Strategies for Crowdsourcing Social Data Analysis

Wesley Willett[⋆], **Jeffrey Heer**[†], **Maneesh Agrawala**[⋆]
[⋆]Computer Science Division, University of California, Berkeley
{willettw, maneesh}@cs.berkeley.edu
[†]Computer Science Department, Stanford University
jheer@cs.stanford.edu

ABSTRACT

Web-based social data analysis tools that rely on public discussion to produce hypotheses or explanations of patterns and trends in data rarely yield high-quality results in practice. Crowdsourcing offers an alternative approach in which an analyst pays workers to generate such explanations. Yet, asking workers with varying skills, backgrounds and motivations to simply "Explain why a chart is interesting" can result in irrelevant, unclear or speculative explanations of variable quality. To address these problems, we contribute seven strategies for improving the quality and diversity of worker-generated explanations. Our experiments show that using *(S1) feature-oriented prompts*, providing *(S2) good examples*, and including *(S3) reference gathering*, *(S4) chart reading*, and *(S5) annotation* subtasks increases the quality of responses by 28% for US workers and 196% for non-US workers. Feature-oriented prompts improve explanation quality by 69% to 236% depending on the prompt. We also show that *(S6) pre-annotating charts* can focus workers' attention on relevant details, and demonstrate that *(S7) generating explanations iteratively* increases explanation diversity without increasing worker attrition. We used our techniques to generate 910 explanations for 16 datasets, and found that 63% were of high quality. These results demonstrate that paid crowd workers can reliably generate diverse, high-quality explanations that support the analysis of specific datasets.

Author Keywords

Information Visualization; Social Data Analysis; Crowdsourcing

ACM Classification Keywords

H.5.3 Group & Organization Interfaces: Collaborative computing.

INTRODUCTION

Making sense of large datasets is fundamentally a human process. While automated data mining tools can find recurring patterns, outliers and other anomalies in data, only people can currently provide the explanations, hypotheses, and insights necessary to make sense of the data [22, 24]. Social data analysis tools such as Sense.us [8], Pathfinder [18] and Many Eyes [30] address this problem by allowing groups of web-based volunteers to collaboratively explore visualizations, propose hypotheses, and seek out new insights. Controlled experiments have shown that groups can use these tools to discover new, unexpected findings [8, 29]. However, eliciting high-quality explanations of the data requires seeding the discussion with prompts, examples,and other starting points to encourage contributions [8, 32].

Outside the lab, in real-world web-based deployments, the vast majority of the visualizations in these social data analysis tools yield very little discussion. Even fewer visualizations elicit high-quality analytical explanations that are clear, plausible, and relevant to a particular analysis question.

We recently surveyed the Many Eyes website and found that from 2006 to 2010, users published 162,282 datasets but generated only 77,984 visualizations and left just 15,464 comments. We then randomly sampled 100 of the visualizations containing comments and found that just 11% of the comments included a plausible hypothesis or explanation for the data in the chart. The low level of commenting may represent a shortage of viewers or may be due to *lurking* – a common web phenomenon in which visitors explore and read discussions, but do not contribute to them [31, 20]. When comments do appear, they are often superficial or descriptive rather than explanatory (Figures 2a, 2b). Higher-quality analyses sometimes take place off-site [5] but tend to occur around limited (often single-image) views of the data curated by a single author. Ultimately, marshaling the analytic potential of crowds calls for a more systematic approach to social data analysis; one that explicitly encourages users to generate good hypotheses and explanations.

In this paper we show how paid crowd workers can be used to perform the key sensemaking task of generating explanations of data. We develop an analysis workflow (Figure 1) in which an analyst first *selects charts*, then uses crowd workers to carry out *analysis microtasks* and *rating microtasks* to generate and rate possible explanations of outliers, trends and other features in the data. Our approach makes it possible to quickly generate large numbers of good candidate explanations like the one in Figure 2c, in which a worker gives several specific policy changes as possible explanations for changes in Iran's oil output. Such analytical explanations are extremely rare in existing social data analysis systems.

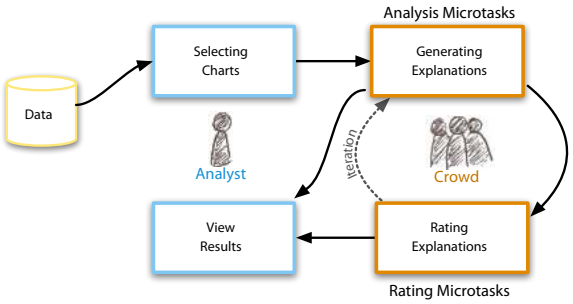

Figure 1. Our workflow for crowdsourcing data analysis.

To appear at CHI 2012.

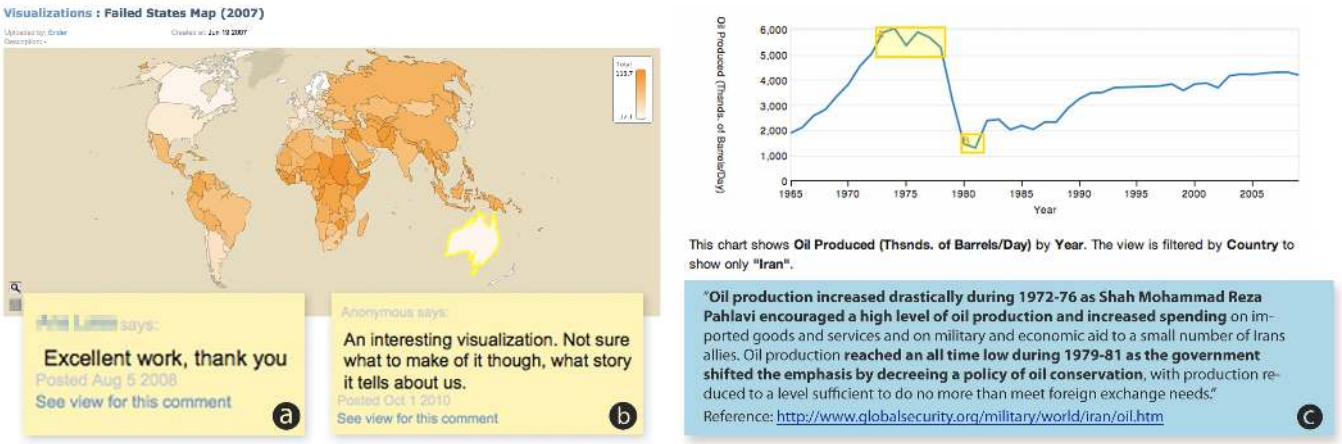

Figure 2. Comments on social data analysis on sites like Many Eyes (a,b) often add little value for analysts. We show that crowd workers can reliably produce high-quality explanations (c) that analysts can build upon as part of their broader analyses. (Emphasis added.)

