

Tools & Strategies for Social Data Analysis

Wesley Jay Willett



Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-224

<http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-224.html>

December 3, 2012

Copyright © 2012, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Tools & Strategies for Social Data Analysis

by

Wesley Jay Willett

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Maneesh Agrawala, Chair
Professor Joseph M. Hellerstein
Assistant Professor Björn Hartmann
Professor Greg Niemeyer

Fall 2012

Tools & Strategies for Social Data Analysis

Copyright 2012

by

Wesley Jay Willett

Abstract

Tools & Strategies for Social Data Analysis

by

Wesley Jay Willett

Doctor of Philosophy in Computer Science

University of California, Berkeley

Associate Professor Maneesh Agrawala, Chair

Data analysis is often a complex, iterative process that involves a variety of stakeholders and requires a range of technical and professional competencies. However, in practice, tools for visualizing, analyzing, and communicating insights from data have primarily been designed to support individual users.

In the past decade a handful of research systems like sense.us and Many Eyes have begun to explore how web-based visualization tools can allow larger groups of users to participate in analyses. Commercial data visualization tools such as Tableau and Spotfire have also begun to embrace the increasingly social web with support for sharing, discussion, and embedding for wider audiences. Social data analysis tools like these mark the beginning of a great sea change in the way we think about data, its impact on our lives, and the ways in which we interact with it. These systems point towards a future in which large teams, communities, and crowds can participate in the collection, discussion, and analysis of data, and benefit from it. Collaborative tools will also improve the quality of analyses by allowing analysis teams to work together more closely—sharing ideas, hypotheses, and findings—and allowing groups with heterogeneous expertise to bring their individual strengths to bear to solve data-driven problems.

However, tools for collaboratively authoring, sharing, and exploring visualizations remain embryonic. The design space of tools for collaborative visual analysis is still largely unexplored and models for understanding the collaboration between analysts, domain experts, and novice participants are limited. This thesis contributes a suite of systems and experiments that explore key aspects of social data analysis and investigate how collaborative data analysis tools can support multiple classes of stakeholders.

First, we explore the design of asynchronous tools for team-based collaboration and analysis and examine how they can facilitate more productive collaboration. We present an interactive tool, CommentSpace, that allows analysts to discuss visualizations and other analytic content. Using CommentSpace, we explore how lightweight collaboration mechanisms like tagging and linking can help collaborators organize their findings and build common ground.

The growing ubiquity of sensing and analysis tools also opens the door to a range of new non-traditional participants in data analysis. We explore the role of social data analysis tools in citizen science—a domain where novice community members are increasingly engaged in data collection and have the potential to contribute to analysis as well. We examine how analysis tools can be tailored to scaffold novice users into the process of data analysis, encouraging participation and understanding while contributing valuable local insights.

Finally, we explore mechanisms for scaling and parallelizing data analysis, even in the absence of a dedicated community or team of analysts. We investigate how individual analysts can crowdsource pieces of social data analysis tasks using paid workers in order to leverage the collective effort of many participants. We demonstrate how large groups of workers can perform cognitively complex tasks like generating and rating hypotheses, and provide tools to help analysts manage the results of this process.

These tools and strategies, along with our evaluations of them, highlight the potential of social data analysis in a variety of settings with different kinds of stakeholders. Moreover, our findings suggest leverage points for future social data analysis systems.

To Jeannine & Kyler!

Contents

Contents	ii
List of Figures	vi
List of Tables	viii
1 Introduction	1
2 Related Work	5
2.1 Sensemaking	5
2.2 Social Data Analysis	7
2.3 Designing for Collaboration and Analysis	9
2.4 Citizen Science and Environmental Monitoring	11
2.5 Crowdsourcing	11
3 Structured Support for Collaborative Visual Analysis	13
3.1 CommentSpace	15
3.2 Tags and Links	18
3.3 Design Details	19
3.3.1 State Saving and Visualization Support	20
3.3.2 Social Sharing and Filtering	20
3.4 Evaluation	21
3.4.1 Study 1: Tagging and Linking in Analysis Subtasks	22
3.4.2 Live Deployments and Exploratory Analysis	27
3.4.3 Study 2: Exploration, Organization and Synthesis	28
3.5 Discussion	32

