Interface Design and Task Difficulty Impact ML-Assisted Visual Data Foraging

Shayan Monadjemi\textsuperscript{1}, Sunwoo Ha\textsuperscript{1}, Quan Nguyen\textsuperscript{1}, Henry Chai\textsuperscript{2}, Roman Garnett\textsuperscript{1}, Alvitta Ottley\textsuperscript{1}

\textsuperscript{1} Washington University in St. Louis, St. Louis, MO
\textsuperscript{2} Carnegie Mellon University, Pittsburgh, PA

Abstract
Data foraging routinely involves sifting through a large amount of irrelevant information in search of relevant data. In machine learning, the related task of active search considers the automated discovery of rare, valuable items from large data sets – a setting that maps directly onto data foraging. Although there has been a long history of integrating similar assistive technologies into the visual analytics pipeline, we do not fully understand how these technologies impact human behavior or what factors might impact the machine partners’ effectiveness. We frame data foraging as a sequential decision-making process and propose using active search as an assistive technology for accelerating discovery. We conduct a crowd-sourced user study to evaluate this human–machine partnership in data foraging and show that our approach results in higher throughput and more meaningful interactions during interactive visual exploration and discovery. Furthermore, we present evidence from a follow-up user study that the impact of incorporating assistive technology in visual tasks varies with interface design and task difficulty.

CCS Concepts
• Human-centered computing → Visual analytics; Empirical studies in visualization; • Computing methodologies → Active learning settings;

1. Introduction
Many real-world analytic scenarios rely on humans exploring an extensive data collection for a small set of relevant data points. This process, known as data foraging, can be time-consuming, overwhelming, and unnecessarily costly due to a large number of uninformative data points. For example, an intelligence analyst may spend substantial time reviewing unrelated documents while unravelling a terrorist attack plot. Likewise, a scientist may test numerous molecules – incurring high cost – while searching for a new drug candidate, many of which may prove useless in a medical setting. Researchers have identified data foraging as a leverage point, where a human–machine partnership can improve sensemaking [PC05].

The visual analytics community has already made significant strides in developing systems that enable the interplay between humans and machines in exploratory data analysis and sensemaking [KAF\textsuperscript{08}, EFN12, KCD\textsuperscript{19}, JMC\textsuperscript{17}, SFD\textsuperscript{18}, CC12, COC13]. Meanwhile, the machine learning community has proposed systems wherein a machine learning model can work alongside a human “in the loop” to accelerate the process of laborious data labeling through careful selection of the most relevant data points. This idea of learning algorithms strategically choosing their training data by querying an oracle (which may be a human) is known as active learning [Set09]. Under the umbrella of active learning, the machine learning community has developed the paradigm of active search to efficiently search large datasets for rare, valuable items [GKX\textsuperscript{12}, JMC\textsuperscript{17}]. Active search has proven more effective than traditional active learning algorithms in settings such as scientific discovery [JMC\textsuperscript{17}]. Although there has been a long history of integrating machine learning (ML) into visual analytic tools, the impact of these technologies on human behavior and contributing factors to the effectiveness of these technologies are open for investigation.

We investigate the impact of ML on analytic behavior by drawing parallels between data foraging — an essential step in visual analytic sensemaking [PC05] — and active search. This formulation allows us to leverage the leading ML method for data discovery to examine how such assistive technology shapes the way humans interact with intelligent visualization tools. Next, we augment an interactive visualization with an active search algorithm. In this setting, the interactive visualization is the medium of communication between the active search algorithm and the human. The user inspects data points sequentially and determines their relevance to the task at hand. Simultaneously, the algorithm translates the observed interactions into labels for the underlying machine learning models. The active search algorithm then assists the user by recommending the most promising data points for future investigation.

Using a prototype system, we conduct two crowd-sourced user studies to investigate the impact of the active search algorithm in
assisting users during visual data exploration and discovery. We choose a dataset published in the Visual Analytics Science and Technology (VAST) 2011 Challenge which was designed to mimic a realistic analytic scenario [SWPG12, CGW14, GCH11]. Participants saw a visualization of geotagged tweet-like messages, which provided information about the spread of various disease symptoms. The task was to assist authorities by searching through social media posts to identify individuals who may be impacted by the potential epidemic. In each experiment, we randomly assign participants to one of two groups: (1) an active search group that performed an information foraging task with visual cues provided by the active search algorithm and (2) a control group who performed the same task without assistance. Our quantitative analysis of the first user study indicates that users assisted by the active search algorithm make more relevant discoveries while interacting with fewer irrelevant data points. However, we found that a non-trivial percentage of the active search group ignored the recommendations, and an analysis of the subjective responses revealed that these participants reported a low level of perceived “trust” in the system.

We conduct a follow-up user study to further examine how task difficulty and explanations may influence user engagement with active search recommendations. Participants in this follow-up study performed the same information foraging tasks but with a dataset containing a much larger percentage of irrelevant points. Additionally, to encourage suggestion usage and perceived trust ratings, we added explanations to the visual cues. Our findings revealed that users are more engaged with system suggestions when presented with an explanation and that the marginal benefit of incorporating active search in the visual data foraging workflow is significantly higher in a more difficult task.

A summary of our contributions is as follows:

- We frame data foraging as sequential decision-making process and integrate active search into the sensemaking pipeline. Our framework and prototype system can enable the visualization community to integrate active search — a more powerful machine learning technique — into future visual analytic tools.
- Using findings of our crowd-sourced study, we highlight the critical implications for this human–machine partnership. Although the active search suggestions resulted in more discoveries of relevant points and fewer interactions with irrelevant points, many people ignored the suggestions completely.
- We provide guidance on how to improve suggestion usage. In particular, we demonstrate with a follow-up study how an explainable interface design increases user engagement. Furthermore, we show that utility of incorporating ML into the visual analytic pipeline increases for more challenging tasks.

2. Background

Pirolli et al. [PC05] characterized the cognitive process behind intelligence analysis in terms of two major loops: first is the foraging loop in which analysts search and gather relevant information, and second is the sense-making loop in which analysts form hypotheses, reason based on gathered evidence, and make decisions. Analysts may approach a problem in a bottom-up fashion in which they make hypotheses according to discovered data or in a top-down fashion in which they search for evidence given an existing hypotheses. Regardless of the approach, the foraging process of searching for relevant data points in a large dataset is time consuming and costly. Thus, analysts aim to maximize their information intake while minimizing the effort needed.

