Mesh: Scaffolding Comparison Tables for Online Decision Making

Joseph Chee Chang **Nathan Hahn** **Aniket Kittur**
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{josephcc, nhahn, nkittur}@cs.cmu.edu

ABSTRACT
While there is an enormous amount of information online for making decisions such as choosing a product, restaurant, or school, it can be costly for users to synthesize that information into confident decisions. Information for users' many different criteria needs to be gathered from many different sources into a structure where they can be compared and contrasted. The usefulness of each criterion for differentiating potential options can be opaque to users, and evidence such as reviews may be subjective and conflicting, requiring users to interpret each under their personal context. We introduce Mesh, which scaffolds users to iteratively build up a better understanding of both their criteria and options by evaluating evidence gathered across sources in the context of consumer decision-making. Mesh bridges the gap between decision support systems that typically have rigid structures and the fluid and dynamic process of exploratory search, changing the cost structure to provide increasing payoffs with greater user investment. Our lab and field deployment studies found evidence that Mesh significantly reduces the costs of gathering and evaluating evidence and scaffolds decision-making through personalized criteria enabling users to gain deeper insights from data.

Author Keywords
sensemaking, search, note-taking, ecommerce

CCS Concepts
•**Information systems** → **Search interfaces;** •**Human-centered computing** → *Graphical user interfaces; Web-based interaction;*

INTRODUCTION
Whether figuring out which products to purchase or where to eat in an unfamiliar city, consumers today have instant access online to enormous amounts of information on which to base their decisions. Research in consumer behavior has found online information such reviews to be a major factor for online research [16, 35], with the potential to help consumers make informed decisions about how well each option satisfies their various criteria [15]. For example, a coffee drinker looking to buy a new espresso machine might read reviews aiming to evaluate how easy it is to use for a novice barista, how well it steams milk, how likely it is to break down, and so on.

However, users can also be overwhelmed by the number of potential options, the criteria they should use to compare those options, and the number of information sources to collect evidence from [45, 42]. For example, the electronics section of Amazon alone contained more than 1.3 million reviews in 2013 [34], and Yelp has accumulated more than 200 million reviews [47]. Such online reviews can be conflicting, biased, subjective and scattered across many sources [20, 38, 50, 11], requiring users to evaluate and interpret each piece of evidence based on their personal context [43]. The highly bimodal skew of review ratings can lead to compression of ratings in a narrow band [22], and the increasing number of fake reviews (which now may be in the majority for some categories such as electronics and beauty [1]) means that solely relying on automatic aggregation such as averaged ratings or summarization can be inaccurate or uninformative. Automated approaches to addressing these issues, such as aspect extraction [31, 49], review summarization [21, 28], and direct recommendation [6], can be insufficient due to the long tail of usage contexts [4], the need for nuanced contextualization when reading reviews [8], and the challenge of discovering and learning new criteria along the way [27].

Consumers doing this task manually must go through the various reviews and sources, pulling together scattered information, learning about what criteria are useful for picking or ruling out options, evaluating evidence on those criteria, keeping track of their judgments, and weighing them depending on what's most important to make a final decision. To assist with the process, consumers utilize techniques such as building comparison tables with spreadsheets or notepads. However, transferring information between information sources and spreadsheets or notepads can be prohibitively time-consuming [3]. Furthermore, as a user encounters and adds new options, they must gather information for each of their criteria in the table in order to evaluate that feature. Similarly, encountering and adding new criteria requires gathering information for all previously added options. This iterative construction is common in unfamiliar domains [32] and creates an increasing cost the more options and criteria are added to the table.

UIST '20, October 20–23, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7514-6/20/10 ...$15.00.
http://dx.doi.org/10.1145/3379337.3415865

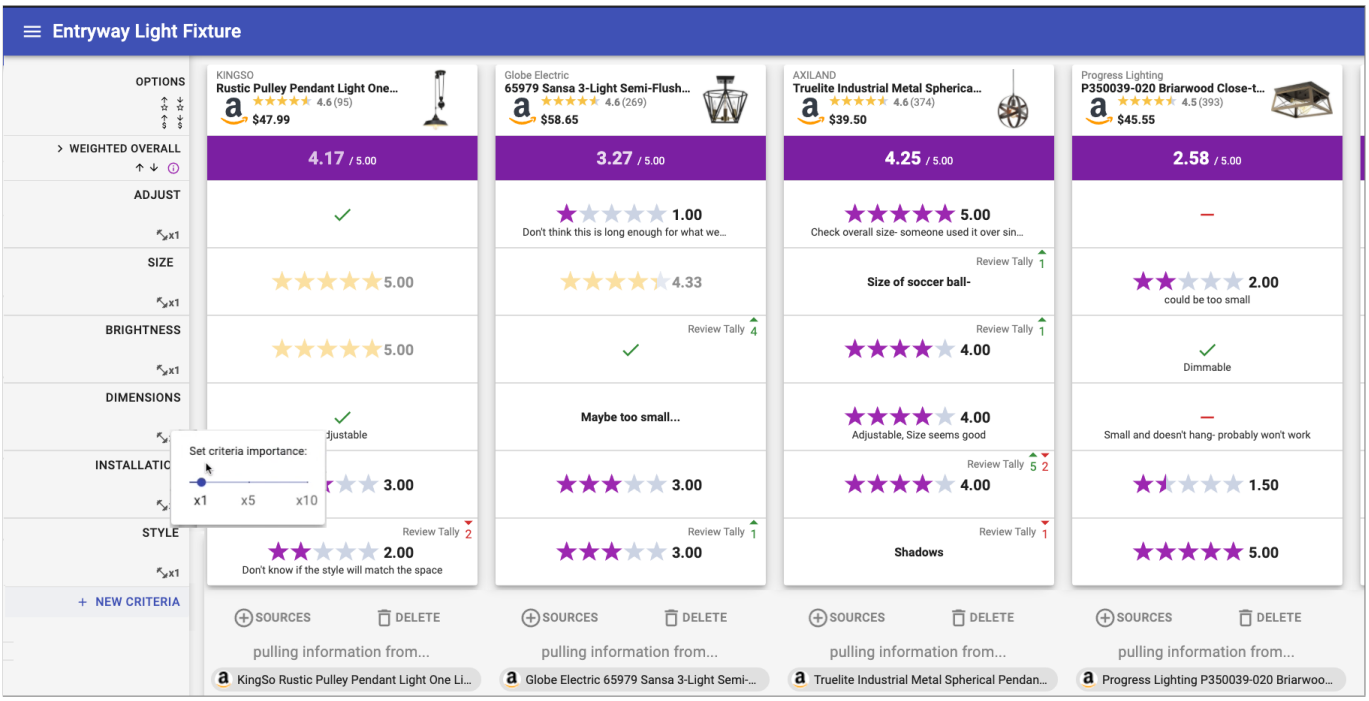

Figure 1. The Table View. Users can create Option Columns by importing Amazon project pages opened in their browser tabs and create Criteria rows to see the average review ratings that mentioned each criterion across their options (in yellow). To explore the reviews more deeply, users can click on the criteria to see the Evidence View (shown in Figure 3), where users can overwrite the default Amazon ratings with their own (in purple) based on their own interpretation of data. To prioritize the criteria, users can also adjust the weight to see a weighted average rating across their criteria for each option. This image is an actual project made by P5 in the field deployment study.

Instead of fully automated or manual approaches, we introduce Mesh, a hybrid approach aimed at scaffolding decision making by helping users progressively build up a comparison table that reflects their personal criteria and evaluation of evidence. Mesh lowers the cost of pulling in information, organizing it by users' criteria, and helping users keep track of their judgments as they evaluate evidence. Importantly, by auto-filling the cells when new criteria or options are added throughout the process, Mesh makes adding to the table stay at a constant cost as the table grows, changing the cost structure to provide an increasing payoff with greater user investment. Finally, Mesh helps keep users on track by prioritizing where to look, which criteria are most important, and reflecting their current beliefs for each option through an overall weighted average.

