Finding Similar People to Guide Life Choices: Challenge, Design, and Evaluation

Fan Du Catherine Plaisant Neil Spring Ben Shneiderman
University of Maryland
{fan,plaisant,nspring,ben}@cs.umd.edu

ABSTRACT
People often seek examples of similar individuals to guide their own life choices. For example, students making academic plans refer to friends; patients refer to acquaintances with similar conditions, physicians mention past cases seen in their practice. How would they want to search for similar people in databases? We discuss the challenge of finding similar people to guide life choices and report on a need analysis based on 13 interviews. Our PeerFinder prototype enables users to find records that are similar to a seed record, using both record attributes and temporal events found in the records. A user study with 18 participants and four experts shows that users are more engaged and more confident about the value of the results to provide useful evidence to guide life choices when provided with more control over the search process and more context for the results, even at the cost of added complexity.

Author Keywords
Similarity; temporal event analytics; temporal visualization; decision making; visual analytics.

ACM Classification Keywords
H.5.2 User Interfaces: Graphical user interfaces (GUI)

INTRODUCTION
People often seek to use examples of similar individuals to guide their own life choices. For instance, patients may want to receive the treatments that work for others with similar physical conditions and disease symptoms, or new students may wish to follow the trajectory of former graduates who had similar backgrounds and academic performances and ended up with a successful career. In the era of big data, where electronic health records and electronic student records are commonplace, exploring the data of similar individuals to receive advice on life choices and foresee potential outcomes is becoming possible. However, finding the records of similar individuals from databases is an important yet difficult step, often overlooked or existing in some analytical applications only as a black-box process [7, 10, 36].

CHI 2017, May 06–11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025777

Imagine a patient suffering from a knee injury who wants to understand if people like her chose surgery first then physical therapy or a more conservative treatment, and wants to know how long before they return to normal use of their knees. But what data should be used as evidence for people like her? As a light-weight woman in her thirties, will she trust results based on data from older women? From strong athletes? From those with prior knee injuries? Or with several unrelated medical conditions? To narrow the results, more information from the medical record could be used to tailor the set of similar patients, such as the degree of everyday physical activities and previous knee conditions. Specifying such a query using standard tools is incredibly complex as a large number of rules need to be specified, and since every person is unique, the result set of specific and complex queries is likely to be empty.

To understand users' needs, we reflected and built on experience accumulated from working with case study partners (medical researchers, doctors, marketing and transportation analysts, etc.) for more than a decade while developing tools and interfaces for the exploration of personal records. Searching for similar records was requested by many users. Our long-term goal is to support prescriptive analytics interfaces that guide users as they make plans informed by the history of similar people [7, 10, 15, 18, 36]. Searching for similar records is the focus of this paper.

After summarizing the challenges in finding similar people, we report on the results of 13 interviews that informed our design effort. We implemented PeerFinder, a visual interface that enables users to find and explore records that are similar to a seed record (either their own record or the record of a person they intend to counsel). PeerFinder uses both record attributes and temporal event information. To encourage engagement and inspire users' trust in the results, PeerFinder provides different levels of controls and context that allow users to adjust the similarity criteria. It also allows users to see how similar the results are to the seed record. Intermediate results are displayed and users can iteratively refine the search.

Our contributions include:

- A clarification of the challenges in finding similar people to guide life choices and a need analysis with 13 interviews.
- A flexible prototype, PeerFinder, which allowed us to explore different levels of controls and context, and interface styles to refine the results.
- The results of a user study with 18 participants and 4 expert reviewers comparing three interface configurations.

CHALLENGES

Every person is unique, and finding similarities between individuals is a multifaceted and subjective process. This paper focuses on similarity in the context of making important life choices (and not other uses such as eliminating duplications, searching for criminal activity, or finding job applicants).

Trust in the Evidence Contained in the Results

Making life choices based on data found in similar records takes a leap of faith. It implies that users are confident that the found records are similar enough to them to provide personalized evidence to guide their choices, and that decisions that were optimal for similar records will also be optimal for them. This confidence may be based on (1) trust in the source of the data and algorithm (e.g., results coming from one's doctor or NIH may be trusted more than those coming from an unknown source), (2) previous experience (e.g., once results have been found useful, the next result may be more likely to be trusted), and (3) understanding how the results were obtained (e.g., looking under the hood and being able to adjust the search parameters) [14,28]. Increased knowledge may also (appropriately) lead to lower trust when users realize that the results are not really very similar to the seed record [32].

No Natural Computable Distance Measure

Electronic records of personal histories (e.g., patients, students, historical figures, criminals, customers, etc.) consist of multivariate data (e.g., demographic information) and temporal data (time-stamped events such as first diagnosis, hospital stays, interventions) where each event belongs to a category. Intuitively, we can consider a record that is identical to the seed record to be the most similar while a record with all opposite attribute values and no common events can be seen as the most dissimilar, but defining a nuanced similarity measure to rank records by similarity is challenging.

The similarity between numerical values (e.g., age or weight) can be easily assessed by standard distance functions and normalized. Ordinal values also lend themselves to such distance (e.g., student letter grades), but categorical values pose problems. Sometimes the distance between values can be estimated using a standard hierarchical structure, e.g., the ICD-10 codes [39] allows a distance measure between diseases to be computed. However, there are no natural distance measures for categorical attributes in general, such as between races or academic disciplines. Moreover, temporal events add enormous complexity to the similarity measurement: not only are there no natural distance metrics between event categories but there is no generally accepted method to rank differences in sequence patterns. Specifically, what should be the "distance" created by a missing event or a reordering of events?

Nevertheless, it is possible to define an initial similarity for each pair of records as a weighted composite of scores arbitrarily set for all individual measures and possible differences. Hundreds of arbitrary decisions have to be made, but users may be able to adjust those parameters for specific applications.

The Subjective Nature of Similarity

While there is no natural numerical distance between people, patients and students express very strong opinions about records being similar or dissimilar to them based on how they identify or not with the other person, making the notion of similarity very subjective. How people perceive similarity depends on their preferences, experiences, and beliefs, and has been dismissed by some as a slippery notion [6]. Educators may see students of different majors as absolutely dissimilar. Doctors may see as the most similar the patients that are taking exactly the same combination of drugs.

Similar for Which Purpose?

How people evaluate similarity is affected by their goals. Someone looking for medical guidance will most likely ignore the similarity of education or place of residence. We identified the following possible use of a similarity search:

- Compute outcome measures, e.g., to estimate the chance of developing a disease or achieving a desired goal. Here a large number of similar records are needed, and knowledge of which criteria influence the outcome will guide the similarity judgment. Physicians may know that having had children affects certain types of cancer but patients may not. Students may only consider publication activities to estimate the likelihood of getting a postdoc position.
- Identify stories to motivate. A physician may be trying to remember the case of a similar patient who had a good outcome to encourage a patient to follow a specific treatment. Here, gender and age may contribute little to the similarity of the clinical cases, but be required to motivate the patient.
- Make plans for future actions, e.g., to define long-term treatment plans based on the outcomes of similar patients or recommend interventions to retain a customer based on the histories of similar customers. Here the records' temporal information may become more important. For example, a student seeking course planning advice will put more weight on the similarity of the sequence of classes and grades.