Yet, simply asking workers with varying skills, backgrounds, and motivations to "Explain why a chart is interesting" can result in irrelevant, unclear, or speculative explanations of variable quality. We present a set of seven strategies that address these problems and improve the quality of worker-generated explanations of data. Our seven strategies are to: *(S1) use feature-oriented prompts*, *(S2) provide good examples*, *(S3) include reference gathering subtasks*, *(S4) include chart reading subtasks*, *(S5) include annotation subtasks*, *(S6) use pre-annotated charts*, and *(S7) elicit explanations iteratively*. While some of these strategies have precedents in other crowdsourcing systems [17, 21], the main contribution of this work is to demonstrate their impact in the context of collaborative data analysis.

We have applied these strategies to generate 910 explanations from 16 datasets, and found that 63% were of high quality. We also conducted six experiments to test the strategies in depth. We find that together our first five strategies (S1-S5) increase the quality ratings (a combined measure of clarity, plausibility, and relevance) of responses by 28% for US workers and 196% for non-US workers. Feature-oriented prompts (S1) are particularly effective, increasing the number of workers who explain specific chart features by 60%-250% and improving quality by 69%-236% depending on the prompt. Including chart annotation subtasks (S5) or pre-annotating charts (S6) also improves workers' attention to features. Additionally, iterative rounds of explanation generation (S7) can produce 71% new explanations without increasing worker attrition. Finally we show how workers can help analysts identify the best unique explanations – providing quality ratings that correlate strongly with our own and identifying redundant explanations with 72% accuracy. Our results show that by recruiting paid crowd workers we can reliably generate high-quality hypotheses and explanations, enabling detailed human analyses of large data sets.

RELATED WORK

We build on two main areas of related work; asynchronous social data analysis and applications of paid crowdsourcing.

Asynchronous Social Data Analysis

Social data analysis systems such as Sense.us [8], Pathfinder [18], Many Eyes [30], and Swivel [27] were built under the assumption that people can parallelize the work required to analyze and make sense of data. Motivated users can visualize, share, and discuss datasets but, as we've noted, few of the visualizations exhibit high-quality analytical discussion. In fact, many of the commercial websites no longer exist.

Heer and Agrawala [6] discuss a variety of issues in designing asynchronous social data analysis systems to improve sensemaking. They suggest that these systems should facilitate division, allocation and integration of analysis work, support communication between workers and provide intrinsic and extrinsic incentives. Building on these suggestions, Willett et al.'s CommentSpace [32] demonstrates that dividing social data analysis into concrete subtasks can improve the quality of analysts' contributions. In this work, we further break the task of generating explanations into smaller *microtasks* in which paid workers explain features of the data and other workers rate those explanations.

Applications of Paid Crowdsourcing

With the rise of online labor marketplaces such as Amazon's Mechanical Turk (`www.mturk.com`), researchers have focused on the use of paid crowdsourcing to supplement purely computational approaches to problem solving and user testing [12, 23]. In the context of visualization, recent work has used crowdsourced workers to perform graphical perception experiments on the effectiveness of charts and graphs [7, 14]. We also pay crowd workers to make judgments about charts and graphs and to provide graphical annotations, but we focus on analytical sensemaking tasks.

Other work has examined how to incorporate human computation into larger workflows. Soylent [2] uses paid workers to perform document editing tasks within a word processor, using a Find-Fix-Verify pattern to break editing tasks into smaller subtasks. Similarly, our workflow helps an analyst break down complex data analysis operations into *analysis microtasks* that many workers can perform in parallel and *rating microtasks* that help the analyst consolidate the results of the parallel analyses. We also take inspiration from CrowdForge [13], Jabberwocky [1], TurKit [17], and Turkomatic [15] which provide general-purpose programming models for leveraging crowds to perform complex tasks.

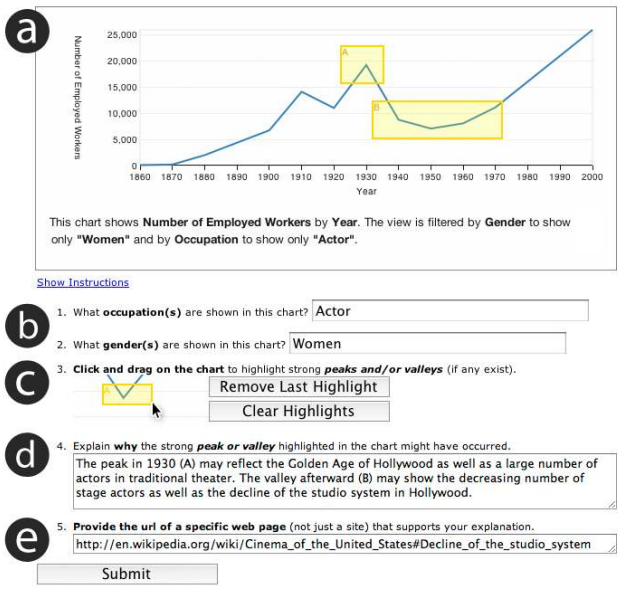

Figure 3. An example analysis microtask shows a single chart (a) along with chart-reading subtasks (b) an annotation subtask (c) and a feature-oriented explanation prompt designed to encourage workers to focus on the chart (d). A request for outside URLs (e), encourages workers to check their facts and consider outside information.

A WORKFLOW FOR CROWDSOURCING DATA ANALYSIS

Hypothesis (or explanation) generation is a key step of Pirolli and Card's [22] sensemaking model and it requires human judgment. Developing good hypotheses often involves generating a diverse set of candidate explanations based on understanding many different views of the data. Our techniques allow an analyst to parallelize the sensemaking loop by dividing the work of generating and assessing hypotheses into smaller *microtasks* and efficiently distributing these microtasks across a large pool of workers.

We propose a four-stage workflow (Figure 1) for crowdsourcing data analysis. An analyst first *selects charts* relevant to a specific question they have about the data. Crowd workers then examine and explain these charts in *analysis microtasks*. Optionally, an analyst can ask other workers to review these explanations in *rating microtasks*. Finally, the analyst can *view the results* of the process, sorting and filtering the explanations based on workers' ratings. The analyst may also choose to iterate the process and add additional rounds of analysis and rating to improve the quality and diversity of explanations.

Selecting Charts

Given a dataset, an analyst must initially select a set of charts for analysis. The analyst may interactively peruse the data using a visual tool like Tableau [28] to find charts that raise questions or warrant further explanation. Alternatively, the analyst may apply data mining techniques (e.g., [10, 16, 33]) to automatically identify subsets of the data that require further explanation. In general, our workflow can work with any set of charts and is agnostic to their source.