We evaluated Mesh through three user studies. In the first study we found evidence that Mesh lowered interaction costs and allowed participants to find answers to objective criteria (such as the *size* and *capacity* of coffee machines) significantly faster and more accurately. In the second study we found similar benefits for subjective criteria (such as *ease of use*) which required additional interpretation of online evidence, resulting in learning summaries rated as more insightful and confident when compared to baseline participants using Google Spreadsheets to conduct the same task. Finally, a field deployment evaluated real-world usage in a week-long study, finding that Mesh increased user satisfaction, confidence and efficiency with actual purchasing decisions.

RELATED WORK

Research in consumer behavior has pointed out numerous difficulties users face when using online evidence to support making purchase decisions. One major challenge is that online evidence, such as consumer or expert reviews, can be messy, subjective, and biased [35, 1]. Furthermore, users may need to go through each piece of evidence in order to interpret them based on their own personal context and unique goals. This process is an important factor in purchase decision making [16, 35], but can incur high cognitive costs as the user tries to keep track of their interpretation of different pieces of evidence [7]. Another challenge is that online evidence is often scattered across many sources due to the distributed nature of the Web. This includes product listing pages on e-commerce platforms, blog and forum posts, and consumer and expert reviews. On the one hand, having multiple information sources can help users to determine the credibility of online evidence [20, 38, 50, 11, 9, 17]. However, cross-referencing multiple sources can be burdensome and costly [37, 33, 48, 5, 18].

Another thread of research has focused on building interactive interfaces that aim to support decision making under multi-criteria and multi-option scenarios, such as faceted interfaces [19, 41] and table-based decision support and visualization systems [14, 39, 46, 29]. While these approaches allow consumers to narrow down their options efficiently by navigating to different subsets of a larger collection or investigate trade-offs through visualizations, the majority of these approaches rely on pre-compiled metadata or require users to manually clip evidence for each source. As a result, they do not support

criteria that require close examination of a large amount of subjective evidence (such as reviews) which are not in the form of structured metadata. For example, to get a sense of how durable an option is a consumer would evaluate many unstructured reviews describing whether and how an item held up over time. In two studies closely related to our work, Chen et al. [13, 12] allowed users to build comparison tables for camera products by allowing them to pick from a list of precompiled common camera criteria and used sentiment analysis of relevant reviews as summaries across different options. While Mesh also allows users to build comparison tables with their own options and criteria, it enables users to use arbitrary search terms as their criteria instead of selecting from a pre-compiled fixed list, allowing it to support the long tail distribution of user needs [4]. Even more importantly, Mesh focuses on helping individuals interpret reviews under their own personal context, and overwrite the summaries generated by the system to better reflect their own views of data. This approach not only provides better support for personal context but can also allow users to recover from errors made by automated summarization approaches.

Instead of automating away the role of the user, our approach focuses on helping users scaffold their decision-making throughout the process, maximizing their ability to apply their personal context and interpretation to evidence while reducing the costs for doing so. This view unlocks a design space in which the interface supports the human in discovering and sharpening their own understanding of what criteria are important to them in the context of the options and evidence available to them; keeping track of their evaluations of that evidence for them; enabling the human to prioritize their attention to the most discriminative evidence; capturing human perceptions of value; and using those perceptions to drive a final decision that integrates values across their personal criteria. At a high level, our work aims to bridge the gap between decision support research in the literature above (which helps people make decisions by imposing a high degree of structure based on metadata or through computation) and the sensemaking process in which users are learning about unknown unknowns to develop personalized context from unstructured data [40, 32, 27, 7, 8].

EXPLORATORY INTERVIEWS AND DESIGN GOALS

To discover common limitations and needs of online product research, we conducted preliminary interviews to inform our design goals. Thirty participants were recruited (age: 3% 19-24, 20% 25-34, 33% 35-40, 23% 41-54, 20% 55+; 22 female, 7 male, and 1 not listed) through posts on social media including Facebook, Twitter and Nextdoor, and interviewed for 60 minutes each. Prior to the interviews, we generated 10 interface design mock-ups addressing various potential issues discussed in the previous sections, ranging from managing information sources to collecting evidence for purchasing decisions (we discuss these design probes in the context of our findings below). During the interviews, we walked through each of the design mock-ups and used them as probes to see how strongly participants identified with the issues they tried to address, as well as how they reacted to the designs. We list below three of the most commonly recurring themes.

Comparing Options with Scattered Evidence

The most common theme mentioned by all participants was the difficulty of managing an overwhelming number of information sources and the amount of evidence scattered across them. Specifically, they pointed to how evidence for options needs to be collected across different webpages, leading to a *stressful number of opened browser tabs* of e-commerce websites (such as Amazon) and expert review websites (such as CNET reviews). When comparing options, participants were especially frustrated by the high interaction cost of *switching back and forth between tabs to compare options on a metric* [criteria] and that it is *not easy to search for [information that mentioned] specific terms across all products*.

Need for Personal Interpretation of Evidence

Consistent with prior work, we also found reading reviews to be a major factor when making purchase decisions [16, 35]. While participants felt overwhelmed by the amount of evidence they needed to process in order to confidently make purchase decisions, they were unenthusiastic about designs centered around automating the process. For example, one design had users answer questions about their preferences and provide personalized product recommendations. Participants were reluctant to trust the output of the automated system, but instead saw it as a way to *get some ideas or guidelines about things they should consider*; in other words, they saw it as an additional source for collecting potential options to conduct further comparisons. Participants further emphasized the importance of seeing raw evidence and making their own judgments such as *reading through reviews to generate a summary of their own opinion*. Participants were enthusiastic about features that would support this process, such as allowing them to easily rate and tally reviews as positive or negative or making a summary rating from reading multiple reviews.

Scaffolding Decision Making

Participants pointed to difficulties in keeping track of their overall research, describing their process as "*erratic*" causing them to "*go down many rabbit holes*" and "*get lost in the weeds*." One central reason cited was the need to constantly make small and personalized judgments throughout, such as interpreting how relevant a review is to their contexts, summarizing how a product fits a criterion, or deciding to keep or rule out an option. Participants were frustrated when "*Sometimes I can't remember why a [product] page was kept opened and had to reread the content*." For this, participants use spreadsheets, scratchpads, and physical notebooks *when things start to get out of hand*, but also pointed to how this process is cumbersome and only used *as a last resort on important purchases*. When asked about the types of information they would typically save, participants described a mix of factual findings (such as product specifications) and their own interpretation of subjective evidence (such as ease of use as described in the reviews). Participants were enthusiastic about designs that would scaffold them in working in a more organized fashion, such as making a comparison table of options they are considering and and being able to compare options side-by-side and ranking them according to their own criteria.

Based on the above, we formulated the following design goals:

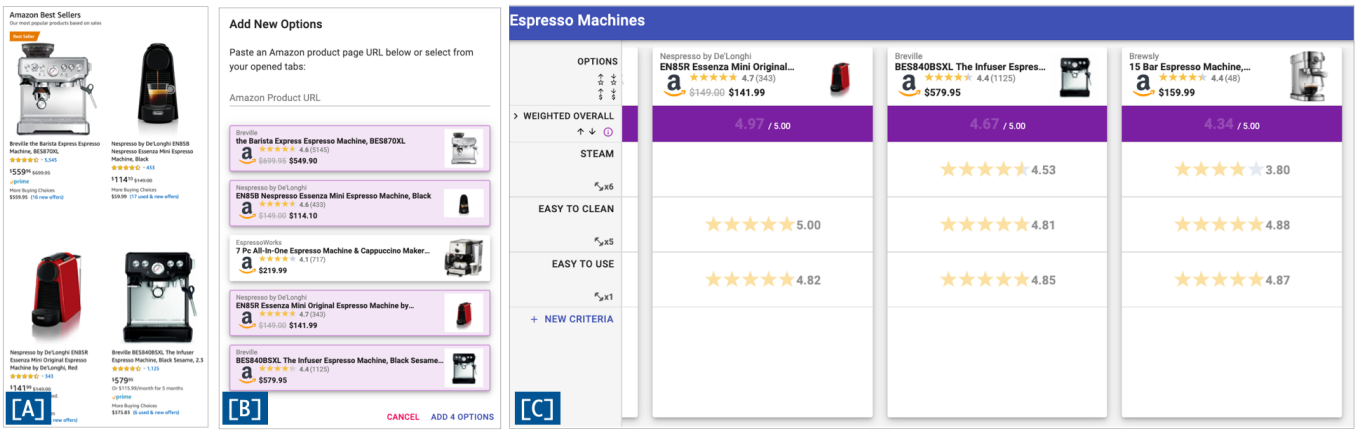

Figure 2. Many products on Amazon are highly rated with thousands of views and it can be difficult for users to differentiate them [A]. Users can open them in browser tabs and import them into Mesh to keep track of them [B]. Mesh automatically fetches reviews relevant to different user criteria for each option to help characterize them [C]. Users can uncover meaningful discrepancies between options based on their own criteria. For example, here seeing a larger difference in the "Steam" criteria, with the first option that lacks this feature returning no reviews that mentioned "Steam".

