Investigating Searchers’ Mental Models to Inform Search Explanations

PAUL THOMAS and BODO BILLERBECK, Microsoft, Australia
NICK CRASWELL and RYEN W. WHITE, Microsoft, USA

Modern web search engines use many signals to select and rank results in response to queries. However, searchers’ mental models of search are relatively unsophisticated, hindering their ability to use search engines efficiently and effectively. Annotating results with more in-depth explanations could help, but search engine providers need to know what to explain. To this end, we report on a study of searchers’ mental models of web selection and ranking, with more than 400 respondents to an online survey and 11 face-to-face interviews. Participants volunteered a range of factors and showed good understanding of important concepts such as popularity, wording, and personalization. However, they showed little understanding of recency or diversity and incorrect ideas of payment for ranking. Where there are already explanatory annotations on the results page—such as “ad” markers and keyword highlighting—participants were familiar with ranking concepts. This suggests that further explanatory annotations may be useful.

CCS Concepts: • **Information systems** → **Users and interactive retrieval**; Page and site ranking;

Additional Key Words and Phrases: Mental models, explanation, ranking, web search

ACM Reference format:
Paul Thomas, Bodo Billerbeck, Nick Craswell, and Ryen W. White. 2019. Investigating Searchers’ Mental Models to Inform Search Explanations. *ACM Trans. Inf. Syst.* 38, 1, Article 10 (December 2019), 25 pages.
https://doi.org/10.1145/3371390

1 EXPLAINING SEARCH

In much of the world, web search is almost ubiquitous and extremely commonly used. However, many searchers have little understanding of how search works and neither engage deeply with search results nor think critically about what a search engine does or does not do [Borgman 1986; Eslami et al. 2016; Hendry and Efthimiadis 2008; Holman 2011; Muramatsu and Pratt 2001; Schultheiß et al. 2018; Zhang 2008]. By helping searchers understand what search does, we can help them be more efficient and effective searchers [Koenemann and Belkin 1996; Muramatsu and Pratt 2001], as well as help them understand any biases in our data, models, and implementations [Baeza-Yates 2018; White and Hassan 2014]. It may also be useful for regulatory purposes, for example, to clarify commercial links between listings and search providers.

System *explanations* have been suggested in other applications [Doshi-Velez and Kim 2017; Lipton 2018; Miller 2017], and “seamful” design [Wenneling 2007] has been used to similar ends in

Authors’ addresses: P. Thomas, Microsoft, Canberra, Australia; email: pathom@microsoft.com; B. Billerbeck, Microsoft, Melbourne, Australia; email: bodob@microsoft.com; N. Craswell, Microsoft, Bellevue, WA; email: nickcr@microsoft.com; R. W. White, Microsoft, Redmond, WA; email: ryenw@microsoft.com.

1046-8188/2019/12-ART10 $15.00
https://doi.org/10.1145/3371390

social media [Eslami et al. 2016] and recommendation systems [Nunes and Jannach 2017]. In our context, it is reasonable to assume that a good explanation, provided either at the time of a search or in advance, would improve searchers' "relatively weak models for how search engines actually work" [Hendry and Efthimiadis 2008]. Some search engines do include such features, for example, by explaining term contributions to a ranking formula,[1] by highlighting query terms where they appear in result listings, or by explaining spelling corrections. An explanation system depends on several factors, including choices in user interface design, and evidence for effectiveness is accordingly equivocal [Ter Hoeve et al. 2017].

In this work, we take a different approach. We do not focus on the *form* of explanations, neither in the way particular explanations are formed nor in the interface and affordances given to searchers (icons alongside results, short texts, links to help pages, etc.). Instead, we ask: *what* should a search engine explain at all? What in the way a search engine operates is familiar and well understood, and what is alien? What in a searcher's mental model is useful, or is wrong, or is missing altogether?

This article describes research to answer these and related questions. In doing so, we make the following contributions:

- Investigate people's understanding of how search engines select and rank web search results, which can directly inform the provision of explanations in search systems.
- Perform a detailed mixed methods study, combining a large-scale survey and face-to-face interviews to learn about nontechnical searchers' understanding of how search engines *currently* operate. This is important since search engine functionality continues to evolve and many of the seminal studies in this area [e.g., Borgman 1986; Hendry and Efthimiadis 2008; Zhang 2008] are now more than a decade old.
- Identify several noteworthy patterns in people's understanding of search engine operation, from specific algorithmic decisions to include a particular item in the result set to general policies that guide search engine result page (SERP) construction.
- Discuss the implications of our findings for the design of search explanations to help searchers better understand search engine operation and interpret search engine responses.

People's mental models of technology have long been studied [Norman 1983]. In this article, we define /mental model/ as a searcher's mental representation of how a search system works, specifically focused on search result selection and search result ranking. More accurate mental models can be employed by users of intelligent systems to better understand and forecast a system's behavior, resulting in heightened satisfaction [Kulesza et al. 2012] and improved task performance [Kulesza et al. 2015], including on search tasks [Koenemann and Belkin 1996; Muramatsu and Pratt 2001].

The remainder of the article is structured as follows. Section 2 describes related work on explanations, spanning many disciplines and communities. Section 3 outlines the study methods, and Section 4 presents our findings. Section 5 discusses the results and their implications for system design. We conclude in Section 6 with a summary of the findings and pointers to future work.

2 RELATED WORK

Prior work on explanations has been rich and varied, ranging from general studies of explanation to specific applications in areas such as search, artificial intelligence (AI), and machine learning (ML), as well as research on understanding explanations when available in various systems.

[1] For example, Elasticsearch at https://www.elastic.co/guide/en/elasticsearch/reference/current/search-explain.html or Splainer at http://splainer.io.

2.1 Explanations in General

Research in philosophy, psychology, and cognitive science has explored how people define, generate, select, evaluate, and present explanations [Ruben 2015; Sandis 2011]. This research argues that people employ certain cognitive biases and social expectations toward the explanation process. Relevant related research in psychology has focused on mental models and their role in human reasoning [Johnson-Laird 1980]. Norman [1983] suggests that system designers must help computer users form accurate and useful mental models. This has practical implications for the utility of intelligent systems. For example, Kulesza et al. [2012] found that those users of a personalized recommendation system who built sound mental models were more quickly able to use the system effectively. Beyond mental models, well-documented cognitive biases have also been shown to affect how people interpret and use the information they encounter [Tversky and Kahneman 1974]. Such biases can impact people's interactions with search engines [White 2013].

2.2 Explanations in Search

Search engines use sophisticated algorithms to determine which search results to return and which additional elements (e.g., related searches, instant answers) to show alongside these results on SERPs [Bailey et al. 2010]. Singh and Anand [2018] trained interpretable learning-to-rank models. They also developed a system to offer insights on a search engine's understanding of query intent, relative document rankings, and document relevance to the query [Singh and Anand 2019]. Ranking methods also include those based on historic search activity of searcher populations [Agichtein et al. 2006] and personalization based on, for example, short- and long-term searcher interests [Bennett et al. 2012] or searcher geographic location [Bennett et al. 2011]. In addition, how people interpret results can have significant consequences, such as decisions regarding courses of medical treatment [Paul et al. 2015; Pogacar et al. 2017].

Several studies have examined people's mental/conceptual models of search engines. Borgman [1986] sought to train novice searchers to use a Boolean logic-based search system by developing a mental model of search system operation. Subjects with the model-based training were more able to perform complex tasks requiring extrapolation from basic system operations. Grounded in the complexity of human-human, goal-directed dialogue, Belkin [1988] proposed a general model of clarity in human-computer systems, of which explanation by a computer intermediary is one component. Koenemann and Belkin [1996] studied relevance feedback (RF) interfaces that were *opaque* (RF functionality hidden), *transparent* (searchers could see query expansion terms generated by RF), and *penetrable* (searchers could see and adjust the terms generated by RF). They showed that increased opportunity for searcher interaction with and control of RF made the interactions more efficient and make the search engine more usable while maintaining or increasing search effectiveness. Muramatsu and Pratt [2001] conducted a user study to understand people's knowledge and reaction to query transformations performed by search engines. They also developed a system called *Transparent Queries*, which provides searchers with feedback about internal query transformations. Efthimiadis et al. [2004] focused on different aspects of the mental models that people employ during interactions with retrieval systems. This includes the factors affecting model construction, how people react to the search process and information found, and how they conceptualize search engines. Hendry and Efthimiadis [2008] solicited sketches from students on how they believed search engines function. The figures that resulted were diverse and contained many core concepts but were also simplistic. Zhang [2008] performed a mixed methods study of undergraduate students' mental models of the web as an information source. Participants' mental models of search on the web were found to cover several areas: avenues for obtaining information, search engine mechanics, and search tactics. They also found that searchers can learn by "personal observation, communication with others, and class instruction," which can be inefficient learning mechanisms when search systems could be capable of conveying explanations for their actions directly. Holman [2011] studied undergraduate students and found that although no students had strong mental models of search engine mechanisms, those with stronger models constructed more complex queries.

These studies on mental models in search have targeted query formulation or search engines holistically, and not how searchers conceptualize ranking algorithms and the presence of specific SERP elements, as we do in this study. A survey by Nakamura et al. [2007] found that searchers thought that engines made use of visit counts, "relevance," and freshness to rank pages; 17% also believed payment played a role. These results, although a decade old, are fairly consistent with our findings. Our study is more thorough, however, allowing us to capture a broader range of possible explanations, and, more recent, meaning that we can situate our study in the current state of search engine operation, which continues to evolve over time.

Other studies have looked at the decisions searchers make when examining SERPs—for example, decisions on whether to examine an individual result (e.g., see Saracevic [2007a] or Freund [2008] for an overview). Criteria that have emerged include clarity, accessibility, and scope, as well as topicality. Rather than ask how searchers make *their own* decisions, in this work we ask how searchers think that *search engines* make decisions, and we expect different criteria.