In our experience, analysts often know *a priori* that they are interested in understanding specific features of the data such as outliers, strong peaks and valleys, or steep slopes. Therefore, our implementation includes R scripts that apply basic

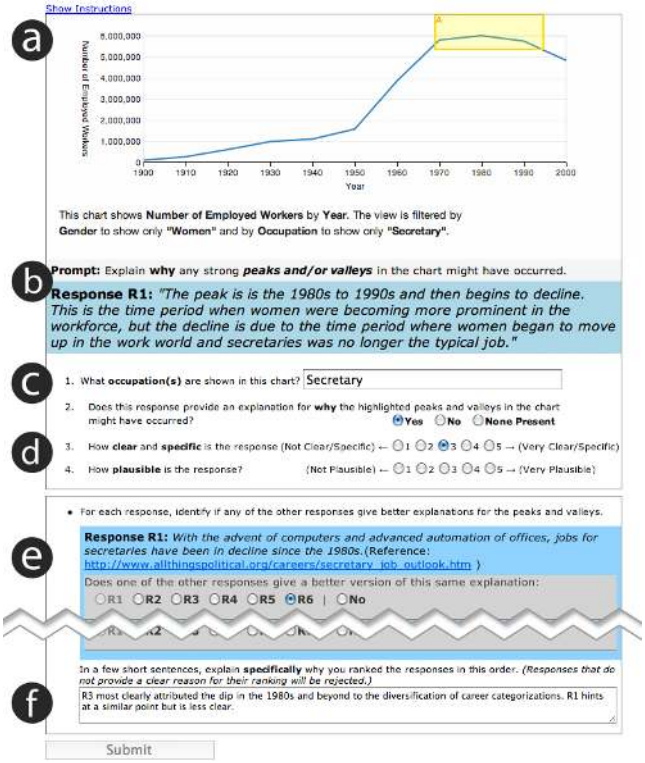

Figure 4. An example rating microtask showing a single chart (a) along with explanations (b) from several workers. The task contains a chart-reading subtask (c) to help focus workers' attention on the charts and deter scammers, along with controls for rating individual responses (d), indicating redundant responses (e), and summarizing responses (f).

data mining techniques to a set of time-series charts in order to identify and rank the series containing the largest outliers, the strongest peaks and valleys and the steepest slopes. The analyst can then review these charts or post them directly to crowd workers to begin eliciting explanations. We leave it to future work to build more sophisticated data mining algorithms for chart selection.

Generating Explanations

For each selected chart, our system creates an analysis microtask asking for a paid crowd worker to explain the visual features within it. Each microtask contains a single chart and a series of prompts asking the worker to explain and/or annotate aspects of the chart (Figure 3). The analyst can present each microtask to more than one worker to increase the diversity of responses.

Rating Explanations

If a large number of workers contribute explanations, the analyst may not have the time to read all of them and may instead wish to focus on just the clearest, most plausible or most unique explanations. In the rating stage the analyst enlists crowd workers to aid in this sorting and filtering process. Each *rating microtask* (Figure 4) includes a single chart along with a set of explanations authored by other workers. Workers rate explanations by assigning each a binary (0-1) *relevance* score based on whether it explains the desired feature of the chart. Workers also rate the clarity (how easy it is to interpret) and plausibility (how likely it is to be true) of each response on a numerical (1-5) scale. We combine these ratings into a numerical *quality* score (0-5) that mea-

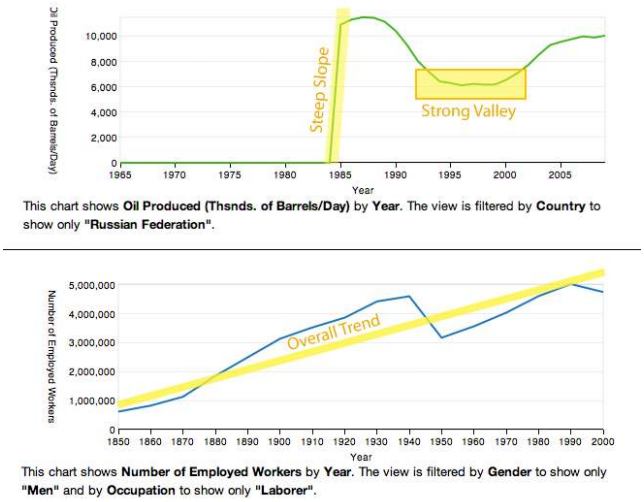

Figure 5. Sample charts from the oil production and US census datasets used in our examples and experiments. Depending on their interests analysts may with to focus workers' attention on a variety of different features of a chart, including *slopes*, *valleys*, and *overall trends*.

sures how well a worker's response explains the feature they were asked to focus on, using the formula:

$$quality = (clarity + plausibility)/2 \times (relevance).$$

Analysts can use these scores to quickly assess the quality of responses and quantitatively identify the best explanations. Workers also mark each redundant response by indicating any other response in the set that provides a better version of the same explanation.

View Results

Once workers have generated explanations, the analyst can view the responses and incorporate them into their own analyses. If the explanations have been rated, the analyst can sort and filter them using the ratings and can hide redundant responses. For example the analyst may examine only the most plausible, unique explanations. Optionally, the analyst can examine and organize the results further using a collaborative visualization environment such as CommentSpace [32]. An analyst may also choose to have workers iterate on a task, generating additional unique explanations or explanations that improve on the best responses from a prior round.

STRATEGIES FOR ELICITING GOOD EXPLANATIONS

Simply asking workers to look at a chart and explain why it is interesting may not produce good results. We consider five types of problems that can reduce the quality of these explanations and discuss strategies (S1-S7) designed to mitigate these problems.

Example datasets. For illustration we focus our discussion of the strategies on two time series datasets (Figure 5); historical data on world oil production by nation from 1965-2010, and US census counts of workers by profession from 1850-2000. We consider other datasets later in the paper.

Problem 1: Irrelevant Explanations

A chart may be interesting for many reasons, but analysts are often interested in understanding specific visual features such as outliers or overall trends. Without sufficiently detailed instructions, workers may explain features irrelevant to the analyst. For example, workers may comment on the visual design of the chart rather than the features of the data.

S1. Use feature-oriented explanation prompts. Refining the prompt to focus on the specific features the analyst is interested in increases the likelihood that workers will provide relevant explanations. Consider the line charts in Figure 5. An analyst may be interested in *peaks and valleys* or *steep slopes and flat regions* in the oil production chart because such features indicate significant events in the oil market. Alternatively, the analyst may be interested in longer-term tendencies of the labor market as indicated by the *overall trend* of the census chart. For other charts, analysts may be interested in more complex features such as clusters, repeating patterns, and correlations between dimensions.

A feature-oriented prompt might ask workers to *"explain the peaks and/or valleys in the chart (if any exist)"*. A specific prompt like this can increase the chance that workers will refer to peaks and valleys in their explanations, and also makes it easier for workers to note the absence of these features. Such negative explanations can be just as informative as explanations of the features themselves.

Problem 2: Unclear Expectations

Workers may not know what typical and atypical charts look like or what kinds of explanations they are expected to produce. Similarly, they may not know how to identify specific features like peaks or slopes.

