Kurator: Using the Crowd to Help Families with Personal Curation Tasks

David Merritt[1], **Jasmine Jones**[2], **Mark S. Ackerman**[1,2], **Walter S. Lasecki**[1,2]
Computer Science & Engineering[1], School of Information[2]
University of Michigan – Ann Arbor
{afdavid,jazzij,ackerm,wlasecki}@umich.edu

ABSTRACT
People capture photos, audio recordings, video, and more on a daily basis, but organizing all these digital artifacts quickly becomes a daunting task. Automated solutions struggle to help us manage this data because they cannot understand its meaning. In this paper, we introduce Kurator, a hybrid intelligence system leveraging mixed-expertise crowds to help families curate their personal digital content. Kurator produces a refined set of content via a combination of automated systems able to scale to large data sets and human crowds able to understand the data. Our results with 5 families show that Kurator can reduce the amount of effort needed to find meaningful memories within a large collection. This work also suggests that crowdsourcing can be used effectively even in domains where personal preference is key to accurately solving the task.

Author Keywords
Crowdsourcing; mixed-expertise; hybrid intelligence; personal curation; digital audio; digital curation

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION
People have much more digital content than they can manage, even when it comes to relatively narrow subsets of content, such as digital photos. Researchers have called out the need for a strategy for forgetting, preserving, and remembering personal digital content (e.g., [38]), but this requires families to significantly shift their habits, which is often impractical. Instead, a focus on using systems to help manage familial digital information may be valuable [17].

Crowdsourcing might be one approach, but deciding about digital content can be highly subjective. Although there has been some crowdsourcing research in subjective problems (e.g., word processing in [4], itinerary planning in [51], and managing email in [26]), some researchers doubt crowdsourcing is a useful approach to problems having personal or highly subjective aspects to them, as summed up by Simko and Bieliková [47]: "Automated or crowdsourcing approaches are inapplicable in [the] case of personal content or content of a small social group (e.g. family). (p. 45)" The pervasiveness of this sentiment is unclear, but Organisciak et al. [42] acknowledged that researchers in crowdsourcing have only recently begun focusing more on problems with a "subjective aspect to them." We believe personal digital content curation falls into this class of subjective problems where crowds may be helpful but have not yet been leveraged.

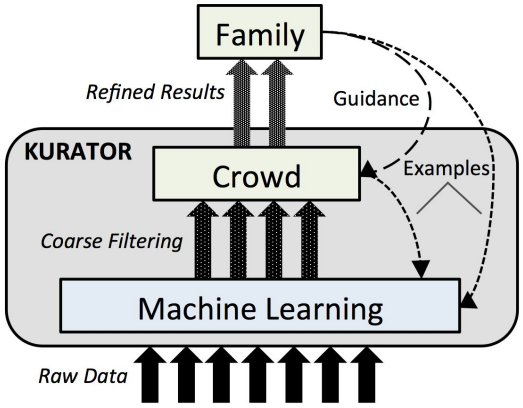

Figure 1. Kurator system diagram. Kurator starts with a collection of digital media content. A machine learning tier reduces the amount of content by filtering, based on criteria for that media type (such as no volume for audio). The crowd tier then does further refinement, producing a candidate set for the family, who is the ultimate judge for family memory. Feedback from the family can guide the improvement of the machine learning tier and the crowd tier.

This paper introduces Kurator, a system designed to help families curate their own digital audio recordings. Kurator uses mixed expertise crowds as part of a hybrid intelligence system to reduce the curation burden on families. It is a hybrid intelligence system because it uses inputs from machine learning and crowds (see Figure 1). Kurator also leverages the mix of expertise levels between “crowds”: families (experts) and paid web workers (variable expertise). Kurator has a tiered approach whereby a machine learning (ML) classifier performs coarse-grained filtering on a family’s entire digital audio collection, and the crowd refines the classifiers’ output into a smaller, more

CSCW '17, February 25-March 01, 2017, Portland, OR, USA
© 2017 ACM. ISBN 978-1-4503-4335-0/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2998181.2998358

manageable set of higher quality recordings that can be presented back to the family.

As well, Kurator allows the family to provide natural language feedback to crowd workers ("Guidance" in Figure 1). The family's labels, as well as the crowd's labels, can be used to further train the ML ("Examples" in Figure 1). While hybrid intelligence systems have used the crowd's labels to train machine learning so as to eventually replace the crowd (e.g., Tohme [19], Zensors [29], and Guardian [20]), we are providing a new, more efficient way to accomplish this. Kurator, in our field study, demonstrates that experts can train crowds effectively. Although our approach could use expert feedback to train the machine learning until the ML is good at the task, we found it was more effective and took less effort for the families to train the crowd. The crowd can then provide high-quality labels in large numbers to train the ML.

We evaluated Kurator through a user study with five families. We found that not only is the resulting curation useful but also that crowdsourcing can be applicable to an important class of subjectively-based problems. For these problems, we show that Kurator is able to effectively leverage crowds to provide useful assistance. We believe our work demonstrates an important new problem setting in which crowds can benefit users.

We make the following contributions in this paper:

- We show that crowd workers are effective at predicting whether specific children's digital audio recordings will be valuable to a family, and their ability to assess this value is improved when the family conveys context and preferences to them in natural language feedback.
- Through this problem of digital curation, we show there is a class of subjective problems where crowds may be helpful but have not yet been leveraged.
- Specifically, we introduce the Kurator system, a hybrid intelligence system which uses mixed-expertise crowds in a tiered architecture to synthesize inputs from multiple layers of contributors, such as machine learning, the crowd, and the family, to reduce the burden on a family with family curation tasks.
- Through the Kurator system, we show that adding an expert layer to a tiered architecture adds new capability to hybrid intelligence because of how the training and communication between layers is done. As mentioned, we show that it is more effective to leverage time-constrained experts (in our case, the families) to train the crowd than train the ML. The crowd can then provide high quality labels, in large number, that can be used to train the ML. Conveying context to the crowd to help them assess future importance and to provide additional training to the ML is new.

After we walk through the background literature for our work, we then present Kurator and its major design assumptions and features. We then present the results of our user study where we examined whether Kurator and its features were effective. We conclude with a brief discussion and future work.

RELATED WORK

Curation in HCI and CSCW

The practices and processes of selecting, organizing, and maintaining a collection of material is broadly considered as "curation." Curation has been extensively studied in institutional archives and library science, and it has recently been extended to data and digital curation [49]. The problem of curating personal digital content is a difficult one that remains difficult and unsolved. Marshall [35] noted that "digital material accumulates quickly, obscuring those items that have long-term value (p. 5)." Marshall [35], as well as Marshall et al. [36], found that almost all users do not do an adequate job of curation.

Early work on digital curation in CSCW/HCI focused on studies of and systems for sharing digital photos. Recent curation research in CSCW/HCI has focused on the work of curation in social media sites. Chang et al. [8] examined the curation work taking place a social curation site, Pinterest. Zhao and Lindley [52] examined how the use of a social media site leads to a curated archive of personal digital content. Very recently, there have been studies of people's perceptions and understandings of algorithmic curation on Facebook News Feed and how it affects their use of the system [43, 13]. As well, a recent study on the "modern day baby book" investigated new mothers' photo sharing activities on Facebook [28]. These studies did not investigate how families might curate digital content when they do not want to share or keep private.

Automating personal digital media curation

It is clear that people can hand select digital content for preservation and use. Relatively little work has investigated how digital curation might be done through systems, either machine-learning based or crowdsourced.

Obrador et al. [40] inferred user preferences for "style" using social cues from their online photo albums, but it did not allow for explicit end user feedback into the system. Other work that builds on the idea of inferring user preferences includes Guldogan et al. [14], which required a profiling task to be performed by the end user. This, however, required training, and a usability goal might be to be effective "out of the box" without requiring user tasks before being useful, as we do with Kurator.

