specified to the participants target attributes to interact with, this provides us with a notion of ground-truth bias. In their user study, Ottley et al. [38] captured mouse click data as participants interacted the visualization. Consistent with Ottley et al. [38], we filtered the data to hold only the interactions of participants who successfully answered the task questions and created three data subsets: 28 user sessions for location-based tasks, 23 for typed-based, and 27 for mixed.

5.2 Vastopolis as seen in [33]

We selected the *Vastopolis* dataset because of the attributes in the dataset. **It is the only user interaction dataset in this study that includes unstructured text.**

Inspired by the 2011 VAST Challenge, Monadjemi et al.’s [33] study describes a major epidemic that started in the fictitious city of Vastopolis. Participant’s interacted with a map of Vastopolis with 3000 geolocated tweet-like data. In addition to the social media posts, the map display major roadways, waterways, and landmarks in the city. They recruited 130 participants from Amazon Mechanical Turk and tasked each participant to search through the dataset of microblogs via an interactive map and bookmark as many posts containing illness-related information as possible. Following the same pre-processing step as the authors, 74 user interaction logs remained from the control group of the study. The only interaction type we considered in the interaction logs were intentional “bookmarks” of the microblogs.

5.3 Political Committee as seen in [46]

The *Political Committee* user study examined cognitive and exploration bias as participants selected their picks for a hypothetical

committee. Thus we selected this dataset as a candidate for examining exploration bias.

Wall et al. [46] generated a dataset of 180 fictitious politicians, representing the composition of the Georgia General Assembly. The dataset contains three discrete and six continuous attributes that characterizes each politician. In the study conducted by Wall et al. [46], the authors had users interact with an visualization system that supports the exploration of the fictitious political committee data. There were two versions of the visualization system: a Control version of the interface, and an Intervention version of the interface, which was modified to visualize traces of the user's interactions with the data in real-time. To assess the effectiveness of visualizing the user's interaction traces, the authors performed an in-lab study with 24 participants. They were tasked with selecting a committee of 10 candidates to review public opinion in Georgia on the controversial bill that bans abortion after 6 weeks. In order to observe the raw, inherent biases of the users, we selected the interaction logs of those in the Control group (total of 12 participants). The authors recorded several types of interactions the user can make with the interface such as hovers, changing of axes on the scatterplot, committee selection etc. We filtered the interaction log for each user to contain only committee selection interactions with the data points.

5.4 Boardrooms as seen in [16]



The *Boardrooms* study did not involve a particular task, but instructed the participants to freely interact with the visualisation. **We expect this dataset to include the highest levels of exploration noise.**

Feng et al. [16, 17] utilized the visualization, "Inside America's Boardroom," which was published by the *Wall Street Journal* [28] in 2016. This point-based visualization displays seven different attributes to the user: market capitalization, ratio of unrelated board members, ratio of female board members, average age of board members, average tenure of board members, and median pay of board members. The study adopted an open-ended approach and observed the user's interactions as they explored the dataset through the visualization. The interaction log contains a succession of hovers each user made on different data points. To account for unintentional hovers, we followed the same approach as Monadjemi et al. [32], each user's session was filtered only include hovers that lasted for over one second. Additionally, we only considered sessions that consisted of more than three hovers, resulting in a total of 39 user sessions.

6 EVALUATION MEASURES

In this section, we describe the performances measures we used to evaluate the modeling techniques' accuracy in predicting data interactions and detecting exploration bias with the four user study logs. We note that 2 out of 4 interaction logs included a single interaction type. In particular, the *STL Crimes* captured only mouse clicks and *Boardrooms* recorded mouse hovers alone. *Vastopolis* included bookmarks and hovers, but recording inconsistencies in the logging made it impossible to integrate the two interaction modalities. Thus, our evaluation focused solely on predictions based on the more intentional "bookmark" interaction. Finally, the *Political Committee* dataset included the most diverse interface and data interactions (e.g., hovers, reconfiguring, committee selections). For consistency, we select a single interaction observation (committee selections), and the algorithms' predictions are limited to their observations. Suppose, for example, the algorithm's observations are limited to mouse clicks. Then the algorithm can only predict the next click or detect data exploration bias among the clicked data points. In the case of *Political Committee*, this means that the algorithms can only observe when the user selects a committee member and must make predictions about the next committee selection or estimate selection biases solely from the limited observations.

