Answering *What If*, *Should I* and Other Expectation Exploration Queries Using Causal Inference over Longitudinal Data

Emre Kıcıman
Microsoft Research
emrek@microsoft.com

Jorgen Thelin
Microsoft Research
jthelin@microsoft.com

ABSTRACT

Many people use web search engines for expectation exploration: exploring what might happen if they take some action, or how they should expect some situation to evolve. While search engines have databases to provide structured answers to many questions, there is no database about the outcomes of actions or the evolution of situations. The information we need to answer such questions, however, is already being recorded. On social media, for example, hundreds of millions of people are publicly reporting about the actions they take and the situations they are in, and an increasing range of events and activities experienced in their lives over time. Here, we show how causal inference methods can be applied to such data to generate answers for expectation exploration queries. This paper describes a system implementation for running ad-hoc online causal inference analyses. The analysis results can be used to generate pros/cons lists for decision support, timeline representations to show how situations evolve, and be embedded in many other decision support and planning applications. We discuss potential methods for evaluating the fundamental quality of inference results and judge the short-term and long-term usefulness of information for users.

1 INTRODUCTION

Everyone, at some point in their lives, finds themselves in an unfamiliar situation, considering what they should do, and trying to understand what to expect of the future. We see such *expectation exploration* occurring in web searches, with people exploring possible consequences of their choices and the outcomes of situations. These explorations cover both consequential topics, such as life-changing education and career choices (e.g., "Should I join the military?") or major financial and personal decisions (e.g., "Should I move to California?"); as well as more quotidian topics, such as the consequences of purchase decisions, athletic training regimens and dating rituals.

The answers to these questions are not readily available in a knowledge base or Wikipedia. But, the information necessary to answer these questions is already being recorded on social media, where hundreds of millions of individuals regularly and publicly report their personal experiences, including the situations they are in, the actions they take, and the experiences they have afterwards. For example, people talk about work or relations [12, 15] health and dietary practices [1, 38], and even log information about their illnesses and coping strategies [8, 13]. People report and share this information for many reasons: keeping in touch with friends, gaining social capital, diary-keeping, or even helping others. And with increasing use of personal sensors and devices, from exercise trackers to health monitors, such data streams are becoming more regular, more detailed and more reliable [4, 26, 32]. These longitudinal data streams, in aggregate, capture a rich set of relationships between the situations in which people find themselves, the actions they choose to take, and the outcomes they experience.

We describe *Outcomes Engine*, a system for analyzing such large-scale longitudinal data to characterize how situations evolve over time, and to capture the consequences of people's actions. Given a query representing some target action T, Outcomes Engine identifies individuals who have reported doing T, and compares their subsequent experiences to peers who did *not* report doing T. This comparison results in an *expectation map* detailing "what changes to expect" over time due to T. A key aspect of Outcomes Engine is its use of causal inference methods to compare the two sets of individuals so as to isolate the specific consequences of T from subsequent experiences that are correlated with, but not due to T.

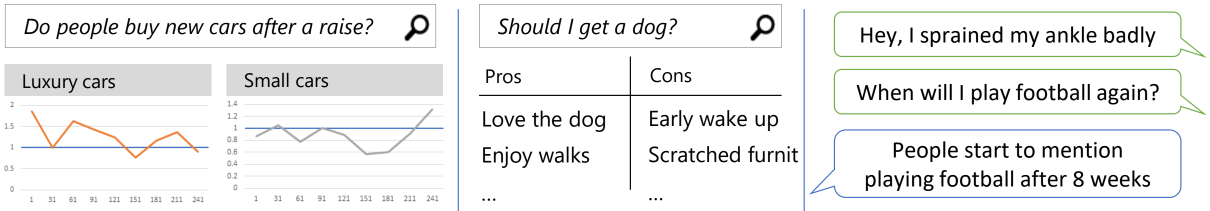
The expectation maps generated by Outcomes Engine are an important building block for a wide variety of data-driven search and decision-support applications—from automatically generating decision aids, such as pros and cons lists, to helping individuals ground their experiences in how a situation is likely to evolve over time (cf Figure 1). In addition, expectation maps may be useful for policy makers' and scientists' explorations across a variety of domains. In this paper, we discuss our approach and prototype system, several application scenarios, as well as evaluation challenges and strategies.