S2. Provide good examples. To introduce workers to a dataset or feature type, analysis microtasks can include example charts showing several representative views. Similarly, including example responses may help to establish expectations and calibrate workers to the style and level of detail expected in their response [3]. In our implementation, analysts can generate examples by selecting a small set of charts (typically 2-3) and performing the analysis microtask themselves. We then package the example charts with the analyst's responses and present them to workers before they begin their first microtask. To reduce the amount of work an analyst needs to do before launching a new dataset, the examples may come from different datasets analyzed earlier. However, the data, chart type, and desired features should be similar to the new dataset.

Problem 3: Speculative Explanations

Explanations of data invariably depend on outside information not present in the data itself. Often interpretations are speculative or based on assumptions from prior experience.

S3. Include reference gathering subtasks. To encourage validation, an analysis microtask can require workers to provide references or links to corroborating information on the web (Figure 3e). Requiring such links may encourage workers to fact-check more speculative answers and may also uncover useful resources that the analyst can use later in the analysis process. However, asking workers to gather outside references may increase the time and effort associated with a microtask, and may increase worker attrition.

Problem 4: Inattention to Chart Detail

In an effort to increase their payment, workers may proceed quickly through the microtask without thoughtfully considering the prompt. They may also attempt to scam the task by entering junk responses. Even well-intentioned workers may not attend to the chart features specified in the instructions.

S4. Include chart reading subtasks. Chart reading questions (Figure 3b) can focus workers by requiring them to inspect axes, legends or series (*"What country is shown in this chart?"*), to extract a value from the chart (*"In what year did the number of workers peak?"*), or perform a computation based on the chart (*"How many more workers were there in 2000 than in 1900?"*). Such questions force workers to familiarize themselves with the data and can draw attention to important aspects of a particular chart like missing data or a non-zero axis. Additionally, because "gold standard" answers to such chart reading questions are known a priori, we can automatically check workers' answers and eliminate responses from spammers or workers who do not understand the instructions. Including such gold standard questions is a well known technique for improving result quality in crowdsourcing tasks [21, 26]. In our case these questions also help direct workers' attention to chart details.

S5. Include annotation subtasks. Requiring workers to visually search for and mark features in the chart can further focus their attention on those details. For example, the microtask may ask workers to first annotate relevant features of a chart and then explain those features (Figure 3c). Such annotations encourage attention to details and support deixis [9], allowing workers to ground their explanations by pointing directly to the features they are explaining. In our implementation each annotation is labeled with a unique letter ("A","B","C",...) so workers can refer to them in their text explanations. The worker-drawn annotations are also amenable to further computation. For example, when summarizing responses, a system could aggregate marks from multiple workers to highlight hot spots on a particular chart, or to calculate a collective "best guess" for the overall trend of a time series [6].

S6. Use pre-annotated charts. Alternatively, the analyst can pre-annotate visual features in the chart (Figure 5) so that workers pay attention to those details. Such annotations help focus workers on specific chart details and also reduce irrelevant explanations (Problem 1). Although pre-annotating charts greatly reduces the possibility that workers will attempt to explain the wrong feature, creating such annotations may require the analyst to perform additional data mining or manual annotation on the dataset.

Problem 5: Lack of Diversity

Multiple workers may generate similar explanations while leaving the larger space of possible explanations unexplored.

S7. Elicit explanations iteratively. As with other human computation tasks [17], analysis microtasks can be run in multiple, sequential stages, in which workers see a chart along with the best explanations generated in prior iterations. The analyst may elicit more diverse responses by asking workers to generate explanations that are different from the earlier ones. Alternatively, the analyst can increase explanation quality by asking workers to expand and improve upon the earlier explanations.

DEPLOYMENT

We have deployed our crowdsourced data analysis workflow on Amazon's Mechanical Turk and used workers to generate 910 explanations for 64 different charts drawn from 16 different datasets.

Our deployment included the census and oil datasets described earlier, as well as data on world development (UN food price indices, life expectancy data by nation), economics (US foreign debt, employment and housing indices for major US cities, return on investment data for US universities), and sports (team winning percentages from the NBA and MLB, historical batting averages of professional baseball players, olympic medal counts by nation, and Tour de France standings). As a proof-of-concept, we generated a set of 2 to 5 charts for each dataset that exhibited a particular characteristic, such as sharp peaks, valleys or steep slopes. In some cases we selected charts by hand, while in others we used our data-mining scripts to automatically select the charts.

We (the authors) examined and rated all 910 responses generated by workers and scored them using the quality metric described earlier in the Workflow section. We assigned $quality >= 3.5$ to 276 of the 435 responses (63.4%) that used our strategies but were not part of our experiments, indicating that most explanations were very good. Throughout the deployment, we found that workers consistently generated high-quality explanations for all datasets and provided numerous explanations that we had not previously been aware of. For example, one worker who examined the US debt dataset suggested that a large spike in British purchases of US debt might be due to Chinese purchases through British brokers. In another case, five different workers examining a chart of baseball player John Mabry's batting average (Figure 6c) independently attributed a prominent valley to a mid-season trade that reduced his at-bats. Other novel insights are shown in Figures 2, 4, and 6.

EXPERIMENTS

A full factorial experiment to evaluate all seven strategies would be prohibitively large. Instead we evaluated the strategies as we developed them. We first tested five initial strategies (S1-S5) together to gauge their overall impact. We then examined the effects of S1, S2, and S5 in a factorial experiment. Based on these results, we added three additional experiments to compare reference gathering (S3), annotation strategies (S5 and S6), and iteration (S7). Finally, we examined the results of our rating microtasks.

Experiment 1: Strategies S1-S5 in Two Worker Pools

To evaluate the cumulative impact of the first five strategies (S1-S5) we had one pool of workers complete analysis microtasks that included all of them (*strategies* condition) while a second pool completed the same microtasks but without the strategies (*no-strategies* condition).