2.1. Learning from User Interactions and Model Steering

A significant body of work in visual analytics has sought to enable human–machine partnerships in which machines are informed by low-level user interactions with interactive visualizations [BCS16, OGW19, DC16, BOZ14, BWN19]. Here, the technique of semantic interaction is key, where user interactions with visualization tools translate into observations for underlying models, integrating user knowledge in the analysis process and informing intelligent response by the visualization system [EFN12]. Semantic interactions have been used to learn expert knowledge [BLBC12], improve visual projection [IHG13], improve text analysis models [KCD19, KPSK17, SEAHJ18], reduce visualization latency [BCS16], and mitigate selection bias [GSC16].

Learning from user interaction to steer machine learning procedures has shown success in the visual analytics community. For example, Brown et al. [BLBC12] proposed Dis-Function, a technique to represent expert knowledge as a distance function which is interactively learned by drag/drop interactions with a 2D visualization. Extending interactive learning to topic modeling, Kim et al. [KCD19] proposed TopicSifter, an interactive system with the primary purpose of building models with high recall on text documents. Another example of similar work is ALVA by Kucher et al. [KPSK17] which steers a stance classification model over text documents by strategically querying the human for labels.

In addition to using interactions to inform machine learning models, researchers have also developed systems where machines take actions to guide users in the analytic process [GIC19, GDM19, JCG20, DCCE19]. Researchers in the visual analytics community have defined guidance as a computer-assisted process aimed at resolving users’ knowledge gap during an interactive session [CGM16]. Returning to the work by Kim et al. [KCD19], TopicSifter is an example of a system which guides the user by recommending potentially relevant keywords to include in the search space. In the context of our work, the machine assists the user by recommending data points to investigate and learns the user’s latent interest by observing interactions. We refer to the survey by Xu et al. [XOW19] for a more comprehensive review of learning from user interactions.

The evaluation of these mixed-initiative systems often relies on case studies or user studies with a small sample. These evaluations often include measures such as task accuracy and speed to determine the usefulness of the system [BLBC12, BCS16, GSC16]. Battle et al. [BCS16] observed that based on the type of task, users favored certain interactions more to gather insights with the assistive technology. Dabek et al. [DC16] found in the evaluation of
their system that after being shown a suggestion, users ultimately performed the action that had been suggested to them 20% of the time. From the qualitative survey, users stated that they found the suggestions from the assistive technology useful, but did not always need them to solve the task. Lee et al. [LST*21] proposed a system, Frontier, which recommends new ways to visualize a given dataset. They observed that users followed recommendations while exploring unfamiliar attributes of the data or when they did not know what to explore next. Additionally, their findings suggest that interpretability of the recommendations positively impacted recommendation usage. From previous work we can see that assistive technologies impact user behavior in different ways. We aim to better understand human–machine partnerships by observing human behavior during data foraging.

2.2. Active Learning and Interactive Visualization

Active learning is a subfield of machine learning where learning algorithms are allowed to choose their training data strategically by querying an oracle for labels [Seo09, LGG*04]. Some examples of using active learning for model improvement are in domains of video and text analysis [HNH*12, TOO9, WSBD10]. Visual analytics researchers have identified the success of active learning algorithms as an opportunity for enabling human–machine collaborations. Most notably, Bernard et al. [BHZ*17] conduct an experimental study to compare active learning policies with interactive labeling based on visualizations. Additionally, Bernard et al. [BZSA18] unify the process of active learning and visual-interactive labeling into one conceptual framework (called VIAL) which can be applied to data exploration and labeling tasks. Lin et al. [LGG*17] present RCLens, a visual analytic system for identifying and exploring rare categories in a dataset. RCLens is powered by an active learning-based algorithm called rare category detection [HC09]. While the algorithmic component of RCLens was validated through a set of simulations, the evaluation pertaining to the visual component was limited to a small case study with five participants and primarily subjective feedback. We aim to build upon their work to investigate the impact of incorporating active search into analytic workflows by examining throughput, interface design, and task difficulty.

More specific to our work is active search, a realization of active learning where the goal is to iteratively search a large dataset for members of a rare, valuable class. When investigating a given point is expensive — e.g. it requires querying a human expert — we wish to direct the search process strategically to maximize discovery throughput. Garnett et al. [GKX*12] formalized active search as a sequential decision making process and showed how one can leverage a machine learning classifier trained on observed data to design queries maximizing the expected number of discoveries. Furthermore, Garnett et al. [GKW*11] demonstrated that active search outperforms traditional active learning policies (such as those used in previous visual analytic studies [KPSK17, LGG*17, LSDN19, BHZ*17]) when the goal is to discover rare, valuable data points. This technique has shown promise on a range of scientific discovery tasks including drug discovery and materials design [JMC*17, JMMG18]. The efficient nonmyopic search (ENS) algorithm presented by Jiang et al. represents the state-of-art in active search algorithms in terms of empirical discovery throughput [JMC*17].

In this work, we use active search for interactive data 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 [SF*18] and drug discovery [GGVB15], the closest work on combining active search and HCI is by Klyuchnikov et al. [KM19], which uses hypothetical human input and product reviews for improved product recommendation. In another related attempt in the field of chemistry, Shields et al. [SSL*21] conducted an experiment where human experts were compared to the Bayesian optimization algorithm in their abilities to search a space with the goal discovering a chemical with optimal characteristics. Although their findings suggest that Bayesian optimization outperforms human experts, the outcome of creating a human–machine partnership in this area is still an open question. To the best of our knowledge, this is the first time that active search has been applied in an interactive setting and evaluated on human subjects.

3. Defining Visual Data Foraging as an Active Search Problem

Human–machine collaboration refers to the process of two or more agents working towards a shared goal where at least one agent is a human and at least one agent is a computer [Ter95]. Using this conceptual framework, we assume the human–machine collaborative workflow depicted in Figure 1. Starting with a dataset, we create an interactive visualization with which the analyst interacts in order to inspect individual data points. As the user sequentially inspects and discovers relevant data points, we create a cycle where user interactions with the interface train a classifier on the relevance of unobserved data points and the active search algorithm picks a set of promising points to present to the analyst for investigation. To integrate active search into an interactive visualization, we need to define each of the components in Figure 1. To specify this human–machine framework, we need:

- a probabilistic model for inferring relevance of data points (A1),
- a querying policy to provide strategic recommendations (D), and
- an interactive visualization of the data (A2) and means of communication between human and the machine (B, C, E).

