Active Visual Analytics: Assisted Data Discovery in Interactive Visualizations via Active Search

Shayan Monadjemi
monadjemi@wustl.edu
Washington University in St. Louis
St. Louis, MO

Quan Nguyen
quan@wustl.edu
Washington University in St. Louis
St. Louis, MO

Henry Chai
hchai@wustl.edu
Washington University in St. Louis
St. Louis, MO

Roman Garnett
garnett@wustl.edu
Washington University in St. Louis
St. Louis, MO

Alvitta Ottley
alvitta@wustl.edu
Washington University in St. Louis
St. Louis, MO

ABSTRACT

We propose the Active Visual Analytics technique (ActiveVA), an augmentation of interactive visualizations with active search to aid data foraging and information extraction. We accomplish this by integrating an active search algorithm into the visual interface, which leads users to explore promising data points on a visualization and updates suggestions upon observing user feedback. Using a fictitious epidemic dataset published by the VAST community, we conduct two evaluations. First, we present simulation results that demonstrate the effectiveness of active search in data discovery. Second, we show how a human-computer partnership based on ActiveVA can result in more meaningful interactions during interactive visual exploration and discovery with a crowd-sourced user study. Finally, we outline open research questions regarding human factors in active search and the implications of this line of research in real-world scenarios such as visual drug discovery and intelligence analysis.

CCS CONCEPTS

• **Human-centered computing** → **Visual analytics**.

KEYWORDS

visual analytics, active search, interactive data discovery

1 INTRODUCTION

Many real-world scenarios such as intelligence analysis and drug discovery rely on humans exploring a large data collection for a small set of valuable data points. This process, known as *information foraging*, can be time consuming, overwhelming, and incur unnecessary monetary costs due to the large number of futile data points. For example, an intelligence analyst may spend a substantial amount of time reviewing unrelated documents in an effort to uncover a terrorist attack plot, or a scientist may incur significant monetary costs by examining undesirable chemicals in an effort to discover new drugs.

Motivated by combining unique strengths of machines and humans, visual analytics and machine learning researchers have investigated human-computer partnerships which result in more effective data discovery [10, 18–20, 29]. In particular, the visual analytics community has made significant strides in developing systems that enable the interplay between human and machine analysis [7, 8], and the machine learning community has developed active search as a promising technique capable of identifying as many members of a certain class as possible by intelligently querying oracle(s) for classification labels [14]. With real-world applications relying on human expertise as an acting oracle for active search, we believe the means of effective interaction between the models and humans in an active search setting is critical and worthy of investigation.

In this paper, we particularly consider the augmentation of interactive visualizations with active search, where the human interacting with the visualization acts as the oracle having expertise over the true labels. The primary purpose of this human-computer collaboration is for the human to lead exploration through a visualization of the data and identify as many members of a given class as possible. To the best of our knowledge, this is the first work that considers human factors in data discovery via active search and formally evaluates a working prototype of an interactive visualization with built-in active search.

To investigate the feasibility and impact of this augmentation, we choose a dataset published in the Visual Analytics Science and Technology community (VAST) as a realistic scenario in which an epidemic has initiated in the fictitious city of Vastapolis and authorities are searching through social media posts to identify symptoms and impacted parts of the city. In order to determine feasibility of active search on said data, we build a k-NN model and perform active search simulations to discover relevant data points. Our simulation results indicate that a significantly larger number of relevant data points were discovered via active search in comparison to random search. Intrigued by the promising simulation results, we conduct a crowd-sourced user study to investigate the impact of assisting users during visual data discovery. Our user study results indicate that ActiveVA users make more relevant discoveries while interacting with fewer irrelevant data points. The contributions of this paper are summarized as follows:

- *We introduce Active Visual Analytics (ActiveVA), an augmentation of interactive visualizations with active search for accelerated information extraction.* Using dot-based visualizations and k-NN models, we identify and present data points believed to satisfy users' latent interest.
- *We present a crowd-sourced user study on a prototype of ActiveVA for a fictitious epidemic dataset published by the VAST community.* Our findings suggest that this human-computer

partnership results in more meaningful interactions during interactive visual exploration and discovery.

- *We outline open research questions regarding human factors in active search.* We discuss the implications of active visual analytics in real-world problems and outline open questions involving humans in active search enabled systems.

2 BACKGROUND

In order to understand how humans extract and understand information via visual interfaces, Pirolli and Card [26] presented the sensemaking process as two major loops: first is the *foraging* loop in which users search and gather relevant information, and second is the *sense-making* loop in which users form hypotheses, reason based on gathered evidence, and make decisions. This work introduces a human-computer partnership for interactive information foraging – gathering information relevant to a task – via active search. The related literature to this work covers two main areas: systems that *passively learn* by observing user interaction, and systems that *actively search* for a class of data points by querying the user. For the first time to our knowledge, we will merge the two fields to investigate how active search algorithms can utilize passive user interaction to assist in data foraging.

2.1 Learning from Passive Interactions

A significant body of work in visual analytics has investigated human-computer partnerships in which machines are informed by low-level user interactions with interactive visualizations [2, 6, 9, 24]. These techniques utilize *semantic interaction*, where natural user interactions with the dataset translate into observations for underlying models, integrating user knowledge in the analysis process. Semantic interaction was first introduced by Endert et al. [10] in ForceSPIRE, an interactive system for visual text exploration. ForceSPIRE users were presented with a 2D force-directed graph of text documents and interacted with the documents by highlighting terms, re-positioning documents, and making annotations. In response, the system updated the visual metaphor so that similar documents are visually closer to each other. The purpose of ForceSPIRE was to adapt the visual metaphor to users' cognitive understanding in order to assist them in discovering patterns. Although rich and seamless in interaction space, this technique was limited in support for data types other than text documents and was only validated qualitatively through a case study. In a continued effort to couple cognition and computation, Bian et al. [3] proposed DeepVA, where high-level data features learned by deep learning replace low-level data features. Case studies suggest that DeepVA is more effective in uncovering complex cognitive models with fewer seamless interactions.