2 BACKGROUND AND RELATED WORK

2.1 Expectation Exploration Tasks

Exploring expectations on the Internet plays an important role in people's planning, decision-making, and forecasting for both everyday and extraordinary scenarios. These explorations encompass a broad variety of tasks, including explorations of hypothetical, ongoing or past problems, or seeking informational support, emotional satisfaction, or preparation for a future event. Taxonomies of web search activities classify these as an information gathering task, which encompass 35% to 80% of people's web searches [17, 36]. Expectation exploration may also be considered as a temporal web query, where time is relative to individually experienced timelines, rather than, for example, a calendar date or global event [5, 7].

DESIRES 2018, August 2018, Bertinoro, Italy



(a) Timeline answer (b) Pros / cons list (c) Conversational agent

Figure 1: Interface mockups: Expectation exploration tasks may be satisfied with a variety of information presentations.

Decision-making processes in particular depend critically on such information gathering—especially in unfamiliar situations—where the web augments more conventional information sources such as professional and friends' advice, training, etc. In 2004, Rose and Levinson measured advice-related searches as 2-5% of web search tasks [34]. Bailey et al. find that decision-related tasks—including comparing (9%) and planning (2%)—constitute a significant portion of overall web tasks. Lagan et al. find that even in pregnancy—a scenario with dedicated information infrastructures, related health professionals and care programs—over 80% of women used web search to help make decisions [23].

Though there are online resources and crowdsourced methods for exploring some scenarios, extracting outcomes from aggregated personal data streams has many distinct advantages [22, 25] First, results are grounded in the real experiences of users who have taken an action, potentially leading to more reliable results than simply reading advice from web pages. Second, a question may be too rare for someone to have devoted writing advice about it, but there is still plenty of social data to answer via data mining. For example, someone may ask whether to move to one city vs. another. Web pages may exist to answer such a question for some city pairs, but not for all. In contrast, we need only look at social postings from people who have moved to one city vs. the other and compare their postings to see the relative benefits of each. Third, an answer may be contextually dependent on the asker. The methods presented in this paper can potentially be extended to provide answers personalized to the asker.

Once an expectation map has been extracted for a scenario, it can be embedded in many distinct presentations and applications to provide the asker with a high-level overview of the implications of a choice or evolution of a situation. For example, a timeline view may show how outcomes evolve over time (Figure 1-a). Another application, specifically for decision support, is an automatically generated pros/cons list [20] (Figure 1-b). The resultant data could also be used within a conversational agent (Figure 1-c).

While our work may benefit individuals who wish to understand their situations and the possible implications of their actions, there is also an opportunity to use this kind of analysis to better understand behavioral phenomena of societal importance, third-party interventions and other policy questions. As well, while we focus on analysis of timelines of individual people's experiences, such analyses may also be applied to event timelines of other kinds [2], subject to sufficient data availability and assumptions.

2.2 Causal Inference

In this paper, we propose to analyze individual-level longitudinal datasets with causal inference methods to directly identify what can be expected following some action or individual experience. We believe this can provide a semi-structured representation of expectations that can be used in a wide variety of ways to aid individual's planning, decision-making, and forecasting.

Because we are interested in using our analysis results to aid decision-making—essentially an intervention—our goal is fundamentally one of causal inference. While we do not believe we can achieve the ideal identification of causal relationships, we can use methods borrowed from the causal inference literature to reduce the bias of naïve correlational analyses. Here, we give a brief introduction to potential outcomes, one framework for causal reasoning [35].

In the potential outcomes framework, whether some experience "causes" an outcome is computed by comparing two potential outcomes: one outcome $Y_i(T = 1)$ after a person i has a target experience T [1], and another outcome $Y_i(T = 0)$ when the same person in an identical context does not have the experience. The causal effect of T is then $Y_i(T = 1) - Y_i(T = 0)$. Of course, it is impossible to observe both $Y_i(T = 1)$ and $Y_i(T = 0)$ for the same individual i. Once we observe i having the experience or not, we cannot observe the other, counterfactual outcome.

