

1      **DynamicLabels: Supporting Informed Construction of Machine Learning Label  
2      Sets with Crowd Feedback**

3  
4      ANONYMOUS AUTHOR(S)  
5

6      Label set construction—deciding on a group of distinct labels—is an essential stage in building a supervised machine learning (ML)  
7      application, as a badly designed label set negatively affects other stages such as training dataset construction, model training, and model  
8      deployment. Despite its significance, it is challenging for ML practitioners to come up with a well-defined label set, especially when  
9      no external references are available. Through our formative study (n=8), we observed that even with the help of external references  
10     or domain experts, ML practitioners still need to go through multiple iterations to gradually improve the label set. In this process,  
11     there exist challenges in collecting helpful feedback and utilizing them to make optimal refinement decisions. To support informed  
12     refinement, we present DynamicLabels, a system that aims to support a more informed label set-building process with crowd feedback.  
13     Crowd workers provide annotations and label suggestions to the ML practitioner's label set, and the ML practitioner can review the  
14     feedback through multi-aspect analysis and make refinements from crowd-made labels. Through a within-subjects study (n=16) using  
15     two datasets, we found that DynamicLabels enables better understanding and exploration of the collected feedback and supports a  
16     more structured, flexible refinement process. The ML practitioners were also able to see surfacing conflicts and edge cases that could  
17     have been ignored. In addition, the crowd feedback helped ML practitioners to gain diverse perspectives, spot current weaknesses, and  
18     shop from crowd-generated labels. With DynamicLabels, ML practitioners can successfully gain concrete understanding and evidence  
19     from the crowd and make an informed refinement to iteratively improve the label set.  
20  
21

22  
23      CCS Concepts: • Human-centered computing → Interactive systems and tools.  
24

25      Additional Key Words and Phrases: crowdsourcing, machine learning, label set construction, artifact or system  
26

27      **ACM Reference Format:**

28      Anonymous Author(s). 2018. DynamicLabels: Supporting Informed Construction of Machine Learning Label Sets with Crowd Feedback.  
29      In . ACM, New York, NY, USA, 27 pages. <https://doi.org/XXXXXXX.XXXXXXX>  
30

31      **1 INTRODUCTION**  
32

33      A 'label' or a class, is a word or a phrase that explains a piece of data in a supervised machine learning (ML) model.  
34      A group of distinct labels works together as a 'label set' to provide the model with a set of candidate labels for  
35      classification [1]. For example, in classifying a clothing dataset, a label set may consist of four distinct labels: **top** ,  
36      **bottom** , **outer** , and **accessory** . The label set is provided to the annotators to construct a training dataset and is  
37      utilized as model inputs and outputs. It is utilized often for classification tasks, such as the fashion classification model.  
38  
39

40      Preparing a well-constructed label set is important to build a successful ML application. Building an ML application  
41      involves a multi-stage process, which includes (1) preparing the raw data, (2) building a label set, (3) using the label  
42      set to annotate the training data, (4) implementing and training the model, and (5) deploying the model. Every other  
43      stage in the process is highly interconnected with the label set building stage: an unclearly defined label set affects the  
44

---

45      Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not  
46      made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components  
47      of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to  
48      redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
49

50      © 2018 Association for Computing Machinery.  
51      Manuscript submitted to ACM  
52

53 outcome of the annotation, and an indistinct or low-coverage label set affects the performance of the model, which  
54 subsequently negatively affects the experience of the user in the deployment stage [23].  
55

56 When building label sets for classification models, ML practitioners usually refer to existing labeled datasets or  
57 theories (e.g., referring to existing psychology taxonomies for emotion recognition models) to come up with the label set  
58 [10], and iteratively validate and refine it with additional data [21]. The iterative refinement is essential for constructing  
59 a high-quality label set in many real-world situations. For example, applying a pre-established label set to real-world  
60 data requires revision of the label set to accurately represent the distribution of the data. In addition, building a label  
61 set from scratch for a new domain without a well-defined or representative label set requires a significant amount  
62 of feedback and consensus-building among ML practitioners. With multiple iterations, ML practitioners collect bad  
63 signals (e.g., low coverage, unclear distinction) about the label set and revise based on the signals to prevent possible  
64 downstream issues, which is critical to the success of the ML application [21].  
65

66 To further understand the practices and challenges of building label sets with iterative refinements, we conducted a  
67 formative study with eight ML practitioners with experience constructing label sets from scratch. ML practitioners, even  
68 with existing references or domain experts, found revision cycles to refine the label set to be important and necessary.  
69 During this process, they found it challenging to collect large-scale, fresh-perspective feedback to improve the label set.  
70 They also found it difficult to extract meaningful insights from the feedback and confidently decide on an optimal label  
71 set, due to many different aspects (e.g., clarity of each label, distribution of the data, clear boundary between the labels)  
72 they have to consider along with the uncertainty of each improvement decision.  
73

74 To support collecting meaningful feedback and making informed decisions for refining the label set, we propose the  
75 idea of inviting crowd workers to provide feedback about the label set from varying perspectives inspired by successful  
76 feedback mechanisms in the past [11, 19, 29] and utilize those feedback to design interactions to understand the data  
77 better and make sufficient refinements. With the crowd as potential users of the deployed model, having the crowd's  
78 collective opinions and suggestions on the ML practitioner-built label set will guide the refinements. Providing analysis  
79 support with crowd feedback will help ML practitioners make a more confident and knowledgeable refinement to the  
80 label set.  
81

82 To explore the proposed idea, we propose DynamicLabels, a novel system that supports ML practitioners to iteratively  
83 construct their label set with label feedback collected from the crowd. When an ML practitioner provides an initial  
84 version label set, crowd workers produce feedback by annotating with the ML practitioner's label set and making their  
85 own label set with the assigned data through the *feedback collection interface*. With the collected label feedback and  
86 suggestions, the ML practitioner is provided with multiple-aspect analyses of the feedback and a playground to test and  
87 iterate on their label set in the *label set refinement interface*.  
88

89 We conducted a 2-day within-subjects study comparing DynamicLabels with the baseline annotation system to  
90 examine how DynamicLabels supports an informed label set refinement with crowd feedback. 16 ML practitioners used  
91 two types of datasets (natural scene images and event fliers) for a classification model to construct and refine two label  
92 sets through a round of iteration. The feedback collection interface of DynamicLabels enabled collecting large-scale,  
93 diverse feedback from the crowd, which participants identified as meaningful and useful in understanding the crowd's  
94 perspectives and the weaknesses in their label set to make refinements. The refinement interface of DynamicLabels  
95 enabled a high-level understanding of the feedback, encouraged flexible refinements to be made, and supported a  
96 structured refinement process. In addition, it helped the participants spot possible issues and examine various refinement  
97 options. We also discuss how DynamicLabels can support various types of data as well as the goals of ML practitioners.  
98

In addition, we suggest further utilization of the crowd feedback in making better-informed decisions and discuss how DynamicLabels supports the construction of a user-centered model.

Our contributions are as follows:

- DynamicLabels, a system that supports ML practitioners' label set construction process with crowd feedback and feedback analysis. DynamicLabels consists of a feedback collection interface that collects annotation and label suggestions on the ML practitioner-built label set, and a label set refinement interface that supports ML practitioners to make comprehensive refinement decisions.
- Findings from the formative interview which examines the practice and challenges of the label set construction process.
- Findings from a within-subjects study that compares DynamicLabels with a baseline system using two datasets, which shows that DynamicLabels supports an exploratory and structured refinement process, as well as in-depth analysis of how participants utilized crowd feedback in making label set refinements.

## 2 RELATED WORK

To situate our research, we first investigate challenges and existing approaches in the label set construction process. Then, since this work utilizes crowd feedback to support ML practitioners' label set building, we discuss how crowdsourcing is utilized to support expert work. We also review decision-making supports enabled by large-scale data.

### 2.1 Label set construction for multi-class classification

When training a multi-class classification model, the ML practitioner should prepare an annotated dataset. Without external references, constructing a label set is more challenging as there is no standard practice in categorizing contents for a multi-class label set. One commonly used approach is applying clustering algorithms such as LDA [2] and EM with GMM model [30]. These algorithms work in an unsupervised manner and categorize data points to compose clusters. However, these algorithms are mostly limited to numerically represented structured data. When the data contents are complicated and unstructured, other additional numerically abstracting algorithms or models are required to use these algorithms. Furthermore, they often fail at achieving reliable performance because these algorithms may not work perfectly. In addition, machine-generated clusters may not have appropriate representations or labels for human understanding.

To mitigate the issues from machine-generated label sets, previous work has invited humans to participate in the taxonomy or label set construction process [3, 7, 8, 10]. Cascade [10] presents a crowdsourcing workflow where workers provide suggestions and vote for the best descriptions over iterations to generate reliable categories with the crowds. Alloy [8] suggests a human-machine hybrid workflow to cluster text clips. A machine categorizes the text clips leveraging the salient keywords identified by crowd workers, then they put additional effort into clustering machine-failed clips. However, they rely entirely on the crowd and the machine, which can leave out practical considerations that an ML practitioner could make with intuition and experience. In addition, Revolt [7] leverages disagreement of crowds' annotation on a data instance to build label sets. They motivate that the labels and the descriptions created from disagreements can support an understanding of potential ambiguities in the label set. This creates more opportunities for ML practitioners to review and apply subjective labels, but is only investigated in binary classification scenarios.

In this work, we extend from prior works supporting the label set construction process with human work. DynamicLabels investigates the label set process after considering multi-stakeholders' (the crowd's and the ML practitioner's)

157 perspective and for multi-label classification which is more complex. To the authors' best knowledge, this is the first  
158 work that investigates label set construction from the ML practitioners' perspective, with crowds as feedback givers.  
159

## 160 **2.2 Utilizing the crowd to support expert work**

161 We define the label set construction as an open problem where no one best solution exists, so offering a diverse range  
162 of responses would help ML practitioners find an optimal label set satisfying their needs. Previous studies have shown  
163 that crowd inputs can help expert work as feedback [11, 19, 29] and inspiration for improvements [4, 15, 26].  
164

165 Previous work leveraged crowd input as feedback to expert work. Voyant [29] collected structured crowd feedback  
166 on visual designs by providing five feedback types to the crowd. ProtoChat [11] collected multiple levels of feedback  
167 including utterance-level feedback and overall conversation feedback by asking questions while testing the conversation.  
168 CrowdCrit [19] introduced key sources in visual design for the crowd to refer to in making a critique, to collect detailed  
169 and actionable feedback.  
170

171 Other work emphasizes the importance of incorporating aggregated crowd opinion in high-level concept or design  
172 of a product whose end user is a wide range of the public. Sutton and Lawson [26] proposed democratizing emoji design  
173 and selection by reflecting on how the public recognizes and uses emojis. Brambilla et al. [4] proposed a collaborative  
174 development process of Domain-Specific Modeling Languages in which end users and crowd workers are invited to  
175 provide feedback on diverse concepts of the language. In addition, Perspective [25] provided a set of auxiliary images  
176 and guiding questions to identify a diverse set of atypical images.  
177

178 Our work is inspired by the approach of utilizing the crowds as feedback provider to expert work, successfully  
179 investigated in various domains including designers ([11, 19, 26, 29] and engineers ([4, 25]). DynamicLabels collects  
180 crowd annotations which can illustrate potential problems in the ML practitioner-built label set (e.g., confusion between  
181 labels or limited coverage of the label set). In addition, by asking crowds to design their own label set, DynamicLabels  
182 collects natural feedback with fresh perspectives on the label set. This allows ML practitioners to explore and examine  
183 various opportunities in early-stage.  
184

## 185 **2.3 Data-driven decision-making support**

186 To make an informed decision with crowd feedback, ML practitioners need to understand the feedback thoroughly.  
187 Many previous work has explored ways to present data in a way that users can easily comprehend and utilize [13, 29, 31].  
188 Voyant [29] automatically generates a word cloud with the collected feedback which is more helpful. Decipher [31]  
189 aggregates multiple feedback and provides a visualization tool to help the interpretation. Mudslide [13] helps teachers  
190 interpret the students' muddy points by visualizing students' feedback on lecture slides. Some studies have emphasized  
191 the importance of presenting a multi-faceted data analysis beyond aggregation [17, 28]. OpinionSeer [28] provides  
192 analysts with an interactive opinion visualization to easily explore the mined opinions. Kairam and Heer [17] used  
193 clustering techniques to leverage disagreement between crowd workers and showed that identified patterns could  
194 illustrate the worker characteristics as well as potential task problems.  
195

196 Data-driven decision-making also enables users to consider various alternatives before making a decision. For  
197 conference session scheduling, Cobi [18] uses preference and constraints data on papers and sessions and presents the  
198 preview of changes in the number of conflicts for each move or assignment action that users consider. ConceptVector [22]  
199 supports an interactive construction of lexicon-based concepts by showing relevant documents and keywords regarding  
200 concepts the user considers. In designing a content-based image retrieval system for pathologists, Cai et al. [6] introduced  
201 tools that users can refine the image search by region, example, and concept.  
202

Our label set refinement interface is inspired by previous work that supports a thorough understanding and consideration of various alternatives in decision-making, but is investigated in the unique context of label set construction. The label set refinement interface of DynamicLabels presents varying levels of crowd feedback, ranging from raw crowd annotation to estimated coverage and confusion, so that users can consider diverse aspects of the label set simultaneously to help users comprehend different aspects of the collected feedback. Also, users can preview the consequence of each change before making a refinement and construct and compare multiple versions of the label set.

