# Crowd Synthesis: Extracting Categories and Clusters from Complex Data

**Paul André, Aniket Kittur, Steven P. Dow**
Human-Computer Interaction Institute, Carnegie Mellon University
{pandre,nkittur,spdow}@cs.cmu.edu

## ABSTRACT
Analysts synthesize complex, qualitative data to uncover themes and concepts, but the process is time-consuming, cognitively taxing, and automated techniques show mixed success. Crowdsourcing could help this process through on-demand harnessing of flexible and powerful human cognition, but incurs other challenges including limited attention and expertise. Further, text data can be complex, high-dimensional, and ill-structured. We address two major challenges unsolved in prior crowd clustering work: scaffolding expertise for novice crowd workers, and creating consistent and accurate categories when each worker only sees a small portion of the data. To address these challenges we present an empirical study of a two-stage approach to enable crowds to create an accurate and useful overview of a dataset: A) we draw on cognitive theory to assess how re-representing data can shorten and focus the data on salient dimensions; and B) introduce an iterative clustering approach that provides workers a global overview of data. We demonstrate a classification-plus-context approach elicits the most accurate categories at the most useful level of abstraction.

### Author Keywords
synthesis; clustering; crowd; categorization; classification

### ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI)

## INTRODUCTION
Analysts and researchers commonly cluster and analyze data to uncover themes and comprehend disparate pieces of information as a whole. Whether for insight into interview data or getting an overview of product reviews or a news event, the underlying process of synthesis requires complex cognition. Multiple stages of sensemaking, such as moving from specific to higher-level concepts, clustering, and classification of data [28], draw on aspects of concept learning, abstraction, and schema induction [16, 25]. The synthesis process typically requires significant amounts of time and effort, and expertise in both the domain and process. Qualitative researchers have discussed problems or even bias that might arise when analysis is from a single point of view and open to debate [18].

Tools to support the synthesis process can assist in an analyst's understanding of a dataset. Ongoing work in automatic text clustering and visualization generates automatically grouped words or topics and ways to navigate them [4, 10, 7, 13], but analysts must still perform the "arduous task of inferring meaningful concepts" [11] from the unlabeled groupings. An alternative approach is to try to leverage the human cognition necessary for deeper analysis of qualitative data. Crowdsourcing provides the opportunity to quickly access a large pool of people, but introduces the question of *how* to harness crowd members for this cognitively complex task. Such a process removes key assumptions in traditional analysis, providing several challenges.

- *Transient*. While analysts often immerse themselves in the full dataset to gain a rich understanding of themes, crowd members may only interact for a short period of time with a small portion of the data, meaning high-level similarities or categories might be missed.
- *Inexpert*. Workers may not have domain expertise, resulting in superficial or shallow categories rather than deeper concepts at higher levels of abstraction.
- *Conflicting*. Many workers, looking at complex data, mean potentially multiple (accurate and inaccurate) interpretations and disagreements over validity.

Recent work has begun to examine the use of crowds to cluster and discover partitions in data [17, 34, 41]. Most of this work has focused on creating clusters from image datasets, and has not attempted to elicit labels for newly discovered clusters. Exceptions that consider text are Cascade [9], a workflow for online crowds to generate taxonomies of datasets, and experiments in communitysourcing clusters of conference papers [2]. However, across all prior work, two major challenges for using the crowd to help with synthesis tasks remain unsolved.

One open challenge is how to enforce global constraints when each worker only sees a small sample of the data. Without a global overview of the data, workers could create incorrect or redundant categories. For example, Chilton et al.'s [9] report of running Cascade on a sample of 100 random colors resulted in a taxonomy with top level categories of "green" and "blue", but also "seafoam green" and "aqua"; another top-


*CSCW'14*, February 15–19, 2014, Baltimore, Maryland, USA.

ACM 978-1-4503-2540-0/14/02...$15.00.
http://dx.doi.org/10.1145/2531602.2531653

level category of "pastels" further included shades of blues and greens that could plausibly belong in these other categories as well. We introduce an *iterative clustering* approach in which workers can see all of the current categories when categorizing items in order to provide global context for their judgment. While workers can still create a new category if none of the current ones fit, the goal of the approach is to minimize redundancy across items and promote a consistent and useful level of abstraction for category labels.

Another major challenge we address in this paper is how to extend crowd synthesis to more complex qualitative data in potentially unfamiliar domains. While previous work has focused on datasets such as images or travel tips that most people are already familiar with, supporting expert analysis often involves textual data which can be highly domain-specific and of which the average worker may have little knowledge. This is especially problematic because novices have different mental models than experts and may categorize problems using surface features rather than deep structure [8], may use terms at a different level of abstraction [35], or may even ignore features that would be apparent to experts [27]. Here we compare a variety of methods aimed at *scaffolding expertise for novice crowd workers* so that they can function effectively despite limits on the time and experience they bring. Specifically, we draw from cognitive theory to examine tradeoffs in how items are represented; having the context of multiple items; and whether items are categorized one-by-one or categories are induced by comparing multiple items.

- *Re-representation*. Raw text is long and unfocused on salient dimensions. Re-representing the data will shorten and simplify, reducing distraction or burden of unnecessary information. On one hand, keeping the raw form will preserve all information that may be needed [23], but on the other, summaries help to abstract data and may aid in categorization [16].

- *Context*. Providing context (by showing multiple text items at once) may aid in producing rerepresentations using a standardized common language, and an understanding of what the most salient dimensions are [25]. On the other hand, multiple items at a time may prove more burdensome (in time or effort) than just one.

- *Classification vs. Comparison*. One approach to eliciting information is to predict a category label for each individual text item–a *classification* approach [25]. While context may help, crowd workers do not necessarily know the space they are classifying into. An alternative cognitive approach is to infer a single label for a group of items. While context in classification implicitly asks the worker to draw connections between the presented items, creating a single label for a group of items explicitly enforces mapping and *comparison*, techniques which have been shown to facilitate schema induction [16]. Additionally, making interproperty relationships more salient may result in more nuanced categories [20]. However, such grouping may mean losing less frequent relationships if two similar items are never shown together, or producing overly generic labels if very different items are shown together.

To address challenges in providing a crowd-generated overview of qualitative text datasets, we make the following unique contributions:

1. While prior work used datasets that build on workers' everyday knowledge, we focus on rich text datasets where a worker has no prior knowledge, asking: How can we get a crowd to generate meaningful categories in an unfamiliar domain?

2. We generate clusters as well as semantic labels and reasons for clustered data, distinct from automated approaches and the majority of prior crowd clustering work.