Thus, the problem of causal inference is, in a sense, a problem of missing data, and causal inference techniques attempt to address this challenge by estimating the missing counterfactual outcome for an individual based on the outcomes of other, similar individuals. A common method for estimating missing counterfactual outcomes is to find pairs (generalizing to groups) of individuals in the observational data whose covariates are statistically very similar to one another, but where one has received a treatment and the other has not. Each individual's matched partner then provides the basis for estimating a counterfactual outcome for that individual. We describe our specific method in Section 3.2.

Prior research demonstrates the feasibility of this approach in high-dimensional settings (such as our proposed analysis of social media and sensor data). For example, Eckles and Bakshy reduced bias in an observational study by 97% compared to a naive analysis, as measured against a gold-standard randomized field experiment, by conditioning on high-dimensional covariate data [11].

[1] In medical and social sciences literature, the target experience is often called the *treatment*, and is compared to a *control* or *placebo* experience. Following this convention, we will use the terms *treated group* and *control group* in this paper.

2.3 Social and Online Data Analyses

Longitudinal studies of online data, including social media data and search query logs, have proven effective in helping understand the behaviors of people in various situations. These studies have been targeted to explore and understand how situations evolve over time, identify predictive factors involved in positive and negative outcomes, and help identify at-risk individuals. For example, using search query logs, Paul et al. [31] characterize the information seeking behavior during various phases of prostate cancer. Fourney et al. [14] align search query logs with the natural clock of gestational physiology of pregnant women to characterize their changing information needs. Althoff et al. study 5 years of fitness tracking data to better understand social influence on physical activity [3].

By mining social media, De Choudhury et al. [9] find behavioral cues useful to predict the risk of depression before onset. Similarly, by leveraging these naturalistic data, prior work examined how dietary habits vary across locations [1]; the links between diseases, drugs, and side-effects [27, 30]; links between actions and outcomes [20]; shifts in suicidal ideation [10]; and how alcohol usage in early college affects long-term outcomes [19]. Olteanu et al. demonstrate propensity scored analysis of social media timelines to understand outcomes across a broad set of domains [28]

3 CAUSAL INFERENCE-BASED MAPPING OF EXPECTATIONS

We present our approach to mapping expectations from social datasets. First, we present our basic design data requirements and assumptions, followed by our definition of a query and result representation. Then, we present our method for extracting expectations by applying causal inference over social data sets. We use causal inference for this purpose to remove merely correlated outcomes and focus on outcomes directly caused by an action or treatment. This is particularly important for applications that will be performing interventions (including decision-support applications for individuals and policy makers)

3.1 Basic Design

Data. The fundamental requirements our approach places on data is that they provide a *longitudinal* view of the actions and experiences of *individuals*. Thus, at a minimum, input data observations must include a user id and datetime in addition to observational content (e.g., message text).

We focus our prototype implementation on social media data for several reasons. First, social media data provides high-dimensional and cross-domain coverage, allowing a broad variety of query topics and increasing the likelihood of observing statistical confounders that would otherwise bias an analysis. Secondly, the textual nature of social media data is relatively interpretable. Third, social media data is available at large-scale and captures individual activities over long periods of time. Beyond social media, our framework may be applied to other kinds of data sources. E.g. personal sensors and other services may be supported, though treatment identification and result interpretation in our framework would require adaptation. Search query histories are particularly promising, as past analyses have demonstrated the potential for longitudinal analysis of search histories [3, 14, 29, 33].

Input Query. Asking a question to explore expectations following an action or event requires identification of individuals who have performed a particular action or experienced a particular situation. The pattern for identifying messages about this experience, then, is the fundamental input query we expect. Our prototype relies on explicit textual mentions of actions and situations and, in our design, we allow a boolean phrase query, with some wildcard support, for identifying a targeted experiential phrases.

Expectation Maps. Expectation maps represent the time-varying effects of an experience or treatment over a population of people. An expectation map for a treatment can be represented as a 2D matrix, where each row is an outcome word or topic, each column represents an epoch of time (e.g., hours or days since treatment). Each cell represents the effect of the treatment on a specific outcome during a specific epoch. The effect itself includes measurements of effect size and statistical significance, and can be extended to include details of heterogeneous effects.