Recently, there has been research on automating personal digital media curation using general preferences for photo selections instead of user-specific preferences [5, 6, 38]. Nejdl and Niederee [38] concluded that a coverage-based approach, which attempted to cover multiple events, did not perform as well as a simple reduction-oriented strategy,

which removed duplicate and near-duplicate photos. We follow this, using a similar reduction-oriented strategy in Kurator by filtering out low quality recordings. Additionally, this line of work suggests a potential utility in utilizing family-specific preferences by re-training the machine learning on data from the family or the crowd.

The only direct example of crowdsourced curation of personal digital media, Cusano and Santini [10], proposed a community-sourced method to help users categorize their photos using labels (i.e., tags). This method correlated photos from the public Flickr user community with target users. Public photos with labels are used to predict labels for similar photos from a target user. This method works only when there is an Internet-scale repository of public data, and is appropriate only for some content. Kurator is designed to work with no public repository of similar data.

As an indirect example of crowdsourced curation, Organisciak et al. [42] used profiling tasks to understand user preferences, then they employed two approaches to understand a user: taste-matching and taste-grokking. Taste-matching works by finding workers similar to the user's profiling results, and with taste-grokking, any benefit is limited to when the users train the crowd. As mentioned, we want an approach that is effective immediately but also improves with additional training.

Similar to taste-grokking, Yi et al. [50] leveraged a user's response to pairwise comparisons from a subset of items. They use a matrix completion algorithm, called crowdrank, to infer the user's preferences on the remaining items. This matrix completion approach, however, can be very lengthy, increasing the task time and cost significantly.

Using expertise in crowdsourcing

Prior systems have leveraged collaboration between only expert crowd workers, between experts and non-experts, and among some mix of expertise. Kurator builds upon that work by combining the mutual efforts of the crowd, expert users, and machine-learning agents, in addition to leveraging expert users' feedback.

Some systems focus solely on expert workers. Chilton et al. [9] introduced Frenzy, a conference-planning tool designed for large groups of experts to collaboratively build a conference program. Instead of tasks being routed to experts, experts self-select tasks (papers) based on their topics of expertise. Frenzy provides a way to facilitate experts working simultaneously on a complex task. Similarly, Foundry [44] enables expert flash teams, where multiple experts come together to quickly and collaboratively work on modular tasks that can be linked to other modular tasks. Kulkarni et al.'s Wish system [27] allows expertise to be solicited when a non-expert crowd lacks what it needs for specialized, creative work. Importantly, this work showed that using expert-only crowds for specialized tasks is workable. Unfortunately, using *only* experts (i.e., the end users) to solve the problem of digital curation is impractical because only users are experts, and according to prior literature, most users simply will not take the time to curate their digital content [35, 36].

Alternatively, one can have workflows designed to allow experts to guide less expert contributors. Kittur and Kraut [25] studied Wikipedia, finding that fewer editors initially, establishing an article's structure and cohesion, likely led to higher article quality. Dow et al. [11] used a shepherding metaphor to demonstrate how task-specific, external feedback, provided at the right time, increases the quality of work provided from the crowd. This shepherding concept has been expanded upon by systems using expert facilitators to direct the work and communicate goals to collaborators. This includes problem areas involving idea-generation [7], collaborative story writing [24], citizen science [33], travel itinerary planning [51], and coding behavioral video [31], where experts provided creative inspirations, creative constraints, clarifications on task instructions, simple bounds or time constraints, and training examples, respectively, to crowd workers. We add to the prior literature by allowing experts to give open-ended natural language feedback to crowd workers that they think will scope their interests and add context. This addresses prediction of preference or future relevance, and how to transfer this understanding to the crowd. In the tiered architecture, the expert layer adds a new structure because of how the training and communication between layers is done.

In a slightly different approach, Huang et al. [20] designed Guardian, a crowd-powered spoken dialog system, to use inputs from a non-expert crowd to filter out "unnatural" parameters from various web API's to lower the threshold for programmers (i.e., the experts) to contribute to Guardian. With this arrangement, the non-expert workers are leveraged earlier in the workflow, laying the groundwork for expert workers to contribute more easily on a separate task. We modify this workflow to allow expert crowds to complete the same type of task as non-expert crowds; thus experts assume the roles of requesters and workers. This subtle but important point allows Kurator to obtain expert labels to train the ML, when those expert labels are available (i.e. when users find the time to provide them). Although we found that experts can train crowds more effectively and efficiently than they can train the machine learning, this architecture provides flexibility for future systems to investigate conditions where the paid crowd layer could be bypassed, such as for users who are willing to spend time manually curating their digital content. For such users, interactive machine learning could leverage their domain expertise, where "domain" is their set preferences. Indeed, Amershi et al.'s recent review of interactive machine learning [2] discusses the importance of the role of domain experts in interactive machine learning.

KIDKEEPER BACKGROUND

Kurator currently uses digital audio recordings collected from the KidKeeper system [22]. KidKeeper (KK) is a toy-like device designed for children to spontaneously capture audio recordings of their everyday activities, combined with a simple delivery system to enable parents to enjoy the recordings their children created. The types of content captured using KK are children singing, telling a story, making up sounds or uttering words, screaming, and short phrases.

A study of KidKeeper in use revealed that children generated hundreds of recordings in only a few days, with diverse content and variable value to parents. Over time, the accumulation of digital artifacts could very well become overwhelming, especially if there were multiple capture devices in a home. Thousands of audio clips, even if only a few seconds each, could be too much. Winnowing down the number of clips to be manageable, such as in creating an audio album, would be tedious, and there is evidence that users just will not do it manually [36].

KK revealed that parents enjoyed listening to the audio recordings, but there was a need for a more sophisticated, automated curation system to help parents find audio recordings, for example the "gems" [41], in a large digital audio collection. Next, we explain how Kurator addresses these user goals in curating their personal digital content.

KURATOR

Improving personal digital content curation requires trading off two key factors: scalability and access to specialized knowledge. A family has "expert" knowledge of what is meaningful to them, but their time is a finite resource. Machines can scale to massive data sets, but cannot understand the "meaning" of content. Crowds of online workers are flexible, available on demand, and can be recruited at scale. Furthermore, crowd workers will likely have some level of common social understanding with the family. But the crowd is still separate from the family and does not know the subtler context underlying the content. Additionally, crowdsourcing can often be cost-prohibitive for very large collections.

Kurator is a hybrid intelligence system that reduces the time-and-effort cost of curation for families so as to make collections of digital memories easier to manage. It uses a tiered architecture (see Figure 1) that first filters raw data using machine learning, and then asks the crowd to assess the content on behalf of the family. Finally, the filtered, significantly smaller set of potentially-interesting artifacts is returned to the user for final evaluation. After viewing and (optionally) further refining the set, family members can provide feedback to the crowd and machine learning to improve future results. As a test, we applied Kurator to personal digital audio recordings collected using KK.

Example Scenario

Daniel, hearing the sound of young children running down the hallway of his hotel room, gets a twinge of nostalgia for his own children. He logs onto the Kurator website to listen to some audio recordings of his kids. He notices two things right away. First, he sees there are now over 1,000 recordings in his collection, and a part of him is thankful he hasn't listened to the vast majority of them. The other thing he notices is that his Top 20 list has three recent additions. He listens to the first recording and, enjoying his son's rendition of *Hush Little Baby*, tags and rates the recording accordingly. He enjoys the second recording, of his daughter saying how much she loves her daddy, and tags and rates it. The third recording, the longest of the three, is less enjoyable because the family dog is barking for half of it. He clicks on the feedback link for this recording, and on the subsequent page, he sees all the previous guidance he and his wife have provided up until now. Seeing that they had, somehow, not yet provided guidance about their dog, Daniel submits the following feedback to the system: "It's not as meaningful if the dog is barking for very long."