In another line of work, researchers investigated how semantic interactions with visualizations can translate into information for steering machine learning models. Motivated by the reliance of many machine learning algorithms on a distance metric, Brown et al. [5] proposed Dis-Function, a technique to represent expert user knowledge as a distance function which is interactively learned by drag/drop interactions with a 2D visualization. Through a case study, they suggested that Dis-Function can improve classification accuracy in comparison to a fixed distance function. Extending interactive learning to topic modeling, Kim et al. [20] proposed TopicSifter, an interactive system with the primary purpose of building models with high recall on text documents. Once again, this work utilized seamless user interactions in order to improve machine learning models. See the suvey by Xu et al. [34] for a more comprehensive review of learning from user interactions.

As opposed to prior work where semantic interaction is used to uncover cognitive patterns or steer machine learning models, in this work we investigate how semantic interactions can lead to discovering a (possibly rare) subset of data deemed valuable for a given task via an active search algorithm.

2.2 Active Search for Data Discovery

Active learning is an approach to machine learning where the algorithm can accomplish greater accuracy with fewer training data under the assumption that it can choose its training data and query an oracle for labels [22, 28]. In cases where the oracle is a human expert, active learning approaches rely on occasional human feedback in order to improve model quality. These systems are commonly known as Human in the Loop systems. Some examples of utilizing active learning for model improvement are in domains of video analysis [15], text annotation [30], and research citation screening [32].

Active search is a special branch of active learning where the goal is to identify as many members of a certain class as possible under a limited querying budget [14]. Active search algorithms iteratively query an oracle to investigate data points believed to be valuable and use the oracle's feedback to maintain a posterior belief over labels of unobserved data points. Fomrally, let k be the number of remaining queries available. The Bayesian optimal policy chooses the unlabeled point that maximizes the conditional expected number of valuable points found at termination. Computing this expectation involves k steps of sampling and conditioning on labels of unseen points and is intractable except for very small values of k [14]. As a workaround, Garnett et al. [14] reccomend myopically looking l steps into the future, where l is a small number. This defines the l-step look-ahead policy, which always assumes the search will terminate after l iterations and makes the optimal decision under that assumption. In the simplest case, the one-step look-ahead or "greedy active search" algorithm queries the oracle at each step assuming only one query remains. This policy behaves myopically, prioritizing immediate discoveries and not risking an unsuccessful observation which could potentially lead to more successful observations in the future. Jiang et al. [18] introduced the Efficient Non-myopic Search policy, or ENS, which exhibits a naturally non-myopic behavior, naturally balancing exploration and exploitation in its recommendations (a classic trade-off in sequential decision making settings), while still being tractable to compute. Empirically, ENS outperforms various benchmark policies on a litany active search problems [18].

In this work, we consider active search for interactive information foraging where humans act as oracles for an active search routine through an interactive visualization. While active search has been studied for some real-world problems such as fraud discovery [29] and drug discovery [13]. The closest works to combining active search and HCI are by Iwata et al. [17] and Klyuchnikov et al. [21] which utilize hypothetical human input and product reviews for improved visual encoding and product recommendation respectively. Nevertheless, neither of these works consider human factors or evaluate their systems on human subjects in real time. To the best of our knowledge, the augmentation of active search and interactive visualizations has never been investigated on human subjects. In this study, we mainly investigate how an augmentation of active search with interactive visualizations impacts human information foraging.

3 ACTIVE VISUAL ANALYTICS

We consider an interactive visual metaphor of a dataset, where each data point is represented by an element on the visual metaphor. The objective of users is to search through this dataset via the visualization and identify data points deemed valuable for a given task. Real-world examples of this scenario include an intelligence analyst who searches through a large set of documents to identify valuable intelligence leading to terrorist plot discovery, or a scientist who searches through a large set of chemical compounds to discover new drugs. With the goal of accelerating visual exploration and discovery, we introduce an augmentation of interactive visualizations with active search. In the remainder of this section, we formalize the workflow of an active visual analytics system and give an overview on how practitioners may adapt this technique to their domain-specific datasets.

3.1 Problem Formulation

We assume there is a dot-based visual metaphor for a given dataset, $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, where each data point in $\mathcal{X}$ has a representative on the visualization. We further assume that each data point is classified as either *relevant* or *irrelevant*, and the objective is to recover as many relevant points as possible without getting distracted by irrelevant points. As users begin providing labels by interacting with the visualization, we maintain a set of observations, $\mathcal{D} = \{(x_1, y_1), ..., (x_m, y_m)\}$, where $y_i \in \{0, 1\}$ denotes the binary classification for a point x_i. A label of $y_i = 1$ indicates the point x_i is *relevant* to the task at hand, whereas $y_i = 0$ indicates the point x_i is *irrelevant*. Note that in an active search setting, a very small portion of the dataset is typically labeled (i.e., $m \ll n$). The objective is to recover as many relevant points as possible, defined by the utility function $u(\cdot)$ where:

$$u(\mathcal{D}) \triangleq \sum\nolimits_{y_i \in \mathcal{D}} y_i,$$

which simply is the number of relevant points in $\mathcal{D}$. In order to accomplish the overarching goal of combining active search and visual analytics, multiple design choices should be taken into consideration. For the remainder of this section, we discuss each component of the ActiveVA workflow as shown in Figure 1.