6.1 Evaluating Data Interaction Prediction

We developed a standardized probabilistic output for each algorithm to compare the models to each other. Specifically, the techniques treat every data point as a potential candidate for being most relevant to the user's data interest. Thus, each method outputs the ranking probability

for each data point given its observations. This output represents the technique's unique interpretation of the data points that the user will likely interact with. For example, k -NN assumes that the data points most relevant to the user will minimize some measure of distance while **AF** examines the frequency on *concepts* from the past interactions.

Given the model's probabilities, we can rank each data point based on the algorithm's belief. We report top- κ data retrieval for the algorithms for $\kappa \in \{1, 5, 10, 20, 50, 100\}$ which we will refer to as the *prediction sets*. We considered a variety of set sizes. For example, an alternative is to use percentages instead of a fixed values. However, for large datasets, retrieving a small percentage to the dataset can still be overwhelming to the user if we consider the goal of provided real-time analysis support. This standardized output enables us to evaluate and compare each technique's ability to predict the next data interaction. We record the following measures for every interaction the datasets:

- **SUCCESS:** $\in \{0, 1\}$, binary value for whether the next data interaction was included in the prediction set for all values of κ .
- **RANK:** the ranked position of the next observed interaction such that lower is better.

6.2 Evaluating Exploration Bias Detection

Evaluating each technique's ability to detect exploration bias poses a unique challenge, in that *ground truth* is not always readily available. When evaluating data interaction prediction, we had access to a notion of *success*: we wanted to predict next data points with which the user interacts, and we were able to verify our ability to do so by observing the data interaction point. For bias selection, however, such ground truth is not as easy to access, especially for user study datasets with open-ended tasks. With the *Political Committee* dataset collected by Wall et al. [46], for example, each user can exhibit their own set of biases as they explore the data.

For this set of evaluations, we use two datasets: *STL Crimes* separated into *location-based*, *type-based*, and *mixed* subsets and *Political Committee*. The *STL Crimes* user study involved a directed task, asking participant to interaction with specific portions of the data, inherently resulting in biased data exploration. For example, the type-based task instructed participants to inspect all "Arsons" and note the time of day they occurred. We used the specific task as the ground truth. In our evaluation, an algorithm successfully detects exploration bias if its outputs indicate data attributes related to crime location, type, or a mixture model of type and location for *location-based*, *type-based*, and *mixed* respectively.

For *Political Committee*, two researchers performed qualitative coding on each study participants' committee selections, following best practices [2]. We used an iterative approach. First, the coders reviewed a single participant's selections and density plots showing the attribute distributions for a given set of interaction logs and the original data distribution. Their goal was to identify evidence of data bias in the

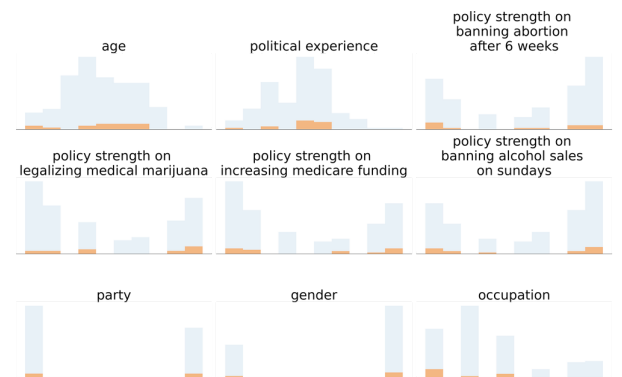


Fig. 2: Comparing participant *Lima*'s interactions (foreground) to the full dataset distribution (background). Both coders reported no significant difference between the two distributions, indicating no bias detected. The result were similar for all 12 study participants from [46].