More formally, we assume there is a dot-based visual metaphor for a given data set, \( \mathcal{X} = \{x_1, x_2, \ldots, x_n\} \), where each data point in \( \mathcal{X} \) has a representative on the visualization (Fig. 1, A2). 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, \( \mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\} \), 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 data set is typically labeled (i.e., \( m \ll n \)). The objective is to recover as many relevant points as possible, defined by the utility function \( u \) where:

\[
u(\mathcal{D}) \triangleq \sum_{y_i \in \mathcal{D}} y_i,
\]

which simply is the number of relevant points discovered.

### 3.1. Probabilistic Model for Inferring Relevance

In an active search procedure, the algorithm relies on a model to predict the relevance of unobserved data points in light of observations (Fig. 1, A1). This model is 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 total number of discoveries at the end of the search process (Fig. 1, D). As suggested by Garnett et al. [GKX*12], we pick a simple k-NN model that provides the posterior probability of an unlabeled point \( x \) being relevant given the observed data: \( \Pr(y = 1 \mid x, \mathcal{D}) \).

This choice of model is non-parametric, fast to update in light of new observations, and is simple in that it only relies on a distance metric between data points. Some examples of distance functions for various structured and unstructured data types include the Euclidean distance for numerical values, Word Mover’s Distance for text documents [KSKW15], and ImageNet for images [DF11]. In scenarios where datasets contain multiple (say \( d \)) attributes, practitioners may build a \( k \)-NN model on each attribute, \( \{M_1, M_2, \ldots, M_d\} \), and merge the predictions via the following weighted sum where the values of \( q_i \in [0, 1] \) are tuned to maximize the likelihood of observed interactions. Then, the posterior belief is:

\[
\Pr(y = 1 \mid x, \mathcal{D}) = \sum_{i=1}^{d} q_i \Pr(y = 1 \mid x, \mathcal{D}, M_i).
\]

### 3.2. Active Search Querying Policy

Active search relies on a policy or strategy to sequentially select unlabeled points with the goal of maximizing discovery under a limited querying budget. In the ideal case, the policy would consider the entire remaining budget when selecting future points for inspection. With the set of data points being the action space and resulting in a large branching factor, this policy becomes computationally intractable. As an alternative Garnett et al. [GKX*12] recommend myopically looking \( \ell \) steps into the future, where \( \ell \) is a small number.

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.

As we increase the look-ahead horizon \( \ell \), the policy becomes less myopic, meaning it will explore the dataset. Since computing this policy for \( \ell > 2 \) becomes intractable quickly, Jiang et al. [JMC*17] propose the efficient non-myopic active search (ENS) algorithm as an approximation of the optimal policy which considers the entire remaining querying budget. ENS exhibits a naturally non-myopic behavior, balancing exploitation and exploration in its recommendations (a classic trade-off in sequential decision making settings), while still being tractable to compute.

### 3.3. Interactive Interface for Human–Machine Communication

Once the model over data relevance and an active search algorithm are in place, the next primary consideration is the communication between humans and the active search procedure. In this workflow, we consider a bidirectional communication channel in which the active search algorithm needs a proxy to receive feedback from user interactions and humans need to be presented with active search queries through user friendly means.

Visual analytics researchers have analyzed low-level interactions to uncover information about users and the task at hand. In particular, they have discovered that analyzing interactions can result in better performance at inferring user expertise [BBT’19], inferring exploration patterns [FPH18, MG020], and modeling the cognitive sense-making process [PJU09]. These successful attempts at analyzing interactions naturally bring us to the following question: can user interactions with a system provide an active search algorithm with a seamless, yet robust, labeling mechanism?

In the simple case which we examine in Section 4, certain low-level interactions can directly map into certain training labels for active search (Fig. 1, B and C). For example, clicking a button to bookmark a data point (or disregard one) can signal a positive (or negative) label. In more complex settings, however, a more ambiguous set of interactions may be used to provide labels seamlessly.

Finally, 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 (Fig. 1, E). 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 the ML-guided queries to be presented to the user in the form of visual cues. In this work, we assume the user leads the analysis and take a non-intrusive approach where active search queries are presented in a distinct color on the visualization. Depending on the visual channels available for a specific application, the risk associated with missing a relevant data point, and the intended degree of human involvement in analysis, practitioners may design other methods of interactive queries.

### 4. Proof of Concept Prototype

The Visual Analytics Science and Technology (VAST) Challenge publishes datasets and analytic tasks annually which are designed to mimic challenges faced in real-world scenarios [CGW14, SWPG12]. We wanted to select a use-inspired dataset and task that might reflect complexities of data foraging for an intelligence analyst. In the 2011 VAST Challenge, the scenario involved an epidemic spread in which professionals at local hospitals have noticed a sudden surge in illnesses. The challenge asked participants to identify the infected areas and characterize the spread.
of the disease, among other tasks. The data set included 1,023,077 microblogs (tweet-like messages) posted on social media from various parts of town during a 21-day period and a satellite image of the city with labeled highways, hospitals, landmarks, and water bodies.

4.1. Probabilistic Model for Inferring Relevance Selection

To model the relevance of unobserved points in light of observations, we build two $k$-NN models over this data set. The first one ($M_L$) 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 ($M_T$) 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 [RS10]. Given some observed data, each of these two models, $M_i$, calculates the probability that an unlabeled data point $x$ is relevant: $Pr(y = 1 | x, D, M_i)$.

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 $1 - q$):

$$Pr(y = 1 | x, D, M_T, M_L) = q \cdot Pr(y = 1 | x, D, M_T) + (1 - q) \cdot Pr(y = 1 | x, D, M_L),$$

where $q$ is chosen using the maximum-likelihood estimation method to maximize the likelihood of the observed data $D$. Prior to deploying this model in our user study prototype, we conducted a cross validation experiment to ensure that this model is appropriate for our data set (results included in the supplemental material).

4.2. Active Search Querying Policy Selection

For this prototype, we select our active search policy based on two criterion: (1) discovery performance and (2) computational runtime. To simulate different querying behaviors in the non-interactive setting, we apply three search policies (random, one-step, ENS-50) to the aforementioned dataset. The details of our simulation are included in the supplemental material. Results of this simulation suggest that active search accelerates discovery of illness-related microblogs in a non-interactive setting given our dataset. Furthermore, it demonstrates its promising performance in for varying incidence rates in our dataset. While ENS-50 policy performs best, we do not choose it for our prototype due to longer runtime. With the goal of generating queries in real-time, we select the one-step (or greedy) active search policy for our prototype.