Lack of Ground Truth Benchmark Data

Well-developed research topics such as face or image recognition, document search or topic classification have a long history and ground truth datasets have been developed to evaluate results of various algorithms and a much lesser extent of user interfaces. Even subjective judgments have been collected and aggregated. In contrast, searching for similar people to guide life choices is a new topic of research and there exists no benchmark dataset to train machine learning models or evaluate prototypes. Besides, since the data structure and perception of similarity vary among domains, it will be difficult to generalize the evaluation results gathered from one domain to others, so various benchmark datasets will be needed.

In summary, searching for similar records is technically easy using arbitrary distance measures, but similarity judgments are subjective and there is no validated measure or established ways to measure the quality of the result set before generating personalized evidence-based recommendations for life choices. Therefore, we believe that providing users with some control over the search and context information about the results is critical to building trust in the recommendations. This paper is a first investigation into the design space of a new research area: personalized search for similar personal records.

RELATED WORK

We discuss related work in similarity measures, similarity search of temporal data, and event sequence visualizations.

Similarity Measures

Psychologists conducted experiments to understand the similarity perceived by people, where they asked participants to compare objects and rate their similarity in a Likert scale [9, 21, 24]. Likewise, data scientists investigated how to measure the similarity between data cubes in multidimensional space. For example, Baikousi et al. [2] explored various distance functions to identify users' preferred measurements between values of a dimension and between data cubes. Spertus et al. [30] present an empirical evaluation of similarity measures for recommending online communities to social network users, where the effects of the measures were determined by users' propensity to accept the recommendation. Sureka and Mirajkar [31] studied similarity measures for online user profiles and discovered that different measures need to be used for different users to achieve the best results.

We extended existing work on similarity measures to temporal data, which is an important component of people's personal histories. Our interviews confirmed that choices of similarity measures rely on users' preferences and analysis goals, and our user studies revealed that providing controls and context will increase users' engagement and trust in search results.

Similarity Search of Temporal Data

To find a group of records with features in common with a seed record, one approach is to specify a query and the results are records that exactly match the query rules. Extensions to standard query languages (e.g., TQuel [29] and T-SPARQL [12]) have been introduced to ease the task of querying temporal data. Temporal queries typically consist of elements such as the required events, temporal relationships between the events, and attribute ranges of the events or records. Precisely formulating temporal queries remains difficult and time-consuming for many domain experts. Visual tools have been developed to further ease the task by enabling users to interactively specify query rules and providing visual feedback to facilitate the iterative refinements of the queries (e.g., (s|qu)eries [40], COQUITO [17], and EventFlow [23]).

The temporal query approach is useful when users have prior assumptions about the data so as to specify query rules. However, it is unsuitable to be applied alone for the task of finding similar records—only a few or zero results will be found if many query rules are specified to fully characterize the seed record, or if only a few rules are used, the results may not be similar to the seed record in aspects outside the query rules.

An alternative approach to finding similar records is to start with the seed record, determine useful patterns, and search for records with similar patterns. Mannila and Ronkainen [20] presented a model for measuring the similarity of temporal event sequences. The model computes an edit distance based on three transformation operations at the event level, including insert, delete, and move. This approach can preserve the order of the matched events and performs better when the number of operations is small. Match & Mismatch measure [38] introduces a similarity score that emphasizes the time difference of matched events and the number of mismatches, which supports matching without preserving the order. Besides, a visual interface was also provided to show a ranked list of similar records and allow users to adjust parameters. Recent work [33, 34] describes more advanced similarity measures for specific domains and problems.

Our work, PeerFinder, extends existing similarity metrics for temporal data by allowing users to flexibly specify patterns of interest based on the seed record. It also enables users to find and explore similar records using both record attributes and temporal event information. To encourage engagement and inspire users' trust in the results, it provides different levels of controls and context for users to adjust the similarity criteria.

Temporal Event Sequence Visualizations

Starting with LifeLines [25, 26], early research on temporal event sequence visualization focuses on depicting the medical history of a single patient (e.g., Bade et al. [1], Harrison et al. [13], and Karam [16]). These tools allow users to visually inspect trends and patterns in a record by showing detailed events. LifeLines2 [35] extends this approach to multiple records but does not scale well when displaying a large number of records in a stacked manner.

Techniques have been introduced to handle large sets of records by offering time or category based aggregations. LifeFlow [37] introduces a method to aggregate multiple event sequences by combining them into a tree structure on an alignment point. Likewise, OutFlow [36] combines multiple event sequences based on a network of states. EventFlow [22] extends LifeFlow's concept to interval events and introduces simplification strategies to deal with large data volumes and pattern variety [8]. DecisionFlow [11] provides support for analyzing event sequences with larger numbers of categories.

PeerFinder's timeline was inspired by prior work and adapted to the needs of showing both detailed histories of individual records and activity summaries of groups. Specifically, we used a simplified table-based timeline design to reduce visual complexity. We also summarized the activities of groups to help users identify unique temporal patterns of the seed record.

INFORMING THE DESIGN

The challenges described above highlight the need to provide users with some level of control over the selection of the criteria to be used in the search. To further understand how users would want to specify which criteria to use and how to present results and context, we conducted a series of interviews.

Interviews

Thirteen potential users were interviewed (4 graduate students, 2 graduate advisors, 2 physicians, a start-up CEO, and 4 researchers working in healthcare or marketing). Each interview lasted approximately one hour, including a semi-structured interview and a ranking task to provoke further discussions. We asked participants about what information they might want to gather from similar records, what criteria they would want to use when searching for similar records, and what information would increase their confidence in the value of the results.

Three separate scenarios were used. A student advising scenario asked participants to imagine a setting where an advisor is meeting with a current student to make plans for the year. For the healthcare scenario, we asked participants to think of a doctor working with a patient to make a treatment plan. For marketing, we asked the participants to imagine that they were designing a series of interventions (e.g., calls, ads, or coupons) to retain an important customer, and could look for similar customers to inform their intervention design. Each participant chose one or two scenarios according to their backgrounds. While most participants could easily identify with the student and healthcare situations, the marketing scenario was used only by three participants. They could assume both user roles: the person expecting to receive guidance or the person hoping to provide guidance to others.

We asked the participants to discuss (1) what they would hope to learn from the data of similar records, (2) what criteria they wanted the tool to consider in the similarity search, and (3) what information they would need to determine if the results were similar enough to provide personalized evidence. We told participants to assume that data privacy concerns had been resolved (e.g., only aggregate data would be available if access to details had not been granted).