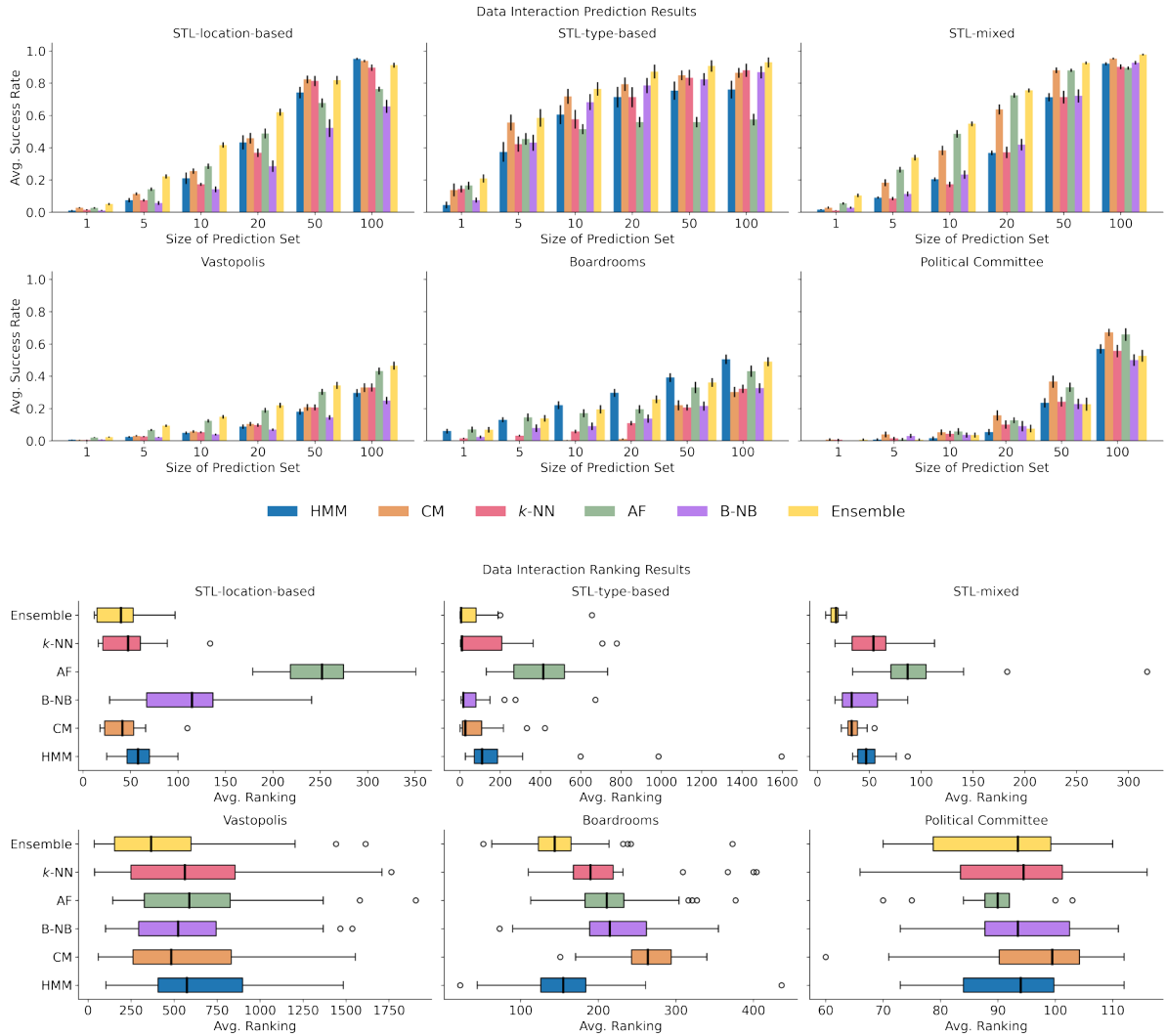


Fig. 3: We evaluated each model’s performance with two metrics: (top) average **success rate** across all the tasks in each user study dataset for varying values of the prediction set size and (bottom) average **rank** of the next data interaction of the user modeling techniques.

observed selections. Evidence of data bias was determined if the attribute distributions for a given set of interaction logs was different than original data distributions. Next, the coders established a codebook from this first round and independently coded all twelve participants’ interaction logs. After inspecting each of the twelve study participants, both coders agreed that there was no observable bias in the any of the interaction logs. For example, Figure 2 compares study participant Lima’s data selection to the individual attribute distribution. No significant difference was observed between the two distributions for each attribute. Although the coding sessions detected no bias, we included this analysis to assess the algorithms’ performance in detecting bias when none exists.

7 RESULTS

7.1 Data Interaction Prediction

Figure 3 shows the models’ accuracy across different values of $\kappa \in \{1, 5, 10, 20, 50, 100\}$, the size of the prediction. We analyzed each algorithms’ success rate at predicting the next observed interaction, calculating the overall success rate for all available user interaction traces for each dataset as $\sum success \div \sum predictions$. For simplicity, the results in this section focused on $\kappa = 100$, i.e., the algorithm picks the 100 points with the highest rankings for the next interaction.