3. We conduct a formal evaluation of techniques within a two-stage workflow where we: (A) re-represent data in a short, salient form more conducive to clustering; (B) introduce a simple iterative clustering mechanism that enforces global constraints.

We demonstrate that a classification plus context approach is the most effective at creating useful categories at the right level of abstraction.

## RELATED WORK

### Topic Modeling and Visualization

Many linguistic and statistical methods to automate text clustering and classification have been proposed. In particular, LDA (Latent Dirichlet Allocation) [4] is a statistical generative model to uncover hidden (latent) topics in a corpus, by learning distributions of words that co-occur. A list of $k$ topics (described by $n$ most common words) is produced, along with distributions of topics for each document. However, such methods still require extensive work by the researcher to infer concepts from the list of words, fine-tune parameters, or manually add and remove words or topics [11].

Visualization of topic models [10, 7], or other forms of visual analytics [13] also offer insight into datasets. An analyst may find these valuable to support understanding and exploration of data (though again, many require knowledge or parameter tuning [13]). Though future work might incorporate such visualizations, our immediate focus is on generating clusters of data using human crowds, and providing descriptive rationale for the clusters.

### Crowd-Based Labeling with Existing Categories

Dataset labeling can be used for direct analysis or as training data before model fitting. Researchers have used crowdsourcing services to perform data labeling against pre-existing categories, e.g., for use in natural language processing [30], computer vision [31], or social media analysis [1]. Willet et al. [40] used crowds to generate hypotheses for features of visual charts. Their work suggests strategies such as prompts and examples that may aid workers in generating useful output, though unlike synthesis, their goal was not to learn a global schema of a dataset.

Semantic attributes have also been elicited through 'games with a purpose' – metadata elicitation of image [38] or music [22] datasets through gameplay. The 'output-agreement'

mechanism (rewarding a player for agreeing with their partner) can produce common and uninformative tags, even with restrictions such as taboo words [26]. To address this, researchers have proposed 'complementary-agreement' [21], which asks players to provide positive or negative examples of a given attribute.

### Crowd-Based Clustering without Existing Categories

Recently there has been a movement to cluster datasets using crowd labeling without providing predefined categories. Tamuz et al. [34] use adaptive triadic comparisons (e.g., "is object $A$ more similar to $B$ or $C$") to create a similarity matrix for supervised learning. Gomes et al. [17] use an algorithm for aggregating worker annotations from partial clusterings of an entire dataset (i.e., multiple overlapping samples of items). Extending this approach, Yi et al. [41] use a matrix completion technique to reduce the number of comparisons needed for partitioning of the entire dataset. Each process was successfully able to uncover meaningful categories within the image data, such as types of indoor/outdoor scene, or features of neckties. Most recently, researchers have experimented with text datasets. Chilton et al. [9] introduced Cascade, a workflow for crowd-generated taxonomies of text datasets such as Quora questions, and André et al. [2] experiment with communitysourced partial clustering and Cascade techniques for clustering conference papers.

We focus on more complex datasets than prior work, extending to data where a worker lacks firsthand knowledge. We also conduct a formal evaluation of cognitively-grounded approaches for re-representing data to shorten and focus on salient dimensions, before utilizing a novel iterative clustering approach that enforces global constraints (helping to eliminate redundant top-level categories or misclusterings). In the experiment below we first discuss challenges of text data, before describing techniques to address those challenges in a crowd synthesis process.

## CLUSTERING HIGH-DIMENSIONAL TEXT DATA

Prior crowd clustering approaches have been successful with image datasets [34, 17, 41], or with text datasets containing mostly everyday knowledge, e.g., travel tips from Quora [9]. However, analysts often work with datasets that use domain-specific, rich text data, posing a number of challenges:

- There are many (even infinite) potential dimensions of text [5], though only higher-level conceptual dimensions are likely to be of value, e.g., "number of vowels" is likely unhelpful, while "emotion" may be an important distinction.
- These dimensions are rarely as salient and obvious as visual features such as shape and color, which we can quickly or pre-attentively assess [37].
- Each data point may be long, complex, and outside the expertise or attention span of a crowd worker.

To address these challenges, we consider the cognitive factors involved in synthesis to inform our workflow. Synthesis involves categorization, concept learning, and schema induction (learning a global pattern or template to determine where to place individual items) [12, 16, 25]. We propose a two-stage process to scaffold aspects of these concepts towards producing an overview of a dataset; we elaborate below but first introduce the overall process (see Figure 1). *Stage A*: we propose to test three key concepts for how to re-represent text to shorten and simplify prior to clustering. *Stage B*: we utilize the Stage A outputs in an iterative clustering process that enforces global constraints by providing workers an overview of existing categories.

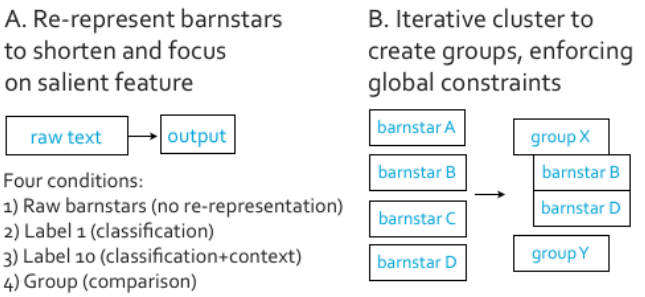


**Figure 1. Two-stage process to cluster and elicit categories from text data. Stage A: four conditions to transform raw data to shortened salient form. Stage B: an iterative clustering technique to group outputs from stage A and extract themes from data.**

First, we describe an experiment empirically examining the effect of re-representation on text items, as input to an iterative clustering technique. Then, in the Discussion, we report the mostly unsuccessful results of two existing techniques (LDA and partial clustering) on our text dataset.

## METHOD

### Text Dataset

To investigate clustering performance, we chose a text dataset with two properties in particular:

1. Sufficiently complex, such that a crowd would not necessarily have prior knowledge or domain expertise. This allows us to test our approach with the type of unfamiliar data an analyst may use.
2. A dataset that had already been analyzed by researchers in order to provide a 'ground truth' for comparison. (Note that while we do not expect to fully recreate an expert analysis, our aim is to evaluate how and where crowd synthesis succeeds and fails.)