### 3 FORMATIVE STUDY

To understand the practice and challenges of ML practitioners in building and refining label sets for multi-class classification models, we conducted an hour-long interview with eight ML practitioners.

The recruitment was done through various AI/ML communities online. The participants consisted of two ML research engineers, three ML engineers, two ML graduate researchers, and one AI planning product manager, and all had one or more experiences building label sets and datasets from scratch. The tasks they worked on were classification and object detection tasks, but the type of dataset varied from OCR tasks (involving business cards, receipts, and invoices), NLP tasks (involving online articles, chat messages, video transcript), to computer vision tasks (involving book covers, supermarket products, objects for autonomous driving). The size of the dataset they constructed ranged from thousands to ten-thousands, and the number of labels in the label set ranged from 10s to even 30s. The participants were compensated with KRW 50,000 (\$40.00) for the interview. During the interview, we asked questions regarding their past label set construction experience, the aspects they consider important in a label set, and the challenges and needs in constructing and refining label sets.

#### 3.1 Practice & Challenges

*Practice.* All participants described label set building as one of the most challenging processes in constructing a training dataset, as the process involves coming up with an entirely new label set (i.e., coming up with a set of clearly described labels as well as detailed descriptions of each label) which involves numerous decisions. Even with relevant taxonomies or experts in the domain, the participants mentioned that *additional modifications are crucial* for the purpose of ML model construction such as granularity or defining each label. For example, P7 (video context classification) referred to information classification taxonomy for initial label set construction, but had to largely modify the label set to fit to video context. P5 (psychological disabilities classification) worked with a professional psychologist but had to redefine the labels with more distinctive criteria considering the model. P8 described such modifications as “a decision area for the model builder to make.”

The participants described their practice of iteratively refining the label set as going through *verify-refine cycles*. They first sample a small proportion of the data to construct an initial version of the label set. Then, they sample a larger amount of data and use the data to annotate with the label set. By looking at the annotations made, they decide whether the current label set is clear and appropriate to construct label sets and models. When the feedback from the verification shows issues with the label set (e.g., too many incorrect annotations, mixed use of labels), the participants make refinements. They mentioned that this cycle continues until no major issues are found in the label set, then proceeds to annotate to build the dataset. When there is more than one person involved in the construction process (e.g., as a team or with external annotators), they would compare and discuss the conflicts in the annotation.

The participants also emphasized the importance of building a robust label set that can prevent latent issues, especially defining each label clearly and distinctly from each other while covering all edge cases. If not, wrong annotations can be

made due to a misunderstanding of the label set, potentially leading to biased dataset construction, poor quality of the model, and bad user experience with deployment. While some issues can be handled on the model side using existing techniques such as data augmentation, sometimes starting again from scratch is costly but inevitable. Considering the cost, they commented that they would rather iterate early in the label set construction (P2). Following the iterative nature of the label set construction, we identified three key challenges that ML practitioners face in the process.

*3.1.1 Lack of helpful feedback to improve the label set.* To improve their label set, it is common for ML practitioners to go through multiple feedback loops. The most prevalent way to collect feedback on the label set is by annotating using the constructed label set and spotting problematic data that cannot be covered or has conflicts. With the spotted data, the ML practitioner would make refinements to the label set until they are convinced that all major issues are resolved. While this approach helps in collecting problematic data and making refinements by adding labels or new descriptions for edge cases, the participants mentioned the limited help that annotations can provide. When the annotation is done by the ML practitioners themselves, it is more difficult to spot uncovered or conflicting data due to biases towards certain labels. P4 mentioned that “even if they go through multiple iterations within the team, there are always unexpected questions asked by the annotators.” To get fresh feedback on the label set, ML practitioners sometimes recruit external annotators. This is more effective than having the team annotate as the problematic data are collected based on the annotator’s perspective. However, often the ML practitioners “end up adding a bunch of rule-based descriptions” (P2), which results in inefficiency and confusion for the annotators in making the training dataset later. In addition, recruiting external annotators to perform annotation can be a troublesome and costly process (P7).

*3.1.2 Difficulty in comprehending meaningful insights from the feedback.* As mentioned previously, constructing an optimal label set is difficult due to the many aspects (e.g., clear distinction of the labels, clear description of the labels) and multiple stakeholders (e.g., annotators, the requester of the model, users of the model) that need to be considered together. While each ML practitioner has a set of criteria they consider important, there is no clear guideline on making an optimal label set, making it difficult for them to decide on the best label set. Thus, they rely heavily on feedback in the construction process.

This complex nature of label set construction makes extracting high-criticality insights from feedback challenging. When feedback about a label set is collected in the form of issues or annotation results, ML practitioners need to examine each piece of feedback and organize them to come up with a concrete revision item. However, it is challenging for them to both find critical feedback from a bunch of collected feedback and group them into a meaningful revision item. More specifically, P2 mentioned that “edge cases that lead to adding a description is relatively easy, while [those] that lead to a change in hierarchical structure and definition is very difficult to spot and decide.” For conflicting labels, the participants mentioned that when annotators use a different label than their original intention, they know that something is confusing. However, it is difficult to decide whether the situation is common and should be prioritized.

*3.1.3 Difficulty in utilizing the insights to make satisfactory changes.* After ML practitioners organize key insights from the feedback, they need to apply the insights to modify the label set. Understanding the insights does not mean that a suitable modification can be made to the label set, as a complex set of criteria must be considered. As a result, ML practitioners are often not sure about their changes and their consequences.

ML practitioners try other approaches to increase certainty in the decision-making process such as discussing with a team of ML practitioners. However, this can be time-consuming and difficult to reach a consensus. Even when a consensus is reached, there is no guarantee that the decision is optimal. The only way to know whether the decision is

313 optimal or not is by getting to the later process of the ML model construction process (e.g., dataset construction, model  
314 training) and seeing if any issues occur. Without being able to check, participants struggled and felt less confident to  
315 choose the right moment to proceed to dataset construction, and was “afraid that the label set will end up creating  
316 issues later in the model building process” (P3).  
317

### 318 3.2 Design Goals

319 Based on the interview results, we came up with the following design goals for a system that addresses the challenges  
320 ML practitioners face in iteratively building and refining label sets.  
321

322 3.2.1 *Collect helpful feedback on the label set from the crowd.* ML practitioners identified a need for collecting nutritious  
323 feedback in their label set construction process to find problems in the label set and make appropriate refinements.  
324 Specifically, during the interview, the participants mentioned the need to receive feedback from fresh perspectives,  
325 that more actively suggest possible changes, and on a larger scale to address as many issues as possible. Through  
326 crowdsourcing, a group of people having fresh, diverse perspectives can be recruited to collect large-scale feedback on  
327 the label set (C1). In addition, the crowd can also provide their own labels as suggestions to support the refinement  
328 process as well (C3).  
329

330 3.2.2 *Provide multi-aspect analysis to derive meaningful insight.* One major characteristic of label set construction is  
331 that there is no best practice for an optimal label set. Thus, ML practitioners face difficulty interpreting and obtaining  
332 meaningful insight from the feedback they collect. During the interview, participants stated that they mainly get a  
333 sense of problematic labels by examining the feedback in varying aspects. Likewise, showing the collected feedback in  
334 multiple aspects (e.g., highlighting conflict, showing edge cases, providing summary) can support ML practitioners to  
335 thoroughly understand the feedback and select ones to prioritize.  
336

337 3.2.3 *Help understand possible changes and consequences in the label set.* Even after extracting meaningful insights  
338 from the feedback, ML practitioners struggle to make confident changes to the label set due to the uncertainty of  
339 their action consequence. Actively supporting ML practitioners with possible label candidates or showing them the  
340 consequence of the label set with the changes will help the refinement process be more informed, and will lower the  
341 barrier to iterate on the label set.  
342

## 343 4 PROPOSED SYSTEM: DYNAMICLABELS

344 We present the design of DynamicLabels, a system that aims to support ML practitioners’ label set construction.  
345 DynamicLabels supports iterative refinement of the label set through two separate interfaces: the **feedback collection**  
346 **interface** and the **label set refinement interface**. The former is provided to the crowd to collect annotations and  
347 label suggestions on the ML practitioner-built label set, and the latter is provided to the ML practitioners for refinement  
348 with multiple analyses of crowd feedback.  
349

350 Overall label set construction workflow and the role of each interface in DynamicLabels are described in Figure 1. In  
351 DynamicLabels, label sets are constructed in tree form (Refer to Fig. 12), consisting of **labels** and **groups** to group  
352 the labels. Each label includes a label name and a description<sup>1</sup>. As for the scope of research, we allow a single label to  
353 be assigned to each image.  
354

355 <sup>1</sup>Following the label format from Google documentation (<https://cloud.google.com/ai-platform/data-labeling/docs/label-sets>)  
356

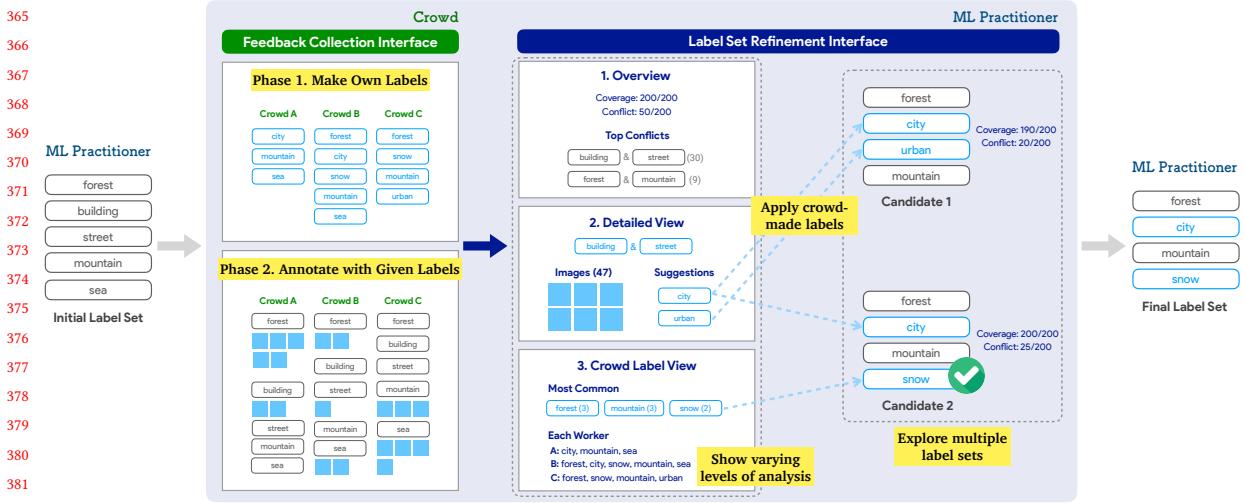


Fig. 1. A label set construction workflow using DynamicLabels. The initial label set of the ML practitioner is given to crowd workers with the feedback collection interface. The collected feedback is presented to the ML practitioner with the label set refinement interface to refine their label set.

#### 4.1 Feedback collection interface

For the crowd workers to use the feedback collection interface, the practitioner needs to have a constructed label set beforehand. This is similar to the practice in real-life settings, where the practitioners first build an initial label set.

The crowd is asked to provide feedback through two phases: (1) providing label suggestions by making the crowd's own label set and (2) annotating with the ML practitioner-built label set (Fig. 1-Footer Collection Interface). Two types of feedback are collected: passive (annotation results) and active (new label creation as suggestions), where an active suggestion aims to collect diverse perspectives on the label set.<sup>2</sup>

**4.1.1 Phase 1 - Providing label suggestions by making the crowd's own label set.** Crowd workers start from the Phase 1 task: creating their own label set (Fig. 2-a). They are first asked to take a look at 30 assigned images (Fig. 2-b) and come up with a set of labels (Fig. 2-a). Then, they are instructed to use the labels to make annotations (Fig. 2-c).