STL Crimes: The top row of Figure 3 shows the three subtasks in the

in *STL Crimes* dataset (*location-based*, *type-based*, and *mixed*). We can observe that most of the techniques can predict the next data interaction point with high success rate at $\kappa = 100$, which represents 5% of the underlying data. For location-based tasks, **HMM** and **CM** are the best performing models, with success rates of 95% and 94% at $\kappa = 100$, respectively. With the type-based task, **ENSEMBLE** performs the best with a success rate of 93% at $\kappa = 100$. For the mixed-based task, **ENSEMBLE** outperforms all the other models, successfully predicting the next data interaction 98% of the time.

Vastopolis: The *STL Crimes* user study was highly directed, asking users to perform specific tasks that lead their exploration. On the other hand, the *Vastopolis* task was less directed, involved a larger data volume, and included textual data (participants interacted with Tweet-like data). The selected models are unable to reach high success rate for data interaction prediction, and success rates were consistently less than 50% at $\kappa = 100$ points. Nonetheless, **ENSEMBLE** and **AF** perform the best with this interaction log dataset.

Boardrooms: The user study sessions with the *Boardrooms* dataset were highly open-ended. Due to the open nature of the task, all of the models had difficulty achieving high success rate for data interaction prediction. **HMM** outperforms all the other modeling techniques with a success rate of 51% at $\kappa = 100$.