We utilized a dataset consisting of Wikipedia 'barnstar' awards, previously analyzed by Kriplean et al. [19]. A barnstar is a token or message of appreciation given to a participant. For example:

> *...Mani, you have contributed a great deal of Estonian articles and done major and useful copyedits in a short time. You are a very productive user and deserve recognition.*

Kriplean's analysis focused on understanding what work was valued in Wikipedia, essentially asking, "What type of work does this barnstar recognize?" Their analysis revealed 7 high-level categories, with 42 total sub-categories (see Table 1).

| | | | |
|---|---|---|---|
| **Editing Work** | | **89** | **31.1%** |
| minor | copy-editing | 13 | 14.6% |
| media | images, audio | 10 | 11.2% |
| initiative | starting articles, stubs | 9 | 10.1% |
| major | substantial textual addition to an article | 5 | 5.6% |
| achievement | shepherding article to a higher quality level | 3 | 3.4% |
| classification | categorizing articles, adding templates | 4 | 4.5% |
| redesign | large-scale refactoring, merging pages | - | - |
| translation | to or from another language | - | - |
| attribution | citing sources, removing unsourceable | 1 | 1.1% |
| general | | 44 | 49.4% |
| **Social and Community Support Actions** | | **87** | **30.4%** |
| commitment | to an article, a wiki-project | 34 | 39.1% |
| teaching | mentorship, question-answering | 11 | 12.6% |
| leadership | of wikiprojects & other initiatives | 7 | 8.0% |
| humor & cheer | being funny, cheering others up | 3 | 3.4% |
| user page design | helping to design another's user page | 2 | 2.3% |
| rewarding | recognizing achievements of others | 1 | 1.1% |
| welcoming | welcoming newcomers | 2 | 2.3% |
| general | | 27 | 31.0% |
| **Border Patrol** | | **35** | **9.1%** |
| vandal fighting | reverting damage to unspecified namespace | 13 | 37.1% |
| deletion | article notability, spam removal | 8 | 22.9% |
| vandal fighting | reverting damage to user pages | 3 | 8.6% |
| vandal fighting | reverting damage to articles | 6 | 17.1% |
| sockpuppets | finding users operating multiple accounts | 2 | 5.7% |
| legal | copyright violations, fair use rationale | - | - |
| general | | 3 | 8.6% |
| **Administrative** | | **9** | **3.1%** |
| privilege granting | helping vet potential administrators | 1 | 1.1% |
| intervention | formal mediation of user conflicts | - | - |
| quality designation | determining article status (e.g. Featured) | 4 | 4.4% |
| technical action | exercise of privileged power | 3 | 3.3% |
| general | | 1 | 1.1% |
| **Collaborative Actions and Disposition** | | **27** | **9.4%** |
| disposition | civility, accepting of criticism, keeping cool | 17 | 26.0% |
| adherence | policy interpretation, integrity | 12 | 44.4% |
| diplomatic action | conflict mediation, consensus-seeking | 2 | 7.4% |
| explanation | rationale for an edit, decision, or standard | 1 | 3.7% |
| general | | 5 | 18.5% |
| **Meta-Content Work** | | **7** | **2.4%** |
| template | design of applicable templates | 2 | 28.5% |
| tool programming | design & support of tools (e.g. bots) | 3 | 42.9% |
| forums/portals | creation & support of help desks | 2 | 28.5% |
| classification | category creation & organization | - | - |
| process & policy | policy authoring & process design | - | - |
| archiving | storing old discussions | - | - |
| general | | - | - |
| **Undifferentiated Work** | | **32** | **11.1%** |
| | **Total work codes applied** | **286** | **100.0%** |

**Table 1. Distribution of work codes for the sample of 200 coded barnstars from Kriplean et al.'s analysis. Top-level work categories are bolded. Work dimensions within each category are given. Global and within-category percentages are given. The general coding represents barnstars that clearly fall into a category, but were not specific enough to be able to identify a specific dimension.**

The full dataset consists of 2,272 barnstars, coded with potentially multiple of the categories. We randomly sampled the dataset for 200 barnstars. Table 1 details the description of the categories, and associated counts in our sample. From now on, we refer to this analysis as the 'expert hierarchy'.

### Study Design

We evaluate a two-stage workflow to re-represent and cluster barnstars. In the first stage, we compare techniques for re-representing barnstars. Drawing on the discussion of re-representation, context, and classification vs grouping in the Introduction, we operationalize these concepts into three tasks, along with a fourth condition that uses raw barnstars. We briefly discuss the motivation, pros and cons, and operationalization of each condition, summarized in Table 2. Images of the the tasks can be seen in Figure 2. Each task asks a similar question to Kriplean's [19] original analysis: "What type of action or work is being rewarded?"

**Stage A: Re-Representation**. We test the three re-representation conditions below, against a fourth: Raw data.

*Label 1*. Shortening and simplifying the text may help to abstract data and aid in categorization [16], but on the other hand, keeping the raw form preserves all information that may be needed [23]. We test simple re-representation by displaying one barnstar and asking the worker to answer the above question. This is a *classification* technique, predicting the category label for one item.

*Label 10*. Showing multiple items at once may aid in understanding what the most salient dimensions are, as well as providing a standardized common language [25]. On the other hand, multiple items at a time may prove more burdensome (in time or effort) than just one. We test the additional context by extending the Label 1 technique, but rather than seeing just one barnstar, ten are shown, each with corresponding textboxes to answer the above question.

*Group*. Rather than classifying each individual barnstar, an alternative approach is to generate a single label for a group of items. This process explicitly asks the worker to consider multiple barnstars, highlighting common aspects of items and promote mapping and *comparison*, processes that have been shown to facilitate schema induction [16]. However, such grouping may mean losing less frequent relationships if two similar items are never presented together. We test this concept by displaying ten items. Workers are asked to create groups of barnstars, and label the groups as appropriate.

**Stage B: Iterative Clustering**. We then apply a second stage to elicit categories from the data. Techniques like partial clustering [17] and Cascade [9] do not explicitly enforce global constraints, potentially leading to incoherent top-level categories or misclusterings (i.e., two redundant high-level groups could be created). We propose a simple iterative clustering technique to test the effect of the re-represented barnstars, while enforcing global constraints. Using the output from the four techniques in Stage A, items are grouped. This is a similar mechanism to the 'Group' task, in that a worker is asked to group similar barnstars (or, re-representations of barnstars). The difference is that after an initial worker performs a task, subsequent workers see existing group names and can either group into them, or create new groups. In this way, the technique enforces global constraints by providing workers an overview of the categories, and the workers perform a complete clustering themselves (without need for a machine clustering step). To reduce any ordering effects of workers or data, and to yield potentially different results, we created five separate threads for each condition. See Figure 2 for examples of Label 1, Group, and iterative clustering tasks.