3.2 Causal Inference Method

In our system, we use a stratified propensity score analysis to estimate missing counterfactual outcomes by identifying matching subpopulations of individuals with similar distributions of covariates, but with differing treatment status. Given a set of social media messages, we apply a preprocessing step to generate a set of per-user timelines. Once a query is issued, we identify the users that have mentioned the treatment experience and place them in a *treated* group, and place all other users in a *control* group. We align user timelines based on when the individual mentioned experiencing the treatment. We align the control users based on a random “placebo” time. To reduce the effects of temporal biases, we assign placebo times to match the distribution of treatment times.

Stratification is achieved by estimating each individual’s likelihood of being in the treated group using a propensity score model. This is a learned function that infers likelihood of being in the treated group as a function of a set of covariates (i.e., individual properties and past tweets that might influence both treated/control status and outcomes). Individuals with similar propensity scores are grouped into strata. In aggregate, individuals within a strata are likely to have similar covariates, allowing us to isolate and estimate the effects of the treatment itself within each strata. Note that the primary purpose of the propensity score model is to identify groups of individuals with similar covariates—the accuracy of predicting group status is secondary. To ensure the quality of counterfactual estimates, the method drops strata that have either too few treated or too few control users. Outcomes are aggregated across remaining strata, weighted by the size of the treatment population in the strata, to estimate the average effect of treatment on the treated population.

The details of our analysis are as follows:

Covariate and outcome features: The content of social media messages from before the treatment (or placebo) time, as well as other user properties (posting frequencies, message lengths, profile information, etc.) are extracted as covariates—potentially confounding features that may influence both treatment status and outcomes. The content of social media messages after the treatment (or placebo) are extracted as the time-varying outcome measures of the treatment.

We represent social media message content in our covariate and outcome features as empirical, unsmoothed word likelihoods. We limit our word distributions to the top 50k unigrams in our corpus. We do not remove stopwords, stem or normalize the text, and use whitespace and punctuation to identify word-breaks. Optionally, given a word-to-topic mapping, we combine outcome word likelihoods to generate the total topic likelihood.

Propensity score modeling: We implement our high-dimensional propensity score analysis as a logistic regression with 10-fold cross-validation. Our analysis divides users into 100 strata, removes strata with either or both too few Treated or too few Control users. In practice, this removes the lowest-propensity strata and the highest-propensity strata, leaving the middle strata in these analyses. The outcome differences in these remaining strata are weighted according to the Treated population distribution and combined to estimate the average treatment effect on the Treated group.

While we borrow propensity score analysis from the causal inference literature, our application of this technique is not a causal analysis, as two key assumptions may not hold: First, all confounding variables must be included in the observed covariates. Yet, while high-dimensional propensity score analyses, such as ours, are more likely to capture those variables correlated with confounding variables, it is difficult to argue that all relevant aspects of individuals' lives are captured in their Twitter streams. Second, the stable unit treatment value assumption (SUTVA) must hold—that is, one person's outcome must be independent of whether another person had the target experience. Additional domain knowledge is required to assert these assumptions.

4 OUTCOMES ENGINE ARCHITECTURE

To execute online ad-hoc causal inference analyses over large-scale datasets, we must provide scalable implementations for treatment identification, covariate and outcome extraction, and propensity score modeling. We use a two-tiered approach to our cluster design: 1) User data is distributed randomly across *data nodes*, with all data from a single user assigned to a single node. Each data node consists of a Treatment Identification server and a Timeline server. 2) A centralized *query node* is responsible for distributing queries across all data nodes, centralized building of the propensity score model, and aggregating stratified outcomes.

Treatment ID Server. The Treatment ID server provides an index over the full text of text messages. Given a query (the treatment identification pattern), the treatment ID server uses the index to return the user ID and treatment time for users who have posted a message matching the query. In addition, the Treatment ID server returns a sample of the remainder of the population to be used as a control group. These user IDs are each returned with an assigned placebo time. The size of the control sample is given as a multiple of the treatment population size. The larger the control population, the more likely that there will be similar users (i.e., better matches) between the treated and control populations. The trade-off is that analyzing a larger control population will require more time.