4	Scaffolding Mobile Sensing and Analysis for Novices	35
4.1	Motivating Fieldwork	37
4.1.1	Methods	37
4.1.2	Personas	37
4.1.3	Design Principles	38
4.1.4	A Framework for Knowledge Generation in Citizen Science	39
4.2	The Common Sense Community Site	44
4.2.1	Collecting Data	44
4.2.2	Applications	45
4.3	Evaluation	51
4.3.1	Methods	51
4.3.2	Scaffolding and Navigation Strategies	52
4.3.3	Usability	53
4.4	Discussion	54
4.4.1	Health and Personal Safety	54
4.4.2	Socializing	54
4.4.3	Exposing Preconceived Notions	55
4.4.4	Visualizations as a Catalyst for Discussion	55
4.5	Additional Design Considerations	56
4.5.1	Qualitative vs. Quantitative data collection	56
4.5.2	Privacy and Security	56
4.5.3	Stakeholder Goals and Competing Interests	57
4.5.4	Importance of Discussion Tools	57
5	Crowdsourcing Social Data Analysis	58
5.1	A Workflow for Crowdsourcing Data Analysis	62
5.1.1	Rating, Clustering, and Checking Explanations	63
5.1.2	Examining and Managing Explanations	66
5.2	Strategies for Eliciting Good Explanations	66
5.2.1	Problem 1: Irrelevant Explanations	66
5.2.2	Problem 2: Unclear Expectations	68
5.2.3	Problem 3: Speculative Explanations	68
5.2.4	Problem 4: Inattention to Chart Detail	69
5.2.5	Problem 5: Lack of Diversity	70

5.3	Assessing Explanation Plausibility	70
5.4	Identifying Redundancy via Crowdsourcing	72
5.4.1	Distributed Comparison	73
5.4.2	Manual Clustering	75
5.5	Explanation Provenance	78
5.5.1	Logging Activity and Sources	79
5.5.2	Supporting Fine-Grained Citations	80
5.5.3	Detecting Copying and Paraphrasing	81
5.6	Deployment	81
5.7	Evaluation	84
5.7.1	Experiment 1: Strategies S1-S5 in Two Worker Pools	84
5.7.2	Experiment 2: Exploring Individual Strategies	88
5.7.3	Experiment 3: Reference Gathering	91
5.7.4	Experiment 4: Annotation Strategies	92
5.7.5	Experiment 5: Iteration	92
5.7.6	Experiment 6: Rating	93
5.7.7	Experiment 7: Redundancy	94
5.7.8	Experiment 8: Copying and Paraphrasing	98
5.8	The Explanation Management Interface	98
5.8.1	Surfacing Explanation Clarity and Specificity	100
5.8.2	Surfacing Explanation Frequency	100
5.8.3	Surfacing Explanation Provenance	101
5.8.4	Surfacing Paraphrasing and Worker Additions	102
5.8.5	Surfacing Corroborating Explanations	102
5.9	Discussion	103
5.9.1	Explanation Segmentation	103
5.9.2	Defining Redundancy	103
5.9.3	Crowd Composition	104
5.9.4	Economics of Crowd Work	104

6	Future Work	105
6.1	Alternate Models for Crowdsourcing Analysis	105
6.2	Engaging Domain Experts	106
6.3	Supporting Ad Hoc Social Data Analysis	108
6.4	Visualization, Presentation, and Storytelling	112
7	Conclusion	115
	Bibliography	117

List of Figures

2.1	Pirolli and Card's sensemaking cycle for intelligence analysts.	6
2.2	Heer et al.'s Sense.us system.	7
3.1	Overview of the CommentSpace user interface.	14
3.2	A CommentSpace discussion showing views from across the web.	19
3.3	Share dialog from a version of CommentSpace with Facebook integration.	21
3.4	Versions of the interface seen in the <i>tag</i> (left) and <i>no-tag</i> (right) conditions.	22
3.5	Interactive visualization of occupation data used in tasks A and B.	23
3.6	Timing of search and filtering operations in Task 1.	26
3.7	Interactive visualization of college return on investment data used in Study 2.	29
4.1	A personal air quality sensor (left). Community members with sensors (right).	36
4.2	Our framework for knowledge generation in citizen science, including personas.	40
4.3	Our framework for knowledge generation in citizen science, including tools.	45
4.4	The Common Sense Community Site showing data collected by a single user.	46
4.5	Two views of the My Exposure application.	47
4.6	The Common Sense Community Hotspots visualization.	49
4.7	The Common Sense Community Comparisons visualization.	50
5.1	Good and bad comments on social data analysis sites.	59
5.2	Our workflow for crowdsourcing data analysis.	60
5.3	An example analysis microtask showing a single chart and prompts.	64
5.4	An example rating microtask showing a single chart and multiple explanations.	65
5.5	Sample charts from the oil production and US census datasets used in our examples.	67
5.6	The <i>distributed comparison</i> interface.	74
5.7	The <i>manual clustering</i> interface.	76