### Hypotheses

We posit that extra context in Label 10 as compared to Label 1 will result in better precision and recall. While classification

| Study Condition | Concept | Benefits | Drawbacks |
|---|---|---|---|
| Raw | — | Retain all information | Long, salient dimension unclear |
| Label 1: Summarize individual item | Classification | Shorten, simplify, focus on salient feature | Lose some detail, little context, requires n workers |
| Label 10: Summarize individual items in groups of ten | Classification plus Context | As above, standardized language, requires n/10 workers | As above, and possible boredom or exhaustion after multiple items |
| Group: infer $m$ groups from $n$ items | Comparison | As above, but different type of context: infer groups rather than classify | Lose detail, bias against less frequent or rare themes |

**Table 2. Summary of Stage A re-representation techniques to transform original raw barnstars. Concepts of context, classification, and inference are operationalized within three mechanisms: Label 1, Label 10, and Group.**

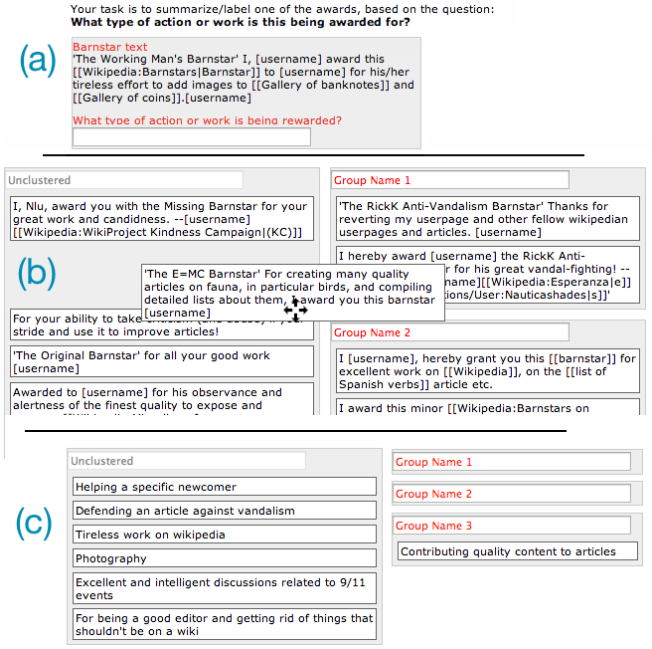


**Figure 2. Examples of tasks in Stage A and B. (a) Stage A, Label 1: A worker sees one barnstar and is provided a textbox to identify the salient work done. (Label 10 task is similar but with 10 instances on the same page.) (b) Stage A, Group: A list of ten items are shown, workers are asked to create groups and label the group. (c) Stage B, Iterative Cluster: input is from one of 4 outputs in Stage A. Here showing summarized forms from 'Label 10' output. Workers see prior group names, and are asked to group into them or create new groups.**

(Label 1 and Label 10) and comparison (Group) both have pros and cons, we believe that the extra context in Label 10, combined with the iterative clustering in Stage B, will result in superior performance for the classification methods. As for re-representation in general, we posit that all methods will outperform Raw, though Label 10 will perform the best.

### Participants

Amazon's Mechanical Turk was used as a source of relatively novice crowd worker. We restricted to US-only workers, with 95% HIT acceptance rate. Each worker could only complete one HIT to ensure no 'domain expertise' was learnt.

### Procedure

After a worker accepted the task entitled 'Label groups of Wikipedia messages (barnstar awards)' or a close variant, they were shown instructions, a brief description of Wikipedia barnstars, and workers were asked to perform a labeling or grouping to answer a similar question to Kriplean et al.'s original paper: "What type of action or work is being rewarded?" In Stage A, one of three re-representation conditions was shown (Label 1, Label 10, Group), and workers had to label, or group and label, raw barnstars. In Stage B, iterative clustering, output from one of the three conditions was shown (or a fourth condition: raw barnstars), and workers were asked to group and label, see Figure 2. After initial workers in Stage B, subsequent workers were able to see existing group names and group into them or create their own.

### Measures

**Stage A: Re-Representation**. We measure the *time* taken to complete the task; *conciseness*: length of the resulting label; *precision*: whether the label captured the most salient code of the original barnstar; and *recall*: how many of the potentially multiple codes of the original barnstar were retained.

**Stage B: Iterative Clustering**. We again measure time and other descriptive statistics such as *number of groups* in a thread, and *number of barnstars* in a group. We calculate *internal consistency*: how many barnstars in the group match the group name and theme; *precision*: whether the group matches an expert code; *recall*: how many concepts from the hierarchy were named in the groups. These are distinct from the Stage A precision and recall measures in that we assess at group-level against the expert hierarchy. We also present sample final outputs from each technique.

## RESULTS

Two raters assessed the outputs. The raters were familiar with Wikipedia having edited numerous articles in the past, one with over 10,000 edits. After an initial period of training, the two raters independently assessed all outputs. Stage A outputs (barnstar labels) were assessed to determine if they matched the most salient code of an original barnstar (as judged by the raters), as well as how many of the potentially multiple codes from the original barnstar. Stage B outputs (groups of barnstars and the group name) were assessed as matching a category from the expert hierarchy or not. There was substantial agreement between raters (Cohen's kappa: 0.74).

Since we are motivated by cognitive differences in conditions, and have specific comparisons of interest, we use a planned comparisons approach to test our hypotheses [36]. This also enables us to combine conditions to test hypotheses; we test the effect of *context* (Label 1 vs Label 10), of *classification*

| Raw Barnstar | Label 1 | Label 10 | Group |
|---|---|---|---|
| Thanks for your comments and inspiration on [[Wikipedia talk:Esperanza]]. I think you've given a very succinct explanation of what Esperanza should be about, even during this difficult time. It might even help me change my mind and come back, despite recent turmoil. | Comments and inspiration | Community interaction | Exceptional work on a topic |
| Thanks for all you help on my user page! May this lead you to more and even greater contributions to Wikipedia. :) Thanks [username] | For making contributions to the user's wikipedia page | Helpfulness | Contributions |
| 'The Working Man's Barnstar' Your tireless work on improving the [[2006 FIFA World Cup controversies]] article is more than worthy of this barnstar - the level head you have shown in dealing with an issue which could have easily got out-of-hand was a lesson for us all. Well done. | Dealing with an issue that could have gotten out of hand | Delicate wording | Hard Work |