Design Considerations

Kurator is designed to address the fact that there is no way for (most) people to keep up with their digital media collections long term. We make a few baseline assumptions about curating digital artifacts, based on the work in [21]:

- *Curation is a process and not a static goal*. It is dynamic over time as tastes, goals, needs, and perspectives change.
- *Everything should be kept.* Digital space is cheap, so curation should no longer be about "keep or throw away" but rather about "what to pay attention to".
- *The primary goal of curation is not to select the single most meaningful artifact*. Even families themselves may not be able to do this. Instead, the goal is to narrow the focus down to a meaningful *set* of artifacts for further processing by the family.

Below, we discuss Kurator's key design features: integrating machine learning, the crowd, and families, and incorporating feedback in the process to improve results.

Integrating machine learning

Design Rationale: We leverage machine learning to reduce decisions for human contributors. Reducing the curation decision space by using automated approaches in a reduction-oriented strategy has been demonstrated on digital photo collections [38]. The automated approach we use needs to handle continually re-training machine learning classifiers over time as the crowd and the family provide inputs to Kurator. For this purpose, Nguyen et al. [39] suggest using logistic regression with gradient descent, which supports incremental training.

System Description: Kurator uses a three-class rating system, where each audio recording is rated as one of three classes. Thus the machine learning classifier (ML) is currently implemented as a multinomial, or multi-class, logistic regression model using gradient descent. This particular ML is meant as a proof of concept, and Kurator is designed to be agnostic to the ML algorithm and even to the use of an ensemble, or a "crowd", of ML algorithms. The core of the ML tier is implemented in Octave [12] scripts, called from a Python script using the oct2py module. When Kurator is initialized for a family, there are no human ratings to use to train the model, so we use regression coefficients from a pre-study as the seed. In practice, regression coefficients could be reused from other families who have already used the system. The ML is re-trained as human ratings become available, and as new artifacts are uploaded to Kurator, the ML predicts ratings for them. Also, our goal for the ML is to remove low quality artifacts, in a reduction-oriented strategy, because low quality audio recordings are likely to have more objective characteristics (e.g., noisy or blank recording).

For feature selection, we analyzed the KidKeeper data. The most common recordings captured with the KidKeeper device were *songs*, *stories*, and *screaming gibberish*. In order to characterize these recordings, we used these general principles: (P1) screaming produces higher average amplitude than talking, (P2) singing or talking has more frequent and dramatic changes in amplitude than constant screaming, random noise, or blank recording, and (P3) longer recordings contain more content and are more likely to have interesting content. The features currently implemented are root mean square (RMS) of the spectrogram (addresses P1), RMS of the peaks in the spectrogram (addresses P2), duration of audio (addresses P3), and ratio of the peaks to the raw RMS (P1 and P2).

Note that our aim is not to create a state of the art ML classifier to eventually replace humans in the loop. There are many other speech classification features and methods we did not incorporate, such as emotion detection (e.g., [46] and [34]), speech activity detection (e.g., [45]), and age and gender detection (e.g., [37] and [18]). Because the problem of *personal* digital content curation is highly subjective, we assume the ML is limited and will eventually fail on some content, no matter how sophisticated. In this study, we wanted to know whether the family could guide the crowd where the ML failed. With our current implementation, we can test this easily and determine whether Kurator is robust to a limited ML. Future implementations can use more sophisticated ML.

Integrating crowd input

Design Rationale: Clearly, machine learning has limits, particularly on highly subjective tasks like personal digital curation, where "personal importance" is a key criterion for users in their decision-making [5]. As discussed above, crowdsourcing has been used on subjective tasks and in personal digital media curation directly [10] and indirectly [42, 50]. Thus harnessing the power of crowdsourcing seemed to be a promising approach to consider.

System Description: Our prototype implementation of Kurator uses Mechanical Turk as its paid crowd. It also uses Amazon's Simple Storage Service (S3) to store the audio files, making them read-accessible only for the duration of crowd tasks. Crowd tasks are automatically generated by Python scripts using Boto3, a Python interface to Amazon Web Services, to allow for API access to Mechanical Turk and S3. We built a crowd-tasking engine to interface with Mechanical Turk. This engine automates the workflow of creating HITs and collecting responses.

A HIT consists of the task description, a link to the audio file, and sections for subjective scoring and free-text feedback. We used, as the description of the task, the question: "*Do you think this audio could be meaningful to the content owner?*" Workers were given three options ("Definitely", "Maybe", "No Way") as well as a free-text feedback section to answer the question: "*Why did you rate it that way?*" We allowed three workers per HIT and used majority voting to determine the crowd's rating for each audio recording. Furthermore, the crowd's ratings were later used to re-train the machine learning classifier.

We investigated the effect of changing the wording of the crowd task question. We tested four questions on the same 30 audio recordings where we had ground truth data, and crowd workers were prevented from working on more than one question. The questions were:

A. *"Do you think this audio could be meaningful to the content owner?"*
B. *"Do you think the content owner would want to hear this again in the future?"*
C. *"Would you want to hear this again in the future?"*
D. *"If this were your child, would you want to hear this again in the future?"*

We found that Question A's ratings were the only ones with at least moderate agreement with the ground truth (kappa > 0.4). Question B showed fair agreement (kappa > 0.3), which suggests that framing the question as a judgment about the content owner's preference would elicit crowd responses that are in line with a parent's responses.

Integrating the family's expertise

Design Rationale: The crowd inherently does not have as much situated understanding as do family members. Therefore the family should add its expertise, but only in a cost-effective manner for them. Our goal was to have family members rate only the content that had been deemed possibly appropriate by the machine learner and the crowd.

System Description: We implemented a family-facing website built with Django and Bootstrap. Referring back to Daniel's activities in the scenario above, he interacts with

the website to access his family's personal content. As he listens to recordings, he provides ratings (similar to how the crowd provides ratings) and occasionally keyword tags.

To solicit ratings and text tags from the family, every audio playback web page contains a task description ("Would you want to hear this again in the future?"), clickable buttons to answer the question ("Definitely", "Maybe", and "No Way"), and a free-text area to submit keyword tags, limited to 140 characters (same as Twitter). The three categories of ratings are the same for the family and the crowd.

Note that Kurator is designed to incorporate family-sourcing as well, where multiple parents, other family members, and family friends can be included. Each person would have his or her own login, and the parent can restrict to whom to give feedback access.

Implementing feedback loops

Design Rationale: Since the family is the end user of Kurator, their subjective preferences need to be considered. As they are the experts for this task, it could benefit Kurator to allow the family to "shepherd" the crowd [11] by being the source of external feedback to crowd workers. Natural language descriptions have proven to be an effective way to guide crowd workers [11, 51, 26].

System Description: The Kurator website allows the user to provide guidance to the crowd. Family guidance is used verbatim in the tasks assigned to the crowds. Furthermore, the family's ratings, as well as the crowd's ratings, are used to re-train the machine learning classifier.

Summary

Kurator is a hybrid intelligence system that uses machine learning alongside mixed-expertise input from people (crowd workers and families) to weed out low quality artifacts. As Kurator's tiered process moves from machine learning to the family, the task requirements are increasingly subjective. The use of machine learning, crowd workers, and experts to collectively label items has been used in active learning [39], where the aim is to optimize labeling cost and accuracy by strategically deciding when to task experts to provide inputs. The design of Kurator differs because we assumed that experts were not "task"able, and their contributions might be sporadic and unpredictable.

USER STUDY AND EXPERIMENTS

To better understand whether Kurator's tiered architecture helped reduce the level of work for end users, we ran a user study with five families. To explore how guiding the crowd and re-training the ML improved system performance, we ran a series of focused follow-up experiments.