3.2 Modeling Data Interrelationships

One of the first steps in active search is to model interrelationships between data points in order to predict how relevant a data point is to a task. This model is later used by a querying policy that, given the current user interactions, suggests unlabeled points to the user for further investigation with the goal of maximizing the utility at the end of the search process. In order to detect promising data points, we need to have access to a classification model that provides the posterior probability that an unlabeled point x is relevant given the observed data: $\Pr(y = 1 \mid x, \mathcal{D})$. In this work, we pick a simple k-NN model which only relies on a distance metric between data points. In order to ensure this is an appropriate model for our dataset, we evaluate its ability to rank data points according to their likelihood of being relevant via the AUC-ROC and precision metrics (included in supplemental material).

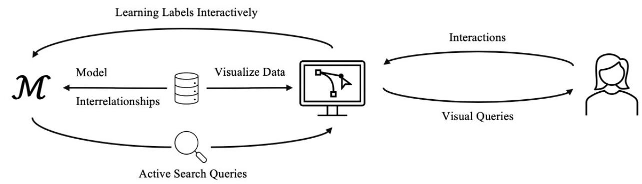

Figure 1: Workflow of an ActiveVA system. We start with a dataset and its interactive visualization. First, we build a model over the dataset to be used by active search to predict relevance of data points to the user. Four major design consideration in an ActiveVA system are: interaction mechanism with the system and how they translate into data labels (Section 3.3), an active search policy to choose data points for querying (Section 3.4), and a visual mechanism for presenting active search queries to the end user (Section 3.2).

3.3 Learning Data Labels Interactively

The communication between humans and an active search algorithm can be viewed as a two-way street: active search needs a proxy to receive feedback from passive user interactions, and humans need to be presented with active search queries through non-intrusive means. In this section we particularly consider how an active search algorithm can simply observe passive human interactions with the visualization and infer whether a data point in relevant or irrelevant to the task at hand. Researchers in visual analytics have analyzed low-level interactions uncover information about users and the task at hand. In particular, they have discovered that analyzing low-level interactions can result in inferring user expertise [4], inferring exploration patterns [12, 23], and modeling the cognitive sense-making process [25]. These successful attempts at analyzing low-level interactions naturally bring us to the following question: can low-level user interactions with a system provide active search with a seamless, yet robust, labeling mechanism? In the simplest case, certain low-level interactions can directly map into certain training labels for active search. For example, clicking a button to bookmark a data point (or disregard one) can signal at positive (or negative) labels respectively. In more complex settings, however, a more ambiguous set of interactions may be used in order to provide labels seamlessly. For example, frequency of hovers on a certain data point and length of hovers may be a more robust approach to uncovering labels from interactions. In this work, we consider interactions that directly map to labels (i.e. bookmarking). However, we identify inferring labels from other means of interactions as a promising future work direction.

3.4 Querying the User Seamlessly

Similar to how well-designed mechanism are needed to translate user interactions into robust labels for active search, we need a mechanism to communicate active search queries to the user effectively. In the most intrusive case, the system would explicitly query the user to provide labels for a given set of points. However, this may cause frustration for the user and undermine the role of humans in leading data exploration. Alternatively, we envision active search queries to be presented to the user if form of visual cues such as color, opacity, and size. In this work, we take a simple approach where active search queries are presented in a distinct color on the visualization. However, we believe there is a need to investigate how various visual cues may impact user involvement and satisfaction with an active search algorithm.

4 PROOF OF CONCEPT APPLICATION

The novelty of this work is the augmentation of interactive visualizations with active search for assisted data discovery. In this section, we present our dot-based visualization prototype for a fictitious epidemic dataset published by the VAST community where the discovery objective is to identify microblogs posted by sick individuals. We discuss the details of the dataset in Section 4.1, the details of the k-NN model over the dataset in Section 4.2, and the details of our visualization interface in Section 4.3.

4.1 Dataset

4.1.1 Overview. The Visual Analytics Science and Technology Community (VAST) published a fictitious epidemic dataset for their annual challenge in 2011. The story involves a terrorist attack in the fictitious city of Vastopolis, where a truck accident over a major river contaminates the water and air with harmful chemicals. The water flow and wind transport these chemicals to two distinct parts of the city, causing citizens to notice symptoms. Those who live by the downstream of the river show waterborne digestive symptoms, whereas those living downwind of the accident show respiratory symptoms. The dataset contains 1,023,077 microblogs posted on social media from various parts of town during a 21-day period (04/30/2011 - 05/20/2011). The fictitious attack occurred on 05/17/2011 and the outbreaks appeared during 05/18/2011-05/20/2011.

4.1.2 Labeling Heuristic. For model evaluation purposes in Section 4.2, we need ground-truth labels for this dataset. Since the relevance of each data point to the spread of epidemic is unknown, we rely on a heuristic for labeling. Specifically, microblogs containing a set of keywords such as 'sore throat', 'diarrhea', and 'pneumonia' are labeled *relevant*, and the remaining microblogs are labeled *irrelevant.* Refer to Appendix A for the full list of keywords. The daily incidence rate (proportion of *relevant* points per day) according to our heuristic is 1% for 04/30/2011 - 05/17/2011 and it increases to 21-27% for 05/18/2011 - 05/20/2011 after the terrorist attack. Although the reliance on this heuristic for labeling introduces false-positives, inspection of the dataset before the epidemic (i.e. the %1 incidence rate on 04/30/2011-05/17/2011) suggests the false-positive rate is less than 3%.