**Table 3. Examples of outputs from re-representing barnstars in Stage A. Label 1 outputs are often more specific. Label 10 tends to capture more context to put into a more standardized and abstract language. Group tends to produce very abstract or even generic labels.**

| | | Label 1 | Label 10 | Group | Raw |
|---|---|---|---|---|---|
| **Stage A** | Time (sum total, min) | 283 | 108 | 67 | — |
| | Conciseness (chars) | 41.02 | 34.42 | 19.57 | 167 |
| | Precision | 73% | 82% | 62% | — |
| | Recall | 63% | 69% | 61% | — |
| **Stage B** | Time (sum total, min) | 312 | 337 | 64 | 392 |
| | # Groups per Thread | 6.6 | 6.8 | 4.2 | 6 |
| | Barnstars per Group (sd) | 30.3 (24.9) | 29.3 (17.5) | 36.7 (24.6) | 34.2 (26.3) |
| | Precision | 45% (15 / 33) | 65% (22 / 34) | 40% ( 6 / 15) | 53% (16 / 30) |
| | Recall | 59% (20 / 34) | 82% (28 / 34) | 35% (12 / 34) | 68% (23 / 34 ) |
| | Cost ($) | 90 | 72 | 72 | 60 |

**Table 4. Results of Stage A: Re-Representation (top), and Stage B: Iterative Clustering (bottom).**

*vs comparison* (Label 1 & Label 10 vs Group), and of *re-representation* in general (Raw vs Label 10). We introduce the results in terms of these comparisons.

### Stage A Results: Re-Representation

Label 1 and Label 10 have a 1-to-1 correspondence, i.e., 200 barnstars result in 200 labels. Since Group does not force all barnstars to be in groups, some do not get grouped; this resulted in 50 groups that contained a total of 155 barnstars.

To illustrate the differences in output, Table 3 presents examples of raw barnstar text and the outputs from the three processing methods. As hypothesized, there is a general trend for Label 1 to more directly capture the barnstar, but without a broader context, the output tends to be specific and potentially not abstract enough to capture the higher-level concepts at work. Label 10, with more context, is able to capture these abstractions (see, for example, the difference between specifics of "comments and inspiration" vs "community interaction", or "contributions to user's wikipedia page" vs "helpfulness".) The Group technique results in labels more abstract and with a slightly different focus, "exceptional work" vs "community interaction", or "hard work" vs "delicate wording". All of these concepts appear in the expert hierarchy, but the clustering may focus on the more prevalent concepts, removing the more subtle ones. It remains to be seen whether these abstractions are helpful in clustering, or whether the loss of detail is harmful to creating groupings. Details for the following measures are summarized in Table 4.

*Conciseness*. Label 1 and Label 10 differed significantly in length of label, Wald test: $\chi^2(1, N = 400) = 8.95$, $p<.01$, with Label 1 labels an average of 41 characters in length, compared to 34 for Label 10. Comparison (Group) and classification (Label 1, Label 10) techniques also differed significantly, Wald test: $\chi^2(1, N = 555) = 75.13$, $p<.01$, with Group outputs an average of 20 characters. The raw barnstars, by comparison, were an average of 167 characters.

*Precision*. Raters assessed the most salient code of the original barnstar (given the coding from the expert hierarchy), and categorized the label as either capturing that most salient code or not. Label 1 outputs captured the salient aspect of the original barnstar 73% of the time, compared to 82% for Label 10, and 46% for Group. Planned contrasts reveal that Label 1 and Label 10 differed, $\chi^2(1, N = 400) = 4.65$, $p<.05$, as did classification (Label 1 and 10) and comparison (Group), $\chi^2(1, N = 555) = 52.14$, $p<.01$.

*Recall*. Raters assessed how many of the potentially multiple codes of the original barnstar were retained in the label. Label 1 outputs retained on average 63% of all codes from the original barnstar, compared to 69% in Label 10, and 61% in Group. Planned contrasts indicate there was marginal difference between Label 1 and Label 10, $\chi^2(1, N = 400) = 3.38$, $p=.07$, while Group differed significantly from the classification conditions (Label 1 and Label 10), $\chi^2(1, N = 555) = 34.37$, $p<.01$.

#### *Summary of Stage A*

Lengths of output from each condition varied, suggesting that the technical and cognitive mechanisms did have an effect on the labels, as seen in the examples. The context of multiple barnstars allowed a higher level of abstraction and shorter la-

| | Label 1 | Label 10 | Group | Raw |
|---|---|---|---|---|
| **Match Expert** | Adding Images | Editing / Fixing | Hard Work | Generating / Improving an Article |
| | Community / Helping Others | Helpfulness / Friendliness | Editing | Dealing with Vandals |
| | Hard Work / Diligence | Community Improvement | Community | Being a Good Person |
| | Other / Unknown | Fighting Vandalism | Behind the Scenes | Art and Media |
| | Editing / Content | Hardworking | Vandalism | Administrative Work |
| | | Administrative Details | | Longterm Accomplishment |
| **Not Match** | How-to's for articles | Driving Content | Work | Make Wikipedia Better! |
| | Unneeded changes to articles | Contributing to a Charity | Contributions | Excellent Contributions |
| | Jibberish | Suspicious Category | Negative | Politics |
| | Continuity | Heroic Behavior | Specialization | Video Games |

**Table 5. Examples of group names from each condition: matching an expert (top), and not matching (bottom).**

bels in Label 10, while the Group technique resulted in labels of just 20 characters, suggesting a higher level of abstraction due to inferring groups that fit multiple barnstars as opposed to classifying individual items. The extra context of Label 10 over Label 1 seemed to enable better precision (more likely to create a label matching the salient feature of the original), with marginal difference in recall. Group performed surprisingly poorly. This is partially inherent in the process—inferring a group name for multiple barnstars is likely to incur some loss of precision, but we had hypothesized better performance. It is possible that workers were overeager to group items, or that with complex, unfamiliar items, classification is a more promising technique than the induction by comparison. We next discuss how these re-representation techniques affected clustering.

### Stage B Results: Iterative Clustering

We used a crowd-based iterative clustering technique to cluster the raw or processed barnstars into mutually exclusive groups. Five separate threads of iterative clustering were performed for each condition, to attempt to control for worker quality or ordering effects. We collapse these threads (since each may result in slightly different information), to report a single measure for each condition. Details, including descriptive statistics, are presented in Table 4.

*Internal Consistency*. Raters assessed whether individual barnstars matched the crowd-generated group name and apparent theme of the group. There was no difference between condition, one-way ANOVA, $F(3,114)=1.23$, $p=.30$.