Model Builder. The Model Builder collects the covariates and treatment/control status of users (or samples of users) from all Data nodes. Then, it applies a supervised algorithm to learn a model of the propensity of users to be treated. This learned model is distributed across all the data nodes.

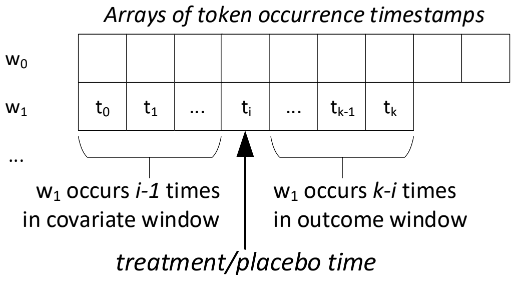

Arrays of token occurrence timestamps
w_0
w_1
...
t_0
t_1
...
t_i
...
t_{k-1}
t_k
w_1 occurs *i-1* times in covariate window
w_1 occurs *k-i* times in outcome window
treatment/placebo time

Figure 2: Timeline data structure

Timeline Server. The Timeline server stores, for each user, a compressed representation of the timeline of token occurrences (the unigrams, bigrams, or phrases mentioned by users). Given a treatment (or placebo) time for a user, the timeline server can quickly return a summary representation of covariates, or a summary representation of outcomes. Figure 2 shows a sketch of the simple timeline data structure. For each token that has been used by a user, we use a binary search to identify the array index of the treatment time, and compute the number of occurrences of the token from the index value. Simple extensions allow us to calculate the number of occurrences within arbitrary time windows.

Outcome Aggregator. The Outcome aggregator is responsible for gathering the partially aggregated outcomes from data nodes, identifying strata to drop due to lack of comparable subpopulations, and performing a weighted aggregation of outcomes across remaining strata. In addition, the Outcome aggregator runs diagnostics on the analysis, such as covariance balance and other validity tests.

Request flow. As shown in Figure 3, when a request arrives from an application to the query node, the query node first forwards the query to all data nodes (step 1), where the Treatment ID server identifies the treated and control groups and individuals' treatment and placebo times (step 2). Then, each Timeline server featurizes the covariates for these users and returns these covariates and their treated/control labels to a Model Builder in the centralized query node (step 3). If the treatment and control groups are very large, they can be downsampled to improve end-to-end performance.

The Model Builder collects these covariate and label data from all the replication nodes, dynamically learns a propensity score model and sends the model to all of the Timeline servers (step 4). Each Timeline server applies the propensity score model to assign users to strata, scan over outcomes experienced by each user and partially aggregate the outcomes. These outcomes are returned to the Outcome Aggregator on the centralized query node (step 5). These outcomes from all data nodes are aggregated and then returned to the app user (step 6).

5 APPLICATIONS AND EVALUATION

Our work can be seen as part of the broader trend in search systems of bridging the online and physicals worlds [6]. Using social media as a longitudinal sensor into people's experiences, we build a digital

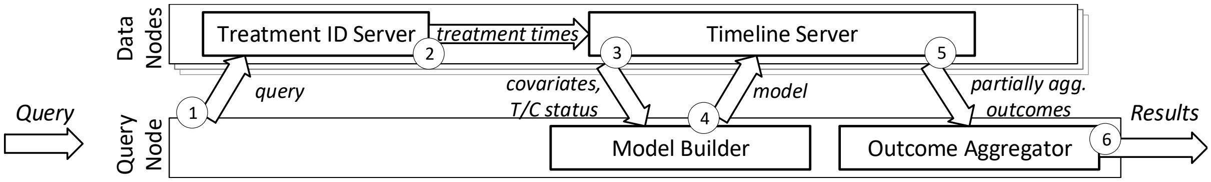


Figure 3: Outcomes Engine

representation of the consequences of actions and situations. A key component to ensuring the interpretabilty and usefulness of this information for improved exploration and decision-making is how and when applications present this information and enable interaction. In this section, we discus some of the considerations for applications and how they might be evaluated.