4.3. Interactive Interface

Using the VAST 2011 challenge dataset, we created an interactive map visualization of the geotagged microblogs. The interface of our prototype is shown in Figure 2. We aimed for a simple interface and clear means of interaction for greater usability. There are two primary components on our interface: an interactive map visualization of microblogs based on their posting location (details to follow), and a side bar containing a list of current bookmarks, time remaining for task completion, and control buttons to report technical issues or leave the experiment. Although a text-based visualization (e.g. t-SNE) may seem more appropriate for the given
task, we assume the primary attribute determining relevance (i.e. microblog text in our case) is not known a priori.

Hovering on data points triggered 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. Experiment 1: Data Foraging Throughput
To investigate the impact of active search on visual exploration and discovery, we designed a crowd-sourced user study† that tasked participants to interact with a map of the fictional city of Vastapolis (Figure 2) which is under a biochemical attack causing an epidemic. We use this scenario as an openly available proxy for studying real-world intelligence analysis sessions. According to Pirolli and Card [PC05], intelligence analysts approach their tasks in either a top-down or bottom-up fashion. In the top-down approach, analysts start with a hypothesis and search through the data to find supporting evidence. On the other hand in the bottom-up approach, analysts start interacting with the data first and form new hypotheses as they discover potentially relevant data points. In our crowd-sourced user study, we narrowed the scope of the experiment to the bottom-up, data foraging phase of intelligence analysis.

5.1. Task
Participants were told that health professionals had reported a spike in reported illnesses with flu-like symptoms, including fever, chills, sweats, nausea and vomiting, and diarrhea. We informed participants that the authorities are interested in identifying the impacted parts of the city by analyzing social media activity, and that we have access to social media posts and their posting locations. Their task was to assist the authorities by searching through a dataset of microblogs using the interactive interface shown in Figure 2 and bookmarking as many posts containing illness-related information as possible. For this study, we narrowed the dataset to a random sample of 3000 points from the approximate start of the epidemic. About 33% of the data points contained illness-related content.

5.2. 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 Section 5.4, there were 46 women, 76 men, and 1 participant with undisclosed sex in our subject pool with ages ranging from 18 to 62 years (μ = 36, σ = 9). About 72% of our participants self-reported to have at least an associate degree. The average completion time (including reading the tutorial, performing the task, and completing the survey) was 12 minutes. The instructions specified that participants will be compensated $1.00 base pay and an additional $0.10 bonus for every relevant microblog they identify (with a maximum of $4.00). 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.

5.3. Procedure
The experiment complied with an approved protocol per Washington University’s IRB. Workers who accepted the HIT followed a URL to the study platform. Our system randomly assigned each participant to one of the following groups: the active search group, which received a batch of 10 active search queries in the form of visual clues that were updated after every bookmark, and the 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 initial “clues,” and in particular the active search group did not receive assistance for selecting their first bookmark. Participants were given at most 10 minutes to identify as many microblogs related to the epidemic as they could using an interactive map visualizing microblogs as dots placed on their posting locations. Hovering on visualized dots triggered a tooltip containing the post, and users could click on a button to bookmark the post if they judged it to contain illness-related content. Once the users were either satisfied with their search for illness-related documents or the 10 minutes were up, they were directed to a post-experiment survey to collect demographic information and general feedback on the system. In case our participants experienced technical difficulties with the system, we provided them with the option to report issues, gracefully exit the session, and receive compensation.

5.4. Data Collection, Cleaning, and Exclusions
We analyze our user study data by focusing on two interactions: inspection of microblogs (hovers) and discovery of relevant posts (bookmarks). These two types of interaction inform us about the speed and accuracy of visual data foraging through the metrics listed in Table 1. The bookmark and hover purity metrics are the proportion of bookmarks and hovers that involved relevant data points, respectively. The bookmarks- and hovers-per-minute metrics inform us about the speed at which users interacted with data points. The relevant hovers and relevant bookmarks-per-minute metrics are the rate at which users interacted with relevant data points, quantifying both speed and accuracy of interactions. Finally, we measure the number of relevant bookmarks discovered by the end of the session and number of unique illness-related keywords contained in the discovered microblogs.

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 based on the following four criterion. We eliminated participants who:

† This experiment was pre-registered on Open Science Foundation.
1. failed the survey attention checks (eliminating 1 subject),
2. reported technical issues with the tool (eliminating 1 subject),
3. hovered on less than 10 data points (eliminating 4 subjects) – we consider a valid hover to be one that lasts at least 500 milliseconds (300 milliseconds for triggering the tooltip, and 200 milliseconds for skimming the text), and
4. did not meet the age qualification (eliminating 1 subject).

A total of 123 subjects remained after filtering (74 in the control group and 49 in the active search group).

5.5. Results

We use a labeling heuristic where microblogs containing a predefined list of illness-related keywords (e.g. flu, fever, diarrhea) are assigned a label of relevant. The full list of keywords used for this heuristic is included in the supplemental material. These labels were hidden from the active search algorithm, which generated suggestions solely based on observed interactions. Thus, the labeling heuristic serves only as a proxy for ground truth in our analysis.

**Suggestion Quality:** We begin our analysis by examining the quality of suggestions provided by the active search algorithm when seeded with users’ interaction data. We define suggestion purity to be the proportion of unique relevant microblogs recommended to the user throughout a given session. On average, active search group participants had a suggestion purity of 79%. We observe a moderate positive correlation between bookmark purity and suggestion purity ($R_{adj}^2 = 0.594$, $p < 0.0001$), suggesting that the active search algorithm provides useful recommendations for participants who interacted with microblogs containing known symptoms.

**Suggestion Usage:** We observed an unexpected pattern in the active search group. As shown in Figure 3, for approximately 24% of participants in the active search group, suggested microblogs accounted for less than 10% of their bookmarks. 9 out of 49 participants did not bookmark any of the suggestions presented to them at all. Further inspection reveals that the 9 active search participants who ignored the suggestions had on average $82\pm9\%$ suggestion purity and $76\pm11\%$ bookmark purity. This compares to the 40 active search participants who did interact with the suggestions, who had on average $79\pm5\%$ suggestion purity and $82\pm5\%$ bookmark purity. Finally, we observed a difference between how subjects reported their trust towards system suggestions on a 1–5 Likert scale in the post-experiment survey ($3.3\pm0.46$ for those who ignored suggestions vs. $4.2\pm0.24$ for those who interacted with the suggestions).