After a period of open-ended discussion, participants were provided with six printed records, among which one was assigned as the seed and the other five were archived records being searched. Participants were encouraged to think aloud as they tried to rank the archived records by similarity to the seed record, and to describe the criteria they considered in the comparison, the difficulties they faced, and any supports they wanted from a visual interface to complete such task.

Results

We summarize the results and present our findings.

What to Learn from Similar Records

In all three scenarios, participants confirmed the expected uses, in particular, the prediction of outcomes. For example, students wanted to know what jobs similar students got after graduation and their salaries; marketing researchers wanted to know the likelihood of a promotion link being clicked. In addition, participants also asked for estimating the effect of an action on the future of the seed record (i.e., "what if" analysis, a simplified action plan recommendation). For example, a student wanted to test if taking an internship in the last year would increase her likelihood of getting a job at Google, and an advisor wanted to answer students asking if taking an extra class in the next semester might drop the GPA, or if giving up a difficult class would delay graduation. A student stated that *"the information I know about my peers would definitely help me make better decisions."* Both advisers and physicians commented that they often used examples from similar records to tell motivational stories to their advisees or patients, but that it is difficult to remember those similar cases.

Similarity Criteria

Participants responded on average with 11 criteria ($SD = 3.92$), using both record attribute criteria and temporal criteria. Record attribute criteria included categorical values (e.g., gender, nationality, major, research topic, diagnosed disease, or membership tier), and numeric values (e.g., age, weight, height, family income, number of chronic problems, or company size). Temporal criteria included the time between events (e.g., between pick advisor and publication, between two painful episodes, or between sending advertisements and clicking on the promotion link), and the pattern of event occurrences (e.g., a change in the number of publications over time, lose weight and then get sick, or search for a product online and then purchase in the store). Most temporal criteria were stated in general terms (e.g., recently, in the past) with some exceptions in the medical domain, where well-defined, specific temporal patterns were mentioned.

Participants did give examples of criteria which should be ignored (e.g., women are rare in computer science so a female participant wanted that criterion to be ignored). Users may also want particular time periods to be ignored as well (e.g., a school semester when the student was ill).

Some criteria were cited as being more important than others, but in many cases, participants were uncertain about how distinguishable a criterion was for the population or how relevant a criterion was for the knowledge they wanted to gain from the similar records. For example, a student advisor said: *"I am sure about certain criteria but not confident about many others. I want to use the tool to decide if a factor is important in the context of my analysis goal."* All participants mentioned their criteria depend on intended use of the similar records. A physician stated *"gender is important for finding similar patients with breast cancer but does not matter for hypertension or diabetes,"* another said *"they are similar for a purpose."*

A common method used to select criteria was to identify unique characteristics of the seed record. For example, a student may have changed advisors three times in a year, or a patient may be uninsured and cannot afford expensive treatment plans. Participants wanted the system to highlight those unique characteristics.

How to Evaluate the Similar Records

The participants proposed five possible strategies for reviewing the results and determining if they are actually similar enough to the seed record: *Sample inspection*, inspecting individual records, especially the most and least similar ones. *Difference between records*, reviewing differences between the seed record and individual similar records. *Distributions*, reviewing histograms of the values of each criterion among similar records. *Statistical information*, reviewing the number of records in the result, the weight of each criterion, and the statistics for each criterion (e.g., min, max, mean, variance). *Context*, comparing and contrasting the set of similar records to the entire population. A student described his reason for choosing such reviewing strategies: *"I picked the criteria, so I just need to confirm if the results reflect my choices."*

System Design Needs

Based on our initial analysis and participants' suggestions, we propose a list of five design needs.

N1. *Dynamic criteria specification:* To see and adjust which criteria are used—or not, and limit acceptable tolerance.

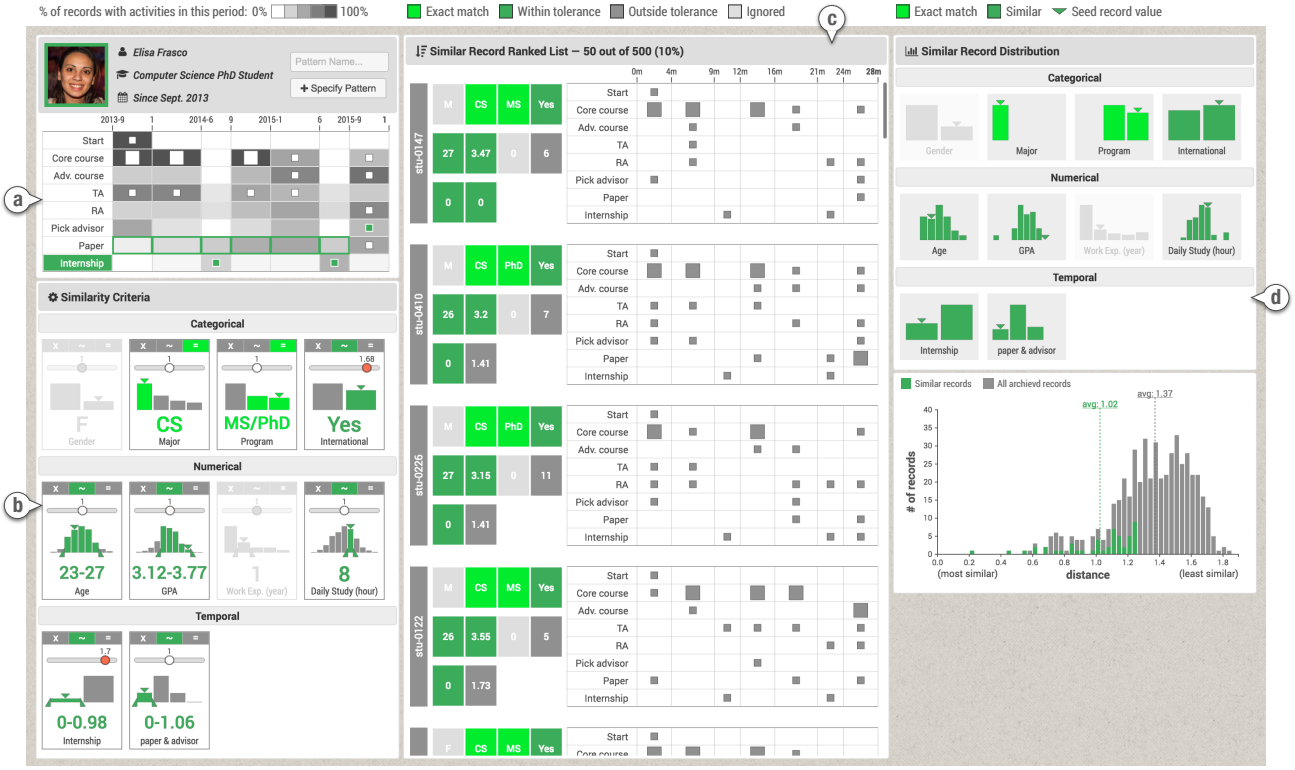

Figure 1. The *Complex* version of PeerFinder, showing all the criteria controls and detailed context. On the left is the seed record attributes and similarity criteria control panel (b). In the center is the ranked list of the similar records with all details (c). On the right is a summary of the results (d). The seed record is a female PhD student in Computer Science. The user chooses to only keep CS student in either MS or PhD program. Tolerance ranges are specified for age and Grade Point Average (GPA). More weight is given to international students. In the timeline (a) two temporal patterns were specified and added to the criteria control panel.