2.3 Explanations in AI and ML

More broadly, there are a range of efforts to achieve transparency and accountability in computer systems. The ACM U.S. Public Policy Council [2017] has highlighted seven core areas: awareness, access and redress, accountability, explanation, data provenance, auditability, and validation and testing. Baeza-Yates [2018], reflecting on bias on the web, focused primarily on awareness regarding the presence of bias and the need to design web-based systems with bias in mind. As a result, search engines may want to surface explanations regarding potential result biases [White and Hassan 2014]. In ML and AI, there has been a long-term desire to help people make sense of both models and their output in specific instances [Michie 1988]. Initiatives to define interpretable ML [Doshi-Velez and Kim 2017; Lipton 2018] and explainable AI [Miller 2017] seek to move from opaque (so-called black-box) models toward a situation more conducive to trust, transferability, informativeness, reliability, and fairness. Lipton [2018] argues for transparent models (where the algorithm can be easily simulated and decomposed) and to help consumers of the model make sense of individual predictions (via text/visual explanations). In pursuit of more scientific rigor in interpretable ML, Doshi-Velez and Kim [2017] proposed a taxonomy for evaluating interpretability and point to a need for interdisciplinary collaboration in this area.

There have been many attempts to develop interpretable and intelligible ML models to tackle specific challenges [Caruana et al. 2015; Jung et al. 2017; Lakkaraju et al. 2017; Letham et al. 2015; Lou et al. 2013]. In many of these and similar cases, the model is used to assist human decision making (e.g., clinical decision support [Bussone et al. 2015]). As such, understandability may be the focus, perhaps at the expense of model accuracy [Caruana et al. 2015]. Explanations are more often at the model level (in a search context, this would be broad policies and mechanisms adopted in constructing the SERP) rather than at the instance level (in search, this corresponds to a specific SERP or individual result). That said, there has been increased attention on methods to generate instance-level explanations for the output of ML models [Moore et al. 2018; Ribeiro et al. 2016; Tamagnini et al. 2017]. Other research has considered people's mental models in the design of explanations in intelligent systems [Eiband et al. 2018; Kulesza et al. 2015; Wang et al. 2019].

2.4 Understanding Explanations

Beyond methods to generate explanations, research has also focused on understanding how people interpret and use explanations [Narayanan et al. 2018]. This accompanies trends toward more explainable, accountable, intelligible systems [Abdul et al. 2018] and more algorithmic transparency in such systems [Rader et al. 2018]. Explanations are common in collaborative filtering/recommender systems [Cheng et al. 2019; Herlocker et al. 2000; McSherry 2005; Nunes and Jannach 2017; Sinha and Swearingen 2002; Tintarev and Masthoff 2011; Vig et al. 2009], where they are frequently offered to explain why particular recommendations were generated, and in context-aware systems [Bellotti and Edwards 2001; Lim et al. 2009], where the role of contextual factors may need to be outlined, and research has focused on the intelligibility of system behavior by, say, visualizing uncertainty [Rukzio et al. 2006] or improving trust by displaying system confidence [Antifakos et al. 2005]. Trust is an important driver in the provision of explanations, especially as searchers grow to rely on systems to automate certain tasks [Lee and See 2004] or, as mentioned earlier, in the case of search engines, inform consequential decisions. Research on trust and explanations has included the effects of transparency on levels of trust [Kizilcec 2016] and the role of explanations to build trust [Glass et al. 2008; Pu and Chen 2006]. In search engines, explanations can assist users, for example, in assessing result credibility on SERPs [Schwarz and Morris 2011; Yamamoto and Tanaka 2011] and in documents [Bountouridis et al. 2018]. Even without explanations, given enough time and use, people can still build sound mental models of intelligent systems. Tullio et al. [2007] performed a 6-week field study to understand how nontechnical searchers use such systems. They found that through using the system, participants addressed their initial misconceptions about system intelligence and could even conceptualize basic ML concepts by the end of the study, even though they had no training. Similarly, given the prevalence and longevity of some search engines, we want to understand whether nontechnical searchers can explain aspects of system operation, even though search engines fail to explain much of what they do. Ideally, users would not need to go through this learning process, and that is where system explanations can help. Tullio et al. [2007] also found that people might be reluctant to change their (possibly incorrect) mental models. Explanations offer a means of possibly correcting wrong mental models.

2.5 Contributions over Previous Work

Most previous studies of search explanations are at least a decade old at the time of this writing (raising concerns about how accurately they reflect searchers' current mental models), they focus primarily on student participants, they target a narrow range of explanations, and/or they consider how searchers make decisions and not how they believe search engines make decisions. We make several contributions over previous work. First, we focus on search engines, an underexplored area in explanations research, and improve the currency of research on search explanations. Second, we learn more about nontechnical searchers' understanding of how search engines rank results in general and how specific result types (e.g., local answers) are selected for inclusion on the SERP. We do this through a mixed methods study, combining the results from a large-scale survey with in-person semistructured interviews to obtain a breadth of opinion from nontechnical users on how they believe ranking and result element selection is performed. Third, we show that there is a broad range of explanations that searchers have for search engine operation, some expected, such as the importance of popularity, and some unexpected, such as payment for promotion in the result ranking. Finally, based on our findings, we offer design implications for the provision of explanations in search engines.

3 METHOD

We are interested in searchers' mental models of search engines, and whether knowing more about these models could inform the explanations that search engines provide. To gather data, we performed two exercises: (1) an online survey with several hundred responses and (2) in-depth face-to-face interviews with 11 participants. We examined survey responses, as well as recordings of the interviews, and coded each according to the distinct concepts offered by our participants: participants offered up to 20 distinct concepts each. Finally, we formed these concepts into a hierarchy to find common elements.

3.1 Overview

We used data from two parallel exercises—a survey and interviews—in a convergent mixed methods design, where observations from one set of data could confirm, disconfirm, or clarify observations from the other [Creswell 2014]. This allowed us to offset the weaknesses in one method against the strengths of the other. The survey gave us a large number of respondents, and hence a large number of concepts and some idea of the distribution of ideas, but no chance to interrogate the respondents or understand their concepts in any depth (Section 3.2); the interviews gave us in-depth responses, and a chance to discuss these, but with many fewer people (Section 3.3). The interviews also allowed us to use a simple prop, described in the following.

Since both exercises, by design, addressed the same constructs (search engines, search results, selection, ranking) and elicited the same sorts of response, we could code them in the same way (Section 3.4). Quantizing the data lets us directly compare the results (Section 4.2).

Core concepts. We are also interested in concepts that participants did not mention, but which are in fact factors in selection and ranking at web scale. Before running the surveys and interviews, the authors agreed a set of core concepts—key ideas that a searcher should understand. This list was based on our own expertise but accords with what is commonly understood in the industry [e.g., see Search Engine Land 2018a, 2018b]. These concepts all play a part in modern search engine operation, although not all are clearly shown on a SERP. Any concepts in this set, if they were infrequently or never mentioned by searchers, should perhaps be explained by a search engine. These concepts were as follows:

(1) *relevance*, especially to common intents or needs among people typing the query;
(2) *recency* of information;
(3) *popularity* of each result;
(4) *authority* of each result, such as reputation or trustworthiness;
(5) *locality*, such as preferring business listings geographically close to the searcher; and
(6) *diversity* of results across the entire SERP.

3.2 Online Survey

The first data comes from an online survey conducted from August through September 2018, with 497 respondents. Survey questions are reprinted in the appendix.

Respondents were recruited through SurveyGizmo,[2] a professional panel service, balanced for age and gender according to the U.S. Census. Each respondent was asked two screening questions. If they indicated they supported or developed information technology (IT) systems, or had an IT qualification, they were removed; similarly, they were removed if they worked in a library or had qualifications in library and information science. We expect such people to be highly sophisticated searchers and therefore of less interest in this study. Post hoc, we also screened out anyone who

[2] https://www.surveygizmo.com/.

gave unambiguously off-topic or spam responses. After screening, we included 497 of 739 respondents. Qualifying participants were compensated with US $3 for a median 2-minute session.

Since experience may be a factor in mental models [Holman 2011], we collected simple demographic data: how long respondents had been using search engines, and how often they used search engines. Besides basic demographics, the instrument included only two questions. These tried to elicit respondents' models of two key aspects of search: *selecting* results to return (which may include ranking, e.g., to form the top 10 results) and *arranging* these results on a SERP. The questions did not refer to explanations—indeed, one goal is to learn what type of explanation people use without prompting—and were deliberately general on what constituted a "result." Our first question asked, What do you think determines what gets shown on the results page? Our second asked, What do you think determines what goes where on the results page? Both questions allowed, and encouraged, long-form answers, but there was no minimum response length.

3.3 Interviews

We also conducted semistructured face-to-face interviews, with 11 people, in August through September 2018. A sample script is included in the appendix. Although this is a very much smaller sample, face-to-face interviews allowed more in-depth questioning, we could follow up on answers that were given, and we were able to use simple props to ground the conversation.

We used an in-house panel service to recruit people from around the Seattle area, and who did not have extensive professional or educational experience with search or library systems. Demographics were collected online at recruitment. Interviewees were compensated US $150 for a session of up to an hour. All interviews were conducted by the first author, and audio was recorded for later analysis.