Non-US workers represent a large portion of the workers on Mechanical Turk [11] and can often provide results more

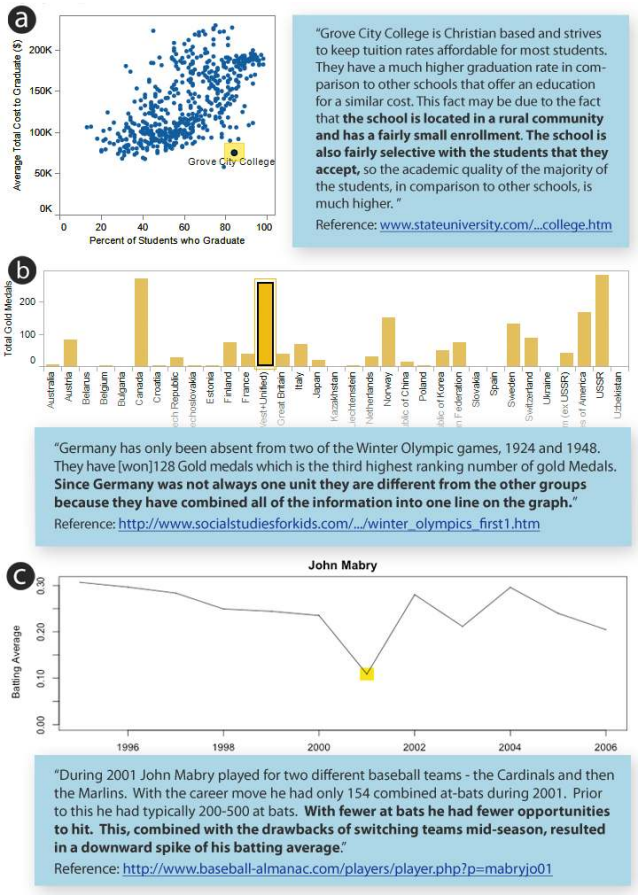

Figure 6. Sample explanations generated for charts showing university tuition and graduation rates (a), olympic medal counts by country (b), and historical batting averages (c). In each case we asked workers to explain a single outlier on a pre-annotated chart. (Emphasis added.)

quickly and cheaply than US-based workers. However, studies of Mechanical Turk have shown that workers from outside the United States exhibit poorer performance on content analysis [25] and labeling tasks [4]. We designed the experiment to determine if a similar performance gap exists for data analysis tasks and whether our strategies could improve results from these workers.

We hypothesized that: (1) Results from *US workers* would be of higher quality than results from *non-US workers*, but (2) employing strategies S1-S5 would increase the quality of explanations produced by workers in both groups.

Methods

Over the course of the experiment, we ran 200 analysis microtasks using Mechanical Turk. We divided these microtasks into 8 experimental conditions:

$$2\,strategy\ variants \times 2\,worker\ pools \times 2\,datasets = 8$$

The microtask in the *no-strategies* condition asked workers to "explain why any interesting sections of chart might have occurred". In the *strategies* condition, the microtask (Figure 3) included a **feature-oriented prompt (S1)** asking workers to "explain why any strong peaks and/or valleys in the chart might have occurred" and an **annotation subtask (S5)** that instructed workers to highlight those peaks and valleys. The microtask was preceded by instructions that included three **example charts (S2)** with annotations and explanations. The *strategies* condition also included a **reference-gathering subtask (S3)** that required workers to provide the URL of a website that corroborated their explanation. To help safeguard against scammers, we included **chart-reading (S4)** subtasks in both conditions. We also asked workers to fill out a demographic questionnaire.

We used both the oil production and US census datasets and selected five charts from each dataset with the largest variance. All of the resulting charts exhibited a range of features including peaks, valleys, slopes, and large-scale trends.

We collected five explanations for each of the charts. We also restricted each worker to a single condition (either *strategies* or *no-strategies*) and allowed workers to explain each chart only once, for a maximum of 10 responses per worker. We paid workers $0.05 per microtask during some early trials, but later increased the pay rate to $0.20 per microtask to reduce completion time. We based these rates on prior studies [7, 19] which have shown that while pay rate impacts completion time, it has little impact on response quality.

Results

Over the course of the experiment, 104 different workers produced responses for the 200 microtasks. To assess how well workers understood the tasks, we (the authors) calculated quality scores for each response (as described in the Workflow section). We also analyzed the content of the responses, labeling each one as either an *"explanation"* if it explained the chart features or a *"description"* if it simply described the features. We also examined whether or not each response referred to *"peaks or valleys"*,*"steep slopes or flat regions"*, or an *"overall trend"*.

We observed no significant difference in response quality, completion time, or length between the census and oil productions datasets in either worker population, indicating that producing explanations was of similar difficulty across both datasets. Thus, we combine the results from both datasets in all subsequent analyses.

Worker Pools. We found that worker pool had a significant main effect on quality ($F_{1,198} = 12.2, p < 0.01$). Response quality was higher for US workers ($\mu = 2.23, \sigma = 1.79$) than for non-US workers ($\mu = 1.37, \sigma = 1.87$) (Figure 7) in part because 83% of responses from US workers contained relevant explanations, while only 42% of responses from non-US workers did so. Instead, 34% of non-US workers described the chart rather than explaining it, and 24% produced responses that were so poorly written we could not classify them. The poor performance of non-US workers may reflect their lack of familiarity with the datasets as well as a language barrier. In our demographic questionnaire, only 35% of non-US workers in the census conditions could accurately describe the US census, versus 100% of US workers. Less than 20% of non-US workers reported English as their native language, versus 95% of US workers.

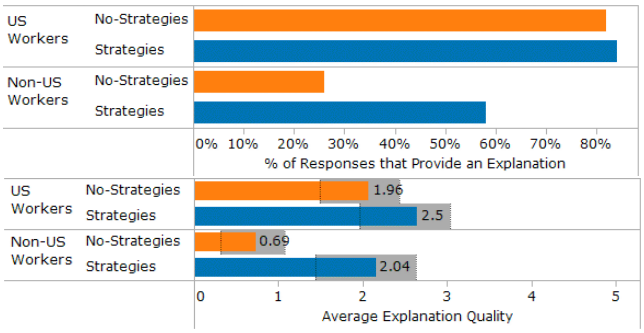

Figure 7. Percent of responses containing an explanation(top) and average explanation quality(bottom), by worker group (*US* / *non-US* workers) and strategy condition (*strategies* / *no-strategies*) in Experiment 1. Error bars give 95% confidence intervals.

We also found that across US and non-US groups, workers in the *strategies* condition produced higher quality responses ($\mu = 2.27, \sigma = 2.00$) than those in the *no-strategies* condition ($\mu = 1.33, \sigma = 1.62$) ($F_{1,198} = 14.5, p < 0.01$). However, the improvement in average quality of responses for non-US workers (196%) was much larger than for US workers (28%). These results suggest that using strategies S1-S5 makes a bigger difference when workers are culturally unfamiliar with the task and/or dataset.

Prompts. The introduction of strategies S1-S5 greatly increased workers' attention to peaks and valleys in the data. Workers in the *strategies* condition, which included a feature-oriented "*peaks and valleys*" prompt (S1) along with examples (S2) and annotation subtasks (S5) that reinforced the prompt, referred to peaks and valleys very consistently (90% of *US* and 68% of *non-US* responses). Workers in the *no-strategies* condition, however, referenced very few of these features (16% of *US* and 6% of *non-US responses*). The *no-strategies* workers often referred to overall trends or slopes in the data or failed to provide an explanation at all.