N2. *Criteria prioritization:* To assign weights to different criteria and highlight criteria with higher importance.

N3. *Uniqueness identification:* To receive assistance in identifying unique characteristics of the seed record compared to all archived records.

N4. *Result review:* To review statistics and distributions of the similar records and detailed information of each individual in the results (if access is granted).

N5. *Goal-driven exploration:* To explore how relevant each criterion is to their analysis goal and identify important criteria depending on that goal.

We hope that providing controls over the search process (**N1-2**) and context for the results (**N3-4**) will reduce the challenges of trust and subjectivity in finding similar records. In an attempt to bound the scope of this paper to a similarity search interface, the last need is not addressed because it depends entirely on the end goal of the overall application. For example, if the goal is to estimate what job is most likely to be attained by a student, the application will need to identify which criteria are correlated to the student job placements. Outcome analysis tools such as DecisionFlow [11] and CoCo [19] could be used.

DESCRIPTION OF PEERFINDER

This section describes the user interface and search algorithm of PeerFinder, a visual interface that enables users to find and explore records that are similar to a seed record.

Interface

PeerFinder has four coordinated views (Figure 1): on the left is the seed record with a timeline (a) and attributes (b), which are also used for criteria control. In the center is the ranked list of similar records (c), and on the right is the overview of the similar records (d). The interface can be configured by advanced users using a control panel that adjusts the visibility of all interface components. Here we describe the *Complex* version of PeerFinder configured to provide maximum control and context. Two simpler versions are described later.

Seed Record Timeline

A simplified timeline of the seed record is shown in a table (Figure 1a), where rows represent event categories and columns represent time periods. Events of the same category and in the same period are aggregated and shown as a square, with the size of the square encoding the number of occurrences. For students' records, time periods can be school semesters (e.g., Spring, Summer, and Fall). Advanced users can specify other time period rules based on specific data and applications. User interviews suggested that temporal criteria use only rough time periods so we chose this table-based design which simplifies the timeline while allowing users to explore how the numbers of event occurrences evolve over time.

Users can select or deselect event categories as criteria or specify temporal patterns by selecting cells in the timeline table. To provide a population overview and help users identify unique

temporal patterns of the seed record, the data from all archived records are shown as a heatmap in the table background. In each table cell, the darkness of the background color encodes the percentage of records that had at least one event in this category and this period. Hovering on a cell shows the details.

Similarity Criteria Controls

Similarity criteria are displayed in three groups (Figure 1b): categorical (e.g., gender or major), numerical (e.g., age or GPA), and temporal. Categorical and numerical criteria are automatically defined based on the available record attributes. Temporal criteria are added when a pattern has been specified on the timeline (e.g., having an internship every summer). Each criterion is represented by a rectangular glyph showing its name and context information (i.e., the value of the seed record attribute and distribution of all archived records), along with controls for tolerance range, matching rule, and weight:

Tolerance range: Users can define a tolerance range to treat multiple categorical values or a range of numerical values equally to the value of the seed record, which will increase the similarity of records with those values. For example, users may decide to treat MS and PhD students equally, and set a value range between 3.1 and 3.7 for GPA.

Matching rule: For each criterion, users can define its matching rule by selecting among "Ignore" ($\times$), "Close Match" ($\sim$), or "Exact Match" ($=$). The default rule for all criteria is "Close Match" where records with smaller differences from the seed record will be considered as more similar and ranked higher. The results could have diverse criteria values since the ranking considers the overall difference between records. To narrow results and explicitly include or exclude certain criteria values, users can switch to the "Exact Match" rule and use the tolerance range selector to specify the criteria values that all records in the results must match (e.g., only keeping Computer Science students who have more than one year of work experience). Users can also set the rule to "Ignore" if they do not want to use that criterion.

Weight: Users can give more importance to certain criteria by adjusting their weights using a slider. Increasing the weight magnifies the differences between each archived record and the seed record while small differences in that criterion become smaller. By default, all criteria have a weight of 1, which can be adjusted to any value between 0 (ignored) and 2 (doubled). The color of the round handle becomes red when the weight is high to help users locate the criteria with higher weight.

Similar Record Ranked List

Each time users add or adjust a similar criterion, PeerFinder automatically re-runs the search and shows the refined list of the top similar records (10% by default) in a ranked list (Figure 1c). Each row in the list represents a similar record, consisting of a record ID, values of specified similarity criteria, and a timeline of temporal events. Specifically, the criteria values are displayed in a table with the same layout as the similarity criteria control panel. Values in a green background are within the specified criteria tolerance range while those with a gray background are outside the range. The criteria values and the timelines provide detailed context of each similar record and enable users to spot check the results.

Similar Record Overview

Criteria value distributions of the similar records are shown at the top of Figure 1d to provide an overview of the results. The colors of the bars are consistent with those in the criteria control glyphs, where green bars represent criteria values within the tolerance range, gray bars represent those outside the tolerance range, and the triangles show the value of the seed record. Our initial design overlaid the distributions of similar records on the distributions of all archived records (Figure 1b) using the same axes. However, the number of similar records is usually very small compared to the entire population, making the bars difficult to see clearly.

The bottom of Figure 1d shows the distribution of the distance scores of all archived records (gray bars) and similar records (green bars). The average distance scores are also marked on the chart. This distribution provides an overview about which records are included in the results and how different they are compared to the entire population.

Other Configurations

Simpler configurations may be needed to satisfy the needs of intermittent users or to be embedded in specific applications. Advanced users or application designers can configure the visibility of all interface components to provide different levels of controls and context. In the user study, we used three configurations: *Baseline*, *Simple* and *Complex*. *Baseline* provides no controls over the criteria, emulating a black-box interface (Figure 3). IDs are only shown to indicate that the search has completed. *Simple* allows turning on and off each criterion and shows distributions of the results (Figure 2).

Search Algorithm

As users add or adjust a similarity criterion, PeerFinder automatically executes the similarity search and updates the results on the display. The search execution consists of two steps. First, a filtering step uses "Exact Match" criteria to eliminate records that do not match. Second, the ranking step uses "Close Match" criteria to sort the records and identify the top most similar records. Details are described below.