5.1 Applications

Applications for Individuals. First, we believe that individuals may benefit from the kind of outcomes we uncover. For instance, prior work on online health communities indicates that new patients seek experience-based information from others in similar situations for advice, or to validate their feeling or life decisions [13, 16] In such a scenario, our work can support users in exploring the type of issues others in similar situations are likely to have experienced as a consequence. Further, even when the outcomes of an action or situation are known, aggregated statistics about their likelihood can prove informative for those seeking information about them. Apart from helping individuals understand new situations, information about potential outcomes can also be used to support them in achieving goals or making decisions.

Figure 1 shows user interface sketches that present expectation maps in different forms. The timeline representation, shown in Figure 1-a, can help users understand how outcomes evolve following an action or experience. A list of pros and cons may be better suited in decision-support scenarios to ensure that the decision-maker is aware of the most important consequences, good and bad, of a choice. Conversational assistants may use expectation maps to aid topical chit chat and banter, as well as provide more direct advice and information support.

Application for Policy-makers & Scientists. While our work is motivated primarily by the desire to help individuals understand their situations and the possible implications of their actions on a need basis, there is also an opportunity to use this kind of analysis to better understand behavioral phenomena of societal importance, third-party interventions and other policy questions. Further, large, quantitative analyses such as ours can complement small-scale qualitative or survey-based studies of social phenomena (e.g., see [8, 18]), and vice-versa. Insights about topics of interest may inform what questions are being asked, while insights on temporal dynamics may be used to align survey answers with time dependent-episodes [14].

Across all of these potential uses of expectation maps by individuals and policy-makers, there are important questions about how searchers interact with this information and how to best support their tasks, their exploration and their understanding of this data. In general, we have found that displaying samples of the underlying supporting evidence—i.e., messages written by individuals who have had an experience and a particular consequence—provides significant help in interpreting results and understanding potential underlying causal mechanisms for an outcome [28]. Beyond these domain-agnostic presentations of textual data, domain-specific applications may utilize additional domain knowledge and context to improve interpretability.

5.2 Evaluation Strategies

We propose three key criteria for evaluating the quality of expectation maps and their use: the *correctness* of the expectations; the *interpretability* of results; and the overall *usefulness* of the information for searchers.

Correctness. In prior work, we measured the surface validity of results of our analysis across a broad variety of domains (including in health, business, and society topics), based on manual annotation of outcomes by crowd-workers [28]. Here we briefly summarize our evaluation method and key results. Each specific expectation—a relationship between a single experience and a single outcome of the experience—was shown to workers, with a question about whether a person who had the given experience would be more likely to talk about the outcome in the future. To aid intepretability, we provided workers with two pairs of text examples of experience and outcome messages, and links to web search results for the experience and consequence. With these annotations, we measured the precision of results @N (ranked by effect size).

Figure 4 shows the precision variation at different cut-offs across experiments. We notice a drop of 10-20% in precision from the top 5 to the top 20 outcomes—with the median precision dropping from close to 80% to about 50%, followed by a slower overall decay. Yet, even after the top 30, the discovered outcomes attain an average perceived precision of over 50%. These results have two main takeaways: overall, the discovered outcomes tend to attain good precision scores across experiences, which correlate with their effect size. Separately, we find that P@10 varies across domains—ranging from over 55% to 100% on average per domain—and that the perceived precision varies strongly with the data volume it was computed on. This partially explains the variance of P@10 across domains. However, other factors, such as errors in the semantic interpretation of words and domain-specific biases in the likelihood of users to mention certain outcomes might also play a factor.