**4.1.2 Phase 2 - Annotating with the ML practitioner-built label set.** Crowd workers then proceed to the next phase and use the ML practitioner-built label set to annotate the same 30 images (Fig. 3). In addition to the ML practitioner-built label set (Fig. 3-a), the workers are provided with an additional `others` label to annotate images that do not fit into the provided label set to spot edge cases. For each image labeled `others`, the workers are asked to provide a brief reason (Fig. 3-d) to justify their choice.

**4.1.3 Post-processing of crowd feedback.** Crowd annotations and crowd-made labels are post-processed to provide meaningful analyses. To avoid suggesting redundant crowd-made labels, we merged multiple crowd-made labels into one if they are identical after stemming and lemmatizing.

As each crowd worker makes their own labels based on 30 images, the number of crowd annotations, or the number of images, for each crowd-made label is limited to 30 at most. To help ML practitioners better estimate the coverage

<sup>2</sup>To prevent biases, the crowd is asked to build their own label set before annotating with the ML practitioner-built label set.

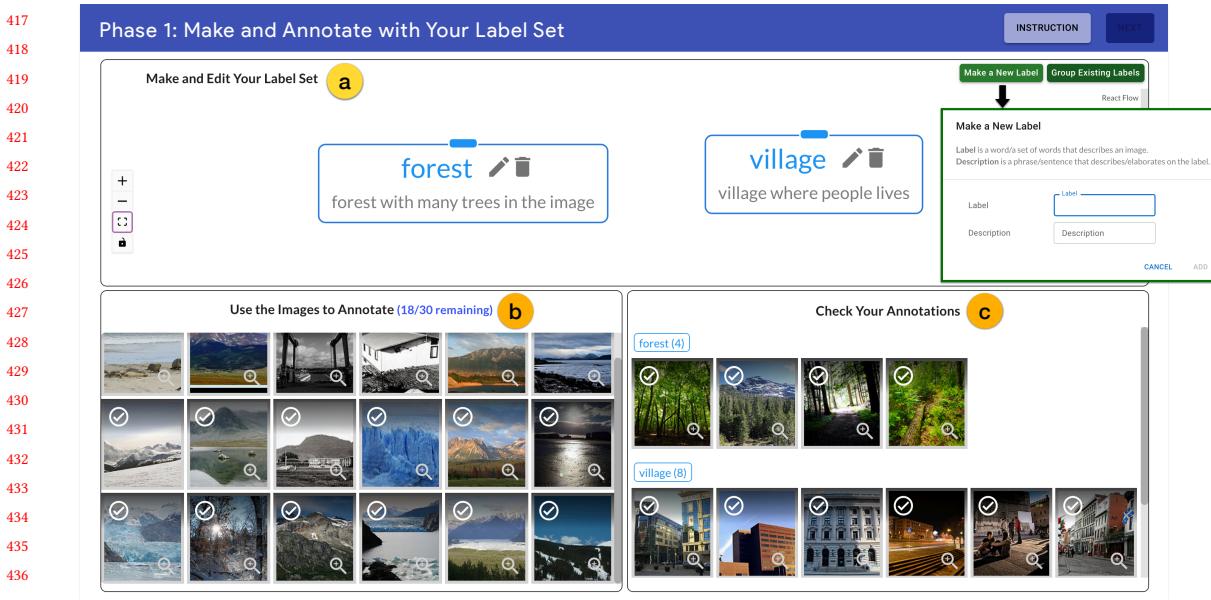


Fig. 2. Phase 1 of the feedback collection interface. The crowd workers are instructed to check the assigned images through (b) a grid of images on the bottom left and make their own label set on the (a) top component by adding, revising, and deleting the labels. For created labels, they can select the images in (b) to annotate the images, which will show up in (c), under each label.

and potential confusion of crowd-made labels, we established extended annotation for crowd-made labels. We first established similarity relationships between crowd-made labels and ML practitioner-made labels. For each crowd worker, we calculated the Jaccard similarity coefficient for each pair of a crowd-made label and an ML practitioner-made label, based on the crowd worker's annotation of 30 images for ML practitioner-made labels and crowd-made labels. For pairs with a similarity higher than 0.8, we assumed that the crowd label and the ML practitioner-made label are similar. Then, for each ML practitioner-made label, we filtered out images whose majority vote (of crowd annotation) match the label and established extended annotation between those images and crowd-made labels with high similarity with the ML practitioner-made label.

#### 4.2 Label set refinement interface

When a sufficient amount of feedback is collected for each image, the ML practitioner can revise their label set through the label set refinement interface (Fig. 4) with the following components: Your label set, Overview, Detailed view, Crowd label view, and Saved label sets. The interface supports reviewing and understanding the feedback with three different analyses (Fig. 4-b,c,d), and adopting the feedback to make changes with crowd-made labels (Fig. 5).

**4.2.1 Showing varying levels of analysis for the collected feedback.** When the ML practitioner enters the label set refinement interface, they can find their initial label set on the top left, under "Your label set". Right next to it is an **overview** (Fig. 4-b) that shows a summary created with the crowd feedback. Inside the overview, we provide four metrics motivated by the formative study, (1) Coverage: number of images with annotation, (2) Conflict: number of images annotated with multiple labels, (3) Top conflicts: top 3 label pairs with the highest number of conflicts, and

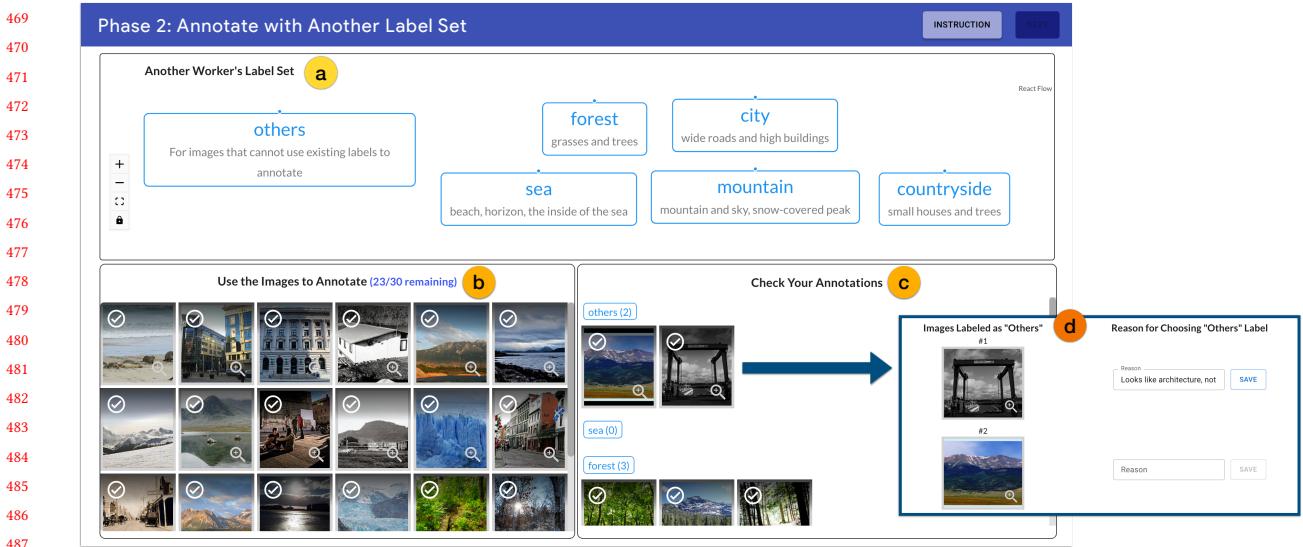


Fig. 3. Phase 2 of the feedback collection interface. The crowd workers are instructed to take a look at the (a) ML practitioner's label set and use the labels to annotate the (b) assigned images. Annotations will show up in (c), under each label. For images annotated using the “others” label, the workers are asked to provide a (d) brief reason each as an additional step.

(4) Unlabeled images: number of images without annotation to spot the main issue in their label set. The metrics are re-calculated when any changes are made to the label set. These metrics help ML practitioners understand how each label would be perceived and understood by the crowd and the coverage of labels as a set.

For the ML practitioner to understand the collected annotation in detail, they can select label(s), top conflicts, or unlabeled images to see a **detailed view** (Fig. 4-c). Here, the ML practitioner can see images annotated with the selected label(s) (current label set), images annotated using the conflicting labels (top conflicts), or images that are not labeled (unlabeled images) on the left. On the right, they can see possible refinement suggestions—a list of crowd-made labels which overlaps the most with the selected set of images.

On the bottom right, the ML practitioner can explore refinement options, through an analysis of the crowd-made labels through the **crowd label view** (Fig. 4-d). DynamicLabels shows crowd-made labels in two different aspects: (1) Most common labels and (2) Labels by each worker. The Most common labels component shows the top 10 frequently-made crowd labels, and the Labels by each worker component shows a list of crowd-made label sets, sorted by the number of labels made.

**4.2.2 Providing refinement support with crowd-made labels.** To support a more informed refinement, we allow ML practitioners to make additions or replacements to their label set using crowd-made labels. The refinement actions can take place from refinement suggestions or labels by each worker, which we illustrate in Figure 5.

On each refinement action, we display the **action consequence modal** (right of Fig. 5), where the change in the overview (the number of labels, coverage, conflict) is shown before making the change. After looking at the modal, the ML practitioner can decide whether to apply the refinement or not.

In addition to the detailed view, the ML practitioner can add crowd-made labels individually through the Crowd label view component (Fig. 4-d). The same action consequence modal is shown for this refinement as well.

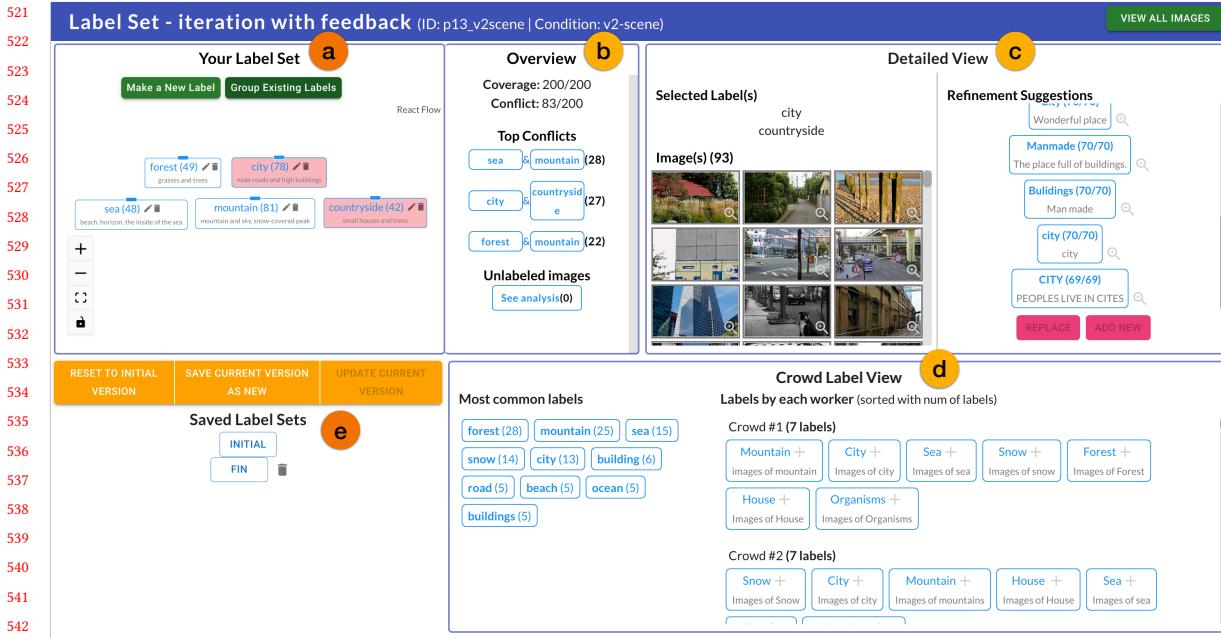


Fig. 4. Overview of the label set refinement interface. The (a) current version label set is displayed at the top left, along with an (b) overview of the collected feedback. By clicking the labels in (a) or top conflicts and unlabeled images in (b), the ML practitioner can see a (c) detailed view. On the bottom right, you can see the (d) crowd label view. During the refinement, you can save different versions of the label sets, which are displayed through the (e) saved label sets.