Filtering

For each criterion marked as "Exact Match" the following process is used: if the tolerance range is not set, only the archived records that have the exact same value (or pattern for temporal criteria) as the seed record will be retained. Otherwise, the records' criteria values need to be within the tolerance ranges to be retained. The tolerance range is represented by a set of values for categorical criteria and by a pair of upper and lower bounds for numerical or temporal criteria.

Ranking

Next, "Close Match" criteria are used to rank the archived records by their similarities to the seed record. A comprehensive distance score is computed for each archived record based on the empirical assumption that the archived records tend to be more different from the seed record if they have (1) nonidentical values for categorical attributes, (2) larger

Figure 2. The *Simple* version of PeerFinder provides basic criteria controls (turning on and off each criterion in timeline (a) and record attributes (b)), and simple context (record IDs (c) and overall distribution of the results (d)).

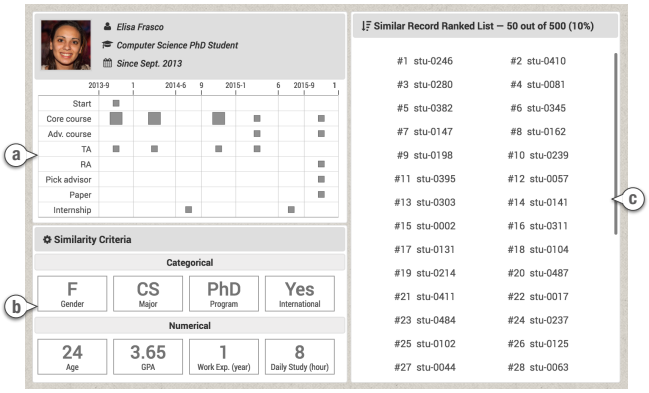

Figure 3. The *Baseline* version of PeerFinder provides no controls over the criteria (users can only see the seed record's temporal events (a) and attribute values (b)) and no context (only a list of IDs as results (c)).

discrepancies in numerical attribute values, and (3) larger deviations in activity patterns. The algorithm first assesses the difference in each criterion and then summarizes them into a single distance score.

Categorical criteria: For each categorical criterion $cc \in C$, we define the difference between an archived record r and the seed record s as:

$$\Delta_C(cc,r,s) = \begin{cases} 0 & v(cc,r) \in t(cc,s) \\ \alpha & v(cc,r) \notin t(cc,s) \end{cases}$$

where $v(cc,r)$ returns cc's value of a given record and $t(cc,s)$ returns the set of values in the tolerance range of cc or $\{v(cc,s)\}$ if the tolerance is not specified. We let $\alpha = 0.5$ to keep a balance between categorical and numerical criteria, but the optimal value depends on the data and analysis.

Numerical criteria: For each numerical criterion $nc \in N$, the difference between an archived record r and the seed record s is formulated as:

$$\Delta_N(nc,r,s) = \begin{cases} |v(nc,r) - t_u(nc,s)| & v(nc,r) > t_u(nc,s) \\ |v(nc,r) - t_l(nc,s)| & v(nc,r) < t_l(nc,s) \\ 0 & otherwise \end{cases}$$

where $v(nc,r)$ returns the nc's value of a given record and $t_u(nc,s)$ and $t_l(nc,s)$ returns the upper and lower bound of the tolerance range of nc, respectively. When the tolerance of nc is not specified, we have $t_u(nc,s) = t_l(nc,s) = v(nc,s)$. Before the computation, values of each numerical criterion are standardized by scaling to range $[0,1]$.

Temporal criteria: For each temporal criterion $tc \in T$, we compute a value $v(tc,r)$ for each archived record r, reflecting its difference from the seed record s in activity patterns:

$$v(tc,r) = \|\mathbf{p}(tc,r) - \mathbf{p}(tc,s)\|$$

where $\mathbf{p}(tc,r)$ returns a two-dimensional vector (x=time, y=event category) representing the activity pattern of r. Since $v(tc,r)$ returns a numerical value, we reuse the difference function for numerical criteria and let $\Delta_T = \Delta_N$.

Finally, we summarize a comprehensive distance score for each pair of archived record r and the seed record s based on weighted Euclidean distance [5]:

$$distance(r,s) = \sqrt{\sum_{cc \in C} w_{cc}\Delta_C^2(cc,r,s) + \sum_{nc \in N} w_{nc}\Delta_N^2(nc,r,s) + \sum_{tc \in T} w_{tc}\Delta_T^2(tc,r,s)}$$

where $w \in [0,+\infty)$ is the weight assigned to a criterion.

EVALUATION

Searching for similar people to guide life choices is still a new research area and many user studies will be needed to evaluate PeerFinder as it gets embedded in applications that use the ranked list of records to provide guidance. Similarity remains subjective (see early section on challenges) and no ground truth dataset exists, so we chose to focus this first lab study and expert interviews on gaining insights into factors that engage users and promote more trust in the results.

User Study

A within-subject user study compared three versions of PeerFinder (Figure 1-3) using different levels of complexity (*Baseline*, *Simple*, and *Complex*), as a combination of

control and context. The goal was to understand how the levels of controls and context affect users' engagement and their confidence in the ability of the results to be useful. We were interested to see if users would defy conventional guidelines and prefer a more complex interface that demanded more time to use. We also wanted to get feedback to improve the interface.

Participants and Apparatus

We recruited 18 university students by email (10 males and 8 females, aged 20–30, $M = 24.67$, $SD = 3.12$). Ten of the participants had technical backgrounds and were experienced in software development, statistics, and data analysis (from the Information School or the Department of Computer Science). The other 8 had limited technical backgrounds but used computers in their study, e.g., web design or print design in the Art Department. None of the participants had prior experience with PeerFinder. Each participant received 10 dollars. A desktop computer was used, with a 24-inch display of resolution 1920×1200 pixels, a mouse, and a keyboard.

Datasets for Evaluation

We constructed three synthetic datasets with realistic but simplified features to test the three PeerFinder designs. Each dataset contained 500 records of archived university students. The records had three categorical attributes: gender (male or female), major (Computer Science, HCI, Math, Art), program (BS, MS, PhD), and international student (yes or no); four numerical attributes: age (when they started school), GPA, previous work experience (year), and average study time per day (hour). Eight categories of temporal events were included, including "start school", "core course", "advanced course", "paper", "TA (Teaching Assistant)", "RA (Research Assistant)", "pick advisor", and "internship". On average each archived record contained 35 events over 5 years. We generated record attributes with normal and binomial distributions. For temporal events, we reviewed real data and included similar patterns with random variations. The names of events and attributes are generic so that all students can conduct the tasks.