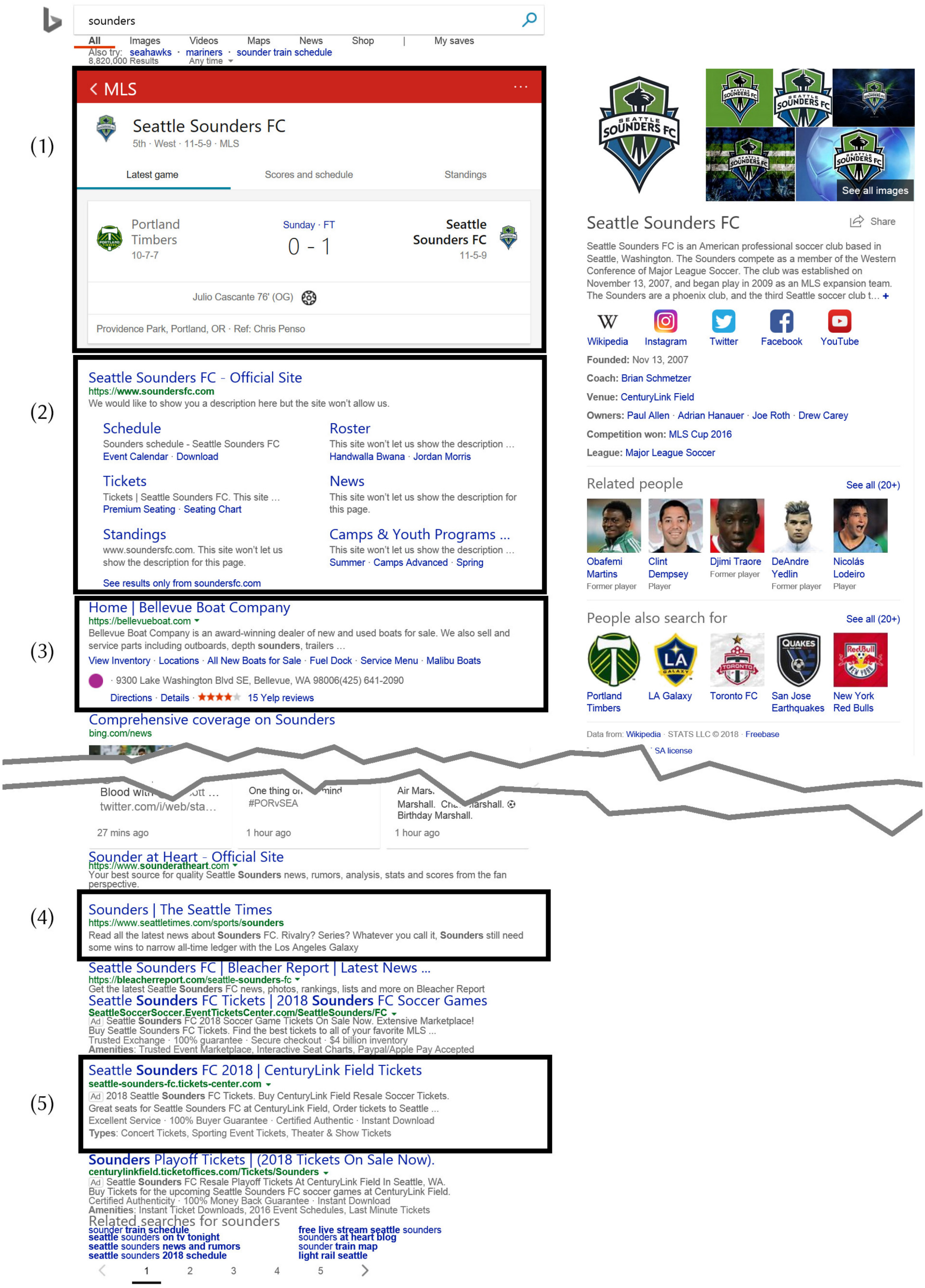
The interviews themselves included variants of the two questions asked in the survey, plus a further "situated" question and a short series of follow-up questions. The first question asked how search engines selected, from among hundreds of billions of possibilities, which results to include, and the second question asked how search engines arranged these results to form a SERP. These questions were worded similarly to the survey.

We expected that these questions might be difficult to answer without a concrete example. We therefore included further questions about a simulated SERP, designed to illustrate the selected concepts described earlier (Figure 1). The SERP was for the query "sounders," where the dominant intent is for a Seattle-based sports team,[3] and we called out five results for discussion:

(1) the scores and statistics for the most recent game (demonstrating common intents and recency);
(2) the team's home page, labeled "official site" (demonstrating authority and popularity);
(3) a fictional company, not related to the sport, but local and matching the team's name (demonstrating location, keyword matching, and topical diversity);
(4) coverage from the local newspaper (demonstrating authority and recency), and;
(5) an advertisement for tickets (demonstrating commercial arrangements).

To prompt further reflection from participants, the interviewer pointed at each of these five results in turn and asked, "Why was this result in particular chosen?" and "Why did it appear where it did on the page?" Although the simulated SERP used all design elements of a major search engine to make it as realistic as possible, a SERP with different results and different types of results may well have elicited a different set of responses.

[3] Seattle Sounders FC is an American professional soccer club based in Seattle, WA.



sounders

All Images Videos Maps News Shop | My saves

Also try: seahawks · mariners · sounder train schedule

8,820,000 Results Any time

‹ MLS ···

Seattle Sounders FC

5th · West · 11-5-9 · MLS

Latest game | Scores and schedule | Standings

Portland Timbers 10-7-7 | Sunday · FT 0 - 1 | Seattle Sounders FC 11-5-9

Julio Cascante 76' (OG)

Providence Park, Portland, OR · Ref: Chris Penso

Seattle Sounders FC - Official Site
https://www.soundersfc.com
We would like to show you a description here but the site won't allow us.

Schedule
Sounders schedule - Seattle Sounders FC
Event Calendar · Download

Roster
This site won't let us show the description …
Handwalla Bwana · Jordan Morris

Tickets
Tickets | Seattle Sounders FC. This site …
Premium Seating · Seating Chart

News
This site won't let us show the description for this page.

Standings
www.soundersfc.com. This site won't let us show the description for this page.

Camps & Youth Programs …
This site won't let us show the description …
Summer · Camps Advanced · Spring

See results only from soundersfc.com

Home | Bellevue Boat Company
https://bellevueboat.com
Bellevue Boat Company is an award-winning dealer of new and used boats for sale. We also sell and service parts including outboards, depth **sounders**, trailers …
View Inventory · Locations · All New Boats for Sale · Fuel Dock · Service Menu · Malibu Boats
· 9300 Lake Washington Blvd SE, Bellevue, WA 98006(425) 641-2090
Directions · Details · ★★★★☆ 15 Yelp reviews

Comprehensive coverage on Sounders
bing.com/news

Blood wi… …ott …
twitter.com/i/web/sta…
27 mins ago

One thing o… mind
#PORvSEA
1 hour ago

Air Mars…
Marshall. Ch… arshall.
Birthday Marshall.
1 hour ago

Sounder at Heart - Official Site
https://www.sounderatheart.com
Your best source for quality Seattle **Sounders** news, rumors, analysis, stats and scores from the fan perspective.

Sounders | The Seattle Times
https://www.seattletimes.com/sports/**sounders**
Read all the latest news about **Sounders** FC. Rivalry? Series? Whatever you call it, **Sounders** still need some wins to narrow all-time ledger with the Los Angeles Galaxy

Seattle Sounders FC | Bleacher Report | Latest News …
https://bleacherreport.com/seattle-sounders-fc
Get the latest Seattle **Sounders** FC news, photos, rankings, lists and more on Bleacher Report

Seattle **Sounders** FC Tickets | 2018 **Sounders** FC Soccer Games
SeattleSoccerSoccer.EventTicketsCenter.com/SeattleSounders/FC
Ad Seattle **Sounders** FC 2018 Soccer Game Tickets On Sale Now. Extensive Marketplace!
Buy Seattle Sounders FC Tickets. Find the best tickets to all of your favorite MLS …
Trusted Exchange · 100% guarantee · Secure checkout · $4 billion inventory
Amenities: Trusted Event Marketplace, Interactive Seat Charts, Paypal/Apple Pay Accepted

Seattle **Sounders** FC 2018 | CenturyLink Field Tickets
seattle-sounders-fc.tickets-center.com
Ad 2018 Seattle **Sounders** FC Tickets. Buy CenturyLink Field Resale Soccer Tickets.
Great seats for Seattle Sounders FC at CenturyLink Field, Order tickets to Seattle …
Excellent Service · 100% Buyer Guarantee · Certified Authentic · Instant Download
Types: Concert Tickets, Sporting Event Tickets, Theater & Show Tickets

Sounders Playoff Tickets | (2018 Tickets On Sale Now).
centurylinkfield.ticketoffices.com/Tickets/Sounders
Ad Seattle **Sounders** FC Resale Playoff Tickets At CenturyLink Field In Seattle, WA.
Buy Tickets for the upcoming Seattle Sounders FC soccer games at CenturyLink Field.
Certified Authenticity · 100% Money Back Guarantee · Instant Download
Amenities: Instant Ticket Downloads, 2016 Event Schedules, Last Minute Tickets

Related searches for sounders

sounder **train schedule**
seattle sounders **on tv tonight**
seattle sounders **news and rumors**
seattle sounders **2018 schedule**
free live stream seattle sounders
sounders **at heart blog**
sounder **train map**
light rail seattle

‹ 1 2 3 4 5 ›

See all images

Seattle Sounders FC Share

Seattle Sounders FC is an American professional soccer club based in Seattle, Washington. The Sounders compete as a member of the Western Conference of Major League Soccer. The club was established on November 13, 2007, and began play in 2009 as an MLS expansion team. The Sounders are a phoenix club, and the third Seattle soccer club t… +

Wikipedia Instagram Twitter Facebook YouTube

Founded: Nov 13, 2007

Coach: Brian Schmetzer

Venue: CenturyLink Field

Owners: Paul Allen · Adrian Hanauer · Joe Roth · Drew Carey

Competition won: MLS Cup 2016

League: Major League Soccer

Related people See all (20+)

Obafemi Martins, Former player · Clint Dempsey, Player · Djimi Traore, Former player · DeAndre Yedlin, Former player · Nicolás Lodeiro, Player

People also search for See all (20+)

Portland Timbers · LA Galaxy · Toronto FC · San Jose Earthquakes · New York Red Bulls

Data from: Wikipedia · STATS LLC © 2018 · Freebase

SA license

Fig. 1. The sample SERP used in the interviews (part of the SERP removed for space). Some of the results—marked above with numbers and black frames—were called out for further discussion with interviewees. These highlighted results illustrate issues such as common intents, recency, authority, popularity, location, keywords, topical diversity, and commercial arrangements. Numbering matches the text in Section 3.3.