4.1.3 Data Selection. In order to demonstrate active search on datasets with varying rates of relevant points, we choose two subsets of the dataset: (1) the first two days of the epidemic (05/18/2011 and 05/19/2011) with 24% of points being related to illness (*High-Incidence*), and (2) the first two available days in the dataset (04/30/2011 - 05/01/2011) with 1% of points being labeled positive by the heuristic (*Low-Incidence*). In the upcoming sections, we use the High-Incidence dataset to demonstrate active search's ability to identify relevant data points in a scenario deemed realistic to the visual analytics community. Furthermore, we use the Low-Incidence dataset to demonstrate active search's exceptional ability to identify rare points of interest. Finally, we use a random sample of 3000 points from the High-Incidence dataset to evaluate ActiveVA in our user study.

4.2 Probabilistic Classifier over Dataset

As mentioned in Section 3.1, performing active search relies on a probabilistic model that computes the posterior probability of an unlabeled point being relevant given the observed data. As suggested by Garnett et al. [14], we choose a simple nearest-neighbors-based classification model to learn interrelationships between data points, where data points with close proximity are considered related. The k-NN classification model is a non-parametric technique that only relies on a distance definition among data points. This flexibility in the choice of distance function offers practitioners from various fields the option to tailor this technique to their domain-specific datasets.

The challenge, however, is that we do not often know the best distance metric explaining user interactions ahead of time. Therefore, we take an approach similar to the one by Monadjemi et al. [23], where we build multiple competing models and utilize user interactions to inform us about the plausibility each model. Specifically, we build two k-NN models over the data. The first one ($\mathcal{M}_L$) is based on the posting *location* of microblogs, where the distance between two data points is the Euclidean distance between locations from which they were posted. The second one, ($\mathcal{M}_T$) is based on the microblog *texts*, where the distance between two data points is the cosine distance between the vector representation of their texts. We define the vector representation of a microblog to be the normalized average over *word2vec* representation of its individual tokens (after removing numerical values, punctuation, and stop words) trained on a large set of news articles [27]. Given some observed data, each of these two models, $\mathcal{M}_i$, calculates the probability that an unlabeled data point x_i is relevant: $\Pr(y_i = 1 \mid x_i, \mathcal{D}, \mathcal{M}_i)$. To combine these two predictions, we use a parameter $q \in [0, 1]$ as the weight of the text-based prediction (the location-based prediction thus has a weight of $1 - q$):

$$\begin{aligned}\Pr(y_i = 1 \mid x_i, \mathcal{D}, \mathcal{M}_T, \mathcal{M}_L) =& q \Pr(y_i = 1 \mid x_i, \mathcal{D}, \mathcal{M}_T) \\ &+ (1-q)\Pr(y_i = 1 \mid x_i, \mathcal{D}, \mathcal{M}_L),\end{aligned}$$

where q is chosen using the *maximum likelihood estimation* method to maximize the likelihood of the observed data $\mathcal{D}$. The performance of this model with respect to the high- and low-incidence dataset is summarized in Table 1 of the supplementary material.

4.3 Visualization Interface

We implemented a prototype of ActiveVA as shown in Figure 2. We aimed for a simple interface and natural means of interaction for greater usability. The majority of the screen was covered by a map of Vastapolis and a random sample of 3000 microblogs from the High-Incidence dataset shown in their corresponding posting location on the map. Most user interactions occurred on the map, where users hovered on data points to see a tooltip containing the microblog (Figure 2, C). The tooltip allowed user feedback in one of three ways: (1) if the hovered data point was suggested by the active search algorithm, the user could either *add bookmark* or report an *irrelevant suggestion*; (2) if the hovered data point was already bookmarked, the user could *remove bookmark*; (3) if the hovered data point was not already bookmarked nor suggested by active search, the user could only *add bookmark*. We utilized three distinct colorblind-safe colors to distinguish between suggested dots, discovered dots, and the remaining dots. To make potential feedback modifications easier, we displayed a list of bookmarks on the sidebar along with an option to *remove bookmark* (Figure 2, A).

5 EVALUATION

We evaluate ActiveVA from two perspectives: (1) we conduct a crowd-sourced user study to demonstrate how user interactions and information foraging throughput are impacted when active search is present in the system, and (2) we conduct simulations to investigate the ability of active search in identifying data points related to the epidemic when queries are made via different policies and when the incidence rate is significantly smaller.

5.1 Crowd-sourced User Study [1]

To investigate the impact of active search in visual exploration and discovery, we designed a crowd-sourced user study in which participants interact with a map of the fictional city of Vastapolis (Figure 2) which is under a biochemical attack initiating an epidemic.

5.1.1 Task. We told participants that the authorities are interested in identifying the impacted parts of the city by analyzing social media activity, and we have access to social media posts and their posting location. Their task was to assist the authorities by searching through a dataset of microblogs via an interactive map and bookmarking as many posts containing illness-related information as possible.

5.1.2 Participants. We recruited 130 participants via Amazon's Mechanical Turk platform. Participants were 18 to 65 years old, from the United States, and fluent in English. Each participant had a HIT approval rating of greater than 98% with more than 100 approved HITs. After data cleaning steps outlined in 5.1.5, there were 45 women, 76 men, and 1 participant with undisclosed sex in our subject pool with ages ranging from 18 to 62 years ($\mu = 36$, $\sigma = 9$). Seventy-one percent of our participants self-reported to have at least an associate degree. The instructions specified that participants will be compensated $1.00 base pay and an additional $0.10 bonus for every relevant microblog they identify (maximum $4.00 bonus). Although the advertised payment structure was designed to incentivize participants to complete the task, we ultimately decided to pay everyone the maximum bonus of $4.00 for fairness. Our results will show that the bonus for the control group was on average lower than the active search group.