The ML practitioner can also directly add new labels, edit existing labels, or delete existing labels (Fig. 4-a), while the overview metrics stay the same.

**4.2.3 Creating and exploring multiple label set candidates.** On the bottom left corner, there is a **saved label sets** component (Fig. 4-e), which enables exploration of potential candidates with version control.

## 5 STUDY DESIGN

We conducted a 2-day study with 16 ML practitioners to investigate how DynamicLabels assists the ML label set construction process, through a within-subjects study comparing DynamicLabels to the baseline system (Described in Section 5.2.2). We chose this study design because of a high variance in ML practitioners' ML expertise, previous label set/dataset construction experience, and perception of crowdsourcing/crowd workers.

We aimed to answer the following research questions:

- (1) Can crowd workers produce helpful feedback with the feedback collection interface?
- (2) How do ML practitioners use crowd feedback to refine their label sets?
- (3) How do ML practitioners use DynamicLabels to make informed refinement decisions?

For the first RQ, we compare DynamicLabels of the collected feedback with that of the baseline system. For the third RQ, we compare and additionally explore how ML practitioners utilized the refinement interface through our

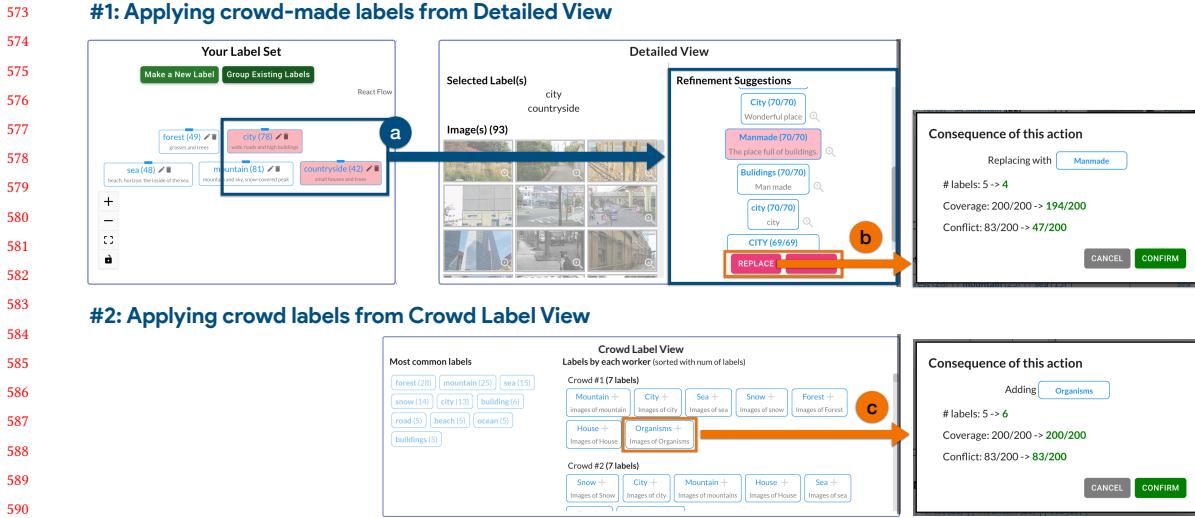


Fig. 5. Two possible ways to apply crowd-made labels to the current label set. In the top example, the ML practitioner can select (a) two labels in the current label set (`city` and `countryside`) to see a detailed view. From the refinement suggestions in the detailed view, the ML practitioner can (b) select crowd-made labels (`Manmade`) and click on the action (`replace`) to trigger the action consequence modal and make refinement decisions. In the bottom example, the ML practitioner can (c) click the plus icon next to the crowd-made label (`Organisms`) to add the label, which will trigger the action consequence modal.

suggested system. For the second RQ, we do not compare DynamicLabels with baseline but derive common patterns of ML practitioners utilizing crowd feedback in both conditions.

## 5.1 Participants

We recruited 16 participants by making an open call in several universities' online communities and social media targeting ML practitioners. Participation was limited to those with experience (1) manually constructing or utilizing label sets for ML models and (2) conducting industry or research projects using multi-class classification models that require label sets with multiple labels.

Among 16 participants, 3 were undergraduate students, 7 were graduate students, and 6 were industry workers. All participants had experience making label sets for classification models. We describe detailed demographics and the task they used for the study in Appendix A.1. Each participant was compensated KRW 120,000 (\$94.00) for a total of 4.5-hour 2-day participation.

## 5.2 Study Setup

**5.2.1 Task and Datasets.** For the study, we asked participants to design a label set for a multi-class classification model. We selected two types of data: natural scene image dataset [33] and event flier dataset (manually collected by the authors). We refer to the natural scene image dataset as *scene* and the event flier dataset as *flier* from below. We chose datasets that do not require a high level of domain expertise for the crowd to understand while having a varied modality, *scene* having images only, and *flier* having images and texts.

625 From each dataset, we randomly sampled 200 images for the study. Among the 200, we randomly selected and used  
 626 50 images for the initial label set construction, and all 200 images for the other two stages. We decided on the two  
 627 numbers (50 and 200) based on the formative practice, where practitioners normally conduct the first iteration with  
 628 hundreds of data.

629  
 630 **5.2.2 Baseline System.** Our baseline system (Figure 6) is designed with reference to the verify-refine feedback loop  
 631 described during the formative study, where (1) crowd workers annotate each image to one of the ML practitioner-  
 632 designed labels, and (2) the raw annotations and majority voting results are presented in the refinement phase (Figure 6).  
 633

| ID | Image | Worker 1  | Worker 2                   | Worker 3  | Worker 4                   | Worker 5  | Majority Label |
|----|-------|-----------|----------------------------|-----------|----------------------------|-----------|----------------|
| 71 |       | street    | buildings                  | house     | buildings                  | buildings | buildings      |
| 72 |       | buildings | buildings                  | buildings | buildings                  | buildings | buildings      |
| 73 |       | snow      | snow                       | snow      | mountain with snow and ice | snow      | snow           |
| 74 |       | house     | buildings                  | street    | street                     | buildings | No Majority    |
| 75 |       | snow      | mountain with snow and ice | snow      | mountain with snow and ice | snow      | snow           |

659 Fig. 6. How the feedback was provided to the practitioners in the refinement interface of the baseline system. On the (a) label view,  
 660 the user can see each label with images whose majority winner is the label and those without majority winners. On the (b) image  
 661 view, the user can see raw annotations and majority voting results for each image. The (c) majority label column on the image view  
 662 can be clicked to sort or filter results.  
 663

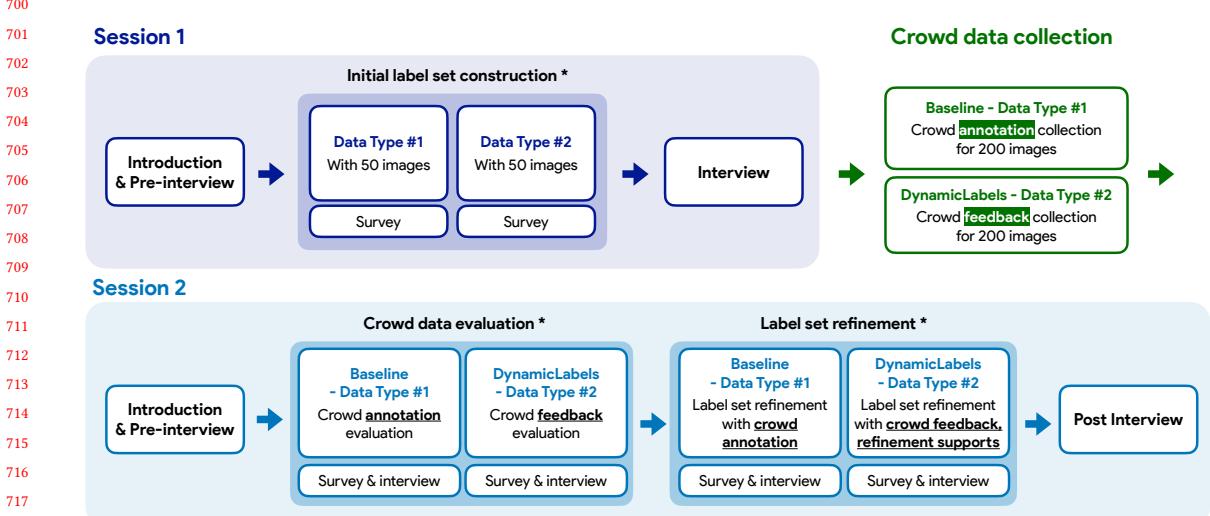
664  
 665 **5.2.3 Procedure.** The study was conducted with two sessions to simulate a single iteration of label set construction  
 666 (Detailed procedure in Figure 7). Each participant conducted the task in the baseline condition for one dataset and  
 667 DynamicLabels condition for the other dataset. The order and the image types assigned to the conditions were  
 668 counterbalanced and randomly assigned.

669 In the first session, the participants first created an initial version of the label set with 50 images. Constructed label  
 670 sets were used to collect crowd feedback (DynamicLabels) or crowd annotation (baseline) for 200 images. After each  
 671 label set construction, we asked the participants to fill out a 7-point Likert scale survey (1 = Strongly disagree, 7 =  
 672 Strongly agree) regarding the construction process. After the participants finished constructing the two initial version  
 673

label sets, we conducted a semi-structured interview regarding the construction process, the challenges in the process, and the participants' expectations of crowd feedback.

In between sessions 1 and 2, we collected crowd feedback (DynamicLabels) and annotations (baseline) on the label sets that participants constructed through Amazon Mechanical Turk<sup>3</sup>. For each label set, we recruited 34 crowd workers, and each image was annotated by five different workers. Each worker was assigned 30 images with a 24-image overlap with the previous worker. The workers were paid \$8.0 per hour for their work. We limited participation to U.S. workers who had completed at least 1000 HITs with an approval rate of at least 97%.

In the second session, the participants were instructed first to look at the raw data of the collected feedback/annotation and rate the data on a 7-point Likert scale to evaluate the helpfulness of the collected data, dimensions including good quality, large-scale, diverse opinions/perspectives, usefulness/meaningfulness of the data. The participants then refined their label set with the crowd feedback (DynamicLabels) or the crowd annotations (baseline) collected. After each refinement task, we asked the participants to fill out a 7-point Likert scale regarding the refinement process, dimensions including whether the process was structured, whether uncertainties were considered, whether alternatives were considered, and whether confident decisions were made (modifying Quality Decision-Making Procedure (QDMP) [5]). Afterwards, we conducted a semi-structured interview with the participant regarding the overall refinement process, utilization of collected crowd feedback/annotation, utilization of features in the refinement interface, and the refinement process and the final label set. After completing the two refinement tasks, we asked additional questions comparing the refinement process using DynamicLabels and the baseline system.



\* The order and data type for each condition were counterbalanced among participants.

Fig. 7. Tasks and procedure for each session. In session 1, each participant creates two initial label sets for each dataset. The label sets are given to the crowd workers to collect annotation or feedback depending on the condition. In session 2, the participant refines their label set with the collected crowd data presented.

<sup>3</sup><https://www.mturk.com>

**729 5.3 Measures**

730 We collected and analyzed the following data: crowd annotations, participants' session 1&2 label sets and refinement logs,  
731 task observations and interview responses, and survey results on crowd feedback/annotation and label set construction  
732 & refinement process.

733 *Crowd annotations.* We measure the collected crowd feedback in terms of diversity, helpfulness, and quality. We use  
734 the total and unique number of crowd-made labels, and the unique number of labels used to annotate a single image to  
735 measure the diversity, and survey and interview responses on crowd feedback/annotation to measure the helpfulness.

736 As a quality measure, we measured the accuracy of crowd annotations. We first filtered out 52 out of 1,073 workers  
737 who showed clear trolling behavior.<sup>4</sup> Then, we randomly sampled 2,000 crowd annotations for participant-made labels  
738 (500 for each dataset and condition) and 1,000 crowd annotations with crowd-made labels (for DynamicLabels condition,  
739 500 for each dataset) among 47973 annotations in total. With 3,000 sampled annotations, two of the authors coded the  
740 accuracy. As the collected labels were subjective, we considered annotations that appropriately describe or represent  
741 each image as correct (e.g., considering both buildings and city correct for the left image in Figure 9).

742 *Refinement actions.* We extracted each participant's refinement actions from session logs and recordings. Then we  
743 categorized each refinement action into seven categories: three label changes (add, revise, delete), one description-level  
744 change, and three group-level changes (add, revise, delete).