Political Committee: With the *Political Committee* interaction logs,

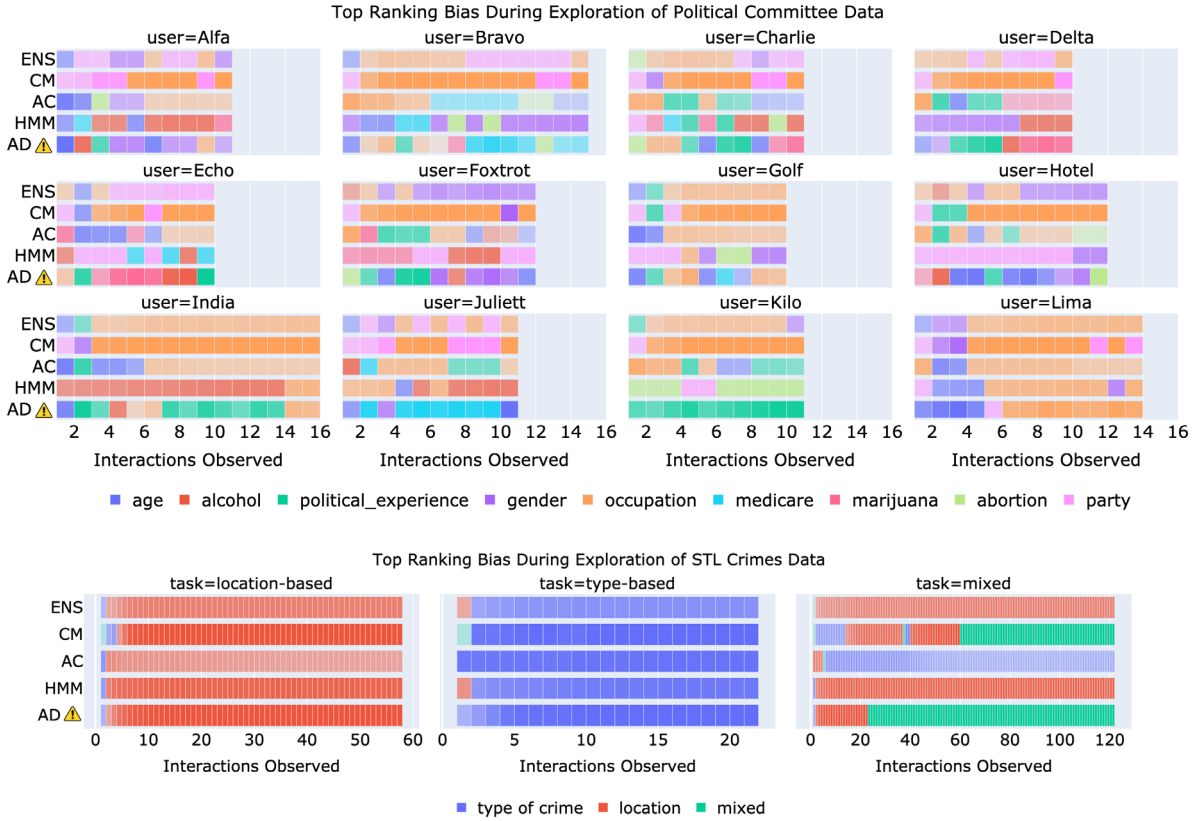


Fig. 4: Each block represents a certain time t within the observed interactions. The color represents the top-ranking attribute and the transparency level represents the technique’s confidence. (⚠ indicates that an assumption of the Chi-Square test was not met.)

we wanted to stress-test the modeling techniques and observe their performance on set of minimal, but *intentional* interactions. However, is it important to note that the dataset set is comparably small, with 180 fictitious politicians. CM and AF showed promising performance with success rates of 67% and 66% at $\kappa = 100$, respectively.

7.2 Exploration Bias Detection

For the evaluation of bias detection, we chose the two following interaction datasets: *STL Crimes* and *Political Committee*. These two datasets were chosen as both studies either aimed to encourage specific biases or elicit the users’ inherent biases during data exploration. Although we include the results of AD in this section, we note that its performance may be unreliable as one of the Chi-Square test assumptions was not met. The Chi-Square test assumes that: *the value of the expected cell should be 5 or more in at least 80% of the cells, and no cell should have an expected of less than one* [25].

Political Committee: Each algorithm outputs a *bias value* for the attributes in the dataset. For example, *Political Committee* has nine attributes including (age, gender, and occupation, etc.) and based on the assigned values, we can rank the attributes by the techniques’ bias belief. For simplicity, Figure 4 and the analysis in this section considers only the attribute with the highest bias belief after each user’s observation. As detailed in section 6.2 our qualitative coding revealed no significant differences between the distribution of values explored by the user and the underlying data distribution, meaning no ground-truth bias was detected for every session. Although there is no evidence of exploration bias, the top of Figure 4 shows the top-ranking bias found by each technique for every user throughout their data exploration. The absence of a ground-truth bias is potentially depicted by the lack of agreement among the techniques for the top-ranking bias. However, we see some agreement for *Lima*, as the techniques believed their exploration was biased towards the occupation attribute.

STL Crimes: The *STL Crimes* user study visually displayed three

attributes (longitude, latitude, and type of crime) and included three subtasks (*location-based*, *type-based*, and *mixed*). Since users were tasked to specifically interact with subsets of the data, we set these attributes as ground-truth biases. To calculate the bias value of location, we multiplied the bias values for longitude and latitude. To calculate the mixed bias value, we multiplied the location bias with the bias value for type of crime. The bottom row of Figure 4 shows the aggregated top-ranking bias among the users over time. For both location-based and type-based tasks, all of the modeling techniques were able to accurately determine the attribute that users were most biased towards. However, for the mixed task, we observed that AD and CM were the only modeling techniques that detected that the users were interested in both type and location attributes. For ENSEMBLE, AC, and HMM, the techniques believed the users were focused on only type-based and only location-based attributes, respectively.

8 DISCUSSION

This study compared eight algorithms for data interaction prediction and exploration bias detection across four user study datasets. *Perhaps the most salient finding is there is no clear ‘winner.’* In particular, we found that no single model excelled with all datasets and the differences between the top-ranking techniques are small.

8.1 Insights from Data Interaction Prediction Results

The *STL Crimes* data subsets demonstrate that the models generally perform relatively well when the tasks are highly-specified and have limited exploration noise. We use *STL Crimes* to represent a best-case modeling scenario. Equally important, our evaluation included datasets that were selected to test the algorithms’ robustness. In particular, *Vastopolis* includes unstructured text, *Boardrooms* involves an open exploration task, and *Political Committee* presents a limited observation window with 10-16 interactions per participant. We can observe that the models’ success rates fell drastically with these less directed and open-ended user study datasets. Still, there are a few noteworthy discoveries

from this analysis. For example, **AF** [49] (developed initially for textual data exploration) achieved the highest overall success rate with the *Vastopolis* dataset, the only dataset with unstructured text. In addition, **HMM** [38], surpassed the others on the only dataset for which the task was entirely open-ended. However, in both instances, the overall success rates were relatively low.