Completion Times and Attrition. Across both pools, workers took significantly longer to complete each microtask in the *strategies* condition (Median=4 minutes 11 seconds) than they did in the *no-strategies* condition (Median=2 minutes 48 seconds) ($t = -3.668, p < 0.01$). We computed attrition as the percentage of participants who began a microtask but quit without completing it and found an attrition rate of 67% for workers in the *strategies* condition. Attrition was 23% in the *no-strategies* condition. These results suggest that workers are less willing to complete analysis microtasks that include additional subtasks like chart reading and reference gathering.

Because *non-US* workers generated such low quality explanations, we used only US workers in our subsequent experiments. Also, because we saw similar results in Experiment 1 across both the oil production and US census datasets, we used only the census dataset in Experiments 2-5.

Experiment 2: Exploring Individual Strategies

Our experience in Experiment 1 led us to believe that three strategies, **feature-oriented prompts (S1)**, **examples (S2)**, and **annotation subtasks (S5)**, had the greatest impact on response quality. To better understand the effects of these strategies, we conducted a factorial experiment that varied each one independently. We hypothesized that: (1) Feature-oriented prompts (S1) would improve quality by increasing the proportion of responses that explained the specified feature; (2) Examples (S2) would improve quality, especially when paired with a feature-oriented prompt, by familiarizing workers with the prompt and chart type as well as the expected length, style, and content of good responses; (3) Annotation subtasks (S5) would encourage workers to mark the prompted feature and thereby improve quality by increasing the number of relevant responses.

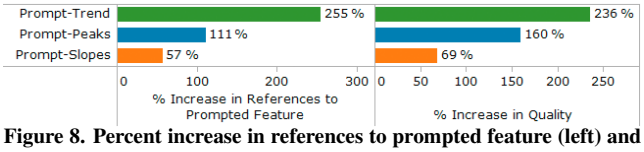

Figure 8. Percent increase in references to prompted feature (left) and quality (right) for each feature-oriented prompt (S1) in Experiment 2.

Methods

In Experiment 2, we ran 160 explanation microtasks divided into 16 conditions:

$$(4\ \ prompts\ ) \times (2\ \begin{matrix} examples \\ variants \end{matrix}\ ) \times (2\ \begin{matrix} annotation \\ variants \end{matrix}\ ) = 16$$

Our 4 prompts included three feature-oriented prompts (S1) *prompt-slopes*, *prompt-trend*, and *prompt-peaks*, and one control prompt, *prompt-control*. In the *prompt-slopes* conditions, we asked workers to "explain why any sharp slopes and/or flat regions in the chart might have occurred", while in the *prompt-trend* conditions we asked workers to "explain why the overall trend in the chart might have occurred". The *prompt-peaks* and *prompt-control* conditions used the same prompts as the *strategies* and *no-strategies* conditions from Experiment 1, respectively.

To test the examples strategy (S2), we included an *examples* condition that showed workers three examples of high-quality explanations and a *no-examples* conditions that provided only short text instructions. To test annotation subtasks (S5), we included a *worker-annotation* condition that required workers to mark features in the charts and a *no-annotation* condition that did not. For consistency with Experiment 1, we included reference-gathering subtasks (S3) and chart-reading subtasks (S4) in all conditions.

Results

Prompts. Including a feature-oriented prompt (S1) increased the percentage of responses that referred to that feature by between 60% and 250% compared to the control condition, depending on the feature (Figure 8). Workers in the *prompt-peaks* ($\chi^2 = 8.455$), *prompt-slopes* ($\chi^2 = 5.952$), and *prompt-trend* ($\chi^2 = 37.746$) were all significantly more likely (all $p < 0.02$) to explain their prompted feature than workers in *prompt-control*. Similarly, including prompts increased response quality by between 69% and 236% compared to *prompt-control*. This increase was significant for workers in *prompt-trend* ($U = 372.0, p < 0.001$) and *prompt-peaks* ($U = 564.5, p = 0.008$). The increase for *prompt-slopes* ($U = 624.5, p = .064$) was not quite significant, probably because *prompt-control* workers were already more likely to explain slopes (35% of responses) than peaks or trends (both 23% of responses).

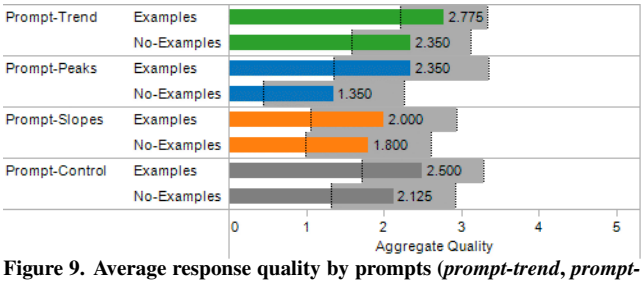

Figure 9. Average response quality by prompts (*prompt-trend*, *prompt-peaks*, *prompt-slopes*, or *prompt-control*) and examples (*examples*, *no-examples*). Error bars show 95% confidence intervals.

Providing Examples. Workers in the *examples* conditions produced higher quality responses ($\mu = 2.41, \sigma = 1.78$) than workers in the *no-examples* conditions ($\mu = 1.91, \sigma = 1.77$) (Figure 9), but the difference in quality was not significant ($U = 2717.5, p = 0.09$). Examples also improved the quality and consistency of annotations. Workers in the *worker-annotation* condition who saw examples of high-quality responses with annotated features, emulated the examples (Figures 2c and 4). Workers who did not see such examples created annotations that were more difficult to interpret and often annotated more features than they explained.

Annotation. In the *worker-annotation* condition, workers annotated chart features that were relevant to the prompt in 60 of the 80 trials. Workers who received a feature-oriented prompt as well as an annotation subtask referred to the feature specified in their prompt more frequently (S1 and S3: 85%) than workers who received a feature-oriented prompt without an annotation subtask (S1 only: 72%), but the difference was not quite significant ($\chi^2 = 3.142, p = 0.076$). Many *worker-annotation* workers also referred to their annotations by letter in their responses, providing clear deictic references to features. Neither the average time to complete the explanation microtask nor the attrition rate were significantly different between the *worker-annotation* and *no-annotation* conditions.

Reference-Gathering. In Experiment 2, we asked workers in all 16 conditions to gather references from the web to support their responses. Out of the 160 responses, 151 included valid URLs, of which 137 were unique. We assigned each reference a quality score from 1-5 based on how well it supported the explanation. Workers in the *examples* condition generated higher quality URLs ($\mu = 2.73, \sigma = 0.96$) than those in the *no-examples* case ($\mu = 2.4, \sigma = 1.0$) but these differences were not significant ($U = 3018, p = 0.08$).

Experiment 3: Reference Gathering

Based on results from Experiments 1 and 2, we hypothesized (1) that **reference gathering (S3)** increased response quality, but (2) the effort required to gather references contributed to high attrition. To test this hypothesis, we ran an experiment with 50 trials split between two conditions; the *gathering* condition was identical to the *strategies* condition in Experiment 1, while the *no-gathering* condition omitted the reference gathering subtask but was otherwise identical.