745 There were cases where multiple refinement actions were made to achieve one high-level refinement (i.e. split and  
746 merge). To analyze such high-level refinement actions made by participants, we grouped refinement actions made for  
747 one high-level split and merge refinement. Three of the authors analyzed three (out of 32) sessions together and then  
748 analyzed the remaining sessions individually.

749 We also recorded whether each refinement action was made based on crowd-made labels. In addition to direct use  
750 (adding or replacing with) of crowd-made labels, we also recorded actions where participants adopted crowd-made  
751 labels or descriptions.

752 *Interview responses and session observations.* Participants' think-aloud and interview responses were transcribed  
753 and analyzed through an open coding process, followed by focused coding. Two authors first individually developed a  
754 set of codes and combined the developed codes. Then, the two authors collapsed the developed codes into themes by  
755 identifying similar codes under each research question.

**756 6 RESULTS**

757 We first present an overview of the study results and then discuss each RQ in detail. The overview presents descriptive  
758 statistics of the label sets that participants made in sessions 1 and 2, refinements made in session 2, and the crowd labels  
759 and annotations collected for each participant's session.

**760 6.1 Descriptive Statistics**

761 *6.1.1 Label set construction and refinements.* In session 1, the participants constructed an initial label set with 50 images.  
762 The median number of labels was 7 (min: 3, max: 11) for scene and 9.5 (min: 4, max: 14) for flier. The initial label set  
763 construction on average took 21.4 minutes ( $\sigma=6.8$ ) for scene and 41.9 minutes ( $\sigma=16.4$ ) for flier.

764 <sup>4</sup>Those who used two or fewer ML practitioner-made labels to annotate 30 images or made out-of-context labels (e.g., making jacket or example1 in  
765 flier) were excluded.

In session 2, the participants refined their label set using the crowd annotations (baseline) or feedback (DynamicLabels) collected with 200 images. Table 1 shows the median number of labels and groups in the initial and revised label sets, and the number of net changes made in the labels and groups for each condition and dataset. Participants generated more labels with *flier* dataset (median: 9.5 with min: 4, max: 14) than with *scene* dataset (median: 7 with min: 3, max: 11). However, within each dataset, there was no statistically significant difference in the number of labels, groups, and changes between conditions.

Participants spent significantly more time refining the label sets with DynamicLabels than baseline and with *event* than with *scene* (two-way repeated ANOVA,  $F=8.38$  with  $p<0.05$  between conditions and  $F=4.84$  with  $p<0.05$  between datasets). With *scene* dataset, participants spent 16.8 ( $\sigma=10.6$ ) minutes for the baseline and 19.5 ( $\sigma=11.4$ ) minutes for DynamicLabels to refine their label sets. For the *flier* dataset, the average time spent was 17.4 ( $\sigma=13.1$ ) minutes for the baseline and 32.4 ( $\sigma=16.3$ ) minutes for DynamicLabels.

Table 1. Median number of labels and groups made in Sessions 1 and 2 and the median number of net changes in the label set, for each dataset and condition. Participants created label sets with more labels and a higher range in *flier* compared to *scene*, while there was no difference in the labels, groups, and net changes between DynamicLabels and baseline.

|               |             |                      | Natural Scene |               |               |               | Event flier   |              |               |               | Session 2      |                |
|---------------|-------------|----------------------|---------------|---------------|---------------|---------------|---------------|--------------|---------------|---------------|----------------|----------------|
|               |             |                      | Session 1     |               | Changes (net) |               | Session 2     |              | Session 1     |               |                |                |
|               |             |                      | Added         | Deleted       | Revised       |               | Added         | Deleted      | Revised       |               |                |                |
| Baseline      | # of labels | Median<br>[Min, Max] | 6<br>[3, 10]  | 1.5<br>[0, 6] | 1<br>[0, 7]   | 2.5<br>[0, 5] | 7<br>[3, 10]  | 9<br>[5, 16] | 1.5<br>[0, 8] | 1.5<br>[0, 9] | 2<br>[1, 6]    | 8.5<br>[6, 13] |
|               | # of groups | Median<br>[Min, Max] | 1<br>[0, 6]   | 0<br>[0, 1]   | 0<br>[0, 1]   | 0<br>[0, 1]   | 1<br>[0, 6]   | 0<br>[0, 4]  | 0<br>[0, 3]   | 0<br>[0, 2]   | 0<br>[0, 3]    | 0<br>[0, 0]    |
| DynamicLabels | # of labels | Median<br>[Min, Max] | 7<br>[5, 11]  | 2<br>[0, 4]   | 3<br>[0, 1]   | 0.5<br>[0, 4] | 7<br>[5, 10]  | 9<br>[4, 14] | 2<br>[0, 6]   | 1.5<br>[0, 4] | 1.5<br>[0, 10] | 10<br>[6, 15]  |
|               | # of groups | Median<br>[Min, Max] | 0.5<br>[0, 4] | 0<br>[0, 2]   | 0<br>[0, 1]   | 0<br>[0, 0]   | 0.5<br>[0, 4] | 0<br>[0, 1]  | 0<br>[0, 3]   | 0<br>[0, 1]   | 0<br>[0, 0]    | 0<br>[0, 3]    |

Table 2 shows the median number of refinement actions made in each condition for each dataset. With the baseline, a median of 8.5 (min: 2, max: 24) and 9.5 (min: 2, max: 23) refinement actions were made by participants *scene* and *flier* datasets, respectively. With the DynamicLabels, participants made a median of 7 (min: 0, max: 25) and 11 (min: 5, max: 15) refinement actions for *scene* and *flier*, respectively. The specific refinement actions made by each participant under each condition are shown in Figure 11. Figure 12 shows an illustrative example of label set refinement made with DynamicLabels.

Table 2. Median number of refinement actions made by participants in each condition for each dataset (Median [Min, Max]).

|               |               | Label         |               |               | Description   |             | Group       |             |             | Total          |
|---------------|---------------|---------------|---------------|---------------|---------------|-------------|-------------|-------------|-------------|----------------|
|               |               | Add           | Delete        | Revise        | Revise        |             | Add         | Delete      | Revise      |                |
| Natural Scene | Baseline      | 1<br>[0, 5]   | 0<br>[0, 6]   | 2.5<br>[0, 8] | 3.5<br>[0, 9] | 0<br>[0, 1] | 0<br>[0, 1] | 0<br>[0, 1] | 0<br>[0, 1] | 8.5<br>[2, 24] |
|               | DynamicLabels | 2<br>[0, 8]   | 2<br>[0, 9]   | 1.5<br>[0, 6] | 0.5<br>[0, 2] | 0<br>[0, 1] | 0<br>[0, 1] | 0<br>[0, 0] | 0<br>[0, 0] | 7<br>[0, 25]   |
| Event Flyer   | Baseline      | 0.5<br>[0, 8] | 1.5<br>[0, 5] | 3<br>[1, 9]   | 1<br>[0, 7]   | 0<br>[0, 3] | 0<br>[0, 4] | 0<br>[0, 0] | 0<br>[0, 0] | 9.5<br>[2, 23] |
|               | DynamicLabels | 2.5<br>[0, 5] | 1.5<br>[0, 8] | 1<br>[0, 5]   | 3<br>[0, 11]  | 0<br>[0, 1] | 0<br>[0, 2] | 0<br>[0, 0] | 0<br>[0, 0] | 11<br>[5, 15]  |

833    6.1.2 *Crowd feedback.* Between sessions 1 and 2, the crowd made annotations (baseline) or feedback—annotations and  
 834    label suggestions (DynamicLabels)—using the feedback collection interface. Table 3 summarizes the crowd feedback  
 835    collected for a single label set. For each ML practitioner-made label set, crowd workers made an average of 1053.63  
 836    and 1023.75 annotations for *scene* and *flier*, and 969.25 and 1023.75 annotations for *scene* and *flier* in the baseline  
 837    system and DynamicLabels, respectively. In addition, for DynamicLabels the workers made 179.75 labels (72.13 unique  
 838    labels) for *scene* and 199.25 labels (106.00 unique labels) for *flier* on average. The accuracy of annotations with ML  
 839    practitioner-made labels was 88.59% (*scene*), 69.18% (*flier*) with the baseline, and 90.52% (*scene*), 72.88% (*flier*) with  
 840    DynamicLabels. The accuracy of annotations with crowd-made labels was 91.60% in *scene* and 73.40% in *flier*.  
 841

842    For the baseline task, the average time spent was 827.57 seconds ( $\sigma = 1044.30$ ) in *scene* and 1770.96 seconds ( $\sigma =$   
 843    1484.30) in *flier*. For the DynamicLabels task, the average time spent was 1269.08 seconds ( $\sigma = 1401.17$ ) in *scene* and  
 844    2181.93 seconds ( $\sigma = 1679.29$ ) in *flier*. While the phase 1 task in DynamicLabels can be more mentally demanding than  
 845    phase 2, requiring workers to create new labels, the time spent is similar to or less than twice the time for that of  
 846    baseline. We presume that this was because the workers utilized the same set of images in phases 1 and 2, decreasing  
 847    the time needed to understand and become familiar with the data in phase 2.  
 848

849    Table 3. Average number of crowd feedback collected per single label set (Mean [Min, Max]): number of crowd labels, number of  
 850    unique crowd labels, number of annotations with crowd-made labels, number of annotations with ML practitioner-made labels, for  
 851    each dataset and condition.  
 852

|   | Natural Scene       |                    | Event flier          |                    |
|---|---------------------|--------------------|----------------------|--------------------|
|   | Baseline            | DynamicLabels      | Baseline             | DynamicLabels      |
| # of crowd labels                                 | -                   | 179.75 [143, 206]  | -                    | 199.25 [143, 229]  |
| # of unique crowd labels                          | -                   | 72.13 [47, 92]     | -                    | 106.00 [83, 137]   |
| # of annotations with crowd-made labels           | -                   | 969.38 [770, 1026] | -                    | 991.00 [813, 1031] |
| # of annotations with ML practitioner-made labels | 1053.63 [990, 1294] | 969.25 [784, 1024] | 1023.75 [1021, 1030] | 989.63 [812, 1028] |

## 864    6.2 RQ1: Can the crowd produce helpful feedback with the feedback collection interface?

865    Feedback from the crowd—both annotation and crowd-made labels—contained diverse viewpoints and helped participants  
 866    understand the various viewpoints of the crowd, some they had never expected before, and found it meaningful  
 867    and useful in making refinements.  
 868

869    6.2.1 *Diversity of Crowd-made Labels.* A single crowd worker on average created 5.58 labels for *scene* and 6.18 labels  
 870    for *flier*, summing up to on average 179.97 labels (*scene*) and 199.25 labels (*flier*) created per ML-practitioner's label set.  
 871    When we counted the number of unique labels after lemmatization, the number of unique labels was 72.13 labels for  
 872    *scene*, and 106.00 labels for *flier*. Considering that the average final number of labels that participants made was around  
 873    7 to 10, around 10x labels were provided to the participants per label set.  
 874

875    Crowd workers' diverse viewpoints were captured with the number of labels they created/utilized for a single image  
 876    (Figure 8, examples in Figure 9). For crowd-made labels, on average 3.77 labels ( $SD=0.43$ , *event*) and 4.41 labels ( $SD=0.47$ ,  
 877    *flier*) were used to annotate a single image, whereas for ML practitioner-built labels, on average 2.07 labels ( $SD=0.39$ ,  
 878    *event*) and 3.32 labels ( $SD=0.46$ , *flier*) were used. The difference between the number of crowd-made labels and ML  
 879    880

practitioner-built labels used was significant for both data types ( $t(398)=40.95, p < .0001$  for *scene*,  $t(398)=23.83, p < .0001$  for *flier*), exhibiting a wider perspective of the crowd in making their own labels.

In follow-up interviews regarding the crowd feedback, participants noted that the crowd-made labels helped them make various interpretations of the crowd workers in perceiving and recognizing the datasets. P15 commented that “through the crowd-made labels, I can see how people perceive and categorize the datasets, which I cannot understand through looking at the consequent annotations.”