We originally wanted to customize the seed record to match the participant's own data and ask them to search for students like themselves, but we decided against this strategy to normalize the task and avoid privacy and confidentiality issues. Instead, we handpicked a record (named Elisa Frasco and illustrated in Figure 3a) that would serve as the seed record: a female international student, majoring in Computer Science and currently in the third year of her PhD study. She is 24 years old and has one year of work experience before starting graduate school. On average, she spends 8 hours on study each day and maintains a relatively high GPA of 3.65. The timeline showed no papers in the first two years, internships in the last two summers, work as a TA all along except for an RA position in the last semester, after picking an advisor.

Hypotheses

Our hypotheses were:

H1. Users' confidence will be the highest with *Complex* and the lowest with *Baseline* in that the result set is similar enough to the seed record to provide evidence to guide making academic plans.

H2. Users will prefer *Complex* and *Simple* over *Baseline*.

H3. Users will spend the longest time using *Complex* and the shortest time using *Baseline*.

H4. Users will make more result refinements using *Complex* than *Simple*.

H5. Users will give higher ratings for ease of learning and ease of use for *Simple* and *Baseline* than *Complex*.

We hypothesized that users would spend longer time (**H3**) and make more result refinements (**H4**) in *Complex*, thus increasing their trust in the results (**H1**) and preference for the interface (**H2**). **H3** and **H4** were also an attempt to capture user engagement. Ease of learning and ease of use (**H5**) was included to replicate prior research showing that added complexity reduces ease of learning and ease of use and contrast the results with preferences [4, 27].

Procedure

After the initial email recruitment, we sent more detailed directions: *"You will be asked to (1) learn about a (hypothetical) close and important friend of yours who needs advice to improve her academic plan, such as when to take advanced classes, whether to intern during the summer, or when to try to publish papers, and (2) use three different user interfaces to search for students similar to that friend. Data from those similar students will be used as evidence to provide guidance for your friend. You will not be asked to provide or review the guidance itself, only to select a set of similar students."* The record of the hypothetical friend was also provided and participants were encouraged to get familiar with it.

In the lab, each session lasted about 60 minutes. In a brief general training (about 5 minutes), the experimenter made sure that participants were familiar with the task and the hypothetical friend, and answered questions. Next, one of the three versions of PeerFinder (*Baseline*, *Simple*, or *Complex*) was used and the participants were shown a short tutorial (max 5 minutes) covering its interface and operations. The experimenter answered questions and encouraged them to think aloud. The participants were reminded to care about their friend and there was no time limit for the task. When satisfied with the results, the participants needed to click a "finish" button and complete a user satisfaction questionnaire using a 7-point Likert scale:

Q1. How easy was it to learn the interface (1=very difficult, 7=very easy)?

Q2. How easy was it to use the interface (1=very difficult, 7=very easy)?

Q3. How confident were you that the records in the results were similar enough to your friend in order to provide evidence to guide her making academic plans (1=not confident at all, 7=very confident)?

The training, task, and questionnaire were repeated with the other two versions using different datasets so that the results varied. Interface order and datasets were counterbalanced. Participants were allowed to see and adjust the subjective rating they gave for previous versions. Task completion times and numbers of result refinements (i.e., the number of adjustments

in criteria controls) were recorded automatically. After using all three versions, participants were asked to rank them based on preference and debriefed to collect feedback.

Results

Repeated Measures ANOVA was applied to compare the completion times (log-transformed) and numbers of result refinements, and paired t-test was used for post-hoc comparisons. For questionnaire ratings, we used Friedman test and pairwise Wilcoxon test. All tests used a significance level of 0.01.

Questionnaire: As reported in Figure 4, *Baseline* was rated the easiest to learn in Q1 followed by *Simple* and *Complex*. Significant differences were found among the ratings ($\chi^2(2) = 28.00$, $p < 0.001$). Follow-up comparisons indicated that all pairwise differences were significant. The average ratings in Q2 showed the same order of the three versions for the ease of use and the differences were significant ($\chi^2(2) = 32.11$, $p < 0.001$). Pairwise comparisons found significant differences between *Complex* and *Baseline* and between *Complex* and *Simple*. These results supported **H5**.

In Q3, *Complex* had the highest confidence rating ($M = 5.89$) followed by *Simple* ($M = 4.11$) and *Baseline* ($M = 1.67$). Significant differences among the ratings were detected ($\chi^2(2) = 32.14$, $p < 0.001$) and all pairwise differences were significant, which supported **H1**.

Completion time: On average, the participants spent 0.65 minutes ($SD = 0.34$) on *Baseline*, 6.16 minutes ($SD = 2.12$) on *Simple*, and 16.03 minutes ($SD = 6.17$) on *Complex* (Figure 5a). Significant differences were found in the log-transformed completion times across these three versions ($F_{2,34} = 248.42$, $p < 0.001$). Post-hoc comparisons showed all pairwise differences were significant, supporting **H3.**

Result refinement: On average, the participants made 16.39 refinements ($SD = 14.08$) using *Simple* and 34.17 refinements ($SD = 16.90$) using *Complex* (Figure 5b), which was a significant increase of 108%, supporting **H4**.

Preference and Feedback

16 out of 18 participants chose *Complex* as their preferred interface. Two picked *Simple*, and *Baseline* was always the least favorite, which confirmed **H2**.

Ease of learning and use: Although *Baseline* was rated as the easiest version to learn and to use, many participants commented on their disappointment, e.g., *"I can do nothing."* 9 participants commented that *Simple* offered a good balance of simplicity and capability, e.g., *"I like the binary controls and clear presentation of the results. I felt more focused."* Another who preferred *Simple* explained that *"the controls satisfy my needs and the Simple interface is easier to explain to the friend I am helping."* As for *Complex*, 11 participants gave a neutral rating in Q1 or Q2 and 4 of them suggested that *"it requires training and practice to become familiar with this interface."* In contrast, one participant who thought *Complex* was easy to learn explained: *"The graphs are the same everywhere. After understanding one, I understand others."*

Confidence: All participants expressed lacking trust in the results generated by *Baseline* and the most common feedback

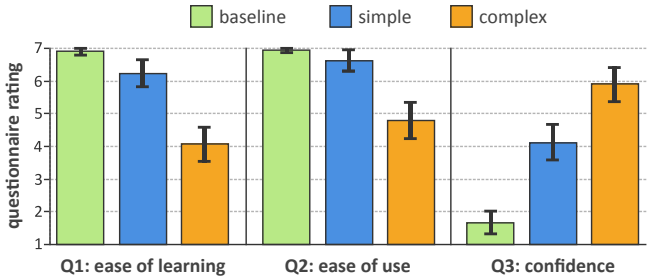

Figure 4. Average ratings for each version of PeerFinder in the user satisfaction questionnaire (error bars show 95% confidence intervals). 1=very difficult and 7=very easy in Q1 and Q2; 1=not confident at all and 7=very confident in Q3.