Results

The 25 responses in the *gathering* condition produced 20 unique URLs and URL quality was similar to Experiment 2 ($\mu = 2.67, \sigma = 1.02$). Surprisingly, however, the *no-gathering* condition produced significantly higher-quality explanations ($\mu = 3.38, \sigma = 1.55$) than the *gathering* condition ($\mu = 2.22$, $\sigma = 1.94$) ($U = 211.5, p = 0.046$). Additionally, the median completion time for *no-gathering* microtasks was only 2 minutes 36 seconds, significantly faster than the 3 minutes 45 second median for *gathering* tasks ($U = 175.5, p = 0.008$), suggesting that while reference gathering produces useful references, it does so at the cost of speed and quality. Given the low number of trials and high variance, further study is necessary to fully understand this relationship.

Experiment 4: Annotation Strategies

In our first two experiments, we found that **annotation subtasks (S5)** helped workers focus on chart features and facilitated deixis. In some cases, however, the analyst may wish to **pre-annotate charts (S6)** to focus workers' attention on specific features. To compare the trade-offs between these two strategies, we conducted another study with 50 trials split between two conditions – *worker-annotation*, in which we asked workers to mark the prompted feature before they explained it, and *pre-annotation*, in which the feature was pre-marked. We hypothesized that workers in the *pre-annotation* condition would generate more responses that explained the prompted feature than those in the *worker-annotation* condition.

Results

We found no significant differences between the two conditions. However the number of response that explained the prompted feature ("peaks and valleys") was high in both the *pre-annotation* (88%) and *worker-annotation* (96%) cases. In 84% of the trials in the *worker-annotation* condition, workers marked the exact same peak or valley that we had highlighted in the *pre-annotation* condition, suggesting we shared a common notion of which peaks or valleys were important.

Experiment 5: Iteration

In our fifth experiment, we tested whether **eliciting explanations iteratively (S7)** could improve the diversity of workers' explanations. First, we asked one group of workers (the *initial* condition) to generate explanations for a dataset. After a second group rated these explanations, we asked a third group of workers (the *iteration* condition) to generate additional explanations that were different from the first set. We hypothesized that (1) the *iteration* condition would produce mostly new explanations, but (2) would have a higher rate of attrition, since later workers might feel unable to author a response that differed from the initial explanations.

We conducted 25 trials in the *initial* round, producing five explanations each for the five US census charts. In the *iteration* round, we conducted 25 more trials, in which we showed new workers the same five charts, along with the initial explanations. We instructed *iteration* workers to generate new explanations that were "different from the explanations already shown". Both conditions included pre-marked charts (S6), but were otherwise identical to the *strategies* condition in Experiment 1.

Results

The 25 trials in the *initial* condition produced 36 distinct explanations, while the 25 trials in the *iteration* condition produced 35 explanations. Of the *iteration* explanations, 71% had not been proposed in the first round. The attrition rate for the *iteration* condition (75.3%) was also slightly lower than the attrition rate in the *initial* round (80.2%), indicating that iteration can increase the diversity of explanations without increasing attrition.

Experiment 6: Rating

In order for rating microtasks to provide an effective means of sorting explanations and identifying duplicates, workers must be able to generate consistent ratings. To test this, we conducted a final experiment in which we asked workers to rate a subset of the explanations generated during our broader deployment. We hypothesized that (1) quality ratings assigned by workers would be similar to our own quality ratings and that (2) workers would consistently detect and label redundant explanations.

Methods

We asked 243 Mechanical Turk workers to rate 192 different explanations across 37 charts. Using the interface shown in Figure 4, workers rated each response according to the criteria (relevance, clarity, and plausibility) described in the Workflow section. We compared these ratings against our own quality ratings for the same results.

Workers also indicated redundancy as follows: for each explanation, workers could mark at most *one* other response as providing a better version of the same explanation. For each worker, we then formed a redundancy graph with the explanations as the nodes. We linked two explanations with an undirected edge if the worker marked them as redundant. To identify groups of redundant explanations we computed the transitive closure of this graph. Each connected component then represented a unique explanation and all explanations within a component were redundant.

To reduce scamming in the rating microtask we also included one "gold standard" explanation on five of the charts. We purposely based the content of each "gold standard" explanation on one of the worker-generated explanations, but modified the language to ensure that workers could not identify it as an exact duplicate. We used these "gold standard" explanation with known redundancy to test whether or not workers could successfully detect redundant explanations.

Results

In total, the workers produced 1,334 individual ratings for the 192 different explanations. We compared these to our own quality ratings for the same responses.

Rating Consistency. A Pearson's chi-square test showed very strong agreement ($\chi^2 = 78.81, p < 0.01$) between workers' *relevance* scores and our own, indicating that workers were good at identifying responses that did not explain the requested feature. A Spearman's rank correlation coefficient showed that workers' overall *quality* scores and our scores were moderately correlated ($\rho = 0.42$).

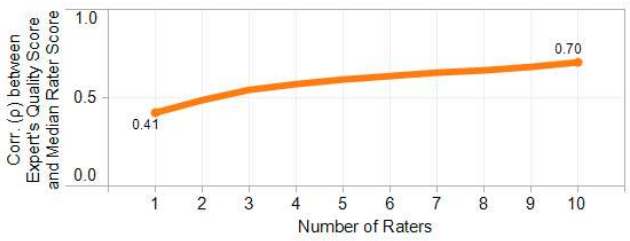

Figure 10. Correlation between our quality scores and the median workers' scores. For each number of raters, we randomly sample from a set of 10 ratings for 25 different explanations. The chart shows average correlations after 10,000 sampling iterations.

Because not all workers' ratings are reliable, an analyst may wish to combine ratings from multiple workers to obtain a more accurate result. To estimate the effect of using multiple raters, we took the set of responses that had been rated by at least ten raters and repeatedly sampled a subset of the ratings for each response.

For example, to estimate the effectiveness of using three raters, we randomly selected three worker's ratings for each response and used the median of them as the response's quality score. We then computed the correlation between the median scores and our own quality scores for all responses. To control for sampling error we randomly sampled and recomputed the correlation 10,000 times for each number of raters (Figure 10). Using the median score from multiple workers produced results that correlated more strongly with our own - increasing steadily from a moderate correlation ($\rho = 0.41$ when using one rater) to a strong one ($\rho = 0.70$ with ten raters).

Redundancy. We tested workers' ability to detect redundant responses by examining the results from the 25 rating microtasks in which we seeded the set of responses with a known redundant explanation. Across these 25 microtasks, workers connected the known redundant explanation to the explanation on which it was based 72% of the time. Workers agreed strongly on 35% of the pairs, with *all* five raters indicating the same redundancy relationship.