Given the high variability in the success rates, we introduced an **ENSEMBLE** approach [9, 44] by averaging the computed probabilities, allowing each model to vote for the data points they believe would best match the users' next data interactions in light of its observations. Then, we rank the data points based on their probability mass and select the top- κ accordingly for predictions. The **ENSEMBLE** model was among the top predictors for all but one dataset. We saw that averaging the predictions resulted in **ENSEMBLE** achieving the second worst overall performance with the *Political Committee* dataset, which we included to represent a limited observation scenario. Although we observe an overall improved performance in data interaction prediction with **ENSEMBLE**, especially when κ is small, we do note that this approach can be costly to compute and may not be suitable for real-time systems.

8.2 Insights from Exploration Bias Detection Results

The *STL Crimes* dataset also provided a convenient means for evaluating exploration bias. In particular, the three data subsets involved tasks in which the study participants either focused on a specified location, data type, or a mixture of the two, allowing us to use these as ground truths. All five tested algorithms successfully identified that participants disproportionately allocated interactions toward location and type-related attributes. However, the most interesting finding relates to the mixed task data subset. Specifically, we can observe that only **CM** [32] and **AD** [45], accurately detected that the task elicited a mixed bias. Furthermore, the exploration bias detection findings are notably insightful for the **CM** and **HMM** techniques – the only two algorithms capable of both data interaction prediction and exploration bias detection. Although both algorithms achieved more than a 90% success rate in predicting the next data interaction at $\kappa = 100$, the exploration bias detection results provide further insight into the models' beliefs and how they produced their data relevance rankings. Markedly, correctly identifying the mixed nature of the task meant that **CM** was one of the best data interaction predictors, second to only **ENSEMBLE** for this data subset. In contrast, misclassifying the task as location-based resulted in an overall lower prediction performance for the **HMM** approach.

Although our qualitative coding failed to detect exploration bias in the committee picks from participants in the *Political Committee* dataset, we included the evaluation results to assess the algorithms' predictions in light of unbiased interaction traces. We observed that all the models presented false positives, often with high certainty, as indicated by the opacity of the blocks in Figure 4. It is noteworthy that the one underlying assumption of some of the models is that the user is always biased. For example, the **HMM** algorithm explicitly includes bias in the algorithm's specification. Specifically, it maintains a *bias probability* for each feature in the dataset, allowing the algorithm to capture the likelihood of bias toward one or more data features. **CM** is the only algorithm that directly encodes the potential for no bias. The algorithm tracks bias by building a model for all possible feature subsets, including the null set, representing no potential exploration bias. However, the null set is only one of the numerous model choices considered in **CM**, making it an unlikely pick.

Altogether, this underlying assumption that the user's interaction is always biased has both positives and negatives for real-world applications. On the one hand, we believe that identifying the features that the user disproportionately explores drives the algorithms' success in data interaction prediction. Also, it is potentially a reasonable assumption for many exploration scenarios. On the other hand, the false positives can erode the user's trust in the algorithms, especially in proposed applications related to helping the user recognize their own exploration bias or recommender systems [21, 45, 46]. That said, a potential solution exists within two models, **AC** [21] and **AD** [45], which leverage statistical methods for hypothesis testing in their bias detection. If used as intended, a failure to reject the null hypothesis

provides evidence of no bias. However, practitioners and researchers should use the appropriate tests for the given data and verify that the data meet their assumptions.

8.3 Challenges and Limitations

There were several challenges we faced when preparing to perform a unified comparison of these models. For example, implementations of the techniques were not readily available. We invested time gaining an understanding of all the techniques and standardizing the implementations to handle various datasets. We note that we are not able to completely replicate the performance of **CM** and **HMM** on data interaction prediction with the *STL Crimes* dataset that we see in Monadjemi et al [32]. We believe that the differences in performance are likely due to our hyperparameters selections. When deploying algorithms in real-world settings, it is important to gather data and tune hyperparameters. However, we excluded this step in our work as the datasets were small, presenting a high risk for overfitting. Instead, we used our domain knowledge to make conscientious decisions about each model's hyperparameters, which are all available and adjustable in our codebase.