Study Design

The study participants were five families who had used the KidKeeper system for about a week to capture recordings of their children. The study was designed to obtain ground truth ratings as well as qualitative data from semi-structured interviews. The ratings were the categorical responses discussed above: "Definitely", "Maybe", and "No Way". Parents also chose their absolute favorites from their list of "Definitely"-rated recordings, with no minimum or maximum number suggested or required. We did this in order to evaluate how well Kurator could find a parent's favorite recordings, which we refer to as Favorites throughout the rest of the paper. The interview consisted of questions about parents' decision-making processes for their ratings, the difficulty of doing the ratings, what they thought about Kurator's selected recordings, and whether they could train others to recognize their preferences.

The size of each family's audio collections varied, ranging from 217 to 620 recordings. Because the limiting factor for the user study was the parents' time, having parents rate their whole collections would have been unreasonably burdensome. As well, there was a much larger percentage of No-Way's than keepers (Maybe's or Definitely's). Therefore we used stratified non-proportional random sampling [3] of a family's collection based on two independent ratings to allow us to oversample recordings in the Maybe's and Definitely's (n=40 for each of the 3 categories). The Kurator website presented one recording at a time to the user; each recording was randomly presented.

Baseline ML parameters. We needed to seed the machine learning classifier with a baseline set of parameters, so we trained the ML on rating data collected from other families not in the user study. Kurator used the same baseline for each family.

Kurator's Top K. The study included a comparison between the parent's Definitely's (including Favorites) and Kurator's selection for the top k recordings. We set $k=12$ (10% of 120) and played those recordings during the interview.

FINDINGS

In this section, we use the ground truth data collected from the user study in follow-on experiments, as well as interview data. In our evaluation, we use precision, recall, and F1 scores to measure Kurator's overall performance as well as the performance of its hybrid intelligence components: the crowd and the ML classifier.

Kurator worked

While every participant found the task of rating their children's recordings enjoyable, this sentiment was in tension with the time burden of listening to and making decisions about every audio recording. One parent remarked, *"it was fun to be reminded of those little clips...they were good to listen to again"*, but said, *"[if] you had to do it every couple of days, it would be annoying."* Another parent characterized this tension diplomatically as: *"It was enjoyable and I think it'd be enjoyable if there was less."* For at least some, then, curating recordings was a burden, even though they enjoyed the task while doing it.

Table 1. ML classifier's precision, recall, and F1 scores

Precision - Machine Learning		Participants					
		A	B	C	D	E	Total
Rating	Definitely	-	1.00	-	-	-	0.43
	Maybe	0.33	0.07	0.04	0.19	0.09	0.14
	NoWay	0.70	0.56	0.94	1.00	1.00	0.76
Recall - Machine Learning		*Participants*					
		A	B	C	D	E	Total
Rating	Definitely	-	0.05	-	-	-	0.03
	Maybe	0.65	0.50	0.75	0.95	1.00	0.76
	NoWay	0.70	0.70	0.31	0.18	0.28	0.38
F1 Scores - Machine Learning Classifier		*Participants*					
		A	B	C	D	E	Total
Rating	Definitely	-	0.09	-	-	-	0.05
	Maybe	0.44	0.13	0.07	0.32	0.16	0.24
	NoWay	0.70	0.62	0.47	0.30	0.43	0.51

Table 2. Crowd's precision, recall, and F1 scores.

Precision - Crowd		Participants					
		A	B	C	D	E	Total
Rating	Definitely	0.58	0.91	-	0.46	0.14	0.58
	Maybe	0.39	0.08	0.05	0.26	0.08	0.17
	NoWay	0.87	0.83	0.94	0.86	0.96	0.89
Recall - Crowd		*Participants*					
		A	B	C	D	E	Total
Rating	Definitely	0.38	0.52	-	0.43	0.67	0.43
	Maybe	0.61	0.38	0.75	0.52	0.50	0.56
	NoWay	0.75	0.80	0.48	0.66	0.50	0.60
F1 Scores - Crowd		*Participants*					
		A	B	C	D	E	Total
Rating	Definitely	0.46	0.66	-	0.44	0.24	0.50
	Maybe	0.48	0.13	0.10	0.35	0.14	0.26
	NoWay	0.80	0.82	0.64	0.75	0.65	0.72

Table 3. Kurator's precision, recall, and F1 scores.

Precision - Kurator		Participants					
		A	B	C	D	E	Total
Rating	Definitely	0.64	0.95	-	0.46	0.15	0.55
	Maybe	0.42	0.09	0.04	0.26	0.09	0.17
	NoWay	0.71	0.59	0.92	0.86	0.97	0.80
Recall - Kurator		*Participants*					
		A	B	C	D	E	Total
Rating	Definitely	0.31	0.31	-	0.43	0.67	0.31
	Maybe	0.45	0.25	0.50	0.52	0.50	0.46
	NoWay	0.87	0.92	0.57	0.66	0.54	0.67
F1 Scores - Kurator		*Participants*					
		A	B	C	D	E	Total
Rating	Definitely	0.42	0.46	-	0.44	0.25	0.39
	Maybe	0.44	0.13	0.08	0.35	0.15	0.25
	NoWay	0.78	0.72	0.70	0.75	0.69	0.73

There were two (not mutually exclusive) sets of preferences that parents followed. Some parents viewed Kurator as a tool to augment their curation work by reducing the overall workload. We called this preference *Best-Of* because the user wanted to hand-curate a reduced set. One parent remarked: *"I don't even go back and look at all 60,000 pictures that I have on my computer. If it's going to send me a smaller sample, I'm more likely to listen to all of them."* Another parent also elaborated on the benefits of working on a reduced collection: *"[Maybe if] it saved 10 minutes worth of samples, where it's small enough that you could sit down and kind of click through them quickly and figure out if you like it or not."* A third parent acknowledged her use for Kurator would depend on the frequency of her curation efforts: *"I would probably let [it] give me the top 20. If I knew this was going to happen once a week, I would let it do it for me. Yeah, I think I would just definitely choose the efficiency over making sure I captured every single moment."*

The benefit of Kurator's utility as a curation tool is further supported by our quantitative data. It was effective in refining families' collections by systematically removing non-keeper recordings. Table 3 shows that Kurator had 80% precision, 67% recall, and a 0.73 F1 when predicting NoWay audio clips. This suggests the tiered refinement approach may be a reliable way to winnow a collection down simply by removing artifacts of the lowest quality. In terms of raw numbers, Kurator removed 342 out of 600 (120 x 5 families) clips, with 78% precision, meaning for every four recordings the system filtered out, three of them were truly non-keepers. This metric suggests Kurator may be a useful tool for reducing the work for families who prefer to hand-curate a reduced set to find the *Best-Of*.

At other times, parents were less interested in winnowing down their collection so they could hand-select interesting clips, and instead were interested in finding the "sonic gems" [41]. They were happy when Kurator found a sufficient number of "gems" even if it did not find all of them, suggesting these clips are in an equivalence class. We called this second preference *Album*. As an example, one parent enjoyed when Kurator returned an audio clip of his young boy reciting the following line from the movie The Princess Bride: "My name is Inigo Montoya. You killed my father. Prepare to die." Another expressed an interest in anything she would find interesting enough to listen to again: "*I pretty much would definitely listen to all the ones that weren't garbage files. I liked all of the ones that were of them talking....*"

Although these findings indicate parents thought Kurator was useful for their purposes, we wanted to understand how well each tier supported the system's effectiveness. Next, we analyze the tiered architecture from several angles, then we analyze where Kurator did and did not work well, and then conclude with what the parents thought about privacy.

ML is effective at filtering non-keepers

As the first tier of our architecture, we wanted to know whether the ML was effective in terms of its quality of predictions (Table 1). Overall, the ML classifier had an F1 score of 0.51 in finding non-keepers (NoWay's). For the three most selective families (those who rated the least number of clips as Definitely's), ML had 100% precision

for two and 94% for the other when predicting NoWay ratings. This is likely due to these families strongly favoring the *Best-Of* approach, meaning they tended to rate a majority of their collections as NoWay. This increases the likelihood that an ML-rated NoWay agrees with the family.