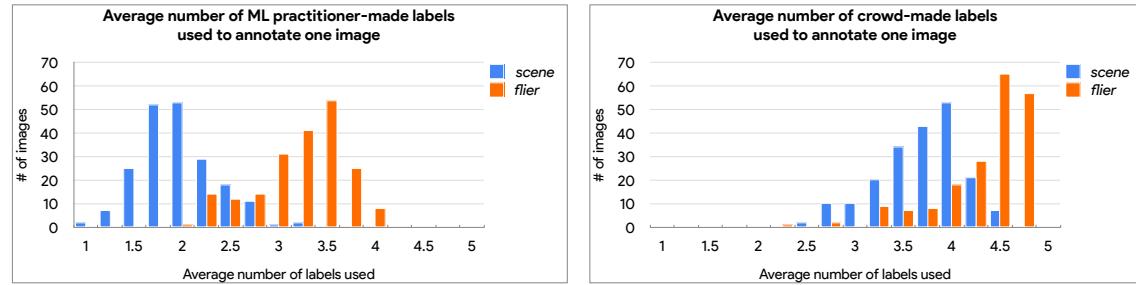


Fig. 8. Overall distribution of the average number of labels used to annotate one image. The left chart shows the average number of ML practitioner-made labels used, and the right chart shows the average number of crowd-made labels used for a single image.

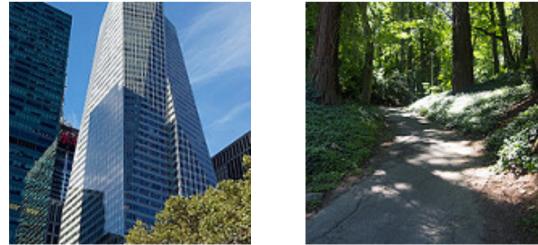


Fig. 9. Two example images in the *scene* dataset. Five different labels (civilization, city, town, buildings, metrocity) were made for the left *scene* image consisting of two tall buildings and a sky in the background, and three labels (forest, hills, trees) for the right *scene* image illustrating a trail inside a forest with trees.

6.2.2 *Perceived Helpfulness*. All participants commented that the crowd feedback was meaningful and useful in making refinements to the label sets. P14 said, “I was able to see a difference in my understanding and the crowd’s understanding of the label, [...] the annotation is different from what I expected. However, this will be helpful [to me in the refinement] as this tells me that my label is poorly defined.” Participants stated that seeing the crowd-made labels for each image in DynamicLabels makes up for the difficult-to-understand annotations, functioning as reasons. For example, P12 was confused about why a running flier was not being annotated as activity rather than sports, but understood the reason by seeing exercise and yoga labels made by the workers.

Participants especially liked the varying granularity of the crowd-made labels in DynamicLabels. P1 perceived different workers’ label sets as “an evolution of the label sets from the most general to the most specific.” They browsed through the crowd-made labels multiple times, thinking about ways to apply them to their own label sets.

**937 6.3 RQ2: How do ML practitioners use crowd feedback to refine their label sets?**

938 Both crowd annotations and crowd-made labels were utilized to help the participants understand and apply the feedback  
939 as refinements. The crowd annotations helped ML practitioners to (1) understand the crowd's general opinions and  
940 perspectives and (2) spot the weakness of their label sets, and the crowd-made labels helped them (3) explore and apply  
941 relevant ones to their label sets.  
942

943     **944 6.3.1 Understanding the crowd's general opinions and perspectives.** Participants mentioned that they could observe both  
945 converging and diverging opinions in the crowd feedback. Collective opinions were visible through *label suggestions*  
946 or *most common crowd-made labels*, and the participants compared their label set with the crowds' labels to see the  
947 similarity of their labels to the general crowd's labels. P3 mentioned "I am usually afraid of bias, especially when building  
948 a label set on my own. Seeing that many of the crowd have made similar labels with mine, I am more confident that [this  
949 is the] right direction." Further, P16 said that "looking at the most common labels helps to deal with the ambiguity in  
950 constructing a label set by oneself, and if there is a label that many crowd workers made, then that shows the necessity  
951 of that particular label."

952     In contrast, participants also observed the diverging opinions of the crowd in the collected annotations and labels.  
953 While there were overlapping crowd-made labels, participants found the labels to be overall diverging, which informed  
954 the participants how an image can be perceived differently by the workers. P13 said, "by looking at the crowd annotation  
955 results for labels and images I was unsure of in session 1, I am more confident that my label set should be defined more  
956 clearly."  
957

958     **959 6.3.2 Understanding the weakness of their label set.** They were also able to realize the weaknesses of their label  
960 set through the feedback. This was mostly done by looking at the actual annotated images using their label set. P2  
961 commented, "I would not have realized how poorly built my label set is without looking at the annotation results",  
962 and made major changes (adding 3, revising 3, deleting 1 label) to their label set in session 2. A common weakness  
963 identified by participants was the lack of good label descriptions, found by looking at the detailed view of each label,  
964 and ambiguous boundaries between labels, found by looking at the top conflicts. P9 commented, in *flier*, that they  
965 "would not have known that yoga fliers could fit into the 'nature' category without crowd annotations." They changed  
966 the description for **nature**, from 'Any poster which the main topic is about nature' to 'Any poster about plants, forest,  
967 and nature' to better specify the boundary of the label.  
968

969     **970 6.3.3 Exploring alternative labels and applying perspectives of the crowd.** Participants were also able to incorporate more  
971 perspectives from the crowd into their label sets. As DynamicLabels made crowd-made labels more visible throughout  
972 the entire refinement process, the participants were able to easily refer to the refinement suggestions, most common  
973 labels, or each worker's labels. A big portion of adding/revising refinements (*scene*: 35.1%, *flier*: 29.0%) in DynamicLabels  
974 utilized crowd-made labels. On the other hand, in baseline, no participants directly referenced crowd expressions (from  
975 'reasons for others') in adding/revising labels.  
976

977     Participants were also able to make more satisfactory refinements by referring to the crowd-made labels in DynamicLabels.  
978 In baseline, participants eventually got a sense of the labels that needed to be revised by extracting summative information  
979 from the annotation results, but the next challenge they faced was in making satisfactory refinements. Participants  
980 struggled to come up with satisfactory label names, which led to more label name changes happening in the baseline system.  
981 For example, in the baseline, P10 revised a single label three times, from **snow** to **snow/glacier without mountain**,

989 extreme cold with snow and glacier , then to extreme cold with snow, glaciers, and mountains and explicitly said  
 990 that the crowd-made labels would have been helpful to decide the label name.  
 991

992 As a result, participants made more changes in label name in the baseline (with a median of 2.5 for *scene* 3 for  
 993 *flier*) than in the DynamicLabels (with a median of 1.5 for *scene* and 1 for *event*) The difference between conditions is  
 994 statistically significant (two-way ART ANOVA [12, 27],  $F=5.34$  with  $p<0.05$ ).  
 995

#### 996 6.4 RQ3: How do ML practitioners use DynamicLabels to make informed refinement decisions?

997 Throughout the study, we were able to observe the distinctive benefits of DynamicLabels over the baseline system in  
 998 making more informed refinement decisions. When asked to compare the two conditions on how they helped their  
 999 refinement decisions, most participants (13/16) rated DynamicLabels better than baseline.  
 1000

1001 The results state that DynamicLabels supports (1) a high-level understanding of the feedback through metrics, (2)  
 1002 better assurance through examining multiple options, (3) a structured refinement process. In addition, DynamicLabels  
 1003 (4) encouraged a more flexible refinement and (5) surfaced issues that might have been missed in comparison to the  
 1004 baseline. Benefits of DynamicLabels supported participants to make more confident, efficient refinements with the  
 1005 crowd feedback.  
 1006

1007 *6.4.1 Metrics support understanding of and refinement from the feedback.* The most frequently identified strength of  
 1008 DynamicLabels was the existence of the metrics (coverage, conflict) in the Overview component. Participants liked  
 1009 how the metrics summarized the collected annotations and described the metrics as an efficient and intuitive way to  
 1010 understand their label set without looking at the raw data. P1 mentioned that, in the baseline, they had to make much  
 1011 more judgments by themselves such as understanding the reason behind images with no majority winner, deciding  
 1012 whether to change the label or not by estimating the expected effect of the change, and verifying whether the changes  
 1013 can fix the issues by going through the images again.  
 1014

1015 Among the metrics, the participants particularly found the conflict metric helpful and many (10/16) aimed at reducing  
 1016 the number of conflicts during the refinement task. They complimented the intuitiveness of the metric, in that “it  
 1017 intuitively shows the labels that are controversial to the crowd (P5)” and utilized the metric to identify which refinements  
 1018 should be prioritized. When a particular label existed in all three top conflicts (e.g., career & socializing , career &  
 1019 volunteering , sports & career for P9 in flier), the participants realized that the label can be confusing to the crowd  
 1020 and began their refinement by clarifying and examining the feedback from that label (e.g., career ).  
 1021

1022 *6.4.2 Examining and experimenting with various refinement options.* With DynamicLabels, participants were able to  
 1023 examine various refinement options before making a decision. With the consequence modal in DynamicLabels that  
 1024 shows the expected changes in the metrics for each refinement action, participants were able to examine each refinement  
 1025 action they considered before applying it. Participants found this consequence modal helpful, as knowing the expected  
 1026 change in the overview helped them make the decision more confidently. In addition, P13 mentioned that “the (action  
 1027 consequence) modal prevented them from making a wrong refinement choice.” When P13 tried to replace the label  
 1028 manmade for the conflict between the labels city and countryside , they saw the rise in the conflict and decided to  
 1029 take back the decision. They later merged the two labels into the label manmade . Even P3, who made no refinements  
 1030 with DynamicLabels, examined two refinements they considered but decided not to apply them.  
 1031

1032 *6.4.3 Establishing a structured refinement process.* The refinement process with DynamicLabels was perceived to  
 1033 be more structured than with the baseline. With DynamicLabels, participants began their refinement process from  
 1034

1041 the overview, then looked into the detailed view for further understanding, and referred to the crowd-made labels  
 1042 whenever they needed more assurance or references when making refinement decisions. Meanwhile, with the baseline,  
 1043 participants went back and forth between the label view and the image view until they identified the need for change.  
 1044 P6 noted that “in [v1 (baseline)], deciding on the starting or ending points was very challenging as I have to check  
 1045 the image and the annotations repeatedly to understand the outcome of the annotations.”. The participants found this  
 1046 implicitly conveyed workflow helpful, as they were able to “prioritize the refinement decisions (P15).” P4 also described  
 1047 DynamicLabels as supporting a more structured process that he could follow to figure out if it was the boundary of the  
 1048 label or the label name that needed to be changed, resulted in a quicker refinement with similar confidence.  
 1049

1050  
 1051 *6.4.4 Encouraging flexible refinement.* Participants noted that having various forms of crowd feedback in DynamicLabels  
 1052 helped them understand the relationship between labels, such as potential conflict or inclusion among labels. During  
 1053 the refinement session, P5 said that “by seeing the number of conflicts between **expo** and **social** and going through  
 1054 images with the conflict, I decided to split those labels into more [specific] ones.”  
 1055

1056 Participants also noted that with DynamicLabels, they could focus on how their label set represent the data,  
 1057 whereas they focused on clarifying each label and description in the baseline system. For example, P16 made three  
 1058 merge refinements (e.g., merging **city street** and **buildings** into **city**) with DynamicLabels whereas no high-level  
 1059 refinements were made with the baseline. P13 also mentioned that “[they] would not have combined **city** and  
 1060 **countryside** if [they] had not seen the label suggestion **manmade** after selecting both labels.  
 1061

1062 With both datasets, more participants made at least one high-level refinement in the DynamicLabels condition (6 out  
 1063 of 8) than in the baseline condition (3 out of 8). With both dataset, the median number of high-level refinements made  
 1064 by participants was 0 (min: 0, max: 3) in the baseline and 1 (min: 0, max: 5) in the DynamicLabels Table 4 summarizes  
 1065 the number of participants who made the split and merge refinement(s) in each condition and dataset.  
 1066

1067  
 1068 Table 4. Number of participants who made high-level refinements in each condition for each dataset  
 1069

|               | Natural Scene |       |       | Event Flyer |       |       |
|---------------|---------------|-------|-------|-------------|-------|-------|
|               | Split         | Merge | Total | Split       | Merge | Total |
| Baseline      | 2             | 1     | 3     | 1           | 3     | 3     |
| DynamicLabels | 2             | 5     | 6     | 3           | 4     | 6     |