A final set of questions explicitly discussed the notion of explanations from search engines, asking whether these would be helpful, how they should appear, whether they would change searcher behavior, and whether alternative page layouts would be more transparent.

3.4 Analysis

We adopted the same approach to analyzing the responses to the online survey and the interview recordings.

In an initial round, one of the authors (Thomas) open coded 50 survey responses for ideas concerning selection or ranking [Corbin and Strauss 2012]. From these fine-grained codes, such as "past purchases" or "my previous searches," we built a hierarchy of concepts. For example, both "past purchases" and "my previous searches" fell under "my history," which further fell under "personalization." This set of concepts formed the basis for recoding both the initial sample and the remainder of the responses, although extra codes were occasionally needed. Of the 497 nonempty, nonspam responses, 79 could not be coded. These included responses like "the things we search" (which could refer at least to keywords, past searches, or topics, and "we" could mean the searcher or a group) or "google decides what you get to see" (indeed, but how?). These responses were excluded from further analysis.

Interviews were coded in a similar way, using a hierarchy that aligned as much as possible with that from the online answers. Since interviewees were prompted with a concrete example of a SERP (and hence they may have been more motivated and had more time to think of answers), we expected them to identify many more concepts, and therefore we analyze their responses separately.

Although we had questions corresponding to selection (choosing results to display) and layout or ranking (choosing how to arrange the SERP), it was evident early in the analysis that most respondents did not have separate models of these two concepts. In other words, a large number of participants answered the second question with variants of "as before" or by repeating themselves. We therefore chose to analyze both answers together.

To check the reliability of our codes, a second author (Billerbeck) used the resulting set of top-level categories, and codebook, to independently code 100 survey responses. (We checked only top-level categories as these form the major part of our analysis in the following.) The coding was consistent: the Jaccard similarity over assigned labels was 0.75 (i.e., the two sets of categories were 75% similar), and the Hamming loss over all labels only 0.03. (The Hamming loss accounts for labels that were not assigned, as well as those that were [Sorower 2010]; i.e., in 97% of cases, we agreed on the application or nonapplication of a code.) We conclude that the following analysis, although based on a single coding, would have been similar with a different person coding the responses.

The final assigned codes gave us a quantization of the responses [Sandelowski 2000; Teddlie and Tashakkori 2009]. Since we used the same questions, and codes, in both the survey and interviews, we were able to compare the two types of response and also able to build a single summary table.

It is not our intention here to measure the correctness, or "soundness" [Kulesza et al. 2012], of our participants' mental models. A modern web search engine makes use of thousands of signals, combined in intricate ways and with weights that vary from query to query, and it is not at all clear how to build an instrument to measure a mental model against this complexity. We do discuss clear errors, as well as coverage of the preselected concepts, in the observations following.

4 OBSERVATIONS

Survey respondents and interviewees provided a range of concepts, although we elicited many more concepts face-to-face than online.

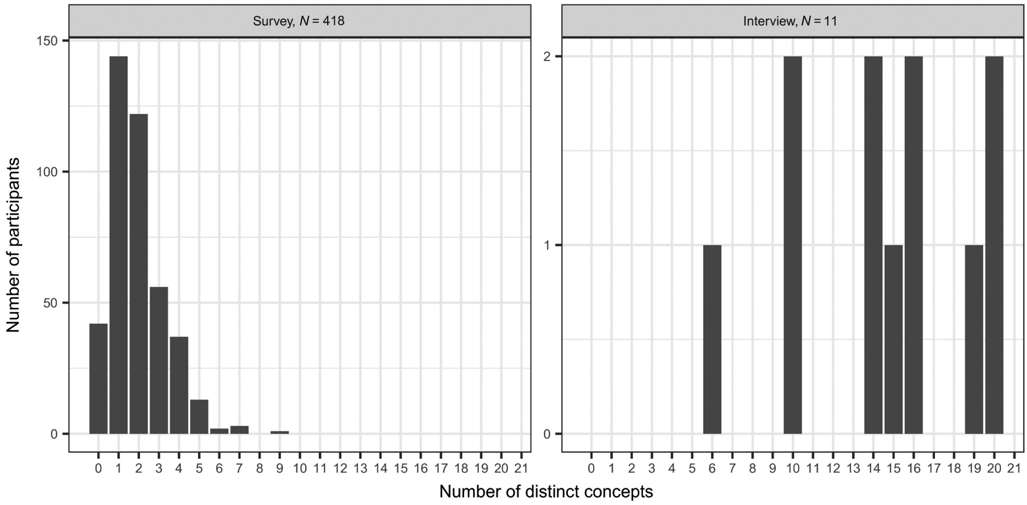


Fig. 2. Number of distinct concepts offered by participants. (N = 418 coded surveys, N = 11 interviews). Participants offering "0" concepts explicitly said they had no idea how search engines work. A further 79 participants whose responses could not be coded are excluded here.

4.1 Text

Respondents to the online survey provided some detail in their answers to the two questions posed (15 words median, 30 words mean, 406 words maximum). There were significant uses of the terms "word" (22% of respondents), "keyword" (9%), and "match" (9%), suggesting an understanding of term matching; "popular" (22% of respondents) and "history" (7%), suggesting an understanding of ideas like logging and using search clicks; "pay" (15% of respondents) and "paid" (12%), suggesting commercial connections (note that "ad" appeared in the prompt); and "relevant" (18% of respondents) and "best" (6%), suggesting a notion of relevance or utility. These ideas are all well represented in the concept analysis that we now describe.

4.2 Concepts

Respondents to the online survey each offered relatively few concepts (Figure 2). A total of 34% volunteered only a single suggestion, the median was two concepts, and only 13% volunteered more than three. This may be partly explained, of course, by the survey mode itself: fixed-size boxes (although large) and little incentive to give a thorough answer. In face-to-face interviews, interviewees suggested many more concepts each, with a minimum of six and median of 15. The prompts were useful here: even in the one case where an interviewee explicitly said she had no idea how search worked, looking at the example SERP prompted many suggestions, including recency, local popularity, and "officialness."

These numbers are broadly in line with those reported by Hendry and Efthimiadis [2008] for freshmen students—for instance, those with an interest in information systems but little formal training. We note, however, a difference in scope: Hendry and Efthimiadis asked a more general question (draw a sketch of how a search engine works). There is also a difference in mode, where Hendry and Efthimiadis asked for pictures rather than text. Our more focused questions demonstrate the detail in searchers' mental models, with a large number of concepts associated with what Hendry and Efthimiadis count simply as "matching" and "ranking."

We were able to distinguish 256 distinct concepts in the responses, which we grouped into a hierarchy with 16 top-level and 116 second-level categories. Table 1 summarizes the major parts of

Table 1. Selection of Concepts Mentioned in the Survey Responses and Interviews

Concept	Survey ($N = 418$)	Interview ($N = 11$)	Notes
None (explicit)	10%	—	Explicit claims of ignorance
Popularity★	38%	91%	Of the result or the topic
Visits/clicks	*18%*	*46%*	
Popularity of the result	*5%*	*36%*	
Popularity given the query	*3%*	—	
Popularity of the query	*3%*	*64%*	
Wording	33%	91%	The form of words in the result, the query, or both
Keywords (in query)	*28%*	*91%*	
—Keywords in page	*10%*	*73%*	
Word matching (not specified)	*4%*	*27%*	
Commercial interests	28%	100%	Search providers' finances, reputation, etc.
Money	*27%*	*100%*	
—Ads	*13%*	*82%*	
—Payment for prominence	*9%*	*46%*	
—Other payment	*3%*	*9%*	
Personalization	24%	82%	Choosing results based on attributes of the searcher(s) themselves
Searcher's history	*17%*	*46%*	
Location★	*8%*	*73%*	
Demographics	*2%*	*9%*	
Relevance or utility	11%	82%	Topicality, subject matter, "aboutness"
Majority intent★	—	*82%*	
Topics	*1%*	*27%*	
Authority★	4%	73%	Of the result, author, host, or information
"Official" pages	—	*64%*	
Search engine optimization (SEO)	4%	—	Attempts to manipulate the ranking
Optimistic answers	3%	—	Assigning magical powers to the engine
Qualities of writing/page layout	3%	46%	Design, style, etc., of the result
Recency★	2%	64%	Time since result publication or update
Inlinks	2%	—	Web hyperlink structure
Specificity of search	2%	—	How well specified the need is
Type of result	1%	27%	Broad type of the result element, e.g., news or images
Properties of a product for sale	1%	—	Price, etc., of an item named in the result
Diversity★	<1%	—	Including fairness, duplication, etc.

Note: Each row includes those indented beneath it. For example "Commercial interests" includes "Money." Participants were counted more than once if they named more than one concept. The concept "None (explicit)" notes those participants who explicitly said they had no ideas. ★ marks the core concepts we identified as important before collecting responses (Section 3.1).

this hierarchy, with the percentage of survey respondents and interviewees offering each. Counts include subconcepts: for example, someone who is counted in both "searcher's history" and "location," both subconcepts of "personalization," would be counted just once in the "personalization" row. That person may also count in other rows if they offered other concepts.