*Precision*. We investigate the proportion of groups matching an expert concept to get a precision measure. There was marginal difference between Label 1 and 10, $\chi^2(1, N = 67) = 3.36$, $p=.07$. There was a significant difference between Group and the Label 1, Label 10 conditions, $\chi^2(1, N = 88) = 4.55$, $p<.05$, with 40% of Group groups, compared to an average of 55% of classification groups, matching an expert concept. There was no significant difference between Label 10 and Raw, $\chi^2(1, N = 64) = 1.37$, $p=.24$.

*Recall*. We measured two versions of recall. First, whether a concept from the expert hierarchy was at all present in the groupings (a fine-grained but overly generous measure), and second, whether a concept from the hierarchy was explicitly named in a group name (a coarse but overly conservative measure). We believe both are valuable—an analyst looking at output is likely to consider the group name as well as the contained barnstars in order to get a feel for the dataset. There was a difference in fine-grained recall between Label 1 and Label 10, $\chi^2(1, N=68) = 4.53$, $p<.05$; Label 1 groups covered 59% of the expert hierarchy, compared to 82% of Label 10 groups. The difference between Group and the Label conditions was also significant, $\chi^2(1, N=102) = 11.66$, $p<.01$, with Group covering 35% of the expert hierarchy. There was no difference between Raw and Label 10, $\chi^2(1, N=68) = 1.96$, $p=.16$. There was no difference in coarse-grained recall.

#### *Abstraction and Usefulness of Output*

Table 5 presents a sample of the group names that did or did not match categories in the concept hierarchy. When a group did match an expert concept, it was for the same reasons across conditions: obviously matching either a high-level or specific concept. There were varying reasons for a group not matching an expert concept. Group is the simplest to characterize, the group names were often too generic or abstract to be of use, due to clustering already abstract names. Within Label 1 we see examples of workers misinterpreting the barnstars or the focus of the group being unclear. It is possible we are seeing a curvilinear effect of conciseness or specificity. Raw barnstars contain a lot of information but are unfocused, while Label 10 output has a lot less information but is more focused. Both conditions performed well. Label 1 falls in-between, and it is feasible there is not enough focused information to be of use in determining categories (part of the motivation for providing more context with a Label 10 option).

While Label 10 and Raw were statistically similar in precision and recall, there were differences in the types of group name generated. Based on our hypotheses around classification and context, we investigated instances where a group name was overly specific or focused on irrelevant surface features as opposed to a more abstract category, or more focused on content rather than the activity being rewarded. We see a difference between Label 10 and Raw, $p<.05$, (Fisher's Exact Test). While both conditions had instances of overly generic or unclear names, Raw groupings were more likely to not be at a useful level of abstraction, for example, 'National Award' (specific countries mentioned), 'Video Games' (specific to video games as opposed to editing), and 'Emotion' (focused on words or feelings in the barnstars, not the reason for awarding the barnstars). Label 10 largely avoided these issues because of the re-representation and context to focus the label on the salient features of the barnstar.

*Summary of Stage B*

In summary, Label 10 and Raw performed similarly in terms of precision and recall, though Label 10 was more likely to produce categories at a correct or useful level of abstraction, not focused on specifics or surface features. We found more nuanced differences for context (Label 1 vs Label 10), and classification vs comparison (Label vs Group). The higher precision of Label 10 over Label 1 in Stage A resulted in a marginal difference in precision in Stage B, though did result in higher recall (coverage of expert categories). Group again performed poorly, outperformed by Label conditions in both precision and recall.

## DISCUSSION

We drew on cognitive theory in categorization and concept learning to motivate three approaches in re-representation of text, hypothesizing advantages compared to raw data, and an iterative clustering step that enforced global constraints. A classification plus context approach (Label 10) performed the best in terms of precision, recall, and eliciting category names at a useful level of abstraction. The Raw barnstars condition also performed surprisingly well on precision and recall measures, but upon closer inspection the category names were less useful in terms of level of abstraction. We begin by expanding on the results according to our planned contrasts, before discussing other issues and future work.

### Effect of Re-Representation on Clusters

*Context: Label 1 vs Label 10*

We posited that the added context of Label 10 in Stage A would produce labels more focused on salient dimensions and in a more standardized common language, aiding in clustering similar items in Stage B. This appears to be partially supported. Label 10 re-representations were more accurate, and resulted in higher coverage of concepts in resulting groups, with marginally more groups matching an expert concept.

*Comparison (Group) vs Classification (Label 1 & Label 10)*

Label 1 and 10 are a classification style of concept learning, albeit with the added advantage of context, compared to Group's comparison style [25, 27]. The classification re-representations matched and retained more concepts in Stage A, and produced more groups matching an expert concept in Stage B, though there was no difference in recall (number of unique concepts matched). Though we had posited superior performance for Label 10 in particular, the generally poor performance of Group was a surprise. The Group output went through two stages of grouping/clustering, and output may have been harmed as a result; relatively rare items were lost in initial re-representation and thus could not be present in Stage B, and the highly concise and perhaps overly generic labels from re-representation meant that subsequent cluster names were even more generic. Since workers could not see the underlying barnstars within those group names, it seems that there was not enough context or detail to create useful and meaningful clusters.

*Raw vs Re-Representation (Label 10)*

We had hypothesized that the length and complexity of the raw barnstars would result in an overload of information for crowd workers, making it harder to judge what was salient or similar, and thus resulting in clusters not as likely to be coherent or match the expert hierarchy. Based on precision and recall measures, this does not appear to have been the case; there were similar numbers of high-quality groupings and coverage of the expert hierarchy compared to Label 10. However, a comparison of the level of abstraction of group names revealed that some Raw groups were more specialized, more focused on surface features, or more focused on content rather than activity. For example, categories such as 'National Award' (specific to a country), or 'Video Games' were present in the Raw but not Label 10 groups. It seems that Label 10 was able to produce group names at a more consistently useful level of abstraction through the contextual re-representation process.