Moving forward, our analyses focus on the impact of suggestions on data exploration and information foraging. Thus, we exclude the 9 participants in the active search group who did not interact with the system’s suggestions, leaving us with 74 participants in control group and 40 participants in the active search group.

**The Effect of Suggestions on Data Foraging:** We performed a series of two-sample $t$-tests to investigate differences in behavior in our two study conditions: control and active search. Table 1 summarizes our findings. We found that participants in the active search group bookmarked ($t(112) = 3.98, p = 0.0001; d = 0.79$) and hovered over ($t(112) = 4, p = 0.0001; d = 0.79$) significantly more relevant microblogs per minute than the control group. Furthermore, our findings show that the active search group performed fewer exploratory hovers per minute ($t(112) = -2.58, p = 0.0112; d = -0.51$) than the control group, implying that the suggestions resulted in a more efficient exploratory analysis. For a more fine-grained analysis, we examine bookmark discoveries as a function of time. Figure 4 shows the average number of bookmarks over time for the active search and control groups. We can observe that the active search group consistently outperformed the control group by bookmarking more relevant microblogs throughout the ten-minute information foraging session. However, it is noteworthy that although the suggestions improved the quantity of the bookmarks, we found no measurable difference in the quality or content of the bookmarked discoveries. Both the active search and control groups collectively examined similar geographical regions and symptom sets (see supplemental material for analysis details).

**Impact of Active Search Suggestions on Usability:** In a post-experiment survey, we asked subjects in both groups three questions on willingness to use, ease of use, and ease of task completion. We performed a Mann–Whitney U statistical test at $\alpha = 0.05$ to determine if there was a significant difference between the control and active search groups. The analysis showed some evidence that the active search group found the system easier to use ($U = 1199.50$, $p = 0.0336$, $r = 0.16$) and were more willing to use the system frequently ($U = 1192$, $p = 0.0362$, $r = 0.16$). However, the effect sizes were small for both. Furthermore, we did not find a significant difference between the control and active group’s response to ease of task completion ($U = 1397.50$, $p = 0.2989$, $r = 0.07$).
6. Experiment 2: Interface Design and Task Difficulty

From the data collected in our post-experiment survey, we observe an encouraging and consistent tendency among the active search group to find the system and task easier and being more willing to use the interface. However, it is noteworthy that a non-trivial percentage of the active search group ignored the suggestions entirely, and those participants reported lower levels of perceived trust in the system. Furthermore, our findings show that the data foraged from the active search and control groups would produce similar conclusions. Both groups yielded a similar set of keywords and geographical coverage. This convergence in data exploration indicates that the baseline task was reasonably manageable without the machine’s assistance. Additionally, participants in the control group generally believed that the task was easy to complete.

We conduct a follow-up user study to further investigate the relationship between the effectiveness of assistive technology and human behavior in visual data foraging. Specifically, we aim to understand:

1. How a more explainable interface design may impact human engagement with computer suggestions, and
2. How task difficulty may impact the marginal gain from a human–machine partnership.

Similar to Experiment 1, the participants’ task was to search through a data set of microblogs via an interactive map and to bookmark posts containing illness-related content. There were three primary differences in this study. First, we displayed a random sample of 2000 microblogs — a random sample of the VAST 2011 data set with ~9% of points being related to illness — which results in a more difficult foraging task due to the smaller number and ratio of relevant data points. In contrast to Experiment 1 which focused on the start of the epidemic, we select the random sample of points from the entire 21-day period, hence relaxing the assumption that the starting point of the epidemic is known. Second, inspired by existing literature on building trust in human–machine partnerships [DPP*, DLW*17], we presented active search recommendations to the user in a more explainable manner by displaying a word with the highest influence on recommending a given data point (shown in Figure 5, right). We define an influential keyword to be one which if eliminated, decreases the probability of a microblog being relevant by the largest amount. Lastly, participants in the active search group received suggestions from the system after bookmarking 5 microblogs whereas participants in Experiment 1 received suggestions after bookmarking only one microblog. This change will allow the algorithm to gain more confidence in users’ latent interest by observing more labels before presenting suggestions, thereby reducing model uncertainty.

6.1. Participants

We recruited 97 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 Section 6.3, there were 30 women, 61 men, and 1 participant with undisclosed sex in our subject pool with ages ranging from 20 to 60 years (µ = 36, σ = 10). About 57% of our participants self-reported to have at least an associate degree.

---

\[\text{Table 1: Experiment 1} \rightarrow \text{The results of two-sample} t\text{-tests on the metrics discussed in Section 5.4}\]

| Metric                          | Control n=74 | Active Search n=40 | p-value | t-statistic | Cohen’s d |
|--------------------------------|--------------|--------------------|---------|-------------|-----------|
| Hovers per Minute              | 16.7 ± 1.19  | 14.3 ± 1.23        | 0.0112  | −2.58       | −0.51     |
| Relevant Hovers per Minute     | 6.7 ± 0.68   | 9.2 ± 1.12         | 0.0001  | 4.00        | 0.79      |
| Hover Purity                   | 0.39 ± 0.02  | 0.63 ± 0.05        | <0.0001 | 9.70        | 1.92      |
| Bookmarks per Minute           | 6.9 ± 0.77   | 9.5 ± 1.41         | 0.0006  | 3.52        | 0.70      |
| Relevant Bookmarks per Minute  | 5.4 ± 0.68   | 8.1 ± 1.26         | 0.0001  | 3.98        | 0.79      |
| Bookmark Purity                | 0.77 ± 0.04  | 0.82 ± 0.05        | 0.2249  | 1.22        | 0.24      |
| Relevant Microblogs Bookmarked | 53.9 ± 6.80  | 73.4 ± 11.50       | 0.0026  | 3.09        | 0.61      |
| Unique Keywords Identified     | 16.1 ± 0.83  | 15.6 ± 1.32        | 0.4980  | −0.68       | −0.13     |

---

\[\text{CI} \rightarrow 95\%\]
The average completion time (including reading the tutorial, performing the task, and completing the survey) was 11.4 minutes. Similar to Experiment 1, study participants received a $1.00 base pay plus a $4.00 bonus upon completion.

6.2. Task, Procedures, and Data Collection
The task, procedures, and data collection for this experiment were similar to Experiment 1 (Sections 5.3 and 5.4), except for the design differences detailed above.