The crowd is effective

As the second tier, the crowd was effective at identifying non-keepers (NoWay ratings), achieving 89% precision and 60% recall (F1=0.72) across all families combined (Table 2). The crowd was only moderately successful at predicting Definitely ratings (58% precision, 43% recall, 0.50 F1), but for one family, the crowd achieved 91% precision (n=35). This family rated significantly more Definitely's (n=62) than the other four families, which drives up the crowd's precision for Definitely's. Note that the recall for four of the families was ranged from 38% to 67%, meaning the crowd was able to uncover a significant subset of the Definitely's. The recall and precision of zero for Family C's Definitely's may have been a consequence of their curation preferences, which we discuss below.

Four crowd workers expressed their enjoyment of the rating task, via unsolicited emails to the research team. Two workers said they "loved" hearing these clips, commenting on the cuteness and hilariousness of the children's utterances. One worker even remarked: *"As mine grow up I wish I had saved so much more audio of them."* The crowd also divulged an interesting array of thought processes and criteria they used to make their decisions in their free-text responses in the tasks. Beyond frequent statements about recordings being "cute", "silly", and "adorable", workers often viewed specific activities they heard as being important to parents, such as singing, playing, and a *"child calling for her daddy...means so much"*. Some workers guessed about possible use cases that would make an audio clip valuable, such as: *"meaningful...if long distance"* or *"to a parent who isn't around at the time this occurred"*, *"put into a musical Christmas card...sent overseas if they have a parent in the military"*, and *"they might want to embarrass their kid when he's older; quite funny"*. Others made judgments, different from their own opinion, based on what they thought the parent would choose: *"I think this audio will only be meaningful to the audio owner...while cute, it doesn't mean a lot to people who do not know the child or have some context to go with audio."* Finally, many workers were willing to share personal thoughts about the audio recordings themselves: *"Reminds me of my kids"*, *"heart breaking child wishing for parents, so moving"*, *"I love kids just being kids"*, and *"children grow up so fast"*.

Kurator's tiered architecture is effective

The goal of Kurator's tiered architecture is to allow for contributor types with different strengths to be traded off. For example, we use machine learning as a scalable, cost-effective way to take a quick pass, but human insight (i.e., from the crowd) to make more accurate judgments.

Table 4 shows the tradeoffs induced by the crowd's performance as well as the tuning of the machine learning system used in our trials. Crowds are able to more accurately assess memories, but can be cost prohibitive. For 10,000 audio clips (~8 months of data for our average family) it would cost $2,100 to have the crowd rate them all. Adding the machine learning tier can reduce the cost by $711 if recall of Definitely's is optimized for (*Album*), or by $2,076 if precision of Definitely's is optimized for (*Best-Of*). This is exactly the intended effect of this architecture.

Since the machine learning component itself can be tuned to trade off precision and recall (along a classifier-specific ROC curve), users can adjust the effect of the classifier to fit their preferences. This feature was not implemented in this prototype, and is left for future work.

Table 4. Curation quality, reduction in user effort, and cost savings caused by the machine learning tier of Kurator. *Album* favors recall of Definitely's, and *Best-Of* favors precision for Definitely's (quality = precision for Definitely's, and %reduction = proportion of collection rated as NoWay)

	ML	
Regime	Album	Best Of
quality	76%	43%
%reduction	34%	99%
savings-600	$ 43	$ 125
savings-10k	$ 711	$2,076

Expert feedback improves the crowd and the ML

To evaluate the feedback mechanisms in Kurator, we analyzed the impact of re-training the ML using the crowd's and families' training examples. Then we analyzed the impact of re-training crowd workers from family-provided natural language feedback.

Training examples from crowds improves the ML

First, we evaluated if training examples provided by paid and expert crowds were beneficial to the ML. We re-trained the ML on each family, using the paid crowd's ratings as training examples, and then again using the family's ratings (note these are separate analyses, not a combined training). We report the averages from the five families.

After re-training on the crowd workers' ratings, the ML had a slight improvement in F1 score (0.51 to 0.54) when predicting NoWay ratings. After re-training the ML on the family's ratings, the ML scored much higher in F1 (0.51 to 0.66) for NoWay ratings. Although the crowd workers' inputs were helpful, the families' inputs had greater impact, on average, in helping the ML filter out the NoWay's.

Expert feedback improves the crowd's performance

Second, we obtained family-specific guidance to the crowd. This guidance came from responses to interview questions where parents were asked what they would tell strangers to help them rate the family's recordings. We selected two families for this experiment because they were able to articulate specific feedback to crowd workers. The other three families were able to come up with guidance, but it

was not as specific. An example from Family A, who provided guidance for their Definitely preferences, was: *"Choose 'Definitely' if it makes you laugh or if it gives you an emotional response."*

For each of the two families, we used their feedback to update the crowdsourcing task descriptions, and then obtained new ratings for all 120 recordings for each family. (Previous workers were prevented from working on these new tasks.) For both families, the crowd's F1 score for Definitely's increased (.46 to .57, and .66 to .78) but decreased slightly for NoWay's (.80 to .73, and .82 to .75). For both families, the guidance to the crowd included specific criteria for when to rate a recording as "Definitely", which seemed to cause the crowd to assign more Definitely ratings for each family than they did without this guidance. This increase in Definitely's led to more Definitely's being identified, and it also led to fewer NoWay ratings. This caused the Definitely recall to increase significantly (0.38 to 0.86, and 0.52 to 0.69) and the NoWay recall to decrease (0.75 to 0.62, and 0.80 to 0.70). This indicates families' natural language feedback can be used to convey context to crowd workers to help them assess personal digital content.

The crowd outperforms the ML before and after re-training

Finally, when comparing the ML's re-training on examples provided by the same two families, the re-training resulted in approximately the same F1 scores for NoWay's (0.70 to 0.68, and 0.62 to 0.65). However, the recall of NoWay's improved (0.70 to 0.78, and 0.70 to 0.90). For these two families, this means the family's training examples caused the ML to trade off precision for recall (i.e., the ML found more NoWay's), but it did not affect its F1 scores.

Comparing the re-trained crowd to the re-trained ML for these two families, the crowd's F1 for NoWay's was higher than the ML's (by 0.05 and 0.10 for Family A and B). More significantly, the crowd's F1 for Definitely's was substantially higher (by 0.50 and 0.35 for Family A and B, respectively). We remind the reader that crowd workers were already more accurate, in terms of F1 for Definitely and NoWay ratings, than the ML even before re-training.

These findings indicate that the crowd is more accurate than the ML, and even after feedback from experts, the crowd is still better. We unpack the implications of this in the Discussion section.

What happened?

To understand more about where Kurator differed from the preferences of the parents, we compared Kurator's selection of Favorites, the top *k* on each family's Definitely list. We used this to understand what criteria parents were using and how they differed from what Kurator determined.

Overall, as indicated before, parents were generally satisfied with Kurator's selection. Although all parents enjoyed listening to their families' recordings, one parent stated she preferred Kurator's list over having to listen to 120 recordings. Another parent echoed the desire to have less to listen to. This indicates Kurator's ability to reduce family burden by reducing the number of clips to rate, is in line with at least some parents' desires to have this burden reduced for them.

Another parent was pleased that Kurator caught an important clip she had overlooked in her ratings. As mentioned, Kurator found cute clips such as the Princess Bride quote mentioned above. For that clip, the crowd rated it as a Definitely, and one worker remarked: "*It's really cute but dark! [It] would make a parent laugh*." Other examples, such as a 2-second recording of a parent's two daughters laughing and making unintelligible, silly sounds, suggest the crowd was able to find content likely to be meaningful to parents even without much linguistic content.