- **[D1]** Minimize effort of comparing evidence for the same criteria across different options
- **[D2]** Allow users to make their own interpretation and summaries of data
- **[D3]** Capture user decisions about options and criteria throughout the process in an organized way

SYSTEM DESIGN

Motivated by the design goals uncovered by our exploratory interviews, we developed Mesh to provide a more organized way to conduct research by allowing users to iteratively build up a product comparison table with their own options and criteria. In a standard spreadsheet, people have to start with a blank table and switch back and forth between information sources to fill out everything manually. In contrast, our system provides an increasing payoff for every criterion and option added by connecting each cell in the table with relevant product information and reviews and summarizing them. One challenge here is that automation and auto-summarizing content go against users' desire for personal interpretation; instead, we carefully constructed interactions that allowed users to both deeply explore the raw evidence and adjust their tables when auto-summarization does not fit their own interpretation of the data. To support this, Mesh was designed to capture users' judgments about data throughout their process with little added effort using light-weight interactions at different levels of granularity. For example, flagging a review as positive or negative after reading it, rating different options based on the same criterion, or sorting different options based on the ratings of different criteria. Altogether the system is designed to feel like scaffolding: helping users gain deeper insights from scattered evidence more efficiently, and capturing their own judgments on data in a structured way.

Example User Experience

Consider an example in which a user wants to purchase an espresso machine for the first time to use in her apartment. She starts by searching on Amazon for popular options to consider, but sees that they all have more than 1000 reviews with average review scores between 4.4 and 4.7, making it difficult for her to discriminate between them (Figure 2 [A]). To understand which is best for her she needs to deeply explore the reviews to see which are *easy to clean*, *compact*, *has great steam for making cappuccinos*, and *don't require a lot of cleaning* – a process that would typically take her hours. Using Mesh she creates a new project and imports the options she had opened from a list Amazon product pages open in her browser tabs (Figure 2 [B]). The system then creates columns for each option and automatically pulls in basic product information such as prices, images, and titles (Figure 2 [C]). She then adds her criteria to the system as rows by clicking on the "+ New Criteria" button, with the system automatically fetching a sample of reviews for each product the newly added criterion and displays their average rating (Figure 2 [C]).

She sees that despite the overall rating being indistinguishable between her options, there are large discrepancies in review ratings for "steam". She clicks on it to see reviews mentioning "steam" for all her products in the Evidence View (Figure 3), including one that had no matching reviews (Figure 2 [C]). Clicking on the image icon of that model to see a full-screen carousel containing multiple larger images, she realizes it does not support steaming milk, allowing her to remove it from her project. As she reads reviews of the remaining options and evaluates how well each meets her goal, it takes her little extra effort to tally that review as positive or negative, reducing her working memory load. Doing so she quickly builds up her judgment for each option, and replaces the average Amazon rating with her own when it does not reflect her view. She iteratively adds her other criteria, the system auto-filling each of them for all her existing options, and finds and adds more options, the system auto-filling all their criteria as well.

As more criteria and options are added, she can scroll vertically to see her own notes, ratings, and review tallies about different criteria, and scroll horizontally to see her different options (Figure 1). To help her compare and contrast she drag and drops to reorder her criteria and options and sorts her options based on their values for a criterion to prioritize them. Finally,

after developing a good understanding of what criteria are important to her goals and discriminative across her options, she changes the weights of her criteria so that the system produces overall scores that reflect her personal opinions and goals in the Table View (Figure 1).

[D1] Comparing Evidence across Options and Sources

As reflected in the scenario above, our first design goal was to lower the costs of managing many information sources and examining evidence scattered across them. A fundamental problem we identified was that users often need to compare evidence for a criterion across their different options, but the evidence was typically organized by options and scattered across sources. For example, a user may need to go through multiple Amazon product pages and CNET reviews to get a sense of how different espresso machines were suitable for novices. One way users currently deal with this is by switching back and forth between browser tabs and searching for relevant evidence on each page; another is to focus on one product at a time and try to remember information from other sources to compare them. Both of these strategies can incur high interaction and cognitive costs. As a result, our exploratory interviews found participants had difficulties in keeping track of previous decisions such as which options they were considering, why they had considered each in the first place, and their criteria for comparing them.

To scaffold this process, Mesh allows users to progressively build out a product comparison table to keep track of their options, sources, and criteria. To keep track of their options and sources, a user can import their browser tabs into Mesh and group the sources into Option Columns in Mesh (See Figure 1 for the Table View). For example, a user could create an option column with an Amazon product page grouped with an expert review article from CNET.com for the same product and its product specification page from the manufacturer's website. In the backend, Mesh populates the header of each column with product names, prices, images, and review ratings from Amazon. To keep track of their different criteria, a user can create a set of Criteria Rows (Figure 1). When a criterion is added, for each option Mesh fetches 60 Amazon reviews by via Amazon's review search end-points as well as sentences in the product description and imported sources that mention the criteria as evidence. Users can click on each row to see all the evidence for their options on that criteria side-by-side for comparison in the Evidence View, reducing the high cost of switching between information sources (Figure 1). Longer reviews are by default collapsed to the three sentences surrounding where the criteria name was mentioned so users can stay focused on the current criteria, but can be expanded when needed for additional context.

By default, Mesh shows the average rating of the 60 Amazon reviews as cell values in the Table View. Our rationale for presenting criteria-specific ratings was to provide users with instant feedback and benefit for externalizing their criteria, which would enable two novel interactions: 1) getting a quick overview of how existing options differ or how a new option compares to existing options and 2) comparing how discriminative their different criteria are for their current options. These have the potential of allowing participants to better prioritize their investigation efforts. One major challenge here is that while the reviews did mention the criteria, they can often be noisy and include comments on things other than the criteria users were focused on.

[D2] Interpreting Evidence based on Personal Context

Both our exploratory interviews and prior work pointed to an important need for users to interpret evidence based on their own personal context [43]. This personalized interpretation of online data could also happen frequently throughout the research process – for example, judging how relevant a review was to user's personal context, users' summative perceptions after reading multiple reviews about a criterion, and how users characterized each option. Mesh addresses this by providing a set of light-weight interactions to capture users' interpretation of data, and reflect them back onto the Table View. Using the Evidence View, where evidence about a criterion is presented side-by-side for each option, users can externalize their interpretation of data at different levels of granularity using interactions that require little cognitive effort. For example, after examining a review, it only requires one click for users to label it as positive or negative using the buttons at the end of each review. As users rate the reviews, Mesh automatically creates a tally of positive and negative reviews for each option, providing immediate payoff to the users for labeling them and reducing working memory load. After examining reviews about a criterion for an option, users can leave the average Amazon rating alone if it matches their own perceived rating, or overwrite it with their own rating (color-coded in purple instead of yellow). This approach aims to reduce the cost of rating to zero when the default ratings generated by the system matches users' own judgments. In addition, users can externalize more nuanced mental context through notes, which are shown in the Table View. Based on user feedback, Mesh also enables users to use check marks and minuses (Figure 1) for criteria that have binary values (e.g, does the espresso machine come with a steam wand).

[D3] Scaffolding Decision Making

As a user iteratively builds up a better understanding of their options and criteria, they gradually progress from investigating and interpreting evidence to making a decision between their options. However, participants in the exploratory interviews described spending redundant effort when they lost track of prior judgments about options and had to revisit webpages and reread their content to remind themselves what they liked and disliked about an option. When using Mesh, participants can see all their previous judgments in the Table View presented as cell values in each Option Column, including review tallies and their own ratings and notes about each criterion. This allows users to have a "bird's-eye view" of their research, seeing which criteria and options contain their own ratings and notes, decide what to focus on next, as well as seeing trade-offs between the options when making purchase decisions.