5.1.3 Procedure. Our system randomly assigned each participant to one of the following groups: *active search group* which received a batch of 10 active search queries in the form of visual clues that were updated after every interaction, and *control group* which did not receive any assistance during exploration. Upon giving consent to participate in our study, participants were given a tutorial on their task and their corresponding system. Both groups initiated their task without any clues, meaning that the active search group did not get assistance for their first bookmark and the following suggestions were as relevant as the initial user bookmark. Participants were given at most 10 minutes to identify as many microblogs related to the epidemic as they could using an interactive map where microblogs are visulizations based on their posting locations. Users hovered on the visualized dots to trigger a tooltip containing the post, and clicked on a button to add the post to bookmarks if it contained illness-related content. Once the users were satisfied with their search for illness-related documents or the 10 minutes were up, users were directed to a post-experiment survey to collect demographics information and general feedback on the system.

5.1.4 Hypotheses. At a high level, we hypothesize that the control group will examine more irrelevant data points when completing the tasks. We will capture a collection of measures (Section 5.1.5) from the interaction data, and we hypothesize that:

- the active search group will hover over more relevant points than the control group.
- the active search group will inspect more relevant points per minute than the control group.
- the active search interface *may* result in more bookmarks and fewer hovers than the control condition. However, there are a large number of points of interest, so the overall throughput may not be statistically significant.

5.1.5 Data Cleaning. In a pre-processing step, we filtered the collected data to exclude participants who did not attempt the task or were unable to finish the experiment. Specifically, we eliminated participants who hovered on less than 10 data points and those who reported technical issues with the interface. Here, we consider a valid hover to be one that lasts at least 0.5 seconds (0.3 seconds for triggering the tooltip, and 0.2 seconds for skimming the text). The filtered dataset contained 124 subjects; 74 and 50 samples for control and active search groups respectively.

5.1.6 Data Analysis. We analyze our collected user study dataset based on interactions with data points (*hovers)* and discoveries (*bookmarks*). These two means of interaction inform us about the *speed* and *accuracy* of data foraging through the six metrics listed in Table 1. The bookmark and hover purity metrics inform us about the proportion of interactions that involved relevant data points, measuring *accuracy* of interactions. The bookmarks and hovers per minute metrics inform us about the overall *speed* at which users interact with data points. Finally, the relevant hovers and relevant bookmarks per minute metrics inform us about the rate at which

[1] This experiment was pre-registered on Open Science Foundation.

Figure 2: A view of the ActiveVA prototype on the epidemic dataset. Green dots on the map indicate *relevant* data points bookmarked by the user (also shown in the left panel, A). Orange dots indicate active search recommendations of *potentially relevant* data points (C). Violet dots indicate the remaining data points. Hovering on data points triggers a tooltip containing the microblog and feedback options (C). A timer for remaining time was shown, and users had the option to exit the experiment at any time or report technical issues (B).

users interacted with relevant data points, quantifying both *speed* and *accuracy* of interactions.

5.1.7 Results. For each these metrics, we perform a two-sample t-test to determine if the presence of active search has a statistically significant impact on how users interact with the data. Furthermore, we evaluate the severity of impact via Cohen's d metric. As shown in Table 1, there is evidence that the active search group differed from the control group in all six metrics outlined above ($\alpha = 0.05$).

Our results indicate that the presence of active search in interactive visualizations is indeed helpful in information foraging and discovery. The HPM and RHPM metrics inform us that participants with ActiveVA were able to hover on fewer data points (per minute), while hovering on more relevant data points (per minute). Moreover, the significant evidence provided by the HP metric suggests that ActiveVA users hovered on a more relevant subset of data points than the control group. The three remaining metrics (BPM, RBPM, and BP) indicate that ActiveVA users make more discoveries than those without the assistance of active search. Overall, the findings in our user study suggest that *ActiveVA assists users to interact with a more relevant subset of data points and make more discoveries.* In a sense, active search users were able to disregard irrelevant points and be more mindful towards the relevant data points.

Furthermore, we perform a Mann-Whitney U test on the survey responses listed in Table 2 on *willingness to use*, *ease of use*, and *ease of task completion.* While the results of our statistical test are not conclusive, we observe an encouraging and consistent tendency among ActiveVA users finding the system and task easier and being more willing to use the interface. We believe this is a promising area of further exploration to gauge how humans may react towards active search suggestions in real-time systems.

5.2 Additional Simulations

In our user study from the previous section, about 24% of the points visualized were illness-related (High-Incidence dataset). Furthermore, users were queried via a greedy approach which is purely based on exploitation. In this section, we take an additional step to show how different querying policies would perform on datasets with varying ratio of points being relevant (1% and 24%). While the simulations in this section make strict assumptions and may not represent how users would interact with the data, we include the results to encourage future work with more challenging datasets and more sophisticated querying policies.

5.2.1 Simulation Setup. To simulate different querying behaviors, we apply the one-step look-ahead active search policy and a variant of the ENS to the two aforementioned subsets of epidemic dataset. Recall that the one-step policy is a myopic, greedy approximation of the Bayesian optimal policy with the assumption that the search will terminate after the current query. ENS, on the other hand, is non-myopic and explores/exploits the data with respect to the number of remaining queries k. It is worth noting that k is generally not

Table 1: The results of 2-sample t-tests on the six metrics outlined in 5.1.5. There is statistically significant evidence that ActiveVA participants interact with *fewer* data points (avg. 14.9 vs. 16.0) and discover more relevant data points (avg. 6.1 vs. 3.9).