Kurator also missed some clips. Some cases where responses were counted as incorrect did not have an impact on the users. For example, when duplicate clips (multiple recordings with the same content) were present, Kurator sometimes included all of them in its top *k*, or it would pick a different one than the family chose as a Favorite. This artificially decreased Kurator's measured performance.

One parent had two clips in her collection of her daughter saying "My name is Allie." Although they sounded almost exactly the same, in one of the recordings, the parent heard her daughter use her "home voice", and thus selected that clip as the Favorite out of the two, although she would have been happy with either:

"One example is Allie had two and they were basically exactly the same, but I picked one because it sounded more like what she sounds like at home. She's very shy and she doesn't talk a lot to other people, so only really us and our family know what she really sounds like."

Similarly, another parent had four clips of her son saying the same thing. She marked them all as "Definitely" in her first pass through. However, when she reviewed her selections, she only marked one as a Favorite: "*I think I had saved Henry saying I love you 4 times, but then I [de-selected] 3 of them. I don't need him saying it 4 times.*" Kurator however classified three of the four "*I love you*" in as top picks. She was not upset about this near-miss, although she only saved one to avoid duplication.

This ability to pick up on meaningful content was a key strength of the crowd. Whereas automated curation strategies depend on surface-level features, and parents had a certainty drawn from their in-depth knowledge of their children and their own preferences, the crowd nonetheless was able to draw on its own experience to guess what might be meaningful quite accurately. We return to this "common understanding" in the discussion section.

At times, Kurator severely missed. Often it was because of very specialized and idiosyncratic knowledge that only the

family possessed. We believed this would be true, although future iterations of Kurator need to take them into account.

The one parent who expressed dissatisfaction with Kurator's list was disappointed that the list she heard included recordings mostly of only one of her three children: "*I didn't hear a lot of Sally or Michael. Yeah, it was mainly Max....I need to hear Sally and Michael. I'm an equal rights mom. All my children get to have one each.*" This comment potentially reflected curation preferences that favored representativeness in what was kept.

Where the ML and the family differed, the clips were unremarkable at the signal level: they were monotonic or quietly spoken. For three of the four clips where the crowd and parents differed, workers had trouble understanding poor articulation of otherwise normal words. The fourth crowd-filtered Favorite was a whistle being blown, which the crowd deemed as "just noise". The parent explained:

"*That was when we went for Bella's birthday and they all got those Chuck-E-Cheese whistles. [The kids blew the whistles] in the car the whole way home and the whole next day. Of course, [KidKeeper] got some..., it was hilarious!"*

The sound of the whistle was a trigger of a particularly sonic memory for this parent, but would be perceived by anyone else as just noise. The parent anticipated the obscure nature of the clip, and did not expect the system to catch it because "it was an inside joke." In these examples, "insider" context is required, and Kurator failed expectedly. This suggests that the ML and the crowd may need to be tuned conservatively for some families, which could decrease the system's precision. This might be a reasonable tradeoff for users preferring recall of these types of recordings. However, for users requiring a technical solution to catch these types of recordings, but who are unable to spend time filtering a much larger set of recordings, Kurator may not help them achieve all their curation goals.

Privacy concerns

Although the findings indicate Kurator was able to reduce the burden on families in curating personal digital audio recordings, they would be moot if the study participants had insurmountable privacy concerns. We found this was not the case, according to the participants. Surprisingly, perhaps, none expressed concerns with the audio recordings that where shared with the crowd.

There were three driving factors that made the parents feel comfortable. First, parents generally viewed their children's recordings as harmless. One parent recalled a specific recording where potentially private information was divulged, but concluded it was not an issue: *"What they said seems very harmless. I mean, besides Andrew saying his full name...but they [the crowd] have no connections. That seems harmless."* Another parent expressed they were *"not worried about anything that [the kids] would say"* but that their feeling *"might change as they get older."* Apparently, the young age of their children (8 years and younger) was a factor for at least one parent. This suggests using the crowd to curate audio content generated by young children is not a concern for at least some parents.

Second, parents were comforted by the way the audio recordings were created and anonymously provided to the crowd. Parents reasoned about privacy by comparing it to more familiar content-capturing technologies and concepts. The most common sentiment was that because audio was captured through manual interaction with KidKeeper [22], it was better than *"all-the-time, always-monitoring"* recording devices. Also, multiple parents worried more *"about cameras than audio"*.

Third, parents' familiarity with audio-based technologies and toys seemed to influence their views. Parents used a comparison to other technologies and concepts as a way to justify their lack of concern about audio-only, manually-captured recordings. One parent compared Kurator to owning an Amazon Echo for the past year, saying *"Just a year later we're more integrated with voice commands and voice accessibility. I think it's easier to accept the more widespread it is...if it was just part of normal life."* Another parent said her lack of concern was due to having similar "other toys", where *"you can send messages...where you can record and send them to each other."*

Most parents did not find recording their children as concerning, but some parents did speculate about privacy violations from inadvertently capturing adults' background conversations. Parents envisioned scenarios where inadvertent disclosure of private information might worry them, such as talking about *"social security numbers"*, *"work strategy stuff...names and dates"*, and *"taxes...stocks...financial stuff."* Future versions of the ML tier could filter adult voices.

Although our parents largely did not find privacy to be a concern, some parents might. One parent speculated she might not feel comfortable having others curate her private recordings: *"If...it was being curated by somebody else...even if it's not a financial thing, I'd still feel sort of uncomfortable.... Just knowing that they had heard that would make me feel uncomfortable."* This finding suggests Kurator will not fit everyone: admittedly, some parents will remain uncomfortable with other people "seeing" their content.

DISCUSSION

Kurator's use of mixed crowds and machine learning, in a tiered refinement approach, was effective at helping families reduce their curation burden. In addition, parents were generally happy with Kurator's selection of the top *k* recordings, even when Kurator did not catch some of their favorite recordings. In this section, we unpack reasons why Kurator did and did not work, implications for designing tiered architectures such as Kurator's, the promise of

leveraging specialized crowds, and the implications of designing for privacy-preserving curation systems.

Leveraging the crowd's common understanding

Curating subjective, semantic content has been theorized to be beyond the current capabilities of automated approaches. Barriers range from inability to make idiosyncratic judgments to the lack of needed contextual knowledge. Further, the criteria that those with "expert" knowledge have are difficult to fully articulate. Yet, with Kurator, the combination of crowd and machine was somehow able to be reasonably successful. The findings indicate that many crowd workers were able to pull from cultural assumptions to help them predict what personal content another person would find valuable.

Kurator had some obvious wins. As we pointed out, some clips were selected to keep by both the system and family. These clips were those that were commonly understood to be good and meaningful to a parent. These common understandings, for example recognizing that a child speaking about a parent was highly likely to be valuable or that "I love you" is worth saving, were important to Kurator's success. The crowd has been under-estimated in its capability to react sympathetically to a subjective task. It is actually a viable source of help for this class of problems.

There were also cases where Kurator consistently failed. In these cases, the value of a clip could not be surmised from its content alone. Sometimes, the value of a clip was relative to its role in a collection rather than just its own content. Other times, the value of a clip was highly contextual (e.g., "inside joke"). In each of these cases, Kurator did not have the necessary context or information to rate effectively. However for most of these cases, Kurator was expected to fail. It was not taken for granted by parents that there were some audio clips that would be impossible to rate effectively by anyone but them.

There were hard cases, however, where the value of a clip was more ambiguous. In these cases, the crowd's common understanding was not nuanced enough to recognize the full value of a clip. Yet, the crowd was not consistently wrong. For cases where the crowd missed the nuance, such as failing to recognize a child's message to her father to not leave for work, there were corresponding cases where the crowd actively recognized non-obvious semantic value and even crafted elaborate narratives to explain why they thought it might be valuable to a parent. The ability to recognize the value of some clips may be tied to workers' experience or ability to create a believable narrative for themselves about the potential value of the clip. We may be able to improve this in future versions.