Participants in the exploratory interviews also described "analysis paralysis" when reaching the decision stage, in which many of their options looked similar on the surface (i.e., highly rated based on hundreds of reviews) and that it can be difficult

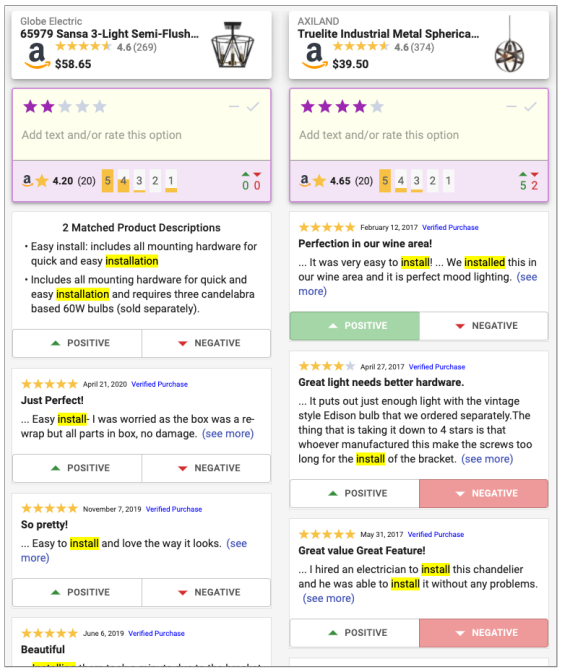

Figure 3. The Evidence View. Users can see evidence that mentioned a criterion side-by-side for their options. To capture their interpretation of evidence, users can also label reviews to build a tally or overwrite the average ratings with their own if they do not reflect their views. This is an actual project made by P5 in the field study.

for them to see clear trade-offs on multiple criteria for their options. Mesh provides several affordances for users to scaffold exploration of the trade-offs between options towards making purchase decisions. Firstly, Mesh computes an overall rating for each option by averaging ratings for its criteria. When averaging, Mesh will use users' own ratings when available and default to the average Amazon review ratings otherwise. Given that participants in the formative studies mentioned the importance of different criteria having differing weights in their decision making, the system also enables users to specify the weight for each criterion which correspondingly alters its impact on the weighted average (e.g., a 5x weight will be counted 5x towards the weighted average more than the default 1x weight). [1] We also supported "soft" prioritization by enabling users to freely reorder rows and columns via drag-and-drop, allowing them to move the most promising options or criteria to the top or the left without altering the overall score. Finally, users can also sort options based on individual criteria ratings or the overall ratings when users click on the sort icon next to the criteria names. This allowed users to quickly explore the best and worst-performing options based on their criteria.

Design Scope and Limitations

In the current implementation, users can group multiple information sources into an option allowing them to search through not only Amazon reviews and product descriptions but also other web pages, such as blog posts or in depth reviews from other sources. However, for each option, one of the sources needs to be an Amazon product page in order for Mesh to auto-fill product names, prices, images, and overall and criterion-specific review scores. In the future, other e-commerce platforms could be supported by implementing additional parsers and/or data connectors to their backend end-points. In theory, yet outside of the scope of this paper, users could also create options with only non-Amazon sources and still create criteria to search across their content and to compare them side-by-side, making Mesh a more general option comparison tool.

Balancing responsiveness and sample size, Mesh makes 3 requests to the Amazon review search end-point to fetch the top 60 most relevant reviews for each criterion. We were concerned about whether users would not trust the system since the reviews we retrieved were not exhaustive (i.e., only the top 60 instead of all reviews that mentioned a criterion) nor perfectly accurate (which was limited by the accuracy of Amazon's review search algorithm). We instead found that people perceived the reviews as a sampling of the distribution about that criteria, and we did not receive any requests for automated summaries of the rest of the reviews as we initially expected. We believe this further accentuates the importance of personalized evaluation of evidence over an exhaustive aggregation, and the value of providing a sample of the distribution as representative of the whole.

During the design phase we explored an alternative design that use sentiment analysis techniques on sentences that mentioned the criterion instead of using average ratings of the whole reviews. A preliminary analysis was conducted where we manually labeled 42 reviews of a popular robot vacuum for the criterion "stuck". Results suggested that searching reviews based the criterion name did retrieve mostly useful results and that the average star ratings represented good overall summaries over the reviews. Specifically, 41 out of the 42 reviews that mentioned the word "stuck" contained useful information about the criterion. We also used two modern sentiment analysis techniques, Vader [24] and Flair [2, 30], on sentences that mentioned "stuck" and found that the average star ratings had a higher Pearson correlation coefficient with the gold-standard labels then sentiment analysis scores (average star ratings: .582, Flair: .352, Vader: .142. N=41). Furthermore, average star ratings can potentially be more transparent and easy for users to understand. Therefore, we chose to use average star ratings over existing sentiment analysis techniques.

Implementation Details

Mesh was implemented in approximately 7,500 lines of TypeScript and 2,500 lines of HTML and CSS. The React library was used for building UI components and Google FireStore for database and user authentication. Firebase and its user account management features were used to allow Mesh users to access their projects across sessions and on different devices. The full version of the system was implemented as a Chrome extension, and a hosted version was ported for conducting Amazon Mechanical Turk user studies in our Evaluation Section. Implementing the system as a Chrome extension

[1] Details of this calculation are explained to users via a hover tooltip. Checks and minuses counted as 5 and 1 stars, respectively.

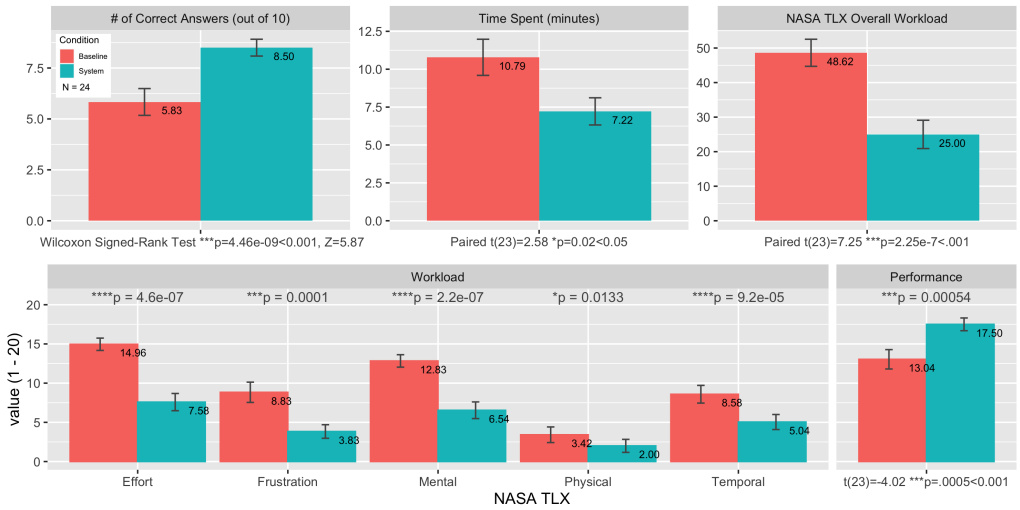

Figure 4. Mean statistic of how participants performed under different conditions in Study 1. Participants who used Mesh were finding more correct answers using a shorter period of time. In addition, they also had lowered perceived workload based on the NASA-TLX survey.

was important for use in the field in order for Mesh to make cross-domain requests for fetching evidence from different information sources. We wrote a custom parser to extract product information from Amazon product pages and fetch reviews using Amazon's review search backend endpoint. Mesh managed a pool of JavaScript Web Workers to query and parse multiple information sources in parallel for responsiveness. The size of the Web Worker pool was determined at run-time to match the number of CPU cores available on users' computers. Finally, implementing Mesh as a Chrome extension enabled it to interact with browser tabs, allowing users to import them into Mesh to build a collection of potential options with lowered effort.