	95% CI				
Metric	Control	Active Search	p-value	t-statistic	Cohen's d
Hovers per Minute	16.9 ± 1.20	14.9 ± 1.21	**0.0225**	−2.31	−0.43
Relevant Hovers per Minute	4.9 ± 0.52	6.9 ± 0.87	< **0.0001**	4.31	0.80
Hover Purity	0.28 ± 0.02	0.46 ± 0.04	< **0.0001**	8.19	1.53
Bookmarks per Minute	6.8 ± 0.77	9.2 ± 1.35	**0.0024**	3.10	0.58
Relevant Bookmarks per Minute	3.9 ± 0.52	6.1 ± 0.94	< **0.0001**	4.21	0.78
Bookmark Purity	0.55 ± 0.03	0.64 ± 0.04	**0.0022**	3.13	0.58

Table 2: Post experiment survey results where users commented on the following statements on a scale of 1 (strongly disagree) to 5 (strongly agree). While the Mann-Whitney U test remains inconclusive, we observe a positive shift in active search group.

	95% CI		
Survey Question	Control	Active Search	p-value
Willingness to use	3.23 ± 0.12	3.49 ± 0.16	0.0935
System ease of use	3.95 ± 0.11	4.14 ± 0.12	0.1068
Task ease of completion	4.01 ± 0.11	4.12 ± 0.12	0.3300

known in our active visual analytics setting. Our workaround is to assume a constant k at each iteration, which defines the policy ENS-k that looks k steps into the future in the same manner as ENS; here we specifically consider ENS-50. As a baseline, we also include the random search policy that queries from the pool of unlabeled points uniformly at random at each iteration.

For each simulation, we randomly select a single relevant data point as the initial observed dataset $\mathcal{D}$. Then, at each iteration, a policy chooses an unlabeled data point to query according to its criterion, and the label of that data point (whether it is relevant or not) is revealed. This information is then added to $\mathcal{D}$, which informs the policy in the next iteration. Each simulation consists of 500 iterations for each policy, and 50 simulations were conducted in total.

5.2.2 Results. The results of these simulations are summarized in Table 3, in which we show the average and 95% confidence interval of the total number of relevant points found by the random policy, one-step, and ENS-50. We see that the two active search policies result in an overwhelming improvement in utility from the random search, especially in the low-incidence dataset where discovery is more challenging. Additionally, ENS-50 significantly outperforms one-step look-ahead in the low-incidence setting, highlighting the benefit of non-myopia in active search that was observed by Jiang et al. [18] in their original work. We further consider two-step look-ahead and ENS-10 using the same experiment and report their performance in the supplementary material. Based on the results from this simulated experiment, we hypothesize that active search

Table 3: Number of relevant points found by various search policies across 50 simulations

	High-Incidence Dataset	Low-Incidence Dataset
Random	142.88 ± 2.81	7.80 ± 0.85
One-step	471.86 ± 8.20	318.62 ± 26.79
ENS-50	478.46 ± 6.00	351.82 ± 18.05

can assist users in analyzing datasets with a varying proportions of relevant points and various querying policies.

6 DISCUSSION AND FUTURE WORK

In this work, we considered an augmentation of interactive visualizations with active search for information foraging. Our results indicate that this computer-human partnership is indeed effective in assisting humans to explore fewer data points and make more discoveries. This observation is inline with the Information Foraging Theory which states humans instinctively want to identify as many relevant data points per amount of effort as possible. While this outcome is encouraging, there are still a number of open research questions on how humans and active search can effectively collaborate in information foraging. In this section, we discuss future research directions and implications of this line of research in real-world problems.

6.1 Human Factors in Active Visual Analytics

Successful incorporation of ActiveVA in real-world application relies on an effective human-computer collaboration. It is therefore critical to take into consideration human-related factors in addition to technical factors involved in an ActiveVA system. Some of these considerations include social issues such as trust, politeness, and communication.

Trust. As an integral component of human interaction, trust allows people to make decisions under uncertainty with the risk of adverse consequences [1]. Consider our prototype of ActiveVA presented in Section 4.3. As a result of observing initial user interactions, some dots are highlighted by the active search algorithm as promising data points. Theoretically, those data points are ones that will maximize the overall data discovery. This outcome, however, will only be feasible if end users trust in highlighted recommendations and interact with them. In a post-experiment survey, 41 out of 50 ActiveVA users indicated they trust visual suggestions made by the system. While this level of trust is encouraging, factors impacting this trust are unknown. Specifically, in the non-myopic setting where the active search algorithm may explore less promising points with hopes of maximizing the overall utility, will humans still be able to still trust system recommendations and collaborate for better outcomes?

Intrusiveness. Human input is integral to the success of active search algorithms. Furthermore, ensuring the introduction of automated means in interactive systems does not hinder usability has is also crucial [11]. For example, Microsoft Clippy is known as being an intrusive character with short memory. Not only could users not trust Clippy's suggestions, they were often disturbed by its frequent and mindless appearances. [9, 16] In a paper on computational politeness, Whitworth [33] argues that perceived politeness in automated assistants depends on factors such as respecting user choices and feedback. In our post-experiment survey 37 out of 50 participants indicated that they did not find active search suggestions to be annoying. This study, however, only had one method of communicating suggestions to the user: by modifying the colors of suggested points on the visual interface. It is, however, unknown how various design choices can result in best outcomes. For example, do less intrusive or more intrusive recommendations result in high overall utility? In line with this question, Wall et al. [31] proposed some design choices in effectively user cognitive biases. Through similar design choices, researchers can study how various means of communication can impact the human-computer partnership in ActiveVA.