With so many concepts, even the most popular categories were mentioned by a minority of people and no single category was mentioned by more than 38% of survey respondents. A "long tail" saw 108 concepts referenced by only one survey respondent each, and even among the top-level categories only five were referenced by more than 10% (popularity, wording, commercial interests, personalization, and relevance). In interviews, we saw much more overlap—partly due to there being more concepts mentioned per interviewee, and partly due to our prompts—and there were a few concepts suggested by 75% or more of interviewees.

Popularity. Appropriately, the most popular concept was that of popularity, mentioned by 38% of survey respondents and 91% of interviewees. "Popularity" was described in many ways, including visits to pages or clicks on links (18% of surveys) and popularity of searches themselves (3%), although a full 13% of survey respondents did not specify in any detail what they meant. Uses in interviews were broadly in proportion.

Responses coded here included, for example, "I think the more times a link gets clicked, those will result in the top of searches" (participant s103, visits/clicks)[4]; "I assume it's by number of people who do those searches" (i3, popularity of searches); and "I think that they are by popular" (s715, popularity without further particulars).

Popularity of various kinds is a well-understood signal in the information retrieval community [Agichtein et al. 2006], and can be seen in earlier work [Zhang 2008], but it is not explicitly highlighted in current search engines, so it is surprising to see it so well represented here. It may be that popularity is well understood in other contexts (television and radio programming, displays in stores, etc.), and this carries over to search.

Wording. A substantial minority of survey responses (33%) mentioned concepts around the wording, or text, of results. Keywords—meaning words in the query—were mentioned in 28% of survey responses, although only 10% explicitly mentioned the idea of matching keywords (either exactly or approximately). For example, participant s240 described "[an] algorithm that considers search terms" (with the concept of search terms or keywords), and participant s143 suggested "the wording seems to determine what I'm shown" (with the concept of wording more generally)—in neither case describing how these terms are used. In another example, participant s222 explicitly mentioned matching with "I think they pull out words and find the top sites with the words."

Fewer than 1% of responses included notions of keyword location (title, position in text, etc.), and an overlapping 1% included ideas of matching terms in tags or other metadata. Very few respondents (<1%) offered concepts of phrase matching, approximate matching, word order, or parts of speech.

The SERP shown to interview participants included a (fabricated) result that was lexically, but not topically, a match for the query—that is, a keyword match for an off-topic document. With this prompt, a good proportion of interviewees were able to mention keyword matching (8 of 11, 73%); however this is still a somewhat low rate, given the matching word was clearly on display and in bold type.

Word matching, exact or approximate, is fundamental to search but received very few mentions. The difference is perhaps precisely because word matching is so fundamental—it may have seemed too obvious to mention—but matching terms are highlighted in web search results, and the rate is remarkably low at 10% (online) or 73% (in person, with a prompt). It is, however, consistent with results from the study by Holman [2011]. In that work, students were observed to "have only a vague sense of keyword matching," and search engines have become less dependent on exact matches in the years since.

Commercial interests. Just over a quarter of survey respondents (28%) suggested that results are chosen with an eye to search providers' commercial interests. Most of these (27% of respondents)

[4] In this discussion we use, for example, "s1" to refer to a survey participant and "i1" to refer to an interview participant.

mentioned money, and in turn, 13% of respondents mentioned advertisements and an overlapping 13% mentioned other payment. Those latter responses may well have been thinking of advertising, but in most cases it was clear that fees-for-listing was meant. For example, "companies pay to come up at the top of the list" (participant s361), "any paid ads or sites get first priority" (s567, note the distinction between paid ads and paid sites), or "the results are probably sorted by paid partnerships. Anyone who's willing to provide money in exchange for visibility better get that visibility" (s714). Again, this is consistent with Zhang [2008], who observes that "it is widely known by students that search engines are often sponsored by certain Web sites." While over a decade ago at least one pay-for-inclusion engine existed [the now defunct overture.com; see Jansen and Molina 2006], to the best of our knowledge, none of the major web search engines currently accept payment this way[5] and did not do so even at the time of the study of Zhang [2008].

The SERP shown in interviews included advertisements, labeled as such, and every interviewee mentioned commercial interests. Advertisements were mentioned by 9 of 11 interviewees (82%). Again, however, a reasonable number (46%) mentioned fees for ranking—for example, "companies usually pay money to get the top few hits, and then there's usually ads" (participant i14).

A much smaller number (1% of survey respondents) mentioned other commercial arrangements, such as affiliation, ownership, or influence, as reasons to include a search result.

Personalization. Slightly fewer participants (24% of survey responses) suggested concepts of personalization. The most common concept here was the use of a searcher's history (17%), including past searches (12%) and tracking their movements on the web (7%, but overlapping with past searches). Examples include participant s177, "I think they choose what gets listed first by what you have previously searched for" (past searches), and participant s503, "I believe that there is a way that search engines collect information on what I personally view online" (tracking).

More survey respondents mentioned search engines using information about individuals' past purchases (1%) rather than past clicks (<0.5%), although the former information would seem much harder to come by. These figures, especially the 7% mentioning tracking, suggest that people are increasingly aware of being followed both online and (purportedly) offline.

Other concepts in this space included searcher location (8%), demographics (2%), interests (1%), and other characteristics with fewer references. Interviews produced many more mentions of location (73%), again due to an element of our sample SERP that mentioned a local company and others that covered a local sports team. Other aspects of personalization were less common.

Relevance or utility. Although relevance is a central notion in information retrieval [Saracevic 2007a, 2007b], it was mentioned by only 11% of survey respondents. As for keyword matching, this might be because it seemed too obvious to mention; but again, 11% is an extremely low rate. Most respondents offered no more concrete concepts of what "relevance" meant—for example, "most relevant things" (participant s260) and "first the most relevant answer/result" (s168). Just over 1% mentioned concepts of topic or page subject matter. Even these subconcepts are still relatively vague, as in "something with the topic you were looking for" (participant s695) or "subject matter relatednees (sic)" (s438).

By comparison, 9 of 11 interviewees (82%) volunteered the notion of majority intent—that is, that a result would be present because it might be useful to most people who issue a query. This was normally prompted by the first item on the example SERP, which showed scores from a recent football game; comments such as "that's probably the top reason people search" (participant i1) or "a lot of people want to go find out what the score and the standings are . . . more than the actual

[5] Google now clarifies this explicitly on their website: https://www.google.com/search/howsearchworks/.

website" (i7) illustrate that interviewees understood that this would be useful to many other people, if not to them.

Other concepts. Many concepts were mentioned only rarely, but we highlight some interesting replies here. Shortly before the interviews, U.S. president Donald J. Trump accused Google of political bias, claims that received widespread media coverage (e.g., see Waterson and Helmore [2018]). This was not mentioned in any interviews, but four survey respondents—writing before President Trump's comments—mentioned political stance as a factor. For example, participant s338: "I can look up [a] conservative website, type it into the search bar and the first thing that will come up will be a couple [of] liberal websites. I don't think this is by accident . . . some are picked simply because some feel that a liberal point of view needs to be there to counteract the conservative." It was not clear from these online responses whether this was seen as a negative (overt bias) or a positive (counteracting louder, conservative, voices).

Only a small number of survey responses (4%) mentioned concepts of authority, but it was described in many different ways. One respondent each referred to spam, legitimacy, validity, reader confidence, expertise, trust, and accuracy; two respondents mentioned ratings and whether sites were well known; and three mentioned reliability. These are different ideas—compare "expertise" to "confidence," for example. It seems that, despite news coverage and despite research and development efforts [Lewandowski 2012], searchers have only a vague notion of authority and bias, and it is not (yet?) a widespread concern. Prompted by the example, interviewees were likely to mention "official" sites—reflecting the wording on the SERP—but unlikely to clarify this or to offer other concepts of authority. That being said, the presence of authoritative terminology such as "official" in SERP captions has been shown to directly influence search behavior, leading to increased click-through rates on certain search results irrespective of rank position [Clarke et al. 2007].

Qualities of writing or layout, of landing pages (linked to from the SERP), were mentioned by 3% of survey respondents. The most common concept was the presence, or quality, of images on the page (e.g., participant s329, "colorful and attractive page, maybe with pictures"); other concepts included color, humor, and (for one respondent) whether the page discusses politics (s29, "items to make you think, or laugh. Items that deal with politics").

Finally, a small set of users gave flatteringly optimistic answers: that search engines find the "best" results, or the sites one is looking for, or the closest meaning. In an online survey, unfortunately there was no way to interrogate these ideas more closely, but in interviews we were able to elicit more concrete concepts.

"None". A full 10% of survey respondents explicitly said they had no idea how search engines select and rank results. Face-to-face, however, the two interviewees who said this were able to suggest several ideas when pressed and when shown our example SERP. This suggests that 10% is likely an overestimate, and in fact very few people have no ideas whatsoever.

Summary. People who responded to our online survey volunteered relatively few concepts, with a full 10% claiming they had no ideas at all. However, it seems that this is an artifact of the online mode, and in face-to-face interviews we could elicit many more ideas. Indeed, it appears that non-expert searchers have a more comprehensive mental model, at least regarding selection, ranking, and layout, than suggested by earlier work [Hendry and Efthimiadis 2008; Holman 2011; Zhang 2008]. There are several possible reasons for this. First, we asked more focused questions and (unlike Hendry and Efthimiadis) included face-to-face interviews. Our use of a sample SERP as a prompt also elicited many more concepts. Second, this study is performed 10 years later than the previous study, and searchers might simply have more experience of web search—or less experience of alternatives—giving them more sophisticated mental models.

Some important concepts in search engine design were recognized by participants in both modes, including relevance, keyword matching, and popularity. It is perhaps surprising that these were reported by no more than 38% of survey respondents; it may be that they were "too obvious," or it may be that they are simply not well recognized. Our interviews suggest the former, even though interviewees had the benefit of a visual aid in the form of the simulated SERP. If we are to explain a search result, or a search engine, we may not need to explain these concepts except if we want to explicitly describe the tradeoffs between (for example) popularity and term matching.