EVALUATION OVERVIEW

We conducted three studies that focused on exploring the following research questions:

- **Study 1**: The usability of our implementation and the benefits of gathering and presenting evidence across sources
- **Study 2**: Whether Mesh enable users to gain deeper insights from data compared to a commonly used baseline (i.e., Google Spreadsheets)
- **Study 3**: The longer-term effects of deploying Mesh to users conducting their own personal tasks

The first two studies were controlled studies comparing Mesh to a baseline condition using predefined tasks to control for task complexity. Participants were recruited from Amazon Mechanical Turk who had more than 100 accepted tasks with above 90% acceptance rate and lived in countries that primarily spoke English. Due to the limitations of running Mechanical Turk studies, we could not install Mesh on their computers as a Chrome extension. We therefore deployed it as a hosted webpage and preloaded and cached necessary Amazon requests for participants to interact with. The third study was a field deployment in which participants installed Mesh on their own computers (as a Chrome extension) and conducted their personal tasks over a period of one to two weeks. Participants for the field study were recruited from the local population primarily by posting to discussion boards on NextDoor, a neighborhood-based social media platform. We used video conferencing and screen sharing software to assist with the installation process and to conduct two rounds of interviews.

STUDY 1 - USABILITY TEST AND INTERACTION COSTS

The main goals of our first study were to verify in a controlled environment the usability of the Mesh and to test if the mechanism of automatically pulling in evidence from different information sources can allow users to work more efficiently and find more accurate information. For this, Mesh was compared to a baseline variant as a within-subject condition where evidence was not automatically pulled in. Objective criteria that had gold-standard answers was utilized in order to measure the accuracy of participants' responses. During the baseline condition, participants could use any strategies based on their own product research experiences, such as searching for keywords on Amazon product pages and/or use search engines to find more sources. In order to measure how effective participants were in finding the right answers, fixed product options (i.e., 5 popular espresso machines on Amazon) and objective criteria were used.[2] One of the authors compiled the gold-standard answers before running the study. Almost all answers were obtained from the manufacturer's website (such as in specification tables and downloading PDF user manuals), with a few resorting to using expert reviews (namely photos or videos that showed a measurement of the portafilters).

The goal of the main task was to find the correct answer for each criterion for the given options. The criteria cells for the first options were filled out to serve as an example. At the beginning of the study, participants were instructed to read

[2]Dimension, Does it have a built-in grinder, Water tank size, Does it use a solenoid valve and Portafilter size

through a brief tutorial to learn the Mesh interface (7 sentences and 4 screenshots). No additional training sessions were performed. The rest of the study was broken down into two segments, and participants worked on two of the four remaining options during each segment with a different condition (counterbalanced for order). During the Mesh condition, evidence was gathered from Amazon reviews and product descriptions, as well as the top two product review webpages, returned from Google when searching with the product names appended with the term "reviews". Links to the same sources were also presented during the baseline condition. During the study, the time each participant spent in the two conditions was recorded as well as their responses. After the study, the NASA-TLX survey was used to collect their perceived workload for each of the two conditions. A total of 24 participants were recruited from Amazon Mechanical Turk (age 21-68 M=36.8; SD=10.5; 15 males and 9 females). Each participant was compensated 3 US dollars for an average of 24.9 minutes (median=22.7, SD=8.3).

Study 1 Results

Results suggest that the 24 participants performed the given task more efficiently when in the system condition than when they were in the baseline condition. Comparing Mesh with the baseline, participants completed their tasks faster when using Mesh that gathered evidence automatically across multiple sources (7.2 vs 10.8 minutes; $t(23)=2.6$, $*p=0.017<0.05$ based on a paired T-test). At the same time, they found information that was more accurate based on gold-standard answers (mean 8.50/10 vs 5.83/10; median: 7/10 vs 9/10; $***p=4.46e-09<0.001$, $Z=5.87$ based on a Asymptotic Wilcoxon Signed-Rank Test). Combining the two metrics we estimated an x2.30 increase in efficiency, where participants were finding 2.23 correct answers on average each minute when using the full Mesh system, compared only 0.97 correct answers per minute on average when using the baseline variant (based on a paired T-test: $t(23)=4.18$, $***p=0.00036<0.001$).

In addition to speed and accuracy, participants also perceived the process to have lowered workload when using the full system across effort, frustration, mental, physical and temporal demands based on the NASA-TLX survey (Figure 4, combined: 25.0/100.0 vs 48.6/100.0; $t(23)=7.25$ $***p=2.25e-7<0.001$ based on a paired T-test) as well as increased perceived performance (17.5/20.0 vs 13.4/20.0; $t(23)=-4.02$, $***p=0.0005<0.001$ based on a paired T-test). This suggests the interface of Mesh can reduce interaction costs when dealing with objective criteria when compared to the baseline where participants relied on their current process, even when they had to learn a new interface.

STUDY 2 - INTERPRETING SUBJECTIVE EVIDENCE

While the first study tested the usability and interaction costs of Mesh when working with objective criteria with gold-standard answers, Study 2 focused on how Mesh can support users when investigating criteria that required subjective and potentially messy and conflicting evidence such as consumer reviews[20]. Unlike looking up the product dimensions in the product description for a coffee machine, investigating its ease of use may require users to read through multiple relevant reviews to get a sense of how previous consumers agreed or disagreed on the criteria while considering how each review fits their personal context. For example, a user buying a robot vacuum who lived in an apartment with wooden floors might down-weight reviews from people who lived in a big house with high pile carpets. For this, we carried out a second study that focused on whether Mesh can provide benefits when researching these types of subjective criteria.

To compare Mesh with people's existing approach, Google Spreadsheets was used as a between-subject baseline. This baseline was chosen because it is a common tool for consumers building product comparison tables and that it is an easily accessible hosted service with APIs that allows us to dynamically create a spreadsheet for each crowdworker. To control for task complexity and the personal preferences of participants, the following persona and task description were used for researching 5 robot vacuum cleaners with the 3 subjective criteria in bold:

> John is looking to buy a robot vacuum for his house. The most important thing for him is that the robot vacuum **doesn't get stuck too often**. It is also important that it is **not too loud**. He also has a dog, so it would be nice if it's also **effective cleaning up dog hair**.
>
> John already narrowed down to 5 final options. Spend around 20 minutes to build up a comparison table to help John research the best option and explain to him why you think it is the best option.

The five options were all popular models on Amazon that had more than 1,000 reviews and above 4 average review ratings (as of April 15, 2020). In both conditions, their tables were populated with the predefined options and criteria to maximize the time participants spent on exploring and learning from data instead of copying and pasting information from the persona (see Figure 6 for the baseline template).

A total of 48 unique participants were recruited from Mechanical Turk for the main study. In which 22 (age 31-58, M=38.7, SD=9.6) were randomly assigned to use Mesh and the remaining 26 participants (age 30-70, M=34.7, SD=11.3) used Google Spreadsheet. Each participant was instructed to conduct the above task for 20 minutes using their assigned systems. It was assumed that participants in the baseline condition were already familiar with a spreadsheet interface and instructed Mesh participants to read through a brief tutorial to learn the interface (13 sentences and 6 screenshots). No additional training sessions were performed. To capture what participants had learned during 20 minutes of research, they were asked to pick one of the options that they recommend and write a short summary for John explaining their choices. This design allowed us to capture the mental models of participants under different conditions through mentions of detailed evidence and how they reasoned and compared the different options, and has shown to be effective for evaluating sensemaking support systems in prior work [25, 36]. Workers who participated in the previous study were excluded from this study to prevent learning effects. Each participant was compensated 3 US dollars.