Feedback from Passive Interactions. One contributing factor to the quality of ActiveVA collaboration is the means through which humans provide labels to the computer. In our prototype, users had to intentionally click on a button to provide labels to the algorithm. Although this mechanism of providing labels worked well in our prototype without overwhelming the user, there are other seamless possibilities to investigate. We can envision a multi-fidelity version of active search in which passive human interactions translate into labeled feedback of various forms. For example, a system could consider hovers to be low-fidelity feedback while more intentional interactions such as clicks and bookmarks are high-fidelity. Future work may investigate how various means of interactions can translate into informative labels for an active search algorithm and how each will impact human exploration.

6.2 Do Users Follow Suggestions?

One interesting observation in our experiment was that about a quarter of users who had access to active search suggestions did not bookmark any of them (Figure 3). Upon manual inspection, we verified that the suggestion were indeed relevant to illness-related microblogs, however, the users did not choose to bookmark them. Whether or not users noticed these data points and hovered on them remains an open question for future work. However, upon comparing the survey results on the two groups (i.e., 9 participants who had active search suggestions and did not bookmark any of them vs. 41 participants who had active search suggestions and bookmarked them), we learn that the group who ignored the suggestions were on average neutral on whether or not they found the suggestions trustworthy or intrusive, where as those who utilized suggestions on average found them trustworthy and non-intrusive (Table 4). This is yet another interesting direction for future research to investigate how to gain user trust and not overwhelm them with suggestions.

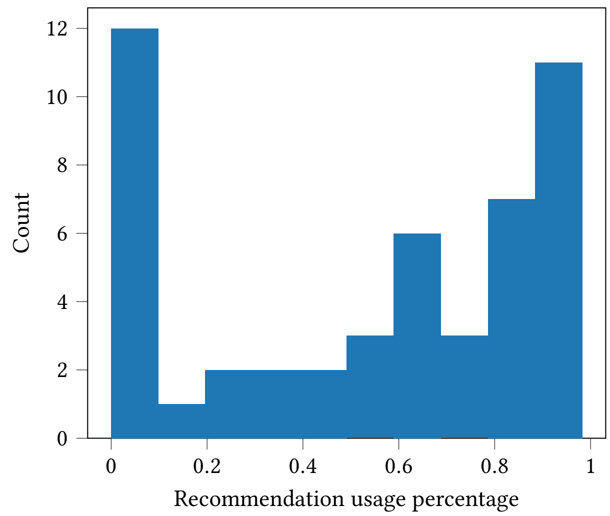

Figure 3: Distribution of percentage of bookmarks resulting from the active search suggestions

Table 4: Post experiment survey results comparing the 9 users who chose not to bookmark the suggestions (Ignore) with the other 41 users in the active search group (Others). Users responses to these questions on a scale of 1 (strongly disagree) to 5 (strongly agree).

	95% CI		
Survey Question	Ignore	Others	p-value
Willing to use	3.00 ± 0.33	3.60 ± 0.18	0.0626
System ease of use	3.78 ± 0.28	4.23 ± 0.14	**0.0452**
Task ease of completion	4.00 ± 0.29	4.15 ± 0.15	0.2849
Suggestion intrusiveness	2.78 ± 0.28	1.58 ± 0.15	**0.0004**
Confused by suggestions	2.33 ± 0.41	1.55 ± 0.14	**0.0244**
Trusted the suggestions	3.33 ± 0.24	4.23 ± 0.12	**0.0006**

6.3 Real-World Application

ActiveVA can be applied to any scenario where human analysts are tasked with searching a large data set for a particular subset of data points with a certain class. In this section we outline two example scenarios in which a collaboration between human and active search can speedup discovery.

Accelerated Visual Drug Discovery. Drug discovery involves identifying promising chemical compounds and performing expensive experiments to determine their effectiveness. In this setup, a human expert decides which chemical compounds are worth lab trials, and mother nature determines the true label after the experiment. We envision a special version of multi-fidelity ActiveVA to be helpful in combining human intuition and nature feedback in an attempt to accelerate visual drug discovery and reduce lab trial expenses.

Intelligence Analysis. Analysts tasked with national security matters often search through a large set of unstructured data points in an attempt to find a subset of points with intelligence value. This process can be frustrating and overwhelming due the the large amount of irrelevant data points in consideration. An implementation of ActiveVA may help analysts discover more promising data points while disregarding irrelevant ones.

6.4 Limitations

One of the main limitations in this work is the assumption that the entire dataset is visualized at once. We intentionally picked a random set of 3000 points in order to avoid overwhelming the visual interface. However, such decision is not reasonable in large real-world datasets. Our approach in this work may be extended to consider large datasets and how active search can mitigate the resulting information overload. One possibility is to start the initial view of the visualization with a random subset of the data, and progressively dismiss irrelevant data points and display more promising ones informed by user interactions and active search. The issues outlined above must be considered in this approach since handling visualization updates and the trade-off between exploration and exploitation may be perceived as intrusive. In future analysis, we encourage researchers to investigate how various active search policies impact exploration coverage.

7 CONCLUSION

We introduced ActiveVA, a novel augmentation of interactive visualizations with active search for accelerated information foraging. We formulated visual data discovery as an active search problem with human input as an acting oracle, and used simulations to demonstrate the ability of active search to identify points of interest in a realistic dataset published by the VAST community. Furthermore, we conducted a crowd-sourced user study to investigate the impact of ActiveVA in information foraging. Our results indicate that ActiveVA users had more meaningful interactions with the dataset in the sense that they discovered more relevant points while interacting with a smaller number of points. Finally, we discussed open research questions regarding human factors in active search.