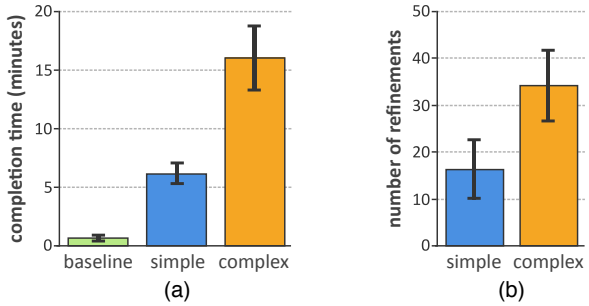

Figure 5. (a) Average completion times and (b) average numbers of result refinements using different versions of PeerFinder (error bars show 95% confidence intervals).

was *"the results look random."* One participant emphasized: *"I want to know how the algorithm gives these IDs."* Another added: *"The results may be good but without any details, I am skeptical about it."* 15 participants gave higher confidence ratings for *Complex* than *Simple*. The most common reason given was control: *"The advanced controls enable me to get more precise results,"* or *"when using the Simple version I can see some flaws in the results but cannot fix them,"* or *"since I have the functionalities to do more, I am more motivated to pay attention and try different settings."* Participants also appreciated seeing the similar records provided by *Complex*: *"it helps me verify the results and correct small mistakes"* or *"seeing concrete students provide inspirations for tuning the controls and specifying temporal patterns."*

Participants expressed concerns about the complexity of the *Complex* version. One described: *"There are many options and data you need to keep track of. It was like piloting a plane."* Another said that *"the [similar] student information were distracting when I was not using it."* One participant who preferred *Simple* commented: *"My trust diminished every time I got lost. I worried about missing anything."*

Search strategies: Most users only briefly reviewed the display of *Baseline*. On the other hand, we observed users repeatedly turning on and off criteria in *Simple* and inspecting the result distributions to see the effects. When using *Complex*, users commonly carefully reviewed the criteria one by one and tried different settings. They kept an eye on the result distributions, and reviewed the details of a few similar records to verify their settings. At the end, many scanned the entire list of similar records looking for problems.

Suggestions: Suggestions included starting simple and allowing users to add controls and details as needed, enabling users to choose colors and interface layout, marking important records. Automatic aids were also requested: recommending criteria settings to save users' effort and detecting outliers in the results for users to review. Usability suggestions included making buttons more noticeable, flipping the layout entirely to show the seed at the top and results below, and merging both distributions (for the population and the seed record) into one.

Expert Review

We conducted one-on-one 45-minute interviews with domain experts whose professional activities involved providing guidance to others: three student advisors (E1-3) and a physician (E4), each having at least 10 years of experiences. We demonstrated the three interfaces using the same datasets as in the user study and asked the experts to explore on their own. We answered questions and recorded comments and preference.

All four experts expressed great enthusiasm for PeerFinder: *"it helps me provide advice based on data and avoid false assumptions"* (E1), *"it provides a new method to make use of the collected student data"* (E2), and *"it provides a faster and data-driven way to quickly profile a student and start the conversation"* (E3). E1 and E3 preferred *Complex*. E1 suggested allowing users to re-arrange and turn on and off each view since different views are used at different stages during the exploration. E3 wanted to look at the "future" activities and outcomes of those similar students. E2 picked *Simple* stating that *"the interface is simpler and helps me communicate the results with other advisors or students."* E4 stated that all three versions have values depending on usage: *"The Complex version could be very useful for patients working on their own for health maintenance. For regular doctor visits the Simple version may work better since the time is very limited."* He mentioned that diseases usually have their own schema which can be used as presets for the criteria settings. The importance of privacy protection was repeated, and ethical issues were mentioned, for example: *"Some students may be demoralized by the worst cases in the results (similar records)."*

DISCUSSION

All hypotheses were confirmed with size effects larger than we expected. For example, we expected more participants would prefer the *Simple* version, but despite the increased complexity, the *Complex* version was preferred by the majority of users. Engagement, as measured by time spent on task and number of interactions, was also higher when using *Complex*. More importantly, confidence was higher when using *Complex*. These findings suggest that users should be provided with controls over the search process when making life choices. The lab study was tied to a particular scenario (student advising) but our research emphasizes that different situations of use require different criteria to be used, reinforcing the importance of customization for the end user and for the application developer. While some of the challenges remain (e.g., no ground truth), we believe that there is value in clarifying those challenges, and that the PeerFinder prototype and evaluation approaches (e.g., measuring trust) will inspire others to develop better solutions to these challenges.

Reviewing ethical issues is important. Bad data that reinforces existing biases may be taken as truth and data that challenges them dismissed. Will a poorly performing student be discouraged when seeing the outcome of similar students? Or will a high achieving "anomalous" student in a poor achievement cohort set her horizon too low? Those issues argue strongly for collaborative use where the advisee is working alongside an experienced advisor who can interpret the results or judge data quality. However, advisors' guidance will not solve all problems since they are also vulnerable to biases [3]. PeerFinder mitigates this issue by giving transparent data access to both advisors and advisees and involving them in the decision-making process.

Our user study had several limitations. We tested only three configurations, omitting alternate versions, for example, one that included no control but provided rich context. Testing all nine configurations will help tease out the separate roles of increased control and increased context. We chose a within-subject design so the *Baseline* may have seemed more disappointing to participants who saw other versions first. Between-group studies may affect the differences in confidence, but then preference cannot be collected. In our study, we made sure that there were records similar to the seed record, but even with "big data" there may be cases where few similar records exist. In those cases, we need to verify that user confidence remains low. We did not evaluate the accuracy of the search algorithm because ground truth is not available. We hope that increased interest in this topic will lead to the development of benchmark datasets. In the meantime, the search algorithm can be improved to handle multi-attribute data, treat ordinal attributes separately, and incorporate refined similarity measures for temporal patterns. Lastly, our study focused on a student advising scenario. Medical scenarios are likely to be more complex unless the tool is customized to a carefully chosen medical specialty and diagnosis. In the future, we also hope to incorporate outcome data and help users identify the similarity criteria that are most correlated to the outcomes of interest.

CONCLUSION

Searching for similar people to guide life choices is a new area of research. After characterizing the challenges facing designers and evaluators of such systems, we described PeerFinder. This prototype interface enables users to interactively find and review records based on similarity to a seed record using both record attributes and temporal event information. While there is still much to do to improve the interface, our study suggests that *users are more engaged and more confident about the results when provided with more control and more context, even at the cost of added complexity*. Further studies of PeerFinder embedded in applications providing guidance and privacy protection mechanisms will advance our understanding of the role of similarity search in guiding life choices.

ACKNOWLEDGMENTS

We thank all the reviewers and study participants for their valuable feedback, in particular, Seth Powsner, Jeff Belden, Evan Golub, and Jennifer Story. We appreciate the partial support for this research from Adobe.