A tiered architecture is flexible

The findings indicate that crowds are effective at predicting whether specific children's digital audio recordings will be valuable family memorabilia. As discussed, leveraging crowds, alone, is not a scalable solution. Combining crowds with machine learning increases scalability and decreases monetary cost, but it comes at a price in terms of system precision. Kurator's ML component was fairly effective, but the crowd was very effective. The combination of the two resulted in a precision that was less than what the crowd could achieve alone. For the tiered architecture, researchers must investigate more deeply what tradeoffs their users want, in terms of price versus precision.

Under conditions where expert users have limited interactivity with the system, as in the case of Kurator, we believe a tiered architecture system is more effective when it leverages expert users' feedback to train crowd workers instead of the ML. This is because crowd workers act as a force multiplier: as they receive expert guidance, they can "convert" it into more training examples for the ML, which improves the ML without requiring training examples directly from expert users. In our evaluation, the family could guide the crowd using natural language feedback, and this guidance improved the crowd workers' accuracy. We also found the crowd workers' training examples improved the ML's performance. This resulted in our expert users not needing to manually rate their digital content often, a useful benefit.

However, for users who have the time and inclination to filter much of their digital content, the tiered architecture approach will still be helpful – expert ratings could improve the ML, which, in turn, would result in the ML layer being more accurate in filtering out unwanted content. Granted, the ML in our prototype system is not state-of-the-art, and we believe more advanced ML mechanisms, such as interactive machine learning (e.g., CueFlik [14], Arnauld [15], and ReGroup [1]) could possibly do just as well as crowd workers. However, interactive ML would likely only be helpful to users who would interact with the system often (or "tightly coupled" with the system [2]), which we believe would not help an important sub-population of users who view this level of involvement as burdensome.

Focus on specialized crowds

Leveraging specialized crowds is an area that shows promise for potential improvement. If a subset of a paid crowd had specialized skills, particularly with how and what to curate for long-term preservation, Kurator could leverage their higher levels of expertise. At least some portions of the crowd workers ostensibly used an in-depth thought process when making their decisions about ratings. We believe there is some amount of common understanding from cultural assumptions[1], such as looking for "cuteness", but the stories the crowd members were telling indicate they were going past "cuteness" per se.

An investigation into specialized crowds would need to, first, identify them and, second, determine how to reliably

[1] We remind the reader that our Turkers were constrained to U.S.-based IP addresses.

go back to those specialized crowds for consistent input into the system. Identifying may be as straightforward as identifying the 45-65 year olds who have older children, or tracking workers who make comments about their own children. Reliably using them may be more difficult, but specialized tests may be helpful in automatically identifying, then using, those with expertise in curating. Future work will explore ways to identify and leverage specialized crowds (e.g., novel workflows).

Privacy

One significant obstacle to deploying crowd-powered content curation systems is privacy. To help partially address this issue, we used a less-identifiable medium (audio) that was captured in short snippets, kept user information private, used large distributed crowds anonymously, and randomly ordered content to prevent workers from "following" certain users or families. However, as the findings suggest, over time, families may share (accidentally or intentionally) content that contain sensitive data, personal information, or private content.

Prior work has shown that there are several means by which workers can access or even reconstruct shared sensitive user information [32]. Obfuscation methods such as audio warping and worker routing that minimizes the amount of information from one family that one worker sees can further improve the chances of safe use of crowd powered systems in our setting. Research has also explored how the intelligent division of content [23, 30] can help reduce the threat of information exposure. Future work should explore how the crowd's ability to assess the sentimental value of content is affected by these filters.

Also, specialized crowds could possibly be leveraged to act as a privacy guard. These specialized crowds would consist of trusted agents, identified through incentive mechanisms (i.e., bonuses for tasks) designed to reward workers for catching potential privacy breaches in advance. Indeed, specialized workers could be identified by seeding content with fabricated private information (e.g., made-up social security numbers, financial information, or names/addresses) in a proactive approach. Although this solution seems technically feasible, a systematic evaluation of this proactive approach is needed to validate its workability. Most importantly, an approach like this would need to assess whether users are comfortable with a specially-selected subcrowd of trusted workers. This notion of trustworthy strangers is not unheard of (e.g., escrow agents or trusted third parties in cryptography), and it may be a step towards addressing the seeming tension between privacy and curation, where some users need help curating their personal digital content but are uncomfortable with other people "viewing" that content.

Limitations

Our study explored the viability of using crowds and our tiered architecture for a curation task with a focused group of participants. A larger scale deployment over a longer period of time is needed to further explore questions about privacy, how people choose to trade off quality and cost, and how assessment of sentiment changes over time.

Another limitation of our study is that we collected rating data from only one parent in each family. A future direction for this work could be collecting ratings from multiple family members to allow for a deeper investigation into the variance of ratings within family, and between an expert crowd and a specialized paid crowd.

Finally, Kurator used a majority voting mechanism to determine the crowd's rating on a particular audio recording. We believe this could be less efficient than weighted voting. A promising future direction for Kurator, particularly in the context of a longitudinal study, would be to track crowd workers' rating accuracies over time and then use a weighted rating when a known worker is involved, yet another way to leverage mixed expertise in the paid crowd to benefit system performance.

CONCLUSION

In this paper, we introduced Kurator, a proof of concept for hybrid intelligence systems that uses mixed expertise between crowds in a tiered architecture to synthesize inputs from multiple layers of contributors, such as machine learning, the crowd, and the family. We applied Kurator to the problem of reducing the burden of curating a family's digital audio memories. Our results demonstrate that crowd workers can accurately assess content that parents find sentimental, and that their assessments are improved by natural language feedback from expert users (i.e., the family). As well, we showed that the expert layer of the tiered architecture adds a new structure to hybrid intelligence because of how training and communication between layers can be done. Specifically, we argued that it is more effective to leverage the time-constrained experts to train the crowd than train the ML. The crowd can then provide high quality labels, in large numbers, that can be used to train the ML.

Our interviews showed that families found Kurator useful because it was able to provide a more tractable set of results while still discovering important memories. The interviews also revealed that parents were not concerned with privacy issues when sharing their children's audio recordings with crowd workers, but parents were able to speculate about potential privacy violations. Our work did not fully address these issues, and a long-term solution would need to more fully account for the scenarios presented by the parents.

Kurator serves as a proof of concept for both the viability of intelligent curation support, particularly when structured as a hybrid intelligence system with a tiered architecture, the potential of using crowdsourcing even in settings where subjective preferences are required to correctly complete some tasks, and the potential of crowdsourcing for curation tasks.

ACKNOWLEDGMENTS
We would like to thank all of our study participants and our reviewers for their assistance.