Some concepts that have seen a lot of research or engineering effort, and which developers see as important, were barely mentioned. Just under 2% of survey respondents mentioned recency, for example, despite this being important across a range of topics and despite significant engineering efforts to update web-scale indexes quickly. Topical diversity has also been considered important for ranking [Clarke et al. 2009; Radlinski et al. 2008], and web search engines use techniques such as removing some results to ensure diversity of domain or promoting different viewpoints [Rosenberg 2018], but diversity of any kind was mentioned by only one interviewee and was mentioned by no survey respondents at all. Other important aspects that garnered few or no mentions include spam and malware removal, authority, mobile friendliness, and duplicated content. A search engine may want to explain these ideas and the impact these signals have on a SERP.

Finally, several people suggested that web search engines rank pages—not just advertisements—according to payment. This is not true of any current search engine, as far as we know, and explaining this may increase trust in individual engines and in the technology generally.

4.3 Types of Model

The great majority of explanations were at the level of mechanism rather than policy—that is, how things are done (counting clicks) rather than why they are done (promoting the results that are most likely useful). There was, however, very little mention of particulars of implementation, which is not surprising given our screening. Several interviewees offered the idea that selection or layout is driven by "algorithms," without offering more detail until they were prompted—as if this one, very vague, notion was somehow a useful explanation.

We looked for common clusters of answers, such as concepts that tended to correlate. Neither hierarchical clustering of people's answers, based on top-level concepts, nor a principal component analysis of a participant/concept matrix, indicated any significant patterns.

Although most survey respondents (and all interviewees) mentioned more than one factor in selection and ranking, very few expressly mentioned any sort of balance, or tradeoff between them, nor any cases where the factors may suggest different answers (2 of 11 interviewees, no survey responses). In a real search engine, of course, tradeoffs abound—for example, a new document is recent but cannot yet be popular—and to create a ranking is to balance many factors. It may be worth explaining some of these tradeoffs, the more so when search engine providers' choices make a big difference to the SERP.

4.4 Demographic Differences

Survey respondents were asked how long they had used search engines, and how often they used them. Both length of use and frequency of use correlated with the number of concepts each respondent offered, consistent with earlier work [Hendry and Efthimiadis 2008].

Length of use correlated somewhat with the number of concepts in the survey responses (approximately 0.04 concepts/year, $r = 0.16$, $p = .001$), although we observed an increase in concepts up to around 20 years of reported use and a slight dip thereafter. (Respondents claiming more than 25 years of use were counted at 25 years, and respondents claiming 0 years were removed for this analysis.) Respondents reporting more frequent searching also offered more concepts, with mean 2 concepts for those searching at least daily ($N = 371$), 1.5 for those searching every few days ($N = 41$), and 1.2 for those searching infrequently ($N = 6$). This difference was statistically significant (ANOVA $F(3, 415) = 2.71, p = .045$; Levene's test for homogeneity of variance $F(3, 415) = 0.63, p > .5$), although we must acknowledge that the near ubiquity of search means there are few people without much search experience or use. Although the number of interviewees is too small to draw any statistically significant conclusions, we did not observe a correlation with the length of use in interviews. No interviewees claimed less than daily searching.

4.5 Further Questions

At the end of each session, we asked interviewees for their opinion on search engines explaining more of the selection and layout process. One interviewee was uninterested (participant i15, "I don't think we care, to be honest . . . as long as it finds the information I want"), but reactions were generally favorable (e.g., participant i10, "definitely, I would like to see [it]"; i9, "for the general person it would be great"). One of the 11 interviewees thought it unnecessary for "professionals" such as themselves but thought it would be useful for the "inept," or for experts to track search engine changes over time. Two interviewees mentioned a desire to know where data (other than links) come from, so they could make their own decisions about quality.

Several interviewees suggested that having better explanations would help them understand a search engine—that is, improve their mental model—and therefore be a better, "smarter," or more efficient searcher either by learning about search features or by learning about term coverage.

A common theme across this discussion was interactivity, in two forms. First, interviewees saw explanations as a way to filter search results, either by media type or by genre, or as an opportunity to skip advertisements. Second, they discussed talking back to a search engine: by ignoring sources, for example, or by correcting the engine's assumptions about what is relevant in a particular query (e.g., participant i1, by saying that in fact location is not important; i3, "be nice if it would let you change the algorithm"). This corresponds to the "correctability" principle of Kulesza et al. 2015. Although there is a small amount of explanation in a current SERP—for example, the phrase "official page" or bolded keywords—these elements are not interactive and do not provide any opening for dialogue.

5 DISCUSSION AND IMPLICATIONS

Despite the ubiquity of web search, searchers are understood to have unsophisticated models of how search engines work. Explaining the result of a search, or explaining the policies that guide the algorithm, may help searchers form better mental models and thus be more efficient and effective.

5.1 Concepts

In this work, we did not focus on the form of explanations, rather asking *what* a search engine should explain at all. In particular, we ask this: what ranking concepts are already familiar and well understood? What is alien? What in searchers' mental models is missing or simply wrong?

Our online survey with almost 500 responses, and face-to-face interviews with a further 11 individuals, investigated what factors searchers think are involved in selecting and ranking objects on a SERP. By focusing on selection and ranking, and by using both online and in-person protocols, we observed many more concepts than expected and many more than observed in past work—257 distinct ideas among 508 people, in 16 major groups. Although there is a "long tail," with 108 concepts getting only a single mention, a few ideas were more frequent.

Common or familiar concepts. The most commonly mentioned concepts were popularity (mentioned by 38% of survey respondents and 91% of interviewees) and wording (33%/91%). These are in fact primary signals used by web search engines, and the effect of wording can be seen in, for example, bolding in the SERP. There is no such feedback for popularity, however, and tracking by search engines is not overt, so it is somewhat surprising to see this concept so well understood. Personalization is also fairly well understood (24%/82%), especially the notions of past search history and of tracking activity on the wider web.

Missing concepts. Some important factors were not mentioned, or mentioned very little. Before collecting responses, we identified six concepts we would like searchers to understand: relevance, popularity, recency, authority, locality, and diversity. For the most part, these concepts were not volunteered by our participants but were occasionally recognized in the context of an example SERP.

Relevance was mentioned by only 11% of survey respondents (although 82% of interviewees, perhaps because of our sample SERP in the latter case). Authority, an important factor and one that we expected would be prominent given media interest in "fake news" and bias, was mentioned by only 4% of online participants. "Official" pages were mentioned by 64% of interviewees, but this reflects the sample SERP and we saw little other mention of authority. Recency and especially diversity were also little mentioned.

Removal of malware, spam, copyright infringements, and similar were mentioned only once despite playing a major part in web search. This is not surprising: a searcher's mental model will of course cover what that have seen, and by design searchers do not see such harmful results. Similarly, very few people discussed the tradeoffs between factors (e.g., between popularity and recency).

Incorrect concepts. Finally, from our observations, it is clear that many searchers believe that web search engines accept payment for prominent listings, not just advertisements (at least 9% of survey responses and 46% of interviewees). We are not aware of any major search engine doing this in the past 20 years. Other misconceptions include engines tracking searchers' purchases, ranking or censoring according to political bias, and researching topics in advance.

5.2 Toward Explanations

The concepts that participants in our study commonly cited, or those that were missing or incorrectly cited, give obvious suggestions for what web search engines could usefully explain and what could perhaps be left out. For example, if a search engine provider believes that diversity is important, the provider may want to explain this. The provider may also want to explain authority and relevance (beyond keyword matching). Further, it may be important to explain that high(er)-rank positions in the search results are not for sale, and to explain that harmful or illegal content has been removed. In the latter cases, it is of course difficult to attach explanatory text to any one part of the SERP. Any such explanations should help the searcher identify the most applicable results with greater ease while providing transparency on why and under what criteria pages were ranked, where this can be explained understandably, concisely, and meaningfully.

We have been deliberately agnostic about what form any explanations might take, but we can make some observations based on our data. Zhang [2008] concluded that undergraduate students "formed mental models of search engines mainly based on system cues and feedback" rather than from explicit instruction. In our work, we see something similar. Interviewees were shown a concrete example of a SERP, which included an "official site," and 64% used the same phrase to explain it. More interviewees (82%) mentioned advertisements, also labeled on the SERP, and 91% mentioned keywords, on the SERP in boldface. This suggests that existing prompts do make a difference

Table 2. Phrasing Explanations: Examples of Objects, Attributes, and Processes That Search Systems Could Explain to a Searcher

	Policy: *why*, desiderata	**Instance: *why*, properties**
Goals:	“We want to be on topic”	“This seems to be about X”
Editorial Principles	“We want to be up-to-date”	“This page is recent”
	“We must remove illegal material”	“This page is illegal”
	…	…
	Mechanism: *how*, algorithms, data	**Detail: *what*, features, values**
Implementation:	“We look at keywords on the page”	“This page has X in the title Y times”
Models and Data	“We subtract X points for each day”	“This page is X days old”
	“We classify pages by text and links”	“This text has features X, Y, and Z”
	…	…
	Abstract: all SERPs or results.	Specific: a particular SERP or result.
	Can produce offline	Must produce online

“We” are the search providers.

and inform searchers’ mental models. Consistent with this, we saw evidence that more use, and more frequent use, of web search engines correlated with more concept mentions in our data. Additional prompts could help explain recency, or diversity, or other notions of authority; perhaps a search interface could even explain some of the tradeoffs made between different factors when generating the ranking.