5.8	Illustration of our algorithm to select good worker clusterings from a larger set of possible clusterings.	77
5.9	Analysis microtask interface with proxy web browser and highlighting tools.	79
5.10	Proxy infrastructure for serving analysis microtasks.	80
5.11	An example of an high-quality explanation generated by a crowd worker for a chart showing changes in foreign holdings of US sovereign debt since 2006.	82
5.12	Sample explanations generated for charts showing university tuition and graduation rates, olympic medal counts by country, and historical batting averages.	83
5.13	Percent of responses containing an explanation and average explanation quality, by worker group (<i>US / non-US</i> workers) and strategy condition (<i>strategies / no-strategies</i>) in Experiment 1.	87
5.14	Percent increase in the number of references to the prompted feature and the average explanation quality score for each feature-oriented prompt (S1) condition in Experiment 2 over the control condition.	89
5.15	Average response quality by prompts (<i>prompt-trend, prompt-peaks, prompt-slopes, or prompt-control</i>) and examples (<i>examples, no-examples</i>).	90
5.16	Agreement between workers' ratings and our own increases if we use the mean or median quality score from multiple workers. Using the mean from 5 or more workers gives strong ($\rho > 0.7$) agreement.	94
5.17	F-measure results for each of our clustering selection methods.	97
5.18	The explanation-management interface.	99
5.19	A closeup of the explanation-management interface introduced in Figure 5.18.	101
6.1	Mockup of a possible design for an embeddable CommentSpace web widget designed to be embedded in Q&A sites, forums, and personal communications.	108
6.2	Several views of the Google Books Ngram viewer.	110
6.3	The CommentSpace slideshow extension.	112

List of Tables

3.1	Average Fleiss's kappa values showing within- and between-group agreement for expert, tag, and no tag groups.	25
4.1	Some of the key personas derived from our initial fieldwork.	38
4.2	Framework phases in detail.	41

Acknowledgments

I owe an immense debt of gratitude to all of the incredible people who have contributed to my experience at Berkeley and who have made an impact on me, both as a scholar and as a individual.

Thanks go first to my committee—Maneesh Agrawala, Joe Hellerstein, Björn Hartmann, and Greg Niemeyer—you have encouraged me to do great work and I have enjoyed the pleasure of working with and learning from each of you. Special thanks go to Maneesh—who for the past six years has tolerated my digressions, called my bluffs, and always pushed me to be more rigorous and to do research for the right reasons. I would also like to thank Jeff Heer, who—both as a senior grad student and junior professor—has been a constant source of mentorship and insight, and has effectively been my second advisor.

I also want to express my appreciation to Clayton Lewis, who impressed upon me the importance of the human side of computing and who, almost singlehandedly, made an HCI researcher out of me.

During my time at Berkeley, I have had the privilege of sharing spaces and conversations with an exceptional cohort. Specifically, I would like to thank all of the current and former citizens of the Berkeley Institute of Design (BiD) and the Visualization Lab, especially Andy Carle, Nicholas Kong, Kenrick Kin, Kenghao Chang, Ana Ramirez Chang, Anuj Tewari, Tye Rattenbury, David Sun, Shiry Ginosar, Valkyrie Savage, Celeste Roschuni, and Mark Fuge. You have been great company!

As a graduate student, I have also had the pleasure of working and publishing with a fantastic set of outside researchers. Early on, at Adobe and Tableau, Jock Mackinlay, Chris Stolte, and David Salesin set me on my course as a researcher and seeded so many insights that would bloom later on. At Intel Research, I was lucky enough to not only work with Rob Ennals but also experience the one-two punch of Allison Woodruff & Paul Aoki, who taught me that great work often comes from the marriage of contrasts. Most recently, at Google, I was fortunate to collaborate with the unflappable Ed Chi, whose mentorship, prescience, and good nature have been invaluable as I consider the road ahead.

Of course, my experience in Berkeley would have been utterly incomplete without the set of phenomenal friends and housemates I lucked into—kindred spirits who have share my enthusiasm for exploration and who value great adventures, good food, and incredible places as much as I do. I want to thank the entire MTB house crew, and especially Anand Varma, Jacob Siegel-Boettner, Robert H. Dahl II, and Jeffrey Hartnett. You—and all other Cal Cycling compatriots with whom I’ve raced, ridden, skied, climbed, rafted, and gallivanted across California—have helped made the past six years routinely incredible.

Perhaps most importantly, I would like to thank my family—all of the Willetts and Joneses who have been my paragons since long before grad school and to whom I owe any success I have achieved. In particular, I want to thank my uncle Kevin Jones, for his thoughtful discussion and for kindling my academic spark, and my grandfather, James Everett Jones, for his perennial inspiration. Jamo, you are the strong, considerate, and honest person I will always aspire to be!

My sincere thanks also go to Lora Oehlberg, my ever-patient, incredibly creative, and utterly enthusiastic sounding board, co-conspirator, and adventure buddy! (She is super-awesome!)

Finally, thank you to my mother and brother, Jeannine Jones Willett & Kyler James Willett, to whom this thesis is dedicated. I cannot imagine a more selfless, warmhearted, and close-knit family than our triumvirate!

Thank you.