Although we selected a diverse collection of datasets, we acknowledge that they do not represent many visual analytics scenarios. Therefore, we take care in reporting our observations and hesitate to make broad statements. Further, we were limited by the small number of interaction datasets that are publicly available, especially those with ground truths. This highlights important considerations for balancing ecological validity where the ground truth is not known or may not exist, and having some measure of success for evaluation. Determining ground truth when evaluating bias in open-ended tasks is not trivial. A call for more guided user studies with built-in biases may be needed so that researchers can further evaluate bias detection algorithms.

9 FUTURE DIRECTIONS

Despite our limitations, our work is a step forward in evaluating user modeling techniques proposed in the visual analytics community and opens possibilities for future work in this area. For example, our evaluations only involved datasets with a single interaction type, even though some of the modeling techniques are able to or have the potential to handle multiple types of interactions [32, 38, 49]. Although our implementations did not include this feature, future work could extend this benchmark study with the added ability to learn from multiple types of interactions with data points. Further, we primarily focused on quantitative measures, but there are other factors beyond accuracy that might affect the appropriateness of a technique for a given user modeling scenario. Factors such as speed, trustworthiness, and flexibility can be important considerations to explore in future evaluations of user modeling techniques.

10 CONCLUSION

In this paper, we present a computational benchmark study that provides a unified comparison of user modeling techniques for data interaction prediction and exploration bias detection. We implemented standardized versions of seven previously proposed user modeling techniques that learns from a user's low-level interactions in real-time. Additionally, we developed an ensemble approach by averaging the models' predictions. We then evaluated the performance of all these techniques across four different interaction logs for data interaction prediction/data ranking and exploration bias detection. We found that there is no clear 'winner' among the modeling techniques, but our analysis highlights the tasks and datasets that elicited the best performance for each technique. Finally, we discuss the open challenges for user modeling and evaluations. This work is a step towards gaining an understanding of user modeling techniques within the visual analytics community and we hope that it encourages further advances in analyzing user interactions and visualization provenance.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under Grant No. OAC-2118201 and IIS-2142977.