A LABELING HEURISTIC KEYWORDS

The following is the full set of keywords used in our heuristic to label microblogs *relevant*: ‘sore throat’, ‘throat’, ‘fever’, ‘fatigue’, ‘cough’, ‘shortness’, ‘breath’, ‘chills’, ‘sick’, ‘pain’, ‘diarrhea’, ‘stomach’, ‘sweats’, ‘pneumonia’, ‘flu’, ‘aches’, ‘nausea’, ‘vomiting’, ‘nauseous’, ‘declining health’.

REFERENCES

[1] Donovan Artz and Yolanda Gil. 2007. A survey of trust in computer science and the semantic web. *Journal of Web Semantics* 5, 2 (2007), 58–71.

[2] 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*. 1363–1375.

[3] Yali Bian, John Wenskovitch, and Chris North. 2020. DeepVA: Bridging Cognition and Computation through Semantic Interaction and Deep Learning. *arXiv preprint arXiv:2007.15800* (2020).

[4] Nadia Boukhelifa, Anastasia Bezerianos, Ioan Cristian Trelea, Nathalie Méjean Perrot, and Evelyne Lutton. 2019. An exploratory study on visual exploration of model simulations by multiple types of experts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[5] Eli T Brown, Jingjing Liu, Carla E Brodley, and Remco Chang. 2012. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 83–92.

[6] 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.

[7] R Jordon Crouser and Remco Chang. 2012. An affordance-based framework for human computation and human-computer collaboration. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2859–2868.

[8] R Jordan Crouser, Alvitta Ottley, and Remco Chang. 2013. Balancing human and machine contributions in human computation systems. In *Handbook of Human Computation*. Springer, 615–623.

[9] Filip Dabek and Jesus J Caban. 2016. A grammar-based approach for modeling user interactions and generating suggestions during the data exploration process. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 41–50.

[10] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 473–482.

[11] Alex Endert, M Shahriar Hossain, Naren Ramakrishnan, Chris North, Patrick Fiaux, and Christopher Andrews. 2014. The human is the loop: new directions for visual analytics. *Journal of intelligent information systems* 43, 3 (2014), 411–435.

[12] 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.

[13] Roman Garnett, Thomas Gärtner, Martin Vogt, and Jürgen Bajorath. 2015. Introducing the ‘active search’method for iterative virtual screening. *Journal of Computer-Aided Molecular Design* 29, 4 (2015), 305–314.

[14] Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, and Richard Mann. 2012. Bayesian optimal active search and surveying. *arXiv preprint arXiv:1206.6406* (2012).

[15] Benjamin Höferlin, Rudolf Netzel, Markus Höferlin, Daniel Weiskopf, and Gunther Heidemann. 2012. Inter-active learning of ad-hoc classifiers for video visual analytics. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 23–32.

[16] Eric J Horvitz, Andy Jacobs, and David Hovel. 2013. Attention-sensitive alerting. *arXiv preprint arXiv:1301.6707* (2013).

[17] Tomoharu Iwata, Neil Houlsby, and Zoubin Ghahramani. 2013. Active learning for interactive visualization. In *Artificial Intelligence and Statistics*. 342–350.

[18] Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, and Roman Garnett. 2017. Efficient Nonmyopic Active Search *(Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 1714–1723. http://proceedings.mlr.press/v70/jiang17d.html

[19] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. Visual analytics: Definition, process, and challenges. In *Information visualization*. Springer, 154–175.

[20] Hannah Kim, Dongjin Choi, Barry Drake, Alex Endert, and Haesun Park. 2019. TopicSifter: Interactive Search Space Reduction Through Targeted Topic Modeling. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 35–45.

[21] Nikita Klyuchnikov, Davide Mottin, Georgia Koutrika, Emmanuel Müller, and Panagiotis Karras. 2019. Figuring out the User in a Few Steps: Bayesian Multi-fidelity Active Search with Cokriging. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 686–695.

[22] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*. Springer, 3–12.

[23] Shayan Monadjemi, Roman Garnett, and Alvitta Ottley. 2020. Competing Models: Inferring Exploration Patterns and Information Relevance via Bayesian Model Selection. (2020).

[24] Alvitta Ottley, Roman Garnett, and Ran Wan. 2019. Follow The Clicks: Learning and Anticipating Mouse Interactions During Exploratory Data Analysis. *Computer Graphics Forum* 38, 3 (2019), 41–52.

[25] Jason Perry, Christopher D Janneck, Chinua Umoja, and William M Pottenger. 2009. Supporting Cognitive Models of Sensemaking in Analytics Systems. *DIMACS, New Brunswick, NY, Tech. Rep* 12 (2009), 2009.

[26] Peter Pirolli and Stuart Card. 2005. 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. McLean, VA, USA, 2–4.

[27] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[28] Burr Settles. 2009. *Active learning literature survey.* Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[29] Md Amran Siddiqui, Alan Fern, Thomas G Dietterich, Ryan Wright, Alec Theriault, and David W Archer. 2018. Feedback-guided anomaly discovery via online optimization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2200–2209.

[30] Katrin Tomanek and Fredrik Olsson. 2009. A web survey on the use of active learning to support annotation of text data. *Proceedings of Active Learning for Natural Language Processing (ALNLP-09)* (2009), 45–48.

[31] Emily Wall, John Stasko, and Alex Endert. 2019. Toward a Design Space for Mitigating Cognitive Bias in Vis. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 111–115.

[32] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. 2010. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 173–182.

[33] Brian Whitworth. 2005. Polite computing. *Behaviour & Information Technology* 24, 5 (2005), 353–363.

[34] Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. 2020. Survey on the analysis of user interactions and visualization provenance. *STAR* 39, 3 (2020).