In our interviews, participants volunteered two ways in which they might use search explanations beyond their own education: filtering results, especially according to media type or source, and correcting assumptions the search engine made. The elements that currently act as explanations—“ad” and “official site” markers, for example—are not interactive and do not provide much information on how to modify the ranking. Elements that were interactive, such as by letting searchers specify that certain ranking criteria were more or less important than the engine assumed, would invite a dialogue and perhaps allow more effective searching. This extends to alternative interfaces—for example, given a small screen or a spoken interface, with limited bandwidth and less scope for diverse results, explanations may be more important and (if they allowed interaction) could provide a type of conversation.

We could phrase explanations several ways (Table 2). Explanations of *policy* or *mechanism* are abstract and would explain how search works in general, without reference to a particular situation: for example, they could take the form of an overview of ranking and presentation [Google 2019, an explanation of mechanism]. They would also let search providers explain why certain actions have been taken, such as results being removed [Microsoft 2019, an explanation of policy] and other policy decisions such as (not) accepting payment in exchange for inclusion in the results. This lets search providers offer searchers an in-depth discussion of the tradeoffs associated with operating search engines—something perhaps not well understood—and since the explanation can be written in advance, it can be carefully crafted to be maximally accessible.

A policy or mechanism explanation does not, however, explain any particular search. An alternative would be explanations of each *instance* or each *detail*, which could call attention to particular results in context and describe their location, popularity, authority, and so on.[6] Explanations of this form could be generated on the fly to include descriptions of how a query was interpreted, for example, and (if the object were a whole SERP) could expose decisions around

[6] Kulesza et al. [2015] call these *in situ* explanations, although they do not distinguish explanations of instances from those of details.

balance and diversity of results. Specific explanations, either per result or per SERP, would be easier to discover and could be made interactive. They do, however, risk overwhelming the searcher; further, per-result explanations cannot, of course, explain what was removed from the SERP.

In our study, participants rarely offered instance-type and policy-type explanations—that is, they rarely offered explanations in terms of editorial principles—and explaining these broad principles may be useful. Implementation details are also likely to be proprietary and are often highly technical, fast-changing, and even difficult for experts to interpret. It seems likely that explanations in the policy or instance domains would be most useful, but this needs to be tested with real searchers.

5.3 A Note on Methods

In this work, we used a mixed methods design with both an online questionnaire and in-person, semistructured interviews. The combination was effective: running a survey online gave us a large number of responses, and with this large dataset we were more sensitive to little-used concepts. We also collected a large range of concepts. However, survey respondents offered few concepts each (including 10% who offered none), and by comparing with interview data we see this is likely to be a distorted account of their true models. Our interviews, although relatively few, relatively expensive, and more prone to experimenter effects, gave us a richer indication of participants' models and indicated the usefulness of existing explanations.

Limitations. We must acknowledge some limitations of our work. Most importantly, we cannot directly observe mental models: we have tried to elicit them here based on general questions (online) or a combination of general questions with particular prompts (face-to-face). Our interviews included a sample SERP that illustrated concepts we thought important, but the SERP could have looked different or illustrated different ideas. Since prompts seemed effective, we expect a different SERP would have led to different responses. However, with any reasonable SERP design, we would still expect participants to notice those explanatory elements (bolding, "official site") that are present and therefore would still recommend investigating more such elements.

Our participants were selected with an attempt to balance for age and gender, but we cannot rule out other sampling biases, and certainly all participants were English speakers from the United States. They were presumably much more familiar with the major U.S.-based search engines (Google, Bing) than the major engines in other markets (e.g., Yandex or Baidu), and our observations may not hold for users of the latter. Similarly, since we screened out professional users, we cannot draw any conclusions about their models; presumably, they would be more sophisticated. Finally, we must note the fairly small set of interviewees, and also note that because of this, and possible biases in coding, we need to run more studies before we can be completely sure of population/ecological validity (i.e., generalizability across people, search engines, and SERPs).

6 CONCLUSION

It is important to understand searchers' perspectives and build experiences that help them better understand search engine operation. In this article, we have studied people's understanding of how search engines select results to return and how they construct SERPs. The goal has been to inform our understanding of the concepts that searchers believe are behind these search engine operations so that we can better explain how search engines work and, in turn, make searchers more efficient and effective.

The findings show that there is a broad range of factors that searchers believe contribute to search engine responses: some that are expected (e.g., popularity, keyword match) and some that are surprising (e.g., paid inclusion in result lists). Searchers do seem influenced by the limited explanations currently available on a SERP, and it would be worth explaining further important concepts such as authority and recency, as well as correcting ideas of paid ranking. We should also see explanations as an invitation to dialogue. In particular, search engines may want to explain the following:

- *payment*, explaining that payment is not used in web ranking, to make search results appear more objective;
- *filtering* (spam, duplicates, or malware), explaining that the search engine has removed inappropriate content to help searchers understand what is missing and why;
- *recency*, displaying dates and/or times of content creation to help searchers make better selections for timely queries;
- *diversity*, flagging different opinions and perspectives to make search engines appear less biased; and
- *credibility*, flagging results from authoritative sources and/or clearly listing source information to boost searcher confidence in result quality.

There are many opportunities for future work. This includes extensions to the analysis. For example, we have only reported top-level concepts; other analysis might focus on smaller parts of the classification. Additional opportunities include studies to more deeply understand common misconceptions such as paid promotion in result rankings, as well as developing and evaluating methods to correct these misconceptions.

Devising explanations, and surfacing them on a SERP, requires considerable attention given the complexity of ranking algorithms and the need for any explanations to be intuitive and easily understood. Explanations could also serve as an entry point into ways for searchers to adjust the results they receive, such as via recourse links or filters, and for searchers to indicate which factors are important to them, for use in personalization and to offer feedback to search providers.

APPENDIX

A PROTOCOL

The protocol for the online survey and interviews were approved by the ethical review panel at Microsoft Research and AI.

A.1 Questions in the Online Survey

The online survey was administered in late August 2018, after a pilot to test the questions and the feasibility of coding responses. Recruits were screened with two questions:

Screener 1. Do you develop or support IT systems, or do you have a qualification in information technology?

Screener 2. Do you work in a library, or do you have a qualification in library studies?

Each screener allowed the options “I develop or support IT systems” (respectively, “I work in a library”), “I have a qualification in information technology” (respectively, “library studies”), or “neither.” Only recruits who answered “neither” to both questions were retained, and after screening we had 497 respondents. Besides basic demographics, they were each asked two questions:

Q1. When you use Google, Bing, or another search engine, they take your query and respond with a set of results. They’re mostly links to other web pages, but there are also maps, ads, videos, calculators, definitions, and all sorts of other things. These things are deliberately chosen out of hundreds of billions of possibilities.

What do you think determines what gets shown on the results page? How do search engines choose out of all those billions of individual things?

One way of thinking of it is to imagine printing every web page there is, and also printing maps, advertisements, pictures, calculations, and all the other things a search engine knows about. If all these printouts were in a big basket, what criteria would Google or Bing use to choose which to pick out for you?

Please use the space below to list as many aspects as you can think of. Feel free to use examples if that helps (but please don't include anything personal).

Q2. As well as choosing things to show, search engines put them in different places on the results page: near the top, in the middle, down the bottom. Sometimes they also put things on the left or on the right of the screen. Of course, everything ends up being above, below, or beside something else.

Again, we are interested in how you think search engines do this.

What do you think determines what goes where on the results page?

Please use the space below to list as many aspects as you can think of. Feel free to use examples if that helps (but please don't include anything personal).

A.2 Questions in Interviews

Interviews were based around three open questions and a set of more focused followup questions. Parts of the script are included in the following. The exact wording varied for each interview, according to what had been discussed. The introductory section served as a warm-up, and to elicit examples for later in the interview, so we have not analyzed the responses here. Again, a pilot phase tested the instruments and the feasibility of coding.

Introduction. We are interested in talking about how you think search engines work. It doesn't matter whether you're right or wrong, I'm interested in what you really think.

To get started, can you tell me a bit about what you do with search engines like Google or Bing? For example, do you tend to search for work or for your own interest? Do you read everything or just grab an answer quickly? What sort of things do you type or say? What sort of device do you normally use?

Part 1. When you give a query to Google, Bing, or another search engine, they respond with a set of things. These things are deliberately chosen out of hundreds of billions of possibilities. They're mostly links to other web pages, but there are also maps, ads, videos, calculators, definitions, and all sorts of other things. What do you think determines what gets shown on the screen? How do they choose out of all those billions?

(Let them articulate as many or as few as they can; stop when they say they're out, or after a long gap, e.g., a minute with no further contribution. In case of a general statement, e.g., "looks useful," push for particular criteria: "how so?" or "how could the search engine tell?")

Part 2. Search engines also put things in different places on the results page: near the top, in the middle, down the bottom. Sometimes they also put things on the left or the right. How do you think they choose what goes where on the results page?

(Let them articulate as many or as few as they can, etc., as previously.)

Part 3. I'm visiting here from Australia, and I was wondering whether there was any sports to see. So I typed in "sounders" to see if they're playing or doing anything interesting. This is the result I got. *(Show printout.)*

For each labeled result: Can you describe this result? What does it offer? Why do you think this result in particular was shown? Why do you think it is where it is on the page [compared to . . .]?

Part 4. What you've just done is explain to me why this page is the way it is. In the future, search engines might be able to explain their choices like you just did. Thinking about what we've just seen but also about the web searches you do, what do you think of this idea? What do you think you'd want from these explanations? What should the search engines explain exactly? How should the explanations be shown or how should they be phrased? If there were explanations such as "this result is here because *(pick one of their concepts)*," what would you want to do in response? Would a different page organization be useful to help achieving more transparent search results? If so, what could help?

The SERP referred to in part 3 is reproduced in Figure 1. Black boxes, which were not in the SERP shown to participants, mark the results we used for discussion. These were chosen to represent recency (scores), majority intents (scores), "official" pages (Sounders FC), authority (Seattle Times), topical diversity and location (Bellevue Boat Company, a fictional company selling depth sounders), and advertising (CenturyLink Field).