1070  
 1071  
 1072 *6.4.5 Surfacing conflicts and edge cases that might have been ignored.* In addition to the refinements made, the participants  
 1073 were also able to spot possible conflicts and edge cases that they might have ignored. When refining with the  
 1074 baseline, most participants made refinements centered around the issues that they expected. P4 mentioned, “I checked  
 1075 that the labels that I assumed to be problematic actually had issues by looking at the annotations, and only revised  
 1076 those labels.” However, when refining with DynamicLabels, participants identified unexpected conflicts, and were able  
 1077 to understand where the conflicts were coming from and make suitable changes. Sometimes, participants appreciated  
 1078 the surfacing of unexpected issues even when they did not make refinements accordingly. P7 commented “If there were  
 1079 no crowd data, I would have made refinements solely based on my subjective opinions. Even if I did not reflect all  
 1080 crowd opinions, being able to see them helps me understand potential issues better.”  
 1081

1082 In DynamicLabels, P4 also created more labels by looking at individual crowd-made labels, commenting that “the  
 1083 crowd helped in detecting edge cases in the 200 images.” They added the labels **cave** and **desert** at the end of their  
 1084 refinement, after seeing images annotated with these labels and realizing their current label set couldn’t cover them.  
 1085

## 1093 7 DISCUSSION

### 1094 7.1 Potential use of DynamicLabels in different domains

1095 We believe that DynamicLabels can be generally expanded to domains that do not require special expertise. Among them,  
1096 we suggest a few domains where the benefit could be further amplified. For subjective domains where rules are decided  
1097 based on collective human judgments (e.g., sentiment classification (P3)), the ML practitioner can effectively understand  
1098 the general crowd's converging opinions and identify a convincing distinction between labels. For complicated domains  
1099 where having a large number of labels is necessary (e.g., receipt information extraction (P15)), the practitioner can  
1100 effectively use DynamicLabels to identify potential edge cases. If they have access to a group of domain experts, they  
1101 could utilize DynamicLabels for a comprehensive understanding of the data with experts' opinions. For example, P7  
1102 mentioned that if they could use DynamicLabels with a group of graduate students, they want to try label set building  
1103 for topic classification of research papers.  
1104

### 1105 7.2 Utilizing large-language models for DynamicLabels

1106 In DynamicLabels, the crowd is in charge of generating the feedback, and then the feedback is automatically filtered,  
1107 linked, and utilized on the review interface for interactive exploration. With the emergence of large-language models  
1108 (LLMs), we see opportunities for to apply them to different parts of DynamicLabels to streamline the process while  
1109 amplifying the benefits of crowd feedback and feedback review interface.  
1110

1111 For crowd feedback, LLMs can collaborate with the crowd to help alleviate the burden of having to generate labels  
1112 and annotate all assigned data by recommending labels or providing initial annotation for the crowd to correct. This  
1113 could support the crowd to focus on generating helpful feedback. However, this must be carefully designed so that the  
1114 initial feedback of the crowd is not hindered by LLM's work [20].  
1115

1116 For label refinements, the LLM could enhance the automatic mechanism we use to provide multiple analyses so  
1117 that the conflicts and the edge cases are better spotted [9]. In addition, refinement consequences could be presented  
1118 after automatic annotation of manually generated labels [32]. P16 wished for a feature that allowed ML practitioners  
1119 to manually add their own annotations to see a more accurate change in the overview, which could be more easily  
1120 enabled through LLMs.  
1121

### 1122 7.3 Providing various forms of crowd feedback and giving more control to ML practitioners over them

1123 In DynamicLabels, we present crowd feedback in various aspects through *overview*, *detailed view*, and *crowd label view*).  
1124 In our study, ML practitioners flexibly utilized these features in combinations to meet their needs, which can change  
1125 over the process, of understanding issues and the evidence behind them. At the same time, some participants expressed  
1126 the need to see raw crowd feedback, such as raw annotations of each image and crowd-made labels and annotations  
1127 before post-processing. Seeing that the participants made use of multiple features of their choice and were not reluctant  
1128 to examine more data with increasing complexity, presenting them with a more dynamic version of analyses that covers  
1129 a wide variety of crowd feedback will be useful in making confident refinements.  
1130

1131 Also, while the collected feedback was provided to the ML practitioners to explore, participants expressed the need  
1132 for more direct control over the collected feedback to reflect in the analysis. For example, participants wanted to filter  
1133 out trolling workers' feedback for the analysis, or give weights to the workers to explore how the conflict changes with  
1134 more reliable workers' opinions valued. Inspired by Jury Learning's approach [14] in allowing the model builder to  
1135 compose a group of juries and their opinions, DynamicLabels can have ML practitioners focus on a particular group's  
1136  
1137

1145 perspectives to construct the label set. Participants also wished to have more control over the crowd annotations by  
1146 fixing annotations they thought were wrongly annotated to indicate that particular labels or images had been examined.  
1147 With this control, ML practitioners can make sure that all feedback is considered and applied.  
1148

1149

#### 1150 7.4 Designing a more human-centered model with the crowd

1151 As the label set constitutes the primary structure of the model, aligning the user's mental model earlier in the label set  
1152 construction stage can be effective in incorporating the potential user's mental model into the model. An interesting  
1153 transition of the participants' behaviors we saw during the study was that they were considering and reflecting more  
1154 on the crowd's opinions, contrary to session 1 where they focused on creating a clear distinction among the labels for  
1155 better model training. P12 commented, "In session 1, I mainly built my label set considering the ML model by focusing  
1156 on the semantic aspects of the images. However, during session 2, I additionally considered the annotators, and ended  
1157 up splitting the labels more and grouping the labels for easier understanding. P12 was worried about the task becoming  
1158 difficult for the model, but still thought of this consideration valuable.  
1159

1160 We believe that DynamicLabels's impact can be enhanced far beyond building label sets with the crowd opinion,  
1161 but further in incorporating the crowd's—or the real users'—opinions in the process of building machine learning  
1162 applications. For example, in the dataset construction stage, the crowd can provide additional explanations or rationales  
1163 for each annotation, which can be utilized to train the model to generate more human-like explanations or logic in a  
1164 similar way to humans. In addition, crowd feedback can be used to collect large-scale opinions about the performance of  
1165 large models, and to come up with human-centered metrics to evaluate those models. When crowd feedback is utilized  
1166 in the later stages of model building (e.g., model building, model evaluation), the crowd can naturally learn about how  
1167 the ML model functions while providing feedback, enabling a natural human-AI interaction.  
1168

1169

### 1170 8 LIMITATIONS & FUTURE WORK

1171 We acknowledge several limitations of this work and discuss possible future work.  
1172

1173

1174 *System design.* While we applied quality control methods by providing the crowd workers with tutorial tasks or  
1175 warnings regarding poor work, the accuracy of the collected feedback was lower than the ML practitioners' expectations.  
1176 Subsequently, some participants lost trust in the crowd feedback, which hindered them from actively utilizing the  
1177 provided interface or ended up making passive refinements. We believe that additional quality control methods such  
1178 as an attention-checking task with a gold standard [16] could improve the feedback quality. Further, a more complex  
1179 crowdsourcing workflow such as worker deliberation methods (as proven effective in [24]) to identify which specific  
1180 cases are more diverse or require additional attention could help the ML practitioner prioritize the core issues.  
1181

1182 We also did not incorporate any machine learning approaches in post-processing or generating the analyses of the  
1183 review interface, as they caused latency in the system. Future works could incorporate such approaches and investigate  
1184 how to optimally utilize automatic techniques for a more efficient yet well-thought-out label set construction.  
1185

1186

1187 *Study setting.* In the study, we did not communicate the purpose of the dataset or the model to the crowd, which  
1188 would have led to the feedback being inherently diverse. Including that information would have resulted in higher  
1189 quality and relevant feedback, but also may have hindered the feedback from containing fresh and natural perspectives.  
1190 Investigating task designs to embed a clear purpose of the ML practitioner but also extract fresh and diverse feedback  
1191 could be valuable as the next steps of this research.  
1192

1193

To measure the effect of crowd feedback and the label set refinement interface, we conducted a comparison study with many factors (e.g., the number of images to look at, amount of crowd feedback) controlled throughout. Since different participants exhibited different patterns in utilizing crowd feedback, we wish to observe the long-term effects of DynamicLabels by testing the system in a more realistic setting.

In addition, the study focused on investigating whether DynamicLabels successfully supported an informed label set construction process, and did not validate the quality of the label set. While the ML practitioners were satisfied with the changes they made, a larger, longer-term study with more iterations to investigate the outcome of the workflow and its performance in the later processes of the ML model construction is a future step.

## 9 CONCLUSION

In this paper, we present DynamicLabels, a system that supports the process of label set construction with crowd feedback and interactive feedback analysis. Our study with 16 participants shows that DynamicLabels enables a more exploratory, flexible, and structured refinement process with fine-grained analysis and crowd-made labels. The crowd feedback helped the participants understand the general crowd's opinions and the weaknesses of their label set, and actively utilized the crowd-made labels for refinements. DynamicLabels suggests a new approach to building label sets for ML models, by incorporating the crowd's—or the general user's—feedback to reflect the user's diverse opinions and perspectives.

## REFERENCES

- [1] [n. d.]. Creating label sets. <https://cloud.google.com/ai-platform/data-labeling/docs/label-sets>. Accessed: 2023-01-15.
- [2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
- [3] Jonathan Bragg, Daniel S Weld, et al. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*.
- [4] Marco Brambilla, Jordi Cabot, Javier Luis Cánovas Izquierdo, and Andrea Mauri. 2017. Better call the crowd: using crowdsourcing to shape the notation of domain-specific languages. In *Proceedings of the 10th ACM SIGPLAN International Conference on Software Language Engineering*. 129–138.
- [5] Magdalena Bujar, Neil McAuslane, Stuart Walker, and Sam Salek. 2022. A process for evaluating quality decision-making practices during the development, review and reimbursement of medicines. *International Journal of Health Policy and Management* 11, 2 (2022), 128.
- [6] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
- [7] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.
- [8] 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.
- [9] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. [arXiv:2306.14924 \[cs.CL\]](https://arxiv.org/abs/2306.14924)
- [10] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1999–2008. <https://doi.org/10.1145/2470654.24466265>
- [11] Yoonseo Choi, Toni-Jan Keith Palma Monserrat, Jeongeon Park, Hyungyu Shin, Nyoungwoo Lee, and Juho Kim. 2021. Protochat: Supporting the conversation design process with crowd feedback. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–27.
- [12] Lisa A Elkin, Matthew Kay, James J Higgins, and Jacob O Wobbrock. 2021. An aligned rank transform procedure for multifactor contrast tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 754–768.
- [13] Elena L Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A spatially anchored census of student confusion for online lecture videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1555–1564.
- [14] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
- [15] Kosa Goucher-Lambert and Jonathan Cagan. 2019. Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation. *Design Studies* 61 (2019), 1–29.

- [1249] [16] Danula Hettichchi, Mike Schaeckermann, Tristan J McKinney, and Matthew Lease. 2021. The challenge of variable effort crowdsourcing and how  
visible gold can help. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–26.
- [1250] [17] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of  
the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.
- [1251] [18] Juho Kim, Haoqi Zhang, Paul André, Lydia B Chilton, Wendy Mackay, Michel Beaudouin-Lafon, Robert C Miller, and Steven P Dow. 2013. Cobi:  
A community-informed conference scheduling tool. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*.  
173–182.
- [1252] [19] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. 2015. Structuring,  
aggregating, and evaluating crowdsourced design critique. In *Proceedings of the 18th ACM conference on computer supported cooperative work &  
social computing*. 473–485.
- [1253] [20] Qianou Ma, Tongshuang Wu, and Kenneth Koedinger. 2023. Is AI the better programming partner? Human-Human Pair Programming vs. Human-AI  
pAI Programming. arXiv:2306.05153 [cs.HC]
- [1254] [21] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin,  
Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI  
Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI ’21). Association for Computing Machinery, New York, NY, USA,  
Article 94, 16 pages. <https://doi.org/10.1145/3411764.3445402>
- [1255] [22] Deokgun Park, Seungyeon Kim, Jurim Lee, Jaegul Choo, Nicholas Diakopoulos, and Niklas Elmquist. 2017. Conceptvector: text visual analytics via  
interactive lexicon building using word embedding. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 361–370.
- [1256] [23] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model  
work, not the data work”: Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [1257] [24] Mike Schaeckermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in  
crowd work. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–19.
- [1258] [25] Shahin Sharifi Noorian, Sihang Qiu, Burcu Sayin, Agathe Balayn, Ujwal Gadiraaju, Jie Yang, and Alessandro Bozzon. 2023. Perspective: Leveraging  
Human Understanding for Identifying and Characterizing Image Atypicality. In *Proceedings of the 28th International Conference on Intelligent User  
Interfaces*. 650–663.
- [1259] [26] Selina Sutton and Shaun Lawson. 2017. A provocation for rethinking and democratising emoji design. In *Proceedings of the 2017 ACM Conference  
Companion Publication on Designing Interactive Systems*. 7–12.
- [1260] [27] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses  
using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.
- [1261] [28] Yingcui Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. 2010. OpinionSeer: interactive visualization of hotel  
customer feedback. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 1109–1118.
- [1262] [29] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: generating structured feedback on visual designs using a crowd of non-experts. In  
*Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1433–1444.
- [1263] [30] Guorong Xuan, Wei Zhang, and Peiqi Chai. 2001. EM algorithms of Gaussian mixture model and hidden Markov model. In *Proceedings 2001  
international conference on image processing (Cat. No. 01CH37205)*, Vol. 1. IEEE, 145–148.
- [1264] [31] Yu-Chun Grace Yen, Joy O Kim, and Brian P Bailey. 2020. Decipher: an interactive visualization tool for interpreting unstructured design feedback  
from multiple providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [1265] [32] Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023. Assessing the potential of AI-assisted pragmatic annotation: The case of apologies.  
arXiv:2305.08339 [cs.CL]
- [1266] [33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition.  
*IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