REFERENCES

1. Amershi, S., Cakmak, M., Knox, W.B. and Kulesza, T., 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*,*35*(4), 105-120.
2. Amershi, S., Fogarty, J. and Weld, D., 2012. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 21-30.
3. Bernard, H.R., 2011. Research methods in anthropology: Qualitative and quantitative approaches. Rowman Altamira.
4. Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D. and Panovich, K., 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313-322.
5. Ceroni, A., Solachidis, V., Fu, M., Kanhabua, N., Papadopoulou, O., Niederée, C. and Mezaris, V., 2015a. Investigating human behaviors in selecting personal photos to preserve memories. In *2015 IEEE International Conference on Multimedia & Expo Workshop* . IEEE, 1-6.
6. Ceroni, A., Solachidis, V., Niederée, C., Papadopoulou, O., Kanhabua, N. and Mezaris, V., 2015b. To Keep or not to Keep: An Expectation-oriented Photo Selection Method for Personal Photo Collections. In *Proceedings of the 5th ACM International Conference on Multimedia Retrieval*. ACM, 187-194.
7. Chan, J., Dang, S. and Dow, S.P., 2016. Improving Crowd Innovation with Expert Facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM.
8. Chang, S., Kumar, V., Gilbert, E. and Terveen, L.G., 2014. Specialization, homophily, and gender in a social curation site: findings from pinterest. In *Proceedings of the 17th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 674-686.
9. Chilton, L.B., Kim, J., André, P., Cordeiro, F., Landay, J.A., Weld, D.S., Dow, S.P., Miller, R.C. and Zhang, H., 2014. Frenzy: collaborative data organization for creating conference sessions. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 1255-1264.
10. Cusano, C. and Santini, S., 2014. With a little help from my friends. Multimedia tools and applications, 70(2), 1033-1048.
11. Dow, S., Kulkarni, A., Klemmer, S. and Hartmann, B., 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM, 1013-1022.
12. Eaton, John W., David Bateman, and Søren Hauberg, 2009. GNU Octave version 3.0.1 manual: a high-level interactive language for numerical computations. CreateSpace Independent Publishing Platform. ISBN 1441413006, http://www.gnu.org/software/octave/doc/interpreter/.
13. Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K. and Sandvig, C., 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 153-162.
14. Fogarty, J., Tan, D., Kapoor, A. and Winder, S., 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 29-38.
15. Gajos, K. and Weld, D.S., 2005. Preference elicitation for interface optimization. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*. ACM, 173-182.
16. Guldogan, E., Kangas, J. and Gabbouj, M., 2013. Personalized representative image selection for shared photo albums. In *2013 International Conference on Computer Applications Technology*. IEEE, 1-4.
17. Gulotta, R., Sciuto, A., Kelliher, A. and Forlizzi, J., 2015. Curatorial Agents: How Systems Shape Our Understanding of Personal and Familial Digital Information. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3453-3462.
18. Hämäläinen, A., Meinedo, H., Tjalve, M., Pellegrini, T., Trancoso, I. and Dias, M.S., 2014. Improving Speech Recognition through Automatic Selection of Age Group-Specific Acoustic Models. In *Computational Processing of the Portuguese Language*. Springer International Publishing, 12-23.
19. Hara, K., Sun, J., Moore, R., Jacobs, D. and Froehlich, J., 2014. Tohme: detecting curb ramps in google street view using crowdsourcing, computer vision, and machine learning. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 189-204.
20. Huang, T.H.K., Lasecki, W.S. and Bigham, J.P., 2015. Guardian: A Crowd-Powered Spoken Dialog System

for Web APIs. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

21. Jones, J. and Ackerman, M.S., 2016. Curating an Infinite Basement: Understanding How People Manage Collections of Sentimental Artifacts. In *Proceedings of the 19th International Conference on Supporting Group Work*. ACM.
22. Jones, J., Merritt, D., and Ackerman, M.S., 2017. KidKeeper: Design for capturing audio mementos of everyday life for parents of young children. In *20th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. Accepted.
23. Kajino, H., Baba, Y. and Kashima, H., 2014. Instance-Privacy Preserving Crowdsourcing. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
24. Kim, J., Cheng, J. and Bernstein, M.S., 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 745-755.
25. Kittur, A. and Kraut, R.E., 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*. ACM, 37-46.
26. Kokkalis, N., Köhn, T., Pfeiffer, C., Chornyi, D., Bernstein, M.S. and Klemmer, S.R., 2013. EmailValet: managing email overload through private, accountable crowdsourcing. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, 1291-1300.
27. Kulkarni, A., Narula, P., Rolnitzky, D. and Kontny, N., 2014. Wish: Amplifying creative ability with expert crowds. In *Second AAAI Conference on Human Computation and Crowdsourcing*.
28. Kumar, P. and Schoenebeck, S., 2015. The modern day baby book: Enacting good mothering and stewarding privacy on facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 1302-1312.
29. Laput, G., Lasecki, W.S., Wiese, J., Xiao, R., Bigham, J.P. and Harrison, C., 2015. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1935-1944.
30. Lasecki, W.S., 2015. Crowd agents: interactive intelligent systems powered by the crowd. Ph.D Dissertation. University of Rochester, Rochester, NY.
31. Lasecki, W.S., Gordon, M., Koutra, D., Jung, M.F., Dow, S.P. and Bigham, J.P., 2014. Glance: Rapidly coding behavioral video with the crowd. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 551-562.
32. Lasecki, W.S., Teevan, J. and Kamar, E., 2014. Information extraction and manipulation threats in crowd-powered systems. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 248-256.
33. Law, E., Dalton, C., Merrill, N., Young, A. and Gajos, K.Z., 2013. Curio: A Platform for Supporting Mixed-Expertise Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*.
34. Le, D., and Mower Provost, E. 2013. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *Automatic Speech Recognition and Understanding*. IEEE, 216-221.
35. Marshall, Catherine C., 2007. How people manage personal information over a lifetime. Personal Information Management, 57-75.
36. Marshall, Catherine C, Sara Bly, and Francoise Brun-Cottan, 2006. The long term fate of our digital belongings: Toward a service model for personal archives. In *Proceedings of the Archiving Conference*, 25-30.
37. Meinedo, H. and Trancoso, I., 2011. Age and gender detection in the I-DASH project. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(4), 13.
38. Nejdl, W. and Niederee, C., 2015. Photos to Remember, Photos to Forget. *IEEE Multimedia*, 22(1), 6-11.
39. Nguyen, A.T., Wallace, B.C. and Lease, M., 2015. Combining Crowd and Expert Labels using Decision Theoretic Active Learning. In *Third AAAI Conference on Human Computation and Crowdsourcing*.
40. Obrador, Pere, Rodrigo De Oliveira, and Nuria Oliver, 2010. "Supporting personal photo storytelling for social albums." In *Proceedings of the International Conference on Multimedia*. ACM, 561-570.
41. Oleksik, G. and Brown, L.M. 2008. Sonic gems: exploring the potential of audio recording as a form of sentimental memory capture. Proc. BCS-HCI 2008, Vol. 1, British Computer Society, 163-172.
42. Organisciak, Peter, Jaime Teevan, Susan Dumais, Robert C. Miller, and Adam Tauman Kalai. 2014. "A Crowd of Your Own: Crowdsourcing for On-Demand Personalization." In *Second AAAI Conference on Human Computation and Crowdsourcing*.

43. Rader, E. and Gray, R., 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, 173-182.
44. Retelny, D., Robaszkiewicz, S., To, A., Lasecki, W.S., Patel, J., Rahmati, N., Doshi, T., Valentine, M. and Bernstein, M.S., 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 75-85.
45. Sadjadi, S.O. and Hansen, J.H., 2013. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. *IEEE Signal Processing Letters*, 20(3), 197-200.
46. Schuller, B.; Batliner, A.; Steidl, S.; and Seppi, D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53(9):1062-1087.
47. Simko, J. and Bieliková, M., 2011. Games with a purpose: User generated valid metadata for personal archives. In *2011 Sixth International Workshop on Semantic Media Adaptation and Personalization*. IEEE, 45-50.
48. Teevan, J., Liebling, D.J. and Lasecki, W.S., 2014. Selfsourcing personal tasks. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2527-2532.
49. Yakel, E., 2007. Digital curation. OCLC Systems & Services: International digital library perspectives, 23(4), 335-340.
50. Yi, J., Jin, R., Jain, S. and Jain, A., 2013. Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. *In First AAAI Conference on Human Computation and Crowdsourcing*.
51. Zhang, H., Law, E., Miller, R., Gajos, K., Parkes, D. and Horvitz, E., 2012. Human computation tasks with global constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 217-226.
52. Zhao, Xuan and Lindley, Sian E., 2014. Curation Through Use: Understanding the Personal Value of Social Media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, 2431-2440.