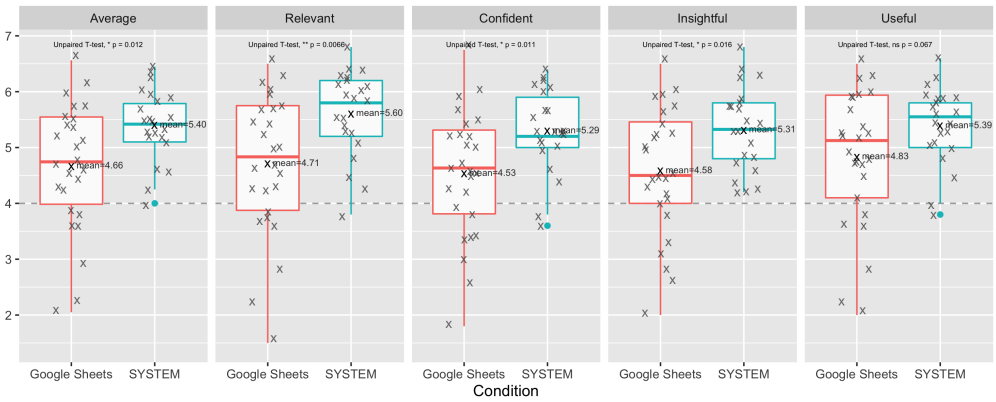

Figure 5. Participants in Study 2 generated learning summaries after 20 minutes of product research. The summaries were rated on 4 statements using 7-point Likert-scales for agreement (7 indicated strong agreement and 4 indicated neutral agreement). A MANOVA was used to correct for multiple comparisons and found a statistically significant difference (F(4, 43)=2.64, *p=0.047<0.05) between the conditions on the combined dependent variables (relevance, confidence, insightful and usefulness).

Name	Amazon Link	Stuck	Loud	Dog Hair
Roborock S4	https://www.amazon.com/dp/B07TXGQS3H			
iRobot Roomba 614	https://www.amazon.com/dp/B00IPEZKBW			
eufy by Anker	https://www.amazon.com/dp/B07DF9GVK9			
iRobot Roomba 960	https://www.amazon.com/dp/B01ID8H6NO			
iRobot Roomba 675	https://www.amazon.com/dp/B07DL4QY5V			

Figure 6. The initial spreadsheet for the baseline condition in Study 2.

To compare summaries collected from the two conditions, each summary was rated by 5 additional crowdworkers. Crowdworkers who participated in the Study were excluded to ensure summaries were not rated by the participants who wrote them. In each rating task, crowdworkers first read the same persona used in the study and one of the summaries. Crowdworkers then rated the following statement using 7-point Likert scales for agreement (a score of 7 indicated a strong agreement, a score of 1 indicated a strong disagreement), and the ratings across 5 workers were averaged as the final ratings:

- I find the summary to be *useful*.
- The summary is *relevant* to the scenario.
- The summary is *insightful*, containing details that may be hard to find.
- I feel *confident* after reading the summary.

The four statements were designed to compare the summaries across conditions on the following aspects: The first statement of usefulness aimed to measure their quality to account for collecting qualitative responses on crowdsourcing platforms [26]. The second statement measured whether participants who used Mesh were able to focus on criteria described in the persona and generate summaries that were more relevant. This is due to the fact that participants in our fact-finding study described their current process as "*a rabbit hole*" and how it can be difficult to "*focus on criteria that really mattered*." The third statement measured how detailed and insightful the summaries were, an important aspect of consumer review helpfulness identified in a prior work [35]. Finally, the fourth statement aimed to explore whether the information in the summaries can support decision making by measuring if they induce confidence.

Workers were paid 0.25 cents for reading the persona and rating summary based on the four statements above. [3]

Study 2 Results

Figure 5 shows the differences between the 22 summaries written by participants using Mesh and the 26 summaries written by participants using Google Spreadsheets for the same task. Averaging across the four aspects, participants who used Mesh generated summaries that were rated higher than participants in the baseline condition (Figure 5, mean 5.40 vs 4.66). A MANOVA was used to correct for multiple comparisons and found a statistically significant difference (F(4, 43)=2.64, *p=0.047<0.05) between the conditions on the combined dependent variables (relevance, confidence, insightful and usefulness). Below are two typical summaries from each of the conditions collected after 20 minutes of product research:

> **Baseline example**: I would pick the Roborock S4 after considering the 3 categories [criteria] that are important to him: how often it gets stuck, noise, and ability to pick up hair. Unfortunately, all of the models he picked do have a tendency to get stuck, which makes it difficult to choose when just using the three factors [criteria]. However, the Roborock was the only model I found, where there weren't many complaints about it being too loud. Additionally, the Roborock is able to pick up dog hair, according to the product description and user reviews.

> **Mesh example**: It seems that while all options do tend to get stuck from time to time, the reviews that the Roomba 675 does somewhat better in that regard. Additionally, many reviews for the Roomba 675 stated how well it picks up pet hair, which was another important consideration that differentiated the Roomba 675 from other options. The Roomba 960 may be marginally better but it costs $200 more and so I didn't think it was worth the

[3] Workers read an average of 124.0 words for each task (range: 50.0-220.0, SD=44.4) and the estimated reading speed of English speakers is 200-300 words per minute [44]. Assuming the lower-bound reading speed of 200 words per minute and 15 seconds was required to answer each of the four Likert-scales. Similar to approach in [23], the estimated the average hourly pay rate was around 9.26 USD.

ID	Tasks
P1	Snow boots. Gourmet cat food.
P2	Backpacks. Pajamas as a gift for his/her sister.
P3	Bread machine. Hair cutting kit.
P4	Running shoes. Printer for learning material for kids.
P5	Entryway light fixture. Toy play-sets for kids.

Table 1. Titles of projects created by participants in the field deployment study based on activity logs.

extra expense. Lastly, there were reviews that found the noise of the Roomba 675 to be acceptable.

While many participants who used Google Sheets mentioned the similarity between options and the difficulty of the task, people who used Mesh point out how they differentiated the options on the given criteria based on multiple pieces of evidence.

	Action Count	P1	P2	P3	P4	P5	M	SD
Options	Add	7	16	22	27	11	16.6	8.1
	Remove	3	9	12	2	2	5.6	4.6
	Drag to reorder	13	1	4	3	17	7.6	7.0
	Sort by criteria	9	16	4	7	16	10.4	5.4
Criteria	Add	9	16	21	29	14	17.8	7.6
	Remove	2	5	5	6	4	4.4	1.5
	Change weight	2	4	4	1	2	2.6	1.3
	Drag to reorder	2	2	5	4	10	4.6	3.3
Cells	Change rating	5	9	8	0	44	13.2	17.6
	Add notes	4	0	7	38	48	19.4	22.0
	Tally review	10	54	8	0	32	20.8	22.0
	Total changes	19	63	23	38	124	53.4	43.1
	# Uniq cells	3	12	9	12	41	15.4	14.8
Total # of Actions		57	132	100	117	200	124.2	48.5
Total minutes spent		70	134	150	82	131	113.4	35.2
Number of sessions		3	3	4	6	5	4.2	1.3

Table 2. Usage statistics about participants in the field deployment study based on the activity logs. Participants utilized a wide range of features provided by Mesh during the 1-2 week deployment.

STUDY 3 - FIELD DEPLOYMENT STUDY

While the first two studies provided quantitative measures on how Mesh affected learning, efficiency, accuracy, and perceived workload when participants were given predefined tasks, we conducted a field deployment to further investigate the longer-term effects of Mesh when participants performed their personal tasks in the wild. Five participants (age: four 25-34 and one 35-40; two females, two males and one non-binary) were recruited by posting to 5 local neighborhood discussion boards on NextDoor (a neighborhood-based social media website). The posts contained a link to an online screener survey, and the responses were used to recruit people who have used a spreadsheet for online research in the past (49.4%, N=89) and prioritized people who had any Chrome extensions installed (49.4 N=89).

Participants were interviewed for one hour at both the beginning and the end of the deployment. Before the initial interview, participant were asked to email us 1 to 3 upcoming online purchases to ensure they have a real task to work on during the initial interview. At the start of the first interview, their demographic information was collected and they were assisted with installing Mesh as a Chrome extension on their computers via screen sharing. Each participant then proceeded to perform a think-aloud session for around 30 minutes using Mesh to conduct one of the tasks they had proposed. After the first interview, participants continued to use Mesh on their own for the same tasks and/or create new tasks. Based on their availability, each participant was interviewed again after 1-2 weeks. Participants shared their screens and retrospectively walked through their projects while they were probed on their experiences, strategies, and issues they had encountered during the deployment. All 5 participants completed the study and were each compensated an Amazon gift card worth 50 US dollars. The interviews were video recorded and transcribed for analysis.