### Utility of Barnstar Synthesis

Although our focus was on the effect of technical and cognitive aspects of re-representation, a question still remains as to whether the overall output would be useful to a Wikipedia analyst. We recruited two Wikipedia-related researchers in our lab, though not experts in barnstars specifically, to provide their informal thoughts on the output. The general consensus was that the groupings, along with an ability to drill down into constituent barnstars, was a useful overview of the data. Their comments highlighted the importance of ensuring an appropriate level of abstraction, with group names such as 'Contribution' eliciting comments such as "*doesn't really tell me much... a context miss,*" while 'Computer Knowledge' was "*too close to the barnstar, doesn't provide much value.*" A question of abstraction was brought up: "*some seem to be directly sampling from barnstar text, not adding value to it.*' Some slightly unclear group names elicited more positive reactions: "*really intriguing that this is encoded, what does it mean?*" (before looking to the raw barnstars). Asked to provide an informal ranking, Group output was clearly last, "*on one hand, these are interesting categories that I can imagine would split types of work. But doesn't give me much information about what is helpful in Wikipedia.*" Label 1 next, "*most are valid, in that I'd like to see them in this list, but not well formulated.*" Label 10 was ranked highest, though Raw was "*almost similar, both provide value.*"

### Comparing to Existing Approaches

As a point of comparison, we also elicited clusters of barnstars using two existing techniques: topic modeling using LDA, and partial clustering (a technique to create a global clustering of items by grouping subsets of the dataset). We used the full 2,272 barnstar dataset for LDA, pre-processed to clean the data [39] before constructing 30 topics using GibbsLDA++. For partial clustering we created Mechanical Turk HITs using an object distribution algorithm used in recent crowd clustering work [17, 32], clustered using an agglomerative clustering tool [14].

While both methods were able to extract some meaningful categories, both had large numbers (around two-thirds) of unhelpful or uninterpretable groupings. LDA produced topics with linguistically similar terms, but the majority were not able to be mapped onto the conceptual groupings of the

expert hierarchy. Some concepts such as vandalism defense were separated into three or four separate groups because of the different terms used to describe the actions. Partial clustering produced 49 second-level clusters, with an average of 4.00 barnstars. There was a lot of redundancy in groupings, and the groups that did not match were not useful, either too specific (e.g., South American related articles), or completely spurious. However, the clustering did have an advantage of producing small groups, some perfectly matched, or at least easily able to be assessed as matching or not matching.

Although future developments may make machine learning approaches more powerful, our experience indicated a number of issues that motivated the approaches tested here.

### Crowd, Novice, and Expert Distinctions

This work focuses on novice crowds, as distinct from merely novices. This meant an explicit focus on crowd challenges such as interchangeable transient workers, leading to the need to coordinate via global constraints, and the iterative clustering mechanism. One or two novices looking at the same data would be able to spend more time and get a better sense of the structure of the data, and mechanisms to aid in that process would likely be different.

In our text dataset of Wikipedia barnstars, some data required specific knowledge of Wikipedia terms, but workers likely had little prior knowledge of the domain. However, most could reasonably be interpreted. We also used a similar prompt to Kriplean [19] in assessing the barnstars – "What type of work is being rewarded?" – thereby focusing the task.

Other datasets may not have the same structure, or analysts might wish to ask a more open-ended question initially to understand different metrics for similarity that people may use, in which case techniques such as confusion matrices to highlight different ways of categorizing may be useful [17]. There is a rich history of novice vs expert categorization and learning, and, e.g., Shafto and Coley [29] discuss how novice generalizations are explained by notions of similarity, while experts tend to use prior knowledge about causal, taxonomic, or conceptual issues to guide them in reasoning. In other words, novices' reasoning is decontextualized. As with our Wikipedia dataset, this may not be an issue, but for biological or medical datasets it may be harder to create categories that match an expert's understanding; though for some areas novice categorization may be useful to understand how a layperson reasons about features.

Qualitative researchers have noted how analysis can be open to debate [18]. One in-depth example comes from a 2010 workshop to discuss software design.[1] Three videos of professionals working on a software design exercise were analyzed by teams drawn from 54 participants, "*resulting in many different perspectives including cognition, representation, ... interaction design, coordination, tools, and design theory*" [3]. The variety of perspectives likely reflects the different backgrounds of the participants, but the analyses were not mutually exclusive: "*sometimes, findings were contradicted, sometimes wholeheartedly affirmed, and sometimes clarifying interpretations resulted in new insight*" [3].

#### *Implications of Utilizing Novices*

In a broader sense, classification systems can be viewed as "*artifacts embodying moral and aesthetic choices that in turn craft people's identities, aspirations, and identity*" [6]. Bowker and Star [6] cite examples of categories linked with social movements, such as homosexuality or postpartum depression included in the *Diagnostic and Statistical Manual* (DSM), or racial categories in the U.S. census. As Suchman notes, "*categories have politics*" [33]. As above, a novice may be able to reason based on perceptual similarity, but they do not have the domain knowledge to make other conceptual judgments, nor the political knowledge to understand the broader impact of the categorization.

### Limitations and Future Work

We restricted workers to one task only to emphasize novice crowds and a lack of domain expertise. However, throwing away workers who gain knowledge is not optimal, and future work should consider how best to utilize workers who wish to do more than one task, as well as workers who have been trained through performing the tasks. A further limitation of our approach was little or no error checking or redundancy, though these mechanisms would increase number of tasks and cost, adaptive sampling [34] or training workers may offset these costs as well as improve output.

Data items presented to workers should be carefully considered, both in terms of scalability and information gain, but also in terms of the order and similarity of items. The order of new features affects learning and generalizations about categories. For example similar items in a sequence might correctly be considered part of the same category, but when separate, the categorizations are often missed [24]. Variability of examples may also affect concept learning: diverse examples can lead to superior categorization (though these findings are based on training data being available), but perceptual categories are learned more slowly when examples are highly variable [15]. These findings suggest that automated methods such as TF-IDF or LDA may be beneficial to explore pre-grouping similar items.

We have discussed results from one text dataset, and next steps are to see if these findings replicate. Particularly interesting might be a different form of data such as longer form product reviews, or interview data, where partitioning of data, temporal dependencies, and tradeoffs of verbatim words vs re-representation may be important. We used Amazon Mechanical Turk as a participant pool, and future work might further investigate the potential of multiple perspectives from a diverse crowd, or look to other crowds with different properties: size of crowd, expertise, incentives, and availability.

Finally, the crowd synthesis process could be more interactive, particularly focused on crowd-requester interaction. For example, an analyst might have an open-ended initial question of the data, or multiple questions, and decide to focus on

[1] `http://www.ics.uci.edu/design-workshop/`

particular subsets of the data, or throw away particular questions or areas of data, as real-time results are explored.