DISCUSSION

We have demonstrated that paid crowd workers can reliably produce high quality explanations and novel insights. In fact, in our deployment we found that 63% of the responses we sampled contained good explanations – far more than in tools like Many Eyes. Moreover, we found that using several basic strategies (S1-S5) can greatly improve explanation quality, particularly when users are unfamiliar with the data. Because paid crowd workers are readily available and can provide good explanations, these results suggest that we may be able to conduct targeted social data analysis at a much larger scale than was possible in previous systems.

In practice, strategies may only be appropriate in certain circumstances. For example, reference gathering (S3) is useful if an analyst requires citations or references for their analy-

ses. However, in our experience, reference gathering causes workers to take longer and can reduce the diversity of explanations, since workers cannot pose hypotheses for which they have no references. Instead, it may be better to make references optional or provide bonuses for good references.

Similarly, while both annotation strategies we tested (S5,S6) improved workers' attention to prompted details, they are useful in different situations. Annotation subtasks (S5) are more useful when the specific features of interest are not yet known, while pre-annotated charts (S6) are useful for directing workers' attention to more subtle features that are relevant to the analyst, but not obvious to workers.

Finally, our analysis of workers' performance on rating microtasks demonstrates that crowd workers can provide high-quality ratings that correlate strongly with our own. However, using multiple workers produces more accurate ratings. Redundancy checking subtasks provide a reliable approach for identifying unique explanations, however, quality-control mechanisms such as "gold standard" questions with known responses may be necessary to make certain that workers understand the task.

While crowd workers generated good explanations for the wide range of public-interest datasets we tested, they may fare less well with domain-specific data. In future work, we plan to apply similar strategies to elicit targeted analysis help from expert communities, volunteers, and enthusiasts.

CONCLUSION

Our work demonstrates how the key sensemaking task of generating explanations can be broken down and performed systematically by paid workers. Relying on paid crowd workers rather than ad-hoc volunteers allows us to marshal the analytic power of hundreds of workers in a systematic way. By packaging simple charts within analysis microtasks and dispatching them en-masse to the crowd, we can solicit large numbers of high-quality explanations much more predictably than we could using existing social data analysis platforms. Moreover, we demonstrate that using a straightforward set of strategies, we can mitigate common problems such as irrelevant explanations, unclear and speculative worker expectations, and inattention to chart detail.

ACKNOWLEDGEMENTS

We would like to thank Björn Hartmann for his helpful input. This work was partially supported by NSF grants CCF-0963922 & CCF-0964173 and a gift from Greenplum/EMC.

REFERENCES

1. Ahmad, S., Battle, A., Malkani, Z., and Kamvar, S. The jabberwocky programming environment for structured social computing. In *Proc. UIST* (2011).
2. Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. Soylent: a word processor with a crowd inside. In *Proc. UIST* (2010), 313–322.
3. Boud, D. *Enhancing learning through self assessment*. Routledge, 1995.
4. Chandler, D., and Kapelner, A. Breaking monotony with meaning: Motivation in crowdsourcing markets. *U. of Chicago mimeo* (2010).
5. Danis, C. M., Viegas, F. B., Wattenberg, M., and Kriss, J. Your place or mine?: Visualization as a community component. In *Proc. CHI* (2008), 275–284.
6. Heer, J., and Agrawala, M. Design considerations for collaborative visual analytics. *Information Visualization 7*, 1 (2008), 49–62.
7. Heer, J., and Bostock, M. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc. CHI* (2010), 203–212.
8. Heer, J., Viégas, F., and Wattenberg, M. Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Comm. of the ACM 52*, 1 (2009), 87–97.
9. Hill, W. C., and Hollan, J. D. Deixis and the future of visualization excellence. In *Proc. of IEEE Visualization* (1991), 314–320, 431.
10. Hurley, C. B., and Oldford, R. W. Pairwise Display of High–Dimensional Information via Eulerian Tours and Hamiltonian Decompositions. *JCGS 19*, 4 (2010), 861–886.
11. Ipeirotis, P. Demographics of mechanical turk. *New York University, Tech. Rep* (2010).
12. Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with mechanical turk. In *Proc. CHI* (2008), 453–456.
13. Kittur, A., Smus, B., and Kraut, R. Crowdforge: crowdsourcing complex work. In *CHI Extended Abstracts*, ACM (2011), 1801–1806.
14. Kong, N., Heer, J., and Agrawala, M. Perceptual guidelines for creating rectangular treemaps. *IEEE TVCG 16* (2010), 990–998.
15. Kulkarni, A., Can, M., and Hartman, B. Collaboratively Crowdsourcing Workflows with Turkomatic. In *Proc. CSCW* (2012).
16. Lee, E.-K., Cook, D., Klinke, S., and Lumley, T. Projection pursuit for exploratory supervised classification. *JCGS 14*, 4 (2005), 831–846.
17. Little, G., Chilton, L., Goldman, M., and Miller, R. Turkit: tools for iterative tasks on mechanical turk. In *Proc. SIGKDD*, ACM (2009).
18. Luther, K., Counts, S., Stecher, K., Hoff, A., and Johns, P. Pathfinder: an online collaboration environment for citizen scientists. In *Proc. CHI* (2009), 239–248.
19. Mason, W., and Watts, D. J. Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl. 11* (May 2010), 100–108.
20. Nielsen, J. Participation inequality: Encouraging more users to contribute. *Jakob Nielsen's Alertbox* (2006). http://www.useit.com/alertbox/participation_inequality.html.
21. Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., and Biewald, L. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Proc. HComp* (2011).
22. Pirolli, P., and Card, S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *The Analyst 2005*, 1–6.
23. Quinn, A. J., and Bederson, B. B. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proc. CHI* (2011).
24. Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. The cost structure of sensemaking. In *Proc. CHI* (1993), 269–276.
25. Shaw, A. D., Horton, J. J., and Chen, D. L. Designing incentives for inexpert human raters. In *Proc. CSCW* (2011), 275–284.
26. Sorokin, A., and Forsyth, D. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops* (2008), 1 –8.
27. Swivel. http://www.swivel.com.
28. Tableau. http://www.tableausoftware.com.
29. Viégas, F., Wattenberg, M., McKeon, M., Van Ham, F., and Kriss, J. Harry potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools. In *Proc. HICSS*, Citeseer (2008).
30. Viégas, F., Wattenberg, M., Van Ham, F., Kriss, J., and McKeon, M. Manyeyes: a site for visualization at internet scale. *IEEE TVCG 13*, 6 (2007), 1121–1128.
31. Whittaker, S., Terveen, L., Hill, W., and Cherny, L. The dynamics of mass interaction. In *Proc. CSCW* (1998), 257–264.
32. Willett, W., Heer, J., Agrawala, M., and Hellerstein, J. CommentSpace: Structured Support for Collaborative Visual Analysis. In *Proc. CHI* (2011).
33. Wills, G., and Wilkinson, L. Autovis: automatic visualization. *Information Visualization 9* (March 2010), 47–69.