Study 3 Results

Table 1 shows the tasks each participant conducted using Mesh based on log data. The first tasks in the table were ones created during the initial interview and the rest created during deployment. There was a wide verity of different tasks such as clothing (P1, P2, P4), appliances (P3, P4), pet supplies (P1) and toys (P5). During the deployment, participants interacted with the Mesh system for 70 to 150 minutes based on the behavior logs (Table 2), and all of them used Mesh in three to six sessions (M=4.2, SD=1.3). Participants saved multiple options and used multiple criteria to compare them. They were also actively removing options and criteria suggesting Mesh allowed them to dynamically decide on which options to consider and based on which criteria. Based on their interpretation of evidence, on average, each participant changed the default values of the cells in their tables 53.4 times (SD=43.1) with different participants preferring different features (i.e., change ratings, type notes and label reviews). Finally, participants also used different Mesh features to help them prioritize information they collected. This included reordering options and criteria via drag-and-drop, and sorting options based on how they were rated on a criterion.

Qualitative findings based on pre- and post- interviews provided deeper insights to how these action benefited the participants. Following an open coding approach based on grounded theory, the first author went through the 10 hours of recordings and transcriptions in three passes, and iteratively generated potential categories from the dialogue until clear themes emerged [10]. Throughout the iterations, inputs from the rest of the research team were also incorporated, including other researchers who also conducted interviews. Our key findings are presented below.

Efficient and Organized

In general, participants responded favorably to using Mesh in the field for their personal tasks, preferring Mesh when asked to compare it against their current online product research process (i.e., using spreadsheets and/or notepads). Specifically, all participants pointed to lowered interaction costs when using the Evidence View to access evidence gathered across information sources to compare their options, as well as lowered cognitive costs from being able to rule out options confidently based on evidence.

It is much better than a spreadsheet... I like that I can really quickly add something and it just pulls in all the

> information, the picture, the price, and [evidence for] all of these different criteria and presents it in a way that is really easy to do comparison across products. I'm able to delete things easily so that I can reduce my cognitive load as I go through my decision-making process. - P3

All participants described how Mesh allowed them to take a more organized and structured approach when managing multiple information sources and collecting evidence. Specifically, P1 and P2 noted that the linear structure of browser tabs can be inefficient when trying to find evidence for a specific criterion across browser tabs for different options. Participants pointed out that while the mechanisms provided by Mesh could be performed manually, the interaction costs of managing many browser tabs and filtering for relevant information to support their criteria amongst them would be prohibitively high in practice.

> In theory, I could do all this myself but it would take 10 times [as] long so I would never do it well. I would say is it technically possible? Yes. But would any person ever do this [manually] for themselves? ... It's nice to have a more organized and systematic approach... Instead of something that right now is very linear. If I pulled up a bunch of boots in different tabs and searched [in] each of them for reviews with the word boar. It's really boring and not a particularly efficient way to look at information. - P1

One participant (P3) described Mesh as providing a more organized scaffold for their process, enabling better support for task resumption and allowing them to make progress on their overall tasks even in shorter sessions.

> I loved being able to come back to this [referring to one project]. It's something we hadn't done in our initial sessions that became so much better when I was using it on my own. I couldn't say, hey, I've got 15 minutes to kill. Let me do some more searching, and then I could say, okay, gotta go to my next meeting. - P3

Analysis of activity logs suggested that participants could effectively use Mesh to suspend and resume tasks, with all participants conducting their product research in three to six separate sessions (M=4.2, SD=1.3) (Table 1).

Prioritizing Effort on Discriminative Criteria

Participants found criteria useful for discriminating between options. All participants saw immediate value when the average Amazon ratings populated automatically for their options when they added a new criterion, allowing them to get an initial overview of how evidence differed between options. Specifically, participants described trying out different criteria as a way to surface meaningful differences (i.e., based on their own criteria) amongst their options. Since participants typically only considered options that were popular and highly rated on Amazon, they described these options as virtually indistinguishable without Mesh:

> Having never purchased it before I literally have no idea what to buy. And so this [task] is what I tried to do [with Mesh] and it's actually like super helpful because [otherwise] every single stupid cat food on Amazon just like looks identical... So, it was really helpful especially [with] this picky criterion. - P1

Conversely, when participants added a new option to a project that had existing criteria, Mesh automatically populated Amazon average review ratings across those different criteria for the new option. Participants used this mechanism to quickly characterize new options and see how they fit with existing options based on their own criteria:

> This new one is pricey, and yet anybody that mentioned cost [a criterion] has given it the full rating. They're more durable [referring to discrepancy between options on the criteria] You know, I could see tangible evidence now. And that makes me want to go – Maybe that's the pair. - P4

Seeing discrepancies between options also influenced participants' process by prompting them to prioritize their effort on investigating criteria that were more discriminative between their options:

> Okay, there wasn't a great difference here in terms of ink [a criterion]. Let me go into what I weighted as more important, and it's this air printing [another criterion] capability... for this middle one [referring to one option], rated pretty poorly... These two have pretty good ratings. So then I went in and started looking [at the evidence] - P4

By focusing first on criteria that were more discriminative amongst the options, participants could rule out options that compared less favorably earlier to shorten their process. All participants described prioritizing their options in the system, either by reordering their options via drag-and-drop or ruling out options completely by removing them.

Scaffolding Decision Making

Participants also described how Mesh supported deep exploration of individual pieces of evidence in the Evidence View that laid out the evidence for specific criteria across their options.

> The second thing that I think is really great for me was the ability to dive into the reviews for specific criteria. It's really nice to be able to open this [the Evidence View] up and have it filter out for all of the products, so I can make this comparison across products. - P3

One participant, in particular, described a sense of relief and progress when removing options in Mesh.

> I don't feel like I would delete things [options] in a spreadsheet. Whereas here it actually feels good to delete it [an option], because I'm like, Great! I've decided that I'm not going to deal with it. - P3

When we introduced the system, we explicitly explained to participants that the average ratings were based on review scores and could be influenced by parts of a review not relevant to their criteria even though the reviews mentioned the criteria. Participants were able to work with this limitation, and replaced the Amazon ratings with their own when they did not reflect how they wished to characterize the evidence. In addition, participants also described creating ratings as a way to keep track of and aggregate how they personally interpreted

evidence and saw benefits in how changing criteria ratings were reflected in the overall weight score of each option.

> I would say in the event that I was going to differ from what's in front of me, I would rate [the criteria]. - P4

> Once I start to make decisions on things like I put my thing [own ratings and notes] in there and say: Okay, this is what my rating is. And now it starts to change the overall ratings, so it would help me make a better decision based on what I think. . . . like, the tool thinks this is a really good value, but maybe I think this value is not enough for me and it's a two because I just think it's two - P3

Four of the participants (P1-P4) also made actual purchases during the deployment based on research they performed with Mesh and expressed how they felt confident in their resulting decisions. P5 wanted to use the project to discuss with a partner and make the purchase decision together. This suggests that their tasks represented real-world user needs, and our participants were able to use Mesh to conduct research for a prolonged period of time and use it to support making their final purchase decisions.

LIMITATIONS AND FUTURE WORK

While all participants' initial responses were positive when adding options and criteria to the Table View, some of them found their first impressions of the Evidence View to be overwhelming. While this suggested a higher learning curve for the Evidence View, all participants were able to complete research with it for their own tasks during the deployment.

> So initially it was like, Whoa, there's a lot going on here. It's a lot of text but I'm kind of over it once I understood what was going on. Now I'm like, Okay, cool. Let's take a look at this [referring to the criteria] across the things [referring to the options] - P3

More commonly, participants expressed a desire to extract evidence from online sources other than Amazon. While the current implementation supports extracting evidence from other sources (by pasting their URLs into the appropriate option), participants pointed to two limitations: 1) extracting and tracking price changes across e-commerce platforms other than Amazon and be notified, and 2) extracting from listicles and forum posts that discussed multiple products:

> Running shoes are kind of discipline-specific. There are other sites solely for this [type of] product that I would go to. [To add a webpage and] track the options to use globally would be cool. But like robot vacuum there's nowhere else [but Amazon] I'm going. Unless I'm tipped off that Target or Bed Bath and Beyond happened to have an incredible sale. - P4

While price tracking could be implemented within Mesh, there are multiple commercial solutions available [4] and we considered it outside the scope of this work. On the other hand, extracting information from sources containing evidence about multiple options presents an interesting research challenge of computationally identifying mentions of products and extracting descriptions about them from the text.