## CONCLUSION

There are many complex datasets and not enough time or expert effort to uncover the potentially valuable categories and insights contained within them. Motivated by recent work in crowd clustering of simple image and text datasets, we explored issues and approaches for utilizing novice crowds to create simple overviews of more complex data than in prior work. In a two-stage process for synthesis of qualitative text data, we provide a theory-grounded approach to re-representation and clustering of data. A classification-plus-context approach (Label 10) performed best in terms of precision and recall of expert categories. Raw data also performed surprisingly well, although the Label 10 approach resulted in group names at a more useful level of abstraction. Further, we utilize a simple iterative clustering approach that differs from prior work by enforcing global constraints, providing workers an overview of data and reducing potential for global misalignments or misclustering.

## ACKNOWLEDGMENTS

We thank Travis Kriplean for the barnstar dataset, and Bob Kraut and Kurt Luther for helpful discussions. This work was supported by NSF grants IIS-1149797, IIS-1217559, OCI-0943148, IIS-0968484, IIS-1111124, IIS-1208382, IIS-1217096, Bosch, Google, and Microsoft.

## REFERENCES

1. André, P., Bernstein, M., and Luther, K. Who gives a tweet?: evaluating microblog content value. In *Proc. CSCW 2012*, 471–474.
2. André, P., Zhang, H., Kim, J., Chilton, L. B., Dow, S., and Miller, R. Community clustering: Leveraging an academic crowd to form coherent sessions. In *Proc. HCOMP 2013*.
3. Baker, A., van der Hoek, A., Ossher, H., and Petre, M. Guest editors' introduction: Studying professional software design. *Software, IEEE 29*, 1 (2012), 28–33.
4. Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.
5. Blum, A. L., and Langley, P. Selection of relevant features and examples in machine learning. *Artificial intelligence 97*, 1 (1997).
6. Bowker, G. C., and Star, S. L. *Sorting things out: Classification and its consequences*. The MIT Press, 2000.
7. Chaney, A. J., and Blei, D. M. Visualizing topic models. In *Proc. ICWSM 2012*.
8. Chi, M. T., Feltovich, P. J., and Glaser, R. Categorization and representation of physics problems by experts and novices. *Cognitive science 5*, 2 (1981), 121–152.
9. Chilton, L. B., Little, G., Edge, D., Weld, D., and Landay, J. Cascade: Crowdsourcing taxonomy creation. In *Proc. CHI 2013*.
10. Chuang, J., Manning, C. D., and Heer, J. Termite: Visualization techniques for assessing textual topic models. In *Proc. AVI 2012*.
11. Chuang, J., Ramage, D., Manning, C., and Heer, J. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proc. CHI 2012*, 443–452.
12. Clapper, J. P., and Bower, G. H. Learning and applying category knowledge in unsupervised domains. *Psychology of Learning and Motivation 27* (1991), 65–108.
13. Endert, A., Fiaux, P., and North, C. Semantic interaction for visual text analytics. In *Proc. CHI 2012*, 473–482.
14. Fernández, A., and Gómez, S. Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification 25*, 1 (2008), 43–65.
15. Fried, L., and Holyoak, K. Induction of category distributions: A framework for classification learning. *J Exp Psychol Learn Mem Cogn. 10*, 2 (1984), 234.
16. Gick, M., and Holyoak, K. Schema induction and analogical transfer. *Cognitive psychology 15*, 1 (1983), 1–38.
17. Gomes, R., Welinder, P., Krause, A., and Perona, P. Crowdclustering. In *Advances in Neural Information Processing Systems (NIPS 2011)*.
18. Kolbe, R., and Burnett, M. Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of consumer research* (1991), 243–250.
19. Kriplean, T., Beschastnikh, I., and McDonald, D. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proc. CSCW 2008*, 47–56.
20. Lassaline, M., and Murphy, G. Induction and category coherence. *Psychonomic Bulletin & Review 3*, 1 (1996), 95–99.
21. Law, E., Settles, B., Snook, A., Surana, H., von Ahn, L., and Mitchell, T. Human computation for attribute and attribute value acquisition. In *Proc. FGVC 2011*.
22. Law, E., Von Ahn, L., Dannenberg, R., and Crawford, M. Tagatune: A game for music and sound annotation. In *Proc. ISMIR 2007*.
23. Marr, D. *Vision*. W. H. Freeman and Company, San Francisco, 1982.
24. Medin, D., and Bettger, J. Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review 1*, 2 (1994), 250–254.
25. Medin, D. L., and Schaffer, M. M. Context theory of classification learning. *Psychological review 85*, 3 (1978), 207.
26. Robertson, S., Vojnovic, M., and Weber, I. Rethinking the esp game. In *Proc. CHI'09 EA*, ACM, 3937–3942.
27. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. Basic objects in natural categories. *Cognitive psychology 8*, 3 (1976), 382–439.
28. Russell, D., Stefik, M., Pirolli, P., and Card, S. The cost structure of sensemaking. In *Proc. InterCHI 1993*, 269–276.
29. Shafto, P., and Coley, J. D. Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *J Exp Psychol Learn Mem Cogn. 29*, 4 (2003).
30. Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP 2008*, 254–263.
31. Sorokin, A., and Forsyth, D. Utility data annotation with amazon mechanical turk. In *Proc. CVPR Workshops 2008*, 1–8.
32. Strehl, A., and Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res. 3* (2003), 583–617.
33. Suchman, L. Do categories have politics? the language/action perspective reconsidered. In *Proc. ECSCW 1993*, 1–14.
34. Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. Adaptively learning the crowd kernel. In *Proc. ICML 2011*.
35. Tanaka, J. W., and Taylor, M. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology 23*, 3 (1991), 457–482.
36. Thompson, B. Planned versus unplanned and orthogonal versus nonorthogonal contrasts: The neo-classical perspective. In *Advances in Social Science Methodology*, B. Thompson, Ed. JAI Press, 1994.
37. Treisman, A., and Gelade, G. A feature-integration theory of attention. *Cognitive psychology 12*, 1 (1980), 97–136.
38. Von Ahn, L., and Dabbish, L. Labeling images with a computer game. In *Proc. CHI 2004*, 319–326.
39. Wang, Y.-C., Burke, M., and Kraut, R. E. Gender, topic, and audience response: an analysis of user-generated content on facebook. In *Proc. CHI 2013*, 31–34.
40. Willett, W., Heer, J., and Agrawala, M. Strategies for crowdsourcing social data analysis. In *Proc. CHI 2012*, 227–236.
41. Yi, J., Jin, R., Jain, A., and Jain, S. Crowdclustering with sparse pairwise labels: A matrix completion approach. In *Proc. HCOMP 2012*.