## A APPENDIX

### 1301 A.1 Participant Demographics

1302 In Figure 10, we describe detailed demographics for each participant.

| 1303 Participant<br>(Occupation) | 1304 Previous Constructed<br>ML Label Set Domain   | 1305 Dataset (Condition, Task Defined)  |
|----------------------------------|--|---|
| P1 (Grad Student)                | Action Classification Task of Video Frames   | (1) Scene ( <i>Baseline</i> , Background video generation from input script) →<br>(2) Flier ( <i>DynamicLabels</i> , Events suggestions from user preference)         |
| P2 (Industry Worker)             | Segmentation Task of Airplane Images   | (1) Scene ( <i>Baseline</i> , Display scene image which suits the atmosphere) →<br>(2) Flier ( <i>DynamicLabels</i> , Reference searching for flier creators)         |
| P3 (Industry Worker)             | Classification Task of English Sentence Structures   | (1) Flier ( <i>Baseline</i> , Searching for similar events in the same category) →<br>(2) Scene ( <i>DynamicLabels</i> , Searching for similar scene images)          |
| P4 (Undergrad Student)           | Classification Task of Footprint Shapes  | (1) Scene ( <i>DynamicLabels</i> , Searching for similar scene images) →<br>(2) Flier ( <i>Baseline</i> , Reference searching for design elements)                    |
| P5 (Industry Worker)             | Text Classification Task of Industry SWOT Analysis<br>Emotional and Stress Level Prediction Task with Bio-Signals        | (1) Flier ( <i>DynamicLabels</i> , Events suggestions from user preference) →<br>(2) Scene ( <i>Baseline</i> , Recommending computer wallpapers)                      |
| P6 (Grad Student)                | Classification Task of IMU Sensor Data for Few-shot Learning<br>Classification Task of Speech Data for Few-shot Learning | (1) Scene ( <i>Baseline</i> , Scene detection for automatic image editing) →<br>(2) Flier ( <i>DynamicLabels</i> , Searching for similar events in the same category) |
| P7 (Grad Student)                | Evaluation Classification Task of Translated Topic Data  | (1) Flier ( <i>Baseline</i> , Events suggestions from user preference) →<br>(2) Scene ( <i>DynamicLabels</i> , Automatic categorization of captured images)           |
| P8 (Grad Student)                | Defect Classification Task for Welding Videos<br>Bounding Box Task of Weld Beads Data                                    | (1) Scene ( <i>DynamicLabels</i> , Data for text-to-image generation model) →<br>(2) Flier ( <i>Baseline</i> , Estimating cost for flier characteristics)             |
| P9 (Undergrad Student)           | Image Classification Task of Kickboard Riding  | (1) Flier ( <i>DynamicLabels</i> , Recommending other events in the same category) →<br>(2) Scene ( <i>Baseline</i> , Searching for similar scene images)             |
| P10 (Undergrad Student)          | Emotion Classification Task of Tweets  | (1) Flier ( <i>DynamicLabels</i> , Searching for similar events in the same category) →<br>(2) Scene ( <i>Baseline</i> , Location detection from scene images)        |
| P11 (Grad Student)               | Trajectory Classification Task of Flight Patterns  | (1) Flier ( <i>Baseline</i> , Searching for similar events in the same category) →<br>(2) Scene ( <i>DynamicLabels</i> , Location detection from scene images)        |
| P12 (Industry Worker)            | Diagnosis Classification Task of ECG Data  | (1) Flier ( <i>Baseline</i> , Searching for similar events in the same category) →<br>(2) Scene ( <i>DynamicLabels</i> , Region identification from scene images)     |
| P13 (Industry Worker)            | Emotion Classification Task of Facial Expression Video<br>Classification Task of Fashion Images                          | (1) Scene ( <i>DynamicLabels</i> , Region identification from scene images) →<br>(2) Flier ( <i>Baseline</i> , Interpretation of events with text and image)          |
| P14 (Grad Student)               | Classification Task of Sensor Data for Diseases Detection<br>Classification Task of Mask Wearing                         | (1) Scene ( <i>Baseline</i> , Recommending computer wallpapers) →<br>(2) Flier ( <i>DynamicLabels</i> , Searching for similar events in the same category)            |
| P15 (Grad Student)               | Emotion Classification Task of Comments Data   | (1) Flier ( <i>DynamicLabels</i> , Searching for similar events in the same category) →<br>(2) Scene ( <i>Baseline</i> , Categorization for scene image searching)    |
| P16 (Industry Worker)            | Intention Analysis Task of In-Company Data   | (1) Scene ( <i>DynamicLabels</i> , Region identification from scene images) →<br>(2) Flier ( <i>Baseline</i> , Categorization of fliers for advertisement display)    |

1341 Fig. 10. Detailed demographic information of each participant including occupation, prior label set construction experience, dataset,  
1342 condition, task defined, and order. We provided the two datasets to the users and asked them to identify a task before the beginning  
1343 of Session 1.

### 1344 A.2 Label Set Refinement Data

1345 In the below table (Figure 11, we describe the number of groups and labels for each participant in sessions 1 and 2, and  
1346 the refinement actions they made in between.

| 1353 | Participant    | Natural Scene Images |     |        |        |             |           |     |        |        |           | Participant | Event Flyers |     |        |        |             |           |     |        |        |           |    |
|------|----------------|----------------------|-----|--------|--------|-------------|-----------|-----|--------|--------|-----------|-------------|--------------|-----|--------|--------|-------------|-----------|-----|--------|--------|-----------|----|
|      |                | # of Group           |     |        |        | # of Labels |           |     |        |        |           |             | # of Groups  |     |        |        | # of Labels |           |     |        |        |           |    |
|      |                | Session 1            | Add | Delete | Revise | Session 2   | Session 1 | Add | Delete | Revise | Session 2 |             | Session 1    | Add | Delete | Revise | Session 2   | Session 1 | Add | Delete | Revise | Session 2 |    |
| 1354 | Baseline       | P1                   | 6   | 0      | 0      | 0           | 6         | 10  | 2      | 2      | 5         | 10          | P3           | 0   | 0      | 0      | 0           | 0         | 13  | 3      | 8      | 2         | 8  |
| 1355 |                | P2                   | 2   | 0      | 0      | 0           | 3         | 8   | 3      | 1      | 3         | 10          | P4           | 2   | 2      | 1      | 0           | 1         | 7   | 1      | 2      | 5         | 6  |
| 1356 |                | P5                   | 0   | 0      | 0      | 0           | 0         | 3   | 0      | 0      | 3         | 3           | P7           | 0   | 0      | 0      | 0           | 0         | 5   | 1      | 0      | 1         | 6  |
| 1357 |                | P6                   | 0   | 0      | 0      | 0           | 0         | 5   | 1      | 0      | 1         | 6           | P8           | 0   | 0      | 0      | 0           | 0         | 9   | 0      | 1      | 1         | 8  |
| 1358 |                | P9                   | 0   | 0      | 0      | 0           | 0         | 5   | 3      | 1      | 2         | 7           | P11          | 4   | 0      | 0      | 0           | 3         | 16  | 4      | 9      | 2         | 11 |
| 1359 |                | P10                  | 2   | 1      | 1      | 0           | 2         | 9   | 6      | 7      | 0         | 8           | P12          | 0   | 0      | 0      | 0           | 3         | 6   | 8      | 3      | 2         | 11 |
| 1360 |                | P14                  | 0   | 0      | 0      | 0           | 0         | 4   | 0      | 0      | 4         | 4           | P13          | 0   | 0      | 0      | 0           | 0         | 9   | 0      | 0      | 6         | 9  |
| 1361 |                | P15                  | 2   | 0      | 0      | 1           | 2         | 7   | 1      | 2      | 2         | 7           | P16          | 0   | 1      | 1      | 0           | 0         | 12  | 2      | 1      | 1         | 13 |
| 1362 | Dynamic Labels | P3                   | 1   | 0      | 0      | 0           | 1         | 7   | 0      | 0      | 0         | 7           | P1           | 0   | 0      | 0      | 0           | 0         | 14  | 3      | 2      | 0         | 15 |
| 1363 |                | P4                   | 2   | 2      | 1      | 0           | 3         | 8   | 4      | 3      | 2         | 9           | P2           | 0   | 3      | 0      | 0           | 3         | 14  | 1      | 0      | 10        | 15 |
| 1364 |                | P7                   | 0   | 0      | 0      | 0           | 0         | 11  | 2      | 3      | 1         | 10          | P5           | 1   | 0      | 1      | 0           | 0         | 7   | 2      | 2      | 0         | 7  |
| 1365 |                | P8                   | 0   | 0      | 0      | 0           | 0         | 7   | 1      | 3      | 1         | 5           | P6           | 0   | 0      | 0      | 0           | 0         | 11  | 2      | 3      | 1         | 10 |
| 1366 |                | P11                  | 2   | 0      | 0      | 0           | 2         | 6   | 0      | 1      | 1         | 5           | P9           | 0   | 0      | 0      | 0           | 0         | 11  | 0      | 0      | 3         | 11 |
| 1367 |                | P12                  | 0   | 0      | 0      | 0           | 0         | 6   | 4      | 4      | 0         | 7           | P10          | 0   | 1      | 0      | 0           | 1         | 6   | 4      | 0      | 2         | 10 |
| 1368 |                | P13                  | 0   | 0      | 0      | 0           | 0         | 5   | 2      | 2      | 0         | 5           | P14          | 0   | 0      | 0      | 0           | 0         | 4   | 6      | 4      | 0         | 6  |
| 1369 |                | P16                  | 4   | 1      | 1      | 0           | 4         | 11  | 2      | 4      | 0         | 9           | P15          | 0   | 1      | 0      | 0           | 1         | 7   | 2      | 1      | 3         | 8  |

Fig. 11. Number of groups and labels in sessions 1 and session 2 for each participant in each data type, and the number of groups and labels added/deleted/revised in between the sessions.

### A.3 Example Label Set Refinement

Figure 12 demonstrates P15's initial and final label set, as well as the refinements.

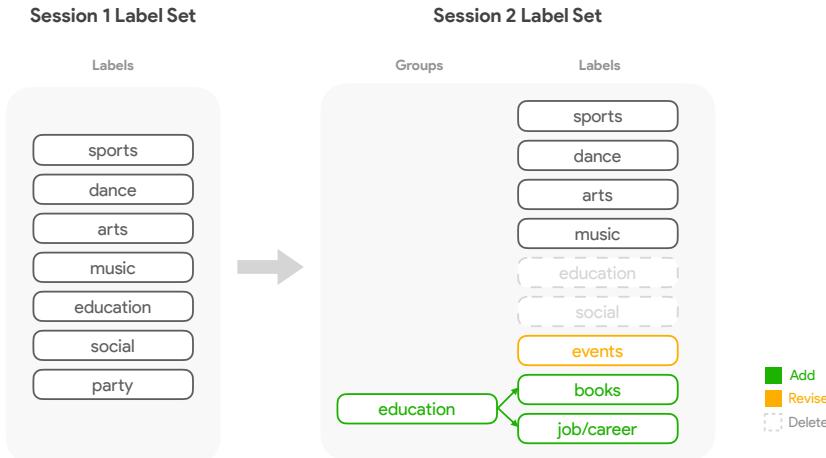


Fig. 12. An illustrative example of how one participant's label set changed in sessions 1 and 2 (P15). Two labels (education, social) are deleted, a label is revised (from social to events), and two labels (books, job/career) are added and grouped with a newly added group (education)

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009