REFERENCES

- [1] L. Battle, R. Chang, and M. Stonebraker. Dynamic prefetching of data tiles for interactive visualization. In *Proceedings of the 2016 International Conference on Management of Data*, pp. 1363–1375, 2016.
- [2] M. Brod, L. E. Tesler, and T. L. Christensen. Qualitative research and content validity: developing best practices based on science and experience. *Quality of life research*, 18(9):1263–1278, 2009.
- [3] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*, 20(12):1663–1672, 2014.
- [4] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):111–120, 2016.
- [5] D. Ceneda, T. Gschwandtner, and S. Miksch. A review of guidance approaches in visual data analysis: A multifocal perspective. In *Computer Graphics Forum*, vol. 38, pp. 861–879. Wiley Online Library, 2019.
- [6] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 116–126. IEEE, 2017.
- [7] K. A. Cook and J. J. Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.
- [8] R. J. Crouser, A. Ottley, and R. Chang. Balancing human and machine contributions in human computation systems. In *Handbook of Human Computation*, pp. 615–623. Springer, 2013.
- [9] T. G. Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pp. 1–15. Springer, 2000.
- [10] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics*, 26(2):1413–1432, 2018.
- [11] A. Doucet, S. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.
- [12] H. Ebbinghaus. *Urmanuskript“ ueber das gedächtniß. Passau: Passavia Universitätsverlag*, 1880.
- [13] H. Ebbinghaus. *Über das gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot, 1885.
- [14] G. Ellis. *Cognitive biases in visualizations*. Springer, 2018.
- [15] C. Fan and H. Hauser. Fast and accurate cnn-based brushing in scatterplots. In *Computer Graphics Forum*, vol. 37, pp. 111–120. Wiley Online Library, 2018.
- [16] M. Feng, C. Deng, E. M. Peck, and L. Harrison. The effects of adding search functionality to interactive visualizations on the web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.
- [17] M. Feng, E. Peck, and L. Harrison. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics*, 25(1):501–511, 2018.
- [18] K. Gadhave, J. Görtler, Z. Cutler, C. Nobre, O. Deussen, M. Meyer, J. M. Phillips, and A. Lex. Predicting intent behind selections in scatterplot visualizations. *Information Visualization*, 20(4):207–228, 2021.
- [19] G. Gigerenzer and H. Brighton. Homo heuristics: why biased minds make better inferences. *Topics in cognitive science*, 1(1):107–43, 2009.
- [20] N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, vol. 140, pp. 107–113. IET, 1993.
- [21] D. Gotz, S. Sun, and N. Cao. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pp. 85–95, 2016.
- [22] D. Gotz and Z. Wen. Behavior-driven visualization recommendation. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pp. 315–324, 2009.
- [23] C. G. Healey and B. M. Dennis. Interest driven navigation in visualization. *IEEE transactions on visualization and computer graphics*, 18(10):1744–1756, 2012.
- [24] M. A. Khan and A. Nandi. Flux capacitors for javascript delorean: approximate caching for physics-based data interaction. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 177–185, 2019.
- [25] H.-Y. Kim. Statistical notes for clinical researchers: Chi-squared test and fisher’s exact test. *Restorative dentistry and endodontics*, 42(2):152–155, 2017.
- [26] P. Langley, W. Iba, K. Thompson, et al. An analysis of bayesian classifiers. In *Aaai*, vol. 90, pp. 223–228. Citeseer, 1992.
- [27] D. J.-L. Lee, H. Dev, H. Hu, H. Elmeleegy, and A. Parameswaran. Avoiding drill-down fallacies with vispilot: Assisted exploration of data subsets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pp. 186–196. Association for Computing Machinery, 2019.
- [28] R. Lightner and T. Francis. Inside america’s boardrooms. <http://graphics.wsj.com/boards-of-directors-at-SP-500-companies/>, 2016.
- [29] Z. Liu, R. J. Crouser, and A. Ottley. Survey on individual differences in visualization. In *Computer Graphics Forum*, vol. 39, pp. 693–712. Wiley Online Library, 2020.
- [30] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.
- [31] M. L. McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.
- [32] S. Monadjemi, R. Garnett, and A. Ottley. Competing models: Inferring exploration patterns and information relevance via bayesian model selection. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):412–421, 2020.
- [33] S. Monadjemi, Q. Nguyen, H. Chai, R. Garnett, and A. Ottley. Active visual analytics: Assisted data discovery in interactive visualizations via active search. *arXiv preprint arXiv:2010.08155*, 2020.
- [34] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009.
- [35] J. M. Murre and J. Dros. Replication and analysis of ebbinghaus’ forgetting curve. *PloS one*, 10(7):e0120644, 2015.
- [36] A. Narechania, A. Coscia, E. Wall, and A. Endert. Lumos: Increasing awareness of analytic behavior during visual data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1009–1018, 2021.
- [37] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220, 1998.
- [38] A. Ottley, R. Garnett, and R. Wan. Follow the clicks: Learning and anticipating mouse interactions during exploratory data analysis. In *Computer Graphics Forum*, vol. 38, pp. 41–52. Wiley Online Library, 2019.
- [39] J. Perry, C. D. Janneck, C. Umoja, and W. M. Pottenger. Supporting cognitive models of sensemaking in analytics systems. *DIMACS, New Brunswick, NY, Tech. Rep.*, 12:2009, 2009.
- [40] W. A. Pike, J. Stasko, R. Chang, and T. A. O’connell. The science of interaction. *Information visualization*, 8(4):263–274, 2009.
- [41] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, vol. 5, pp. 2–4. McLean, VA, USA, 2005.
- [42] 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*, 22(1):31–40, 2015.
- [43] W. Ribarsky, B. Fisher, and W. M. Pottenger. Science of analytical reasoning. *Information Visualization*, 8(4):254–262, 2009.
- [44] L. Rokach. Ensemble-based classifiers. *Artificial intelligence review*, 33(1):1–39, 2010.
- [45] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE conference on visual analytics science and technology (vast)*, pp. 104–115. IEEE, 2017.
- [46] E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE transactions on visualization and computer graphics*, 28(1):966–975, 2022.
- [47] C. Winship and R. D. Mare. Models for sample selection bias. *Annual review of sociology*, pp. 327–350, 1992.
- [48] K. Xu, A. Ottley, C. Walchshofer, M. Streit, R. Chang, and J. Wenskovich. Survey on the analysis of user interactions and visualization provenance. In *Computer Graphics Forum*, vol. 39, pp. 757–783. Wiley Online Library, 2020.
- [49] Z. Zhou, X. Wen, Y. Wang, and D. Gotz. Modeling and leveraging analytic focus during exploratory visual analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.