6.3. Data Cleaning and Exclusions
In an identical process to the one outlined for Experiment 1 (Section 5.4), we filtered the collected data to exclude participants who did not attempt the task or were unable to finish the experiment. Upon excluding 5 participants according to our established criterion, a total of 92 subjects remained (46 in the control group and 46 in the active search group). More details on the number of exclusions per filtering criteria is included in the supplemental material.

6.4. Results
Suggestion Usage To examine H1, we inspect the proportion of bookmarks from suggestions for the active search group in Experiment 2. Figure 2 shows that only a tiny proportion of users ignored the recommendations. Moreover, the system recommendations accounted for more than 80% of the bookmarks for about 70% of the participants. On average, active search suggestions accounted for 53 ± 10% of bookmarks in Experiment 1 and 75 ± 7% of bookmarks in Experiment 2. Overall, participants were significantly more engaged with active suggestions in Experiment 2 than those in Experiment 1 ($t(93) = 3.54, p = 0.0006, d = 0.73$).

The Effect of Task Difficulty on Marginal Benefit of Assistive Technology: To investigate the marginal benefit of assistive technology on discovery throughput, we define speedup to be the increase in the average number of discoveries by the active search group:

$$\text{speedup} = \frac{\mathbb{E}[\# \text{discoveries by active search group}]}{\mathbb{E}[\# \text{discoveries by control group}]} - 1.$$

A speedup of 0 indicates no change in the total number of discoveries between two groups, a speedup of $< 0$ indicates a decrease in the total number of discoveries in the active search group, and a speedup of $> 0$ indicates an increase in the total number of discoveries in the active search group. For convenience, we perform this analysis in the log domain. Taking the logarithm of the first term in speedup gives

$$\log(\mathbb{E}[\# \text{active search}]) - \log(\mathbb{E}[\# \text{control}]) \geq \mathbb{E}[\log \# \text{active search}] - \mathbb{E}[\log \# \text{control}],$$

where we have applied Jensen’s inequality. Thus a confidence interval on the difference between the log number of discoveries provides a conservative (that is, one underestimating the effect of active search) confidence interval on the log speedup, which we can transform to a conservative confidence interval on speedup through exponentiation (which is monotonic). We can compute the desired confidence intervals in the log domain via a two-sample Student $t$ test, noting the log transform has the additional side effect of encouraging normality. The resulting conservative 95% confidence intervals for speedup are ($-13\%, 68\%$) in Experiment 1 and ($80\%, 235\%$) in Experiment 2.

Observations on Reported Perceived Trust: Although the results of Experiment 1 suggest an association between perceived trust and engagement, this relationship is not apparent in Experiment 2. Unexpectedly, the updated interface in Experiment 2 produced higher levels of suggestion usage, but the participants in that study reported significantly lower trust ratings than those in Experiment 1 ($U = 666.5, p = 0.017, r = -0.447$). These results suggest that engagement alone is not a good proxy for trust, and our coarse, one-question assessment does not capture the full complexity of this abstract concept. We uncover one plausible explanation for the low trust ratings in Experiment 2 when we inspect suggestion purity. The active search algorithm produced relevant suggestions 40% of the time in Experiment 2, compared to the 80% observed in Experiment 1.
iment 1 ($t(82) = -14.12, p < 0.0001, d = -3.1226$), which can be explained by the drastically more difficult task in Experiment 2. Therefore, even though the updates to the system design may have improved the explainability of the suggestions, it is possible that the high model uncertainty negatively impacted perceived trust.

7. Discussion

Our results indicate that a human–machine partnership can significantly improve information foraging and data discovery. For example, participants in the active search group hovered on fewer points per minute while hovering on more relevant points per minute than the control group. These findings show that the participants successfully disregarded irrelevant information and were more mindful towards the relevant data points. As a result, users with assistance from active search made more discoveries, a finding that can have high-impact implications for designing visual analytic tools for tasks such as intelligence analysis and scientific discovery.

Overall, we demonstrate that adding active search to the information foraging workflow improves an analyst’s throughput and the quality of interactions significantly. However, let us consider how this partnership may impact the overall downstream hypothesis generation and decision making process often present in intelligence analysis settings. According to our findings, participants from both groups in Experiment 1 arrived at a similar set of discoveries (in terms geographical and symptom coverage). Therefore, we may expect the downstream hypotheses and decisions based on the discovered data by each group to be similar. This convergence indicates that the baseline task was reasonably manageable without the machine’s assistance as reflected by subjective survey results.

Experiment 2 used a dataset where only 9% of the microblogs mentioned illnesses. This noisy data produced a more challenging task for the participants in the study and the machine learning models. We observed that the value added by active search assistance was related to the complexity of the task. This finding leads to the perhaps most obvious conclusion that assistive technology might be most useful when the job is too challenging for the human to perform alone. However, defining task difficulty is not necessarily clear-cut. In our study, we measure difficulty based on the signal-to-noise ratio. Still, there are many other potential metrics for defining task complexity, such as the type of task [BM13] and individual differences in the person’s abilities and expertise [LCO20].

Our findings highlight a vital factor for nurturing the human–machine partnership so that analysts can accelerate the process of data foraging. To reap the benefits, the analysts need to be able to trust the suggestions provided by their machine teammates. In experiment 1, a non-trivial percentage of the active search group ignored the suggestions entirely, and those participants reported lower levels of perceived trust in the system. This raises the question of how explainable suggestions from active search might impact the interaction behavior of participants. Motivated by existing work on trust and design, [DPP*, DLW*17, SSK*16], experiment 2 examined the relationship between engagement, trust, and task difficulty and confirmed that interface design mediates engagement. In the first study, we offered visual cues to aid in discovering relevant data points and observed that some participants ignored the suggestions entirely. In contrast, the interface in the second study provided keyword explanations for the visual cues. Additionally, we delayed the suggestions until the system was more confident in its suggestions. Our findings suggest that these design changes may have encouraged engagement, but the participants in that study reported significantly lower trust ratings than those in Experiment 1. Altogether, this reinforces that trust is a complex and multi-faceted construct, but it is essential for human–machine partnership [SSK*16, HS20] and warrants further investigations.

8. Future Work and Limitations

In our studies, we used bookmark functionality to label data points. 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.