ACKNOWLEDGMENTS

We thank our participants for their time, and Microsoft Research Operations for their help in arranging the survey and interviews. We also thank our anonymous reviewers for their many useful suggestions.

REFERENCES

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 582.

ACM U.S. Public Policy Council. 2017. Statement on Algorithmic Transparency and Accountability. Retrieved November 19, 2019 from https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Eugene Agichtein, Eric Brill, and Susan Dumais. 2006. Improving web search ranking by incorporating user behavior information. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 19–26.

Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the Conference on Human-Computer Interaction with Mobile Devices and Services*. 9–14.

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM* 61, 6 (2018), 54–61.

Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and Seyed M. M. Tahaghoghi. 2010. Evaluating search systems using result page context. In *Proceedings of the Symposium on Information Interaction in Context*. 105–114.

Nicholas J. Belkin. 1988. On the nature and function of explanation in intelligent information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 135–145.

Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction* 16, 2–4 (2001), 193–212.

Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz. 2011. Inferring and using location metadata to personalize web search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 135–144.

Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short-and long-term behavior on search personalization. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 185–194.

Christine L. Borgman. 1986. The user's mental model of an information retrieval system: An experiment on a prototype online catalog. *International Journal of Man-Machine Studies* 24, 1 (1986), 47–64.

Dimitrios Bountouridis, Monica Marrero, Nava Tintarev, and Claudia Hauff. 2018. Explaining credibility in news articles using cross-referencing. In *Proceedings of the 1st International Workshop on ExplainAble Recommendation and Search (EARS'18)*.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the International Conference on Healthcare Informatics*. 160–169.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1721–1730.

Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C. Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable recommendation by leveraging reviews and images. *ACM Transactions on Information Systems* 37, 2 (2019), 16.

Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The influence of caption features on clickthrough patterns in web search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 135–142.

Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2009. Overview of the TREC 2009 web track. In *Proceedings of the Text REtrieval Conference (NIST special publication SP 500-278)*.

Juliet Corbin and Anselm Strauss. 2012. Analyzing data for concepts. In *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd ed.). Sage, Thousand Oaks, CA.

John W. Creswell. 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage, Thousand Oaks, CA.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Efthimis N. Efthimiadis, David G. Hendry, Pamela Savage-Knepshield, Carol Tenopir, and Peiling Wang. 2004. Mental models of information retrieval systems. *Proceedings of the American Society for Information Science and Technology* 41, 1 (2004), 580–581.

Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *Proceedings of the ACM IUI Conference on Intelligent User Interfaces*. 211–223.

Motahhare Eslami, Karrie Karahalios Karrie, Christian Sandvig, Kristan Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I "like" it, then I hide it: Folk theories of social feeds. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 2371–2382.

Luanne Freund. 2008. *Exploring Task-Document Relations in Support of Information Retrieval in the Workplace*. Ph.D. Dissertation. University of Toronto.

Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the ACM IUI Conference on Intelligent User Interfaces*. 227–236.

Google. 2019. How Google search works. Retrieved November 19, 2019 from https://www.google.com/search/howsearchworks/.

David G. Hendry and Efthimis N. Efthimiadis. 2008. Conceptual models for search engines. In *Web Searching: Multidisciplinary Perspectives*, A. Spink and M. Zimmer (Eds.). Springer, 277–307.

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.

Lucy Holman. 2011. Millennial students' mental models of search: Implications for academic librarians and database developers. *Journal of Academic Librarianship* 37, 1 (2011), 19–27.

Bernard J. Jansen and Paulo R. Molina. 2006. The effectiveness of web search engines for retrieving relevant ecommerce links. *Information Processing and Management* 42, 4 (2006), 1075–1098.

Philip Nicholas Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive Science* 4, 1 (1980), 71–115.

Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. Simple rules for complex decisions. arXiv:1702.04690v3.

René F. Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 2390–2395.

Jürgen Koenemann and Nicholas J. Belkin. 1996. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 205–212.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the ACM IUI Conference on Intelligent User Interfaces*. 126–137.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1–10.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable and explorable approximations of black box models. arXiv:1707.01154.

John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

Dirk Lewandowski. 2012. Credibility in web search engines. In *Online Credibility and Digital Ethos: Evaluating Computer-Mediated Communication*, S. Apostel and M. Folk (Eds.). IGI Global, Hershey, PA, 131–147.

Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 30.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 623–631.

David McSherry. 2005. Explanation in recommender systems. *Artificial Intelligence Review* 24, 2 (2005), 179–197.

Donald Michie. 1988. Machine learning in the next five years. In *Proceedings of the European Working Session on Learning*. 107–122.

Microsoft. 2019. How Bing Delivers Search Results. Retrieved November 19, 2019 from http://help.bing.microsoft.com/#apex/18/en-US/10016/0.

Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. arXiv:1706.07269.

Alexander Moore, Vanessa Murdock, Yaxiong Cai, and Kristine Jones. 2018. Transparent tree ensembles. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 1241–1244.

Jack Muramatsu and Wanda Pratt. 2001. Transparent queries: Investigation users' mental models of search engines. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 217–224.

Satoshi Nakamura, Shinji Konishi, Adam Jatowt, Hiroaki Ohshima, Hiroyuki Kondo, Taro Tezuka, Satoshi Oyama, and Katsumi Tanaka. 2007. Trustworthiness analysis of web search results. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries*. 38–49.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. arXiv:1802.00682.

Donald A. Norman. 1983. Some observations on mental models. In *Mental Models*, D. Gentner and A. L. Stevens (Eds.). Lawrence Erlbaum Associates, Mahwah, NJ, 15–22.

Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3–5 (2017), 393–444.

Michael J. Paul, Ryen W. White, and Eric Horvitz. 2015. Diagnoses, decisions, and outcomes: Web search as decision support for cancer. In *Proceedings of the International Conference on World Wide Web*. 831–841.

Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L. A. Clarke. 2017. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR International Conference on the Theory of Information Retrieval*. 209–216.

Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the ACM IUI Conference on Intelligent User Interfaces*. 93–100.

Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 103.

Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. 2008. Learning diverse rankings with multi-armed bandits. In *Proceedings of the International Conference on Machine Learning*. 784–791.

Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

Mir Rosenberg. 2018. Toward a More Intelligent Search: Bing Multi-Perspective Answers. Retrieved November 19, 2019 from https://blogs.bing.com/search-quality-insights/february-2018/Toward-a-More-Intelligent-Search-Bing-Multi-Perspective-Answers.

David-Hillel Ruben. 2015. *Explaining Explanation*. Routledge.

Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 2006. Visualization of uncertainty in context aware mobile applications. In *Proceedings of the Conference on Human-Computer Interaction with Mobile Devices and Services*. 247–250.

Margarete Sandelowski. 2000. Combining qualitative and quantitative sampling, data collection, and analysis techniques in mixed-method studies. *Research in Nursing and Health* 23 (2000), 246–255.

Constantine Sandis. 2011. *The Things We Do and Why We Do Them*. Springer.

Tefko Saracevic. 2007a. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behaviour and effects of relevance. *Journal of the Association for Information Science and Technology* 58, 13 (2007), 2126–2144.

Tefko Saracevic. 2007b. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the Association for Information Science and Technology* 58, 13 (2007), 1915–1933.

Sebastian Schultheiß, Sebastian Sünkler, and Dirk Lewandowski. 2018. We still trust in Google, but less than 10 years ago: An eye-tracking study. *Information Research* 23, 3 (Sept. 2018), 1–13.

Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1245–1254.

Search Engine Land. 2018a. Google Updates Its Search Quality Rating Guidelines. Retrieved November 19, 2019 from https://searchengineland.com/google-updates-its-search-quality-rating-guidelines-302553.

Search Engine Land. 2018b. The Periodic Table of SEO Success Factors. Retrieved September 1, 2018 from https://searchengineland.com/seotable.

Jaspreet Singh and Avishek Anand. 2018. Posthoc interpretability of learning to rank models using secondary training data. In *Proceedings of the 1st International Workshop on ExplainAble Recommendation and Search (EARS'18)*.

Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable search using local model agnostic interpretability. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. 770–773.

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.

Mohammad S. Sorower. 2010. *A Literature Survey on Algorithms for Multi-Label Learning*. Technical Report. Department of Computer Science, Oregon State University.

Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the Workshop on Human-in-the-Loop Data Analytics*. 6.

Charles Teddlie and Abbas Tashakkori. 2009. *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Sage, Thousand Oaks, CA.

Maartje ter Hoeve, Mathieu Heruer, Daan Odijk, Anne Schuth, Martijn Spitters, and Maarten de Rijke. 2017. Do news consumers want explanations for personalized news rankings? In *Proceedings of the FATREC Workshop on Responsible Recommendation at RecSys 2017*.

Nava Tintarev and Judith Masthoff. 2011. Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook*. Springer, 479–510.

Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 31–40.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining recommendations using tags. In *Proceedings of the ACM IUI Conference on Intelligent User Interfaces*. 47–56.

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 601.

Jim Waterson and Edward Helmore. 2018. Trump accuses Google of promoting Obama's speeches over his. *The Guardian*. Retrieved November 19, 2019 from https://www.theguardian.com/us-news/2018/aug/28/donald-trump-google-news-service-is-rigged-against-me.

Oskar Wenneling. 2007. Seamful design—The other way around. In *Proceedings of the Scandinavian Student Interaction Design Research Conference*.

Ryen W. White. 2013. Beliefs and biases in web search. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. 3–12.

Ryen W. White and Ahmed Hassan. 2014. Content bias in online health search. *ACM Transactions on the Web* 8, 4 (2014), 25.

Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. 1235–1244.

Yan Zhang. 2008. Undergraduate students' mental models of the web as an information retrieval system. *Journal of the Association for Information Science and Technology* 59, 13 (2008), 2087–2098.

Received June 2019; revised September 2019; accepted November 2019