REFERENCES

1. Ragnar Bade, Stefan Schlechtweg, and Silvia Miksch. 2004. Connecting time-oriented data and information to a coherent interactive visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 105–112.
2. Eftychia Baikousi, Georgios Rogkakos, and Panos Vassiliadis. 2011. Similarity measures for multidimensional data. In *IEEE International Conference on Data Engineering*. 171–182.
3. Brian H Bornstein and A Christine Emler. 2001. Rationality in medical decision making: a review of the literature on doctors' decision-making biases. *Journal of Evaluation in Clinical Practice* 7, 2 (2001), 97–107.
4. Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 13, 3 (1989), 319–340.
5. Jan De Leeuw and Sandra Pruzansky. 1978. A new computational method to fit the weighted Euclidean distance model. *Psychometrika* 43, 4 (1978), 479–490.
6. Lieven Decock and Igor Douven. 2011. Similarity after goodman. *Review of Philosophy and Psychology* 2, 1 (2011), 61–75.
7. Fan Du, Catherine Plaisant, Neil Spring, and Ben Shneiderman. 2016a. EventAction: Visual analytics for temporal event sequence recommendation. In *Proceedings of the IEEE Visual Analytics Science and Technology*. 61–70.
8. Fan Du, Ben Shneiderman, Catherine Plaisant, Sana Malik, and Adam Perer. 2016b. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. *IEEE Transactions on Visualization and Computer Graphics* PP, 99 (2016), 1–14.
9. Ellen A Ensher and Susan E Murphy. 1997. Effects of race, gender, perceived similarity, and contact on mentor relationships. *Journal of Vocational Behavior* 50, 3 (1997), 460–481.
10. Lyndsey Franklin, Catherine Plaisant, Kazi Minhazur Rahman, and Ben Shneiderman. 2016. TreatmentExplorer: An interactive decision aid for medical risk communication and treatment exploration. *Interacting with Computers* 28, 3 (2016), 238–252.
11. David Gotz and Harry Stavropoulos. 2014. DecisionFlow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1783–1792.
12. Fabio Grandi. 2010. T-SPARQL: A TSQL2-like temporal query language for RDF. In *In International Workshop on Querying Graph Structured Data*. 21–30.
13. Beverly L Harrison, Russell Owen, and Ronald M Baecker. 1994. Timelines: An interactive system for the collection and visualization of temporal data. In *Graphics Interface*. 141–141.
14. Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 241–250.
15. Sandy H Huang, Paea LePendu, Srinivasan V Iyer, Ming Tai-Seale, David Carrell, and Nigam H Shah. 2014. Toward personalizing treatment for depression: Predicting diagnosis and severity. *Journal of the American Medical Informatics Association* 21, 6 (2014), 1069–1075.
16. Gerald M Karam. 1994. Visualization using timelines. In *Proceedings of the ACM International Symposium on Software Testing and Analysis*. 125–137.
17. Josua Krause, Adam Perer, and Harry Stavropoulos. 2016. Supporting iterative cohort construction with visual temporal queries. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 91–100.
18. Christopher A Longhurst, Robert A Harrington, and Nigam H Shah. 2014. A 'Green Button' for using aggregate patient data at the point of care. *Health Affairs* 33, 7 (2014), 1229–1235.
19. Sana Malik, Ben Shneiderman, Fan Du, Catherine Plaisant, and Margret Bjarnadottir. 2016. High-volume hypothesis testing: Systematic exploration of event sequence comparisons. *ACM Transactions on Interactive Intelligent Systems* 6, 1 (2016), 9:1–9:23.
20. Heikki Mannila and Pirjo Ronkainen. 1997. Similarity of event sequences. *TIME* 97 (1997), 136–140.
21. Alaina Michaelson and Noshir S Contractor. 1992. Structural position and perceived similarity. *Social Psychology Quarterly* (1992), 300–310.
22. Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. 2013a. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2227–2236.
23. Megan Monroe, Rongjian Lan, Juan Morales del Olmo, Ben Shneiderman, Catherine Plaisant, and Jeff Millstein. 2013b. The challenges of specifying intervals and absences in temporal queries: A graphical language approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2349–2358.
24. David O'hare. 1976. Individual differences in perceived similarity and preference for visual art: A multidimensional scaling analysis. *Perception & Psychophysics* 20, 6 (1976), 445–452.
25. Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. LifeLines: Visualizing personal histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 221–227.

26. Catherine Plaisant, Rich Mushlin, Aaron Snyder, Jia Li, Daniel Heller, and Ben Shneiderman. 1998. LifeLines: Using visualization to enhance navigation and analysis of patient records. In *Proceedings of the AMIA Symposium*. 76–80.
27. Ben Shneiderman, Catherine Plaisant, Maxine Cohen, Steven Jacobs, Niklas Elmqvist, and Nicholas Diakopoulos. 2017. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson.
28. Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI Extended Abstracts on Human Factors in Computing Systems*. 830–831.
29. Richard Snodgrass. 1987. The temporal query language TQuel. *ACM Transactions on Database Systems* 12, 2 (1987), 247–298.
30. Ellen Spertus, Mehran Sahami, and Orkut Buyukkokten. 2005. Evaluating similarity measures: A large-scale study in the orkut social network. In *Proceedings of the ACM International Conference on Knowledge Discovery in Data Mining*. 678–684.
31. Ashish Sureka and Pranav Prabhakar Mirajkar. 2008. An empirical study on the effect of different similarity measures on user-based collaborative filtering algorithms. In *Pacific Rim International Conference on Artificial Intelligence*. 1065–1070.
32. Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 415–422.
33. Katerina Vrotsou and Camilla Forsell. 2011. A qualitative study of similarity measures in event-based data. In *Symposium on Human Interface*. 170–179.
34. Katerina Vrotsou, Anders Ynnerman, and Matthew Cooper. 2013. Are we what we do? Exploring group behaviour through user-defined event-sequence similarity. *Information Visualization* (2013), 232–247.
35. Taowei David Wang, Catherine Plaisant, Alexander J Quinn, Roman Stanchak, Shawn Murphy, and Ben Shneiderman. 2008. Aligning temporal data by sentinel events: Discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 457–466.
36. Krist Wongsuphasawat and David Gotz. 2012. Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2659–2668.
37. Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. 2011. LifeFlow: Visualizing an overview of event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1747–1756.
38. Krist Wongsuphasawat and Ben Shneiderman. 2009. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *IEEE Symposium on Visual Analytics Science and Technology*. 27–34.
39. World Health Organization. 1992. The tenth revision of the international classification of diseases and related health problems (ICD-10). (1992).
40. Emanuel Zgraggen, Steven M. Drucker, Danyel Fisher, and Robert DeLine. 2015. (s|qu)eries: Visual regular expressions for querying and exploring event sequences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2683–2692.