Another option for optimizing its utility is for the assistive technology to learn what, how, and when to make suggestions, as providing unnecessary help may become bothersome to the human [HJH99]. In a paper on computational politeness, Whitworth [Whi05] argues that perceived politeness in automated assistants depends on factors such as respecting user choices and feedback. Future work can examine whether various levels of human and machine control in guiding exploration may impact the overall utility.

Finally, we envision future work to apply human–machine partnership aided by active search to applications with different characteristics. For example, drug discovery involves identifying promising chemical compounds and performing expensive experiments to determine their effectiveness. A human expert or an active search algorithm decide which chemical compounds are worth lab trials in this setup, and mother nature determines the actual label after the experimentation. We can envision a version of our proposed technique that helps combine human intuition, active search efficiency, and nature feedback to accelerate drug discovery.

9. Conclusion

In this paper, we framed visual data foraging as a sequential decision-making process. We proposed augmenting interactive visualizations with active search to improve the process through a human–machine collaboration. To investigate the impact of this assistive technology on visual data foraging, we conducted two crowd-sourced user studies. Our findings from these experiments indicate that this human–machine partnership could significantly improve information foraging and data discovery, explainable interface design encourages more user engagement with their machine teammate, and task complexity moderates assistive technology’s benefit. Finally, we presented an in-depth discussion of our findings and outlined promising avenues for future investigations.
References

[BBT*19] Boukhelifa N., Bezerianos A., Treflea I. C., Perrot N. M., Lutton E.: 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 (2019), pp. 1–14. 4

[BCS16] Battle L., Chang R., Stonebraker M.: Dynamic Prefetching of Data Tiles for Interactive Visualization. In Proceedings of the 2016 International Conference on Management of Data (2016), pp. 1363–1375. 2

[BHZ*17] Bernard J., Hutter M., Zeppelzauer M., Fellner D., Seldmair M.: Comparing visual-interactive labeling with active learning: An experimental study. IEEE transactions on visualization and computer graphics 24, 1 (2017), 298–308. 3

[BLBC12] Brown E. T., Liu J., Brodley C. E., Chang R.: Dis-Function: Learning Distance Functions Interactively. In 2012 IEEE Conference on Visual Analytics Science and Technology (VAST) (2012), IEEE Computer Society, pp. 83–92. 2

[BM13] Breher M., Munzer T.: A multi-level typology of abstract visualization tasks. IEEE transactions on visualization and computer graphics 19, 12 (2013), 2376–2385. 10

[BOZ*14] Brown E. T., Ottley A., Zhao H., Lin Q., Souvenir R., Endert A., Chang R.: Finding Waldo: Learning about Users from their Interactions. IEEE Transactions on Visualization and Computer Graphics 20, 12 (2014), 1663–1672. 2

[BWN19] Biały Y., Wenskovitch J., North C.: DeepVA: Bridging Cognition and Computation through Semantic Interaction and Deep Learning. Machine Learning from User Interactions for Visualization and Analytics at IEEE VIS (2019). 2

[BZSA18] Bernard J., Zeppelzauer M., Seldmair M., Aigner W.: Vial: a unified process for visual interactive labeling. The Visual Computer 34, 9 (2018), 1189–1207. 3

[CC12] Crouser R. J., Chang R.: An Affordance-Based Framework for Human Computation and Human-Computer Collaboration. IEEE Transactions on Visualization and Compu-ter Graphics 18, 12 (2012), 2859–2868. 1

[CGM*16] Ceneda D., Gschwandtner T., May T., Miksch S., Schulz H.-J., Streit M., Tominski C.: Characterizing Guidance in Visual Analytics. IEEE Transactions on Visualization and Computer Graphics 23, 1 (2016), 111–120. 2

[CGW14] Cook K., Grinstein G., Whiting M.: The vast challenge: History, scope, and outcomes: An introduction to the special issue, 2014. 2, 4

[COC13] Crouser R. J., Ottley A., Chang R.: Balancing Human and Machine Contributions in Human Computation Systems. In Handbook of Human Computation. Springer, 2013, pp. 615–623. 1

[DC16] Dabek F., Caran J. J.: 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. 2

[DCCE19] Dàs S., Cashman D., Chang R., Endert A.: Gaggle: Visual analytics for model space navigation. 2

[DF11] Deselaers T., Ferrari V.: Visual and Semantic Similarity in ImageNet. In CVPR 2011 (2011), IEEE, pp. 1777–1784. 4

[DLW*17] Dasgupta A., Lee J.-Y., Wilson R., LaFranc R. A., Kramer N., Cook K., Payne S.: Familiarity vs trust: A comparative study of domain scientists’ trust in visual analytics and conventional analysis methods. 271–280. 8, 10

[DPP*] Dzindolet M. T., Peterson S. A., Pomranky R. A., Pierce L. G., Beck H. P.: The role of trust in automation reliance. 22. 8, 10

[EFN12] Endert A., Fiaux P., North C.: Semantic Interaction for Visual Text Analytics. In Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems (2012), pp. 473–482. 1, 2

[FPH18] Feng M., Peck E., Harrison L.: Patterns and Place: Quantifying Diverse Exploration Behavior with Visualizations on the Web. In 2018 IEEE Conference on Visual Analytics Science and Technology (VAST) (2018), IEEE Computer Society, pp. 501–511. 4

[GCH*11] Grinstein G., Cook K., Havig P., Liggett K., Nebesh B., Whiting M., Whiteley K., Knoeck S.: Vast 2011 challenge: Cyber security and epidemic. IEEE VAST 2011 (2011), 299–301. 2

[GDM*19] Guo S., Du F., Malik S., Koh E., Kim S., Liu Z., Kim D., Zha H., Cao N.: Visualizing uncertainty and alternatives in event sequence predictions. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019), pp. 1–12. 2

[GGVB15] Garnett R., Gartner T., Vogt M., Bajorath J.: Introducing the ‘active search’ method for iterative virtual screening. Journal of Computer-Aided Molecular Design 29, 4 (2015), 305–314. 3

[GIC*19] Guo S., Jin Z., Chen Q., Gotz D., Zha H., Cao N.: Visual anomaly detection in event sequence data. In 2019 IEEE International Conference on Big Data (Big Data) (2019), IEEE, pp. 1125–1130. 2

[GKC*16] Garnett R., Krishnamurthy Y., Wang D., Schnei-der J., Mann R.: Bayesian optimal active search on graphs. In Ninth Workshop on Mining and Learning with Graphs (2011), Citeseer. 3