Beyond evaluating the surface validity of results, another method for evaluating the correctness of expectation maps is prediction over hold out data. If our predictions are reliable, our treatment effect estimates should match that seen in hold out data. Finally, as a truly end-to-end test of accuracy, we may consider asking searchers

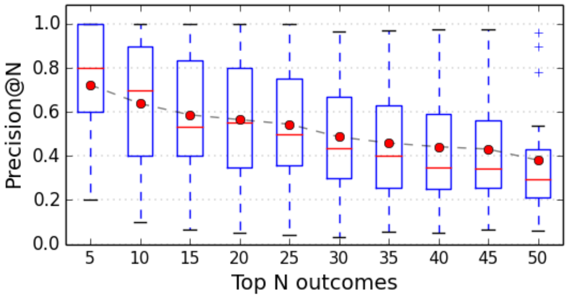


Figure 4: Variations in precision across top N outcomes. The boxplots summarize the precision@N across 39 distinct situations in 9 domains within health, business and society topics. Red lines represent the median, while dots the mean.

to see how their experiences evolved, and how well that matches our mined expectations.

Interpretability. While results may be technically correct, searchers are more likely to be successful if the results they see are quickly and easily interpretable. Methods for improving interpretability can rely on exploration, supporting evidence and context, as mentioned above. While evaluating the interpretabilty of results presents many challenges and is left largely for future work, we believe it will benefit from earlier methods developed for quantitative and qualitative evaluation of search quality [21, 24, 37]

Usefulness. To truly understand the end-to-end benefits of this for end users, however, we must perform end-to-end studies of the usefulness of the results in improving people's outcomes—e.g., are searchers more confident in their choices and making better decisions? For this purpose, we recommend long-running user studies and surveys that capture the situations people are exploring, why they are exploring them (whether for immediate decision-making, for long-term planning, or simply out of curiosity), and later come back to the user and ask them about how this information affected their behavior, choices, and possibly even outcomes.

6 CONCLUSION

As computing devices continue to become more embedded in our everyday lives, they are mediating an increasing number of our interactions with the world around us. From helping people search for the best product to buy, to recommending a restaurant we are likely to enjoy, computing services enable users to evaluate options and take action with "one click". While such services model many facets of the options they present, they do not model the higher-level implications and trade-offs inherent in deciding to take one action instead of another. By aggregating the combined experiences of hundreds of millions of people, our search services have an opportunity to provide significant assistance to individuals in their expectation explorations and decision-making. Integrating causal inference as a fundamental piece of this analysis allows us to capture consequences of actions and situations that enables our search services to be better integrated into interventions, such as decision-support, planning, and advice scenarios, where correlational analyses may be too risky given consequential outcomes.

Acknowledgments

This work builds on ideas developed over several years through collaborations and discussions with many people. We would like to thank many colleagues, including Paul Bennett, Scott Counts, Munmun De Choudhury, Susan Dumais, Adam Fourney, Myeongjae Jeon, Alexandra Olteanu, Matt Richardson, Michael Lowell Roberts, Du Su, James Thomas, Onur Varol, Ryen White, Xinhao Yuan, Lidong Zhou, and Brian Zill.

REFERENCES

[1] Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proc. of ACM CHI.* 3197–3206.

[2] Omar Alonso, Serge-Eric Tremblay, and Fernando Diaz. 2017. Automatic Generation of Event Timelines from Social Data. In *Proceedings of the 2017 ACM on Web Science Conference.* ACM, 207–211.

[3] Tim Althoff, Pranav Jindal, and Jure Leskovec. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *Proc. of ACM WSDM.* ACM, 537–546.

[4] Kat Austen. 2015. What could derail the wearables revolution? *Nature* 525 (2015).

[5] Ricardo Baeza-Yates. 2005. Searching the future. In *SIGIR Workshop MF/IR.*

[6] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. 2018. Here and Now: Reality-Based Information Retrieval: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval.* ACM, 171–180.

[7] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. 2015. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)* 47, 2 (2015), 15.

[8] Wen-Ying Sylvia Chou, Yvonne Hunt, Anna Folkers, and Erik Augustson. 2011. Cancer survivorship in the age of YouTube and social media: a narrative analysis. *Journal of medical Internet research* 13, 1 (2011).

[9] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. In *Proc. of AAAI ICWSM.*

[10] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of ACM CHI.* 2098–2110.