[4] https://camelcamelcamel.com/ and https://www.joinhoney.com/

CONCLUSION

We introduced Mesh, a novel sensemaking system where users build up comparison tables by discovering options and criteria as they explore online information. As options and criteria are added to their tables, evidence about them is automatically gathered across information sources for users to review. When needed, users can also externalize their personal interpretation of data as cell values to keep track of their research progress. This design is novel because it introduces a new process that scaffolds the iterative building up of context, and changes the cost structure from increasing cost to increasing payoffs as the number of criteria and options grow.

Through three rounds of lab and field deployment studies, we uncovered deep insights into how Mesh can benefit online sensemaking in the context of product comparison research. In Study 1, we found evidence that Mesh not only lowered interaction costs (i.e., shorter time spent and lowered perceived effort), but also led to participants finding more accurate information when working with objective criteria (e.g., water tank capacities for espresso machines). In Study 2, when dealing with subjective criteria (e.g., ease of use for espresso machines) we found evidence that participants who used Mesh were more insightful and confident about their choices compared to participants who used a Google Spreadsheet baseline. Finally, in Study 3 we tested Mesh in the wild with participants conducting their own tasks over a longer period of time and found that Mesh allowed participants to better prioritize their effort on criteria that were more discriminative, and was able to capture their interpretations of data to keep track of their progress.

Fundamentally, online evidence can be messy, biased, subjective and conflicting. This requires users to consider many information sources in order to better evaluate both their options and the evidence itself. Providing better scaffolding support when users explore, compare, and interpret online evidence can empower users to gain deeper insights with lowered interaction and cognitive efforts. While Mesh explored this in the context of online product research, we believe the designs introduced here may generalize to other domains where users need to compare options based on online information. For example, travelers could use Mesh to compare different destinations and restaurants, voters could use Mesh to compare different policies and candidates, and patients could use Mesh to compare different hospitals and treatment plans. We believe Mesh represents a first step towards a user-centered sensemaking approach to addressing the subjective and distributed nature of online information today.

ACKNOWLEDGEMENT

This work would not have been possible without Julina Coupland and Bradley Breneisen who generated mock-ups and conducted study recruitment and interviews. This work was supported by the National Science Foundation (PFI-1701005 and SHF-1814826), the Office of Naval Research, Google,

and the Carnegie Mellon University Center for Knowledge Acceleration.

REFERENCES

[1] 2018. How merchants use Facebook to flood Amazon with fake reviews - The Washington Post. https://www.washingtonpost.com/business/economy/how-merchants-secretly-use-facebook-to-flood-\amazon-with-fake-reviews/2018/04/23/5dad1e30-4392-11e8-8569-26fda6b404c7_story.html. (2018). (Accessed on 05/06/2020).

[2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[3] Erik M Altmann and J Gregory Trafton. 2004. Task interruption: Resumption lag and the role of cues. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 26.

[4] Michael S Bernstein, Jaime Teevan, Susan Dumais, Daniel Liebling, and Eric Horvitz. 2012. Direct answers for search queries in the long tail. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 237–246.

[5] Andrea Bianchi, So-Ryang Ban, and Ian Oakley. 2015. Designing a physical aid to support active reading on tablets. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 699–708.

[6] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.

[7] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2016. Supporting mobile sensemaking through intentionally uncertain highlighting. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 61–68.

[8] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: Composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 498–509.

[9] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3180–3191.

[10] Kathy Charmaz and Linda Liska Belgrave. 2007. Grounded theory. *The Blackwell encyclopedia of sociology* (2007).

[11] Chao Chen, Daqing Zhang, Bin Guo, Xiaojuan Ma, Gang Pan, and Zhaohui Wu. 2015. TripPlanner: Personalized trip planning leveraging heterogeneous crowdsourced digital footprints. *IEEE Transactions on Intelligent Transportation Systems* 16, 3 (2015), 1259–1273.

[12] Li Chen and Feng Wang. 2017. Explaining recommendations based on feature sentiments in product reviews. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 17–28.

[13] Li Chen, Feng Wang, Luole Qi, and Fengfeng Liang. 2014. Experiment on sentiment embedded comparison interface. *Knowledge-Based Systems* 64 (2014), 44–58.

[14] EH-H Chi, Phillip Barry, John Riedl, and Joseph Konstan. 1997. A spreadsheet approach to information visualization. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*. IEEE, 17–24.

[15] Bart De Langhe, Philip M Fernbach, and Donald R Lichtenstein. 2016. Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *Journal of Consumer Research* 42, 6 (2016), 817–833.

[16] Qiwei Gan, Qing Cao, and Donald Jones. 2012. Helpfulness of online user reviews: More is less. (2012).

[17] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big picture thinking in small pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2258–2270.

[18] Nathan Hahn, Joseph Chee Chang, and Aniket Kittur. 2018. Bento browser: complex mobile search without tabs. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[19] Marti Hearst. 2006. Design recommendations for hierarchical faceted search interfaces. In *ACM SIGIR workshop on faceted search*. Seattle, WA, 1–5.

[20] Stephen J Hoch and Young-Won Ha. 1986. Consumer learning: Advertising and the ambiguity of product experience. *Journal of consumer research* 13, 2 (1986), 221–233.

[21] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 168–177.

[22] Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.

[23] Chieh-Yang Huang, Shih-Hong Huang, and Ting-Hao Kenneth Huang. 2020. Heteroglossia: In-Situ Story Ideation with the Crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. DOI: http://dx.doi.org/10.1145/3313831.3376715

[24] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[25] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 625–634.

[26] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[27] Aniket Kittur, Andrew M Peters, Abdigani Diriye, Trupti Telang, and Michael R Bove. 2013. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2989–2998.

[28] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. Structure-aware review mining and summarization. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 653–661.

[29] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding Developers' Decision-Making Using the Web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.

[30] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.

[31] Asha S Manek, P Deepa Shenoy, M Chandra Mohan, and KR Venugopal. 2017. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World wide web* 20, 2 (2017), 135–154.

[32] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[33] Catherine C Marshall, Morgan N Price, Gene Golovchinsky, and Bill N Schilit. 1999. Introducing a digital library reading appliance into a reading group. In *Proceedings of the fourth ACM conference on Digital libraries*. ACM, 77–84.

[34] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

[35] Susan M Mudambi and David Schuff. 2010. Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly* (2010), 185–200.

[36] Les Nelson, Christoph Held, Peter Pirolli, Lichan Hong, Diane Schiano, and Ed H Chi. 2009. With a little help from my friends: examining the impact of social annotations in sensemaking tasks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1795–1798.

[37] K O'Hara. 1996. Towards a typology of reading goals rxrc affordances of paper project. *Rank Xerox Research Center, Cambridge, UK* (1996).

[38] Pradeep Racherla and Wesley Friske. 2012. Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications* 11, 6 (2012), 548–559.

[39] Ramana Rao and Stuart K Card. 1994. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 318–322.

[40] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[41] MC Schraefel, Max Wilson, Alistair Russell, and Daniel A Smith. 2006. mSpace: improving information access to multimedia domains with multimodal exploratory search. *Commun. ACM* 49, 4 (2006), 47–49.

[42] Barry Schwartz. 2004. The paradox of choice: Why more is less. Ecco New York.

[43] Steven M Shugan. 1980. The cost of thinking. *Journal of consumer Research* 7, 2 (1980), 99–111.

[44] Eva Siegenthaler, Yves Bochud, Per Bergamin, and Pascal Wurtz. 2012. Reading on LCD vs e-Ink displays: effects on fatigue and visual strain. *Ophthalmic and Physiological Optics* 32, 5 (2012), 367–374.

[45] Herbert A Simon. 1996. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics* 70 (1996), 187–202.

[46] Michael Spenke, Christian Beilken, and Thomas Berlage. 1996. FOCUS: the interactive table for product comparison and selection. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*. 41–50.

[47] Statista. 2019. Yelp: cumulative number of reviews 2019 | Statista. `https://www.statista.com/statistics/278032/cumulative-number-of-reviews-submitted-to-yelp/`. (2019). (Accessed on 05/06/2020).

[48] Craig S Tashman and W Keith Edwards. 2011. LiquidText: a flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3285–3294.

[49] Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 1496–1505.

[50] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 217–226.