[GKX12] Garnett R., Krishnamurthy Y., Xiong X., Schnei-der J., Mann R.: Bayesian Optimal Active Search and Surveying. In Proceedings of the 29th International Conference on Machine Learning (2012). 1, 3, 4

[GSC16] Gotz D., Sun S., Cao N.: Adaptive Contextualization: Combating Bias During High-Dimensional Visualization and Data Selection. In Proceedings of the 21st International Conference on Intelligent User Interfaces (2016), pp. 85–95. 2

[HC09] He J., Carbonell J.: Nearest-neighbor-based active learning for rare category detection. In 21st Annual Conference on Neural Information Processing Systems, NIPS 2007 (2009). 3

[HHH99] Horvitz E. J., Jacobs A., Hovel D.: Attention-Sensitive Alerting. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (1999), pp. 305–313. 10

[HHN*12] Höferlin B., Netzel R., Höferlin M., Weiskopf D., Heidemann G.: Inter-Active Learning of Ad-Hoc Classifiers for Video Visual Analytics. In 2012 IEEE Conference on Visual Analytics Science and Technology (VAST) (2012), IEEE Computer Society, pp. 23–32. 3

[HS20] Han W., Schulz H.-J.: Beyond trust building and calibrating trust in visual analytics. In 2020 IEEE Workshop on Trust and Expertise in Visual Analytics (TREX) (2020), IEEE, pp. 9–15. 10

[IHG13] Iwata T., Houlsby N., Ghahramani Z.: Active Learning for Interactive Visualization. In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (2013), pp. 342–350. 2

[JCG*20] Jin Z., Cui S., Guo S., Gotz D., Sun J., Cao N.: Carepree: An intelligent clinical decision support system. ACM Transactions on Computing for Healthcare 1, 1 (2020), 1–20. 2

[JMC*17] Jiang S., Malkomes G., Converse G., Shofner A., Moseley B., Garnett R.: Efficient Nonmyopic Active Search. In Proceedings of the 34th International Conference on Machine Learning (2017), pp. 1714–1723. 1, 3, 4

[JMMG18] Jiang S., Malkomes G., Moseley B., Garnett R.: Efficient nonmyopic active search with applications in drug and materials discovery. Machine Learning for Molecules and Materials Workshop at NeurIPS (2018). 3

[KAP*08] Keim D., Andrienko G., Fekete J.-D., Görg C., Kohlhammer J., Melançon G.: Visual Analytics: Definition, Process, and Challenges. In Information Visualization. Springer, 2008, pp. 154–175. 1

[KCD*19] Kim H., Choi D., Drake B., Endert A., Park H.: Topic-Site Interactive Search Space Reduction through Targeted Topic Modeling. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST) (2019), IEEE Computer Society, pp. 35–45. 1, 2

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.
[KJO*19] Kery M. B., John B. E., O'Flaherty P., Horvath A., Myers B. A.: Towards effective foraging by data scientists to find past analysis choices. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (2019), pp. 1–13. 2

[KMK*19] Klyuchnikov N., Mottin D., Koutrika G., Müller E., Karras P.: Figuring out the User in a Few Steps: Bayesian Multifidelity Active Search with Cokriging. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2019), pp. 686–695. 3

[KPSK17] Kucher K., Paradis C., Sahlgren M., Kerren A.: Active Learning and Visual Analytics for Stance Classification with ALVA. ACM Transactions on Interactive Intelligent Systems (TiiS) 7, 3 (2017), 1–31. 2, 3

[KSKW15] Kusner M., Sun Y., Kolkin N., Weinberger K.: From Word Embeddings to Document Distances. In Proceedings of the 32nd International Conference on Machine Learning (2015), pp. 957–966. 4

[LCO20] Liu Z., Crouser R. J., Otley A.: Survey on individual differences in visualization. In Computer Graphics Forum (2020), vol. 39, Wiley Online Library, pp. 693–712. 10

[LG94] Lewis D. D., Gale W. A.: A Sequential Algorithm for Training Text Classifiers. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1994), pp. 3–12. 3

[LGG*17] Lin H., Gao S., Gotz D., Du F., He J., Cao N.: Relens: Interactive rare category exploration and identification. IEEE transactions on visualization and computer graphics 24, 7 (2017), 2223–2237. 3

[LSD19] Legg P., Smith J., Downing A.: Visual analytics for collaborative human-machine confidence in human-centric active learning tasks. Human-centric Computing and Information Sciences 9, 1 (2019), 1–25. 3

[LST*21] Lee D. J.-L., Setlur V., Tory M., Karahalios K. G., Parameswaran A.: Deconstructing categorization in visualization recommendation: A taxonomy and comparative study. 3

[MGO20] Monadjemi S., Garnett R., Otley A.: Competing Models: Inferring Exploration Patterns and Information Value via Bayesian Model Selection. In 2020 IEEE Conference on Visual Analytics Science and Technology (VAST) (2020), IEEE Computer Society. 4

[OGW19] Otley A., Garnett R., Wan R.: Follow The Clicks: Learning and Anticipating Mouse Interactions During Exploratory Data Analysis. Computer Graphics Forum 38, 3 (2019), 41–52. 2

[PC05] Pirolli L., Card S.: The Sensemaking Process and Lever age Points for Analyst Technology as Identified Through Cognitive Task Analysis. In Proceedings of International Conference on Intelligence Analysis (2005), vol. 5, McLean, VA, USA, pp. 2–4. 1, 2, 6

[PJIP09] Perry J., Janneck C. D., Umoja C., Pottenger W. M.: Supporting Cognitive Models of Sensemaking in Analytics Systems. Tech. rep., DIMACS, New Brunswick, NY, 2009. 4

[RS10] Rehouré R., Sojka P.: Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (2010), ELRA, pp. 45–50. 5

[SAEH*18] Sebastijanova R., El-Assady M., Hauhti-Janisz A., Kalouli A.-L., Kehlbeck R., Deussen O., Keim D. A., Butt M.: Mixed-Initiative Active Learning for Generating Linguistic Insights in Question Classification. Data Systems for Interactive Analysis at IEEE VIS (2018). 2

[Set09] Settles B.: Active Learning Literature Survey. Tech. rep., University of Wisconsin-Madison Department of Computer Sciences, 2009. 1, 3

[SFD*18] Siddiqui M. A., Fern A., Dietterich T. G., Wright R., Theriault A., Archer D. W.: Feedback-Guided Anomaly Discovery via Online Optimization. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2018), pp. 2200–2209. 1, 3