[11] D. Eckles and E. Bakshy. 2017. Bias and high-dimensional adjustment in observational studies of peer effects. *ArXiv e-prints* (June 2017). arXiv:stat.ME/1706.04692

[12] Kate Ehrlich and N Sadat Shami. 2010. Microblogging Inside and Outside the Workplace. In *AAAI Conf. on Weblogs and Social Media.*

[13] Jordan Eschler, Zakariya Dehlawi, and Wanda Pratt. 2015. Self-Characterized Illness Phase and Information Needs of Participants in an Online Cancer Forum. In *Proc of. AAAI Conf. on Web and Social Media.*

[14] Adam Fourney, Ryen W White, and Eric Horvitz. 2015. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proc. of the ACM CHI.* 737–746.

[15] Venkata Rama Kiran Garimella, Ingmar Weber, and Sonya Dal Cin. 2014. From "I love you babe" to "leave me alone"-Romantic Relationship Breakups on Twitter. In *Conf. on Social Informatics.* Springer, 199–215.

[16] Jina Huh and Mark S Ackerman. 2012. Collaborative help in chronic disease management: supporting individualized problems. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work.* ACM, 853–862.

[17] Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management* 44, 3 (2008), 1251–1266.

[18] Lloyd D Johnston, Patrick M O'Malley, Jerald G Bachman, and John E Schulenberg. 2011. Monitoring the Future national survey results on drug use, 1975-2010. Volume I: Secondary school students. (2011).

[19] Emre Kıcıman, Scott Counts, and Melissa Gasser. 2018. Using Longitudinal Social Media Analysis to Understand the Effects of Early College Alcohol Use. In *ICWSM-18.* AAAI.

[20] Emre Kıcıman and Matthew Richardson. 2015. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *Proc. ACM KDD.* 547–556.

[21] Shirlee-ann Knight and Janice Burn. 2005. Developing a framework for assessing information quality on the World Wide Web. *Informing Science* 8 (2005).

[22] Nicolas Kokkalis, Thomas Köhn, Johannes Huebner, Moontae Lee, Florian Schulze, and Scott R Klemmer. 2013. Taskgenies: Automatically providing action plans helps people complete tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 27.

[23] Briege M Lagan, Marlene Sinclair, and W George Kernohan. 2010. Internet use in pregnancy informs womenâĂŹs decision making: a web-based survey. *Birth* 37, 2 (2010), 106–115.

[24] Dmitry Lagun and Eugene Agichtein. 2011. Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*

Retrieval. ACM, 365–374.

[25] Edith Law and Haoqi Zhang. 2011. Towards large-scale collaborative planning: Answering high-level search queries using human computation.. In *AAAI*.

[26] Andrew Meola. 2016. Wearables and mobile health app usage has surged by 50% since 2014. http://www.businessinsider.com/fitbit-mobile-health-app-adoption-doubles-in-two-years-2016-3. (2016). [Online; Accessed 27-July-2016].

[27] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15, 8 (2013).

[28] Alexandra Olteanu, Onur Varol, and Emre Kıcıman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. of CSCW 2017*. ACM, 370–386.

[29] John Paparrizos, Ryen W White, and Eric Horvitz. 2016. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice* 12, 8 (2016), 737–744.

[30] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health.. In *Proc of. AAAI ICWSM*. 265–272.

[31] Michael J Paul, Ryen W White, and Eric Horvitz. 2015. Diagnoses, decisions, and outcomes: Web search as decision support for cancer. In *Proc. of WWW*. ACM.

[32] Andrew Perrin. 2015. Social media usage: 2005-2015. (2015).

[33] Matthew Richardson. 2008. Learning about the world through long-term query logs. *ACM Transactions on the Web* 2, 4 (2008), 21.

[34] Daniel E Rose and Danny Levinson. 2004. Understanding user goals in web search. In *Proceedings of the 13th Intl. conference on World Wide Web*. ACM.

[35] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

[36] Abigail J Sellen, Rachel Murphy, and Kate L Shaw. 2002. How knowledge workers use the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 227–234.

[37] Diana Tabatabai and Bruce M Shore. 2005. How experts and novices search the Web. *Library & information science research* 27, 2 (2005), 222–248.

[38] Rannie Teodoro and Mor Naaman. 2013. Fitter with Twitter: Understanding Personal Health and Fitness Activity in Social Media. In *AAAI Conf. on Weblogs and Social Media*.