Chapter 1

Introduction

There is an overwhelming amount of data all around us. Big data. Personal data. Data that only people can explain. Data that needs to be examined, considered, and evaluated at a scale larger than that of a single analyst, and that requires more than one individual's expertise. Visualization and statistical analysis tools can augment the process of data analysis and can support cognition and problem solving [17]. However, individuals often don't have the time, knowledge, technical expertise, or diversity of perspectives to tackle large data analysis tasks on their own. As a result, traditional data analysis tools and tools for visual analysis must evolve to support many types of collaborators [46].

The need to collaborate around data can emerge in many different situations. Multiple analysts working on a dataset may need to share questions or views of the data with one another or gather their findings for presentation. Alternatively, communities with local knowledge or a vested stake in the data may wish to engage in the analysis process. Analysts may also wish to call upon domain experts to answer specific questions about their data. For large datasets, analysts may enlist pools of crowd workers to perform specific analysis tasks at scale.

Despite this, current state-of-the-art analysis tools remain targeted mostly at trained data analysts—individuals with statistical and analytical expertise and proficiency with analysis tools and methods. As we enter a world where sensor data, social information, business metrics, and a multitude of other measures and models are ubiquitous, these analysis tools and the strategies with which they are deployed need to transform. Future tools need to support *social data analysis*, in which groups of people come together to explore and make sense of data in a collaborative fashion.

Social data analysis assumes that sensemaking is not only a perceptual and cognitive activity, but also a social one, in which group interpretation and deliberation are essential components of the social data analysis process. As analysts collaborate, they contribute their own contextual knowledge and extend the work of others [48, 104, 82]. Such collaboration distributes the effort required to examine large data sets and helps analysts develop a shared interpretation of the data.

While some recent tools have begun to offer analysts the ability to collaborate around visualizations and share their work via social media, researchers have just begun to explore how this collaboration impacts the outcomes of analysis. As a next step, we need to understand the kinds of activities that take place during social data analysis and design tools that improve analytic outcomes help analysts build common ground and connect important observations.

By focusing primarily on trained analysts, existing tools have also made little effort to engage new and novice stakeholders or explore how analysis tools can support a variety of different users. Collaborative sensemaking tools must support group exploration and evidence gathering tasks by helping users build on one another's findings and pool their efforts to collectively organize and synthesize them. However, current tools do little to support the integration of effort from multiple individuals. Moreover, current systems do not structure the analysis process in ways that make it easy to take advantage of large new pools of collaborators, including novice community members and paid crowds.

Good collaborative analysis requires utilizing collaborators in ways that effectively allocate their abilities. Making sense of datasets requires more than just looking at them. Rather, data analysis is a complex, iterative process in which participants must search for important features in data, generate and test hypotheses, make inferences, and evaluate relationships between sources and datasets [83]. As such, the next generations of analysis tools must make it easier for collaborators to engage in the analysis process in ways that leverage their relative strengths and encourage good analytic outcomes.

This thesis focuses on supporting and understanding the process of collaboratively analyzing data. We explore this process through the design of a series of social data analysis systems and via experiments involving a variety of different user groups. To set the stage, we revisit prior work in collaborative visual analysis. Next we explore each of our core research questions by iteratively designing, building, and testing a set of collaborative visual analysis tools. We evaluate each of these systems via experiments and live deployments using communities and crowds and we present new tools and strategies for social data analysis based on our experience. This work is focused around three

core research thrusts, each of which examines the design of social data analysis tools through a different lens:

1. *Providing social tools that let analysts organize findings and facilitate deeper analytic reasoning.*

We first consider the design of social data analysis tools intended to support teams of analysts. We introduce *CommentSpace*, a collaborative system in which analysts comment on visualizations and websites. In *CommentSpace*, we introduce comment tags and links—lightweight organization tools that analysts can use to organize their findings and identify others' contributions. We then describe experiments that explore how tag and link structure can facilitate productive collaboration. In a pair of studies comparing *CommentSpace* to a system without support for tags and links, we find that a small, fixed vocabulary of tags (*question*, *hypothesis*, *to-do*) and links (*evidence-for*, *evidence-against*) helps analysts more consistently and accurately classify evidence and establish common ground. We also find that managing and incentivizing participation is important for analysts to progress from exploratory analysis to deeper analytical tasks. Finally, we demonstrate that tags and links can help teams complete evidence gathering and synthesis tasks and that organizing comments using tags and links improves analytic results.

2. *Scaffolding analysis in novice communities to enable participation and cultivate local insights.*

As data collection and visualization tools become more widespread, non-professional users and novice community members are also increasingly becoming involved in sensing and analysis activities. This emergent “citizen sensing” movement has the potential to engage vast communities of novice users and tap these communities' unique local knowledge and experience. We explore how collaboration tools can be tailored to support novice communities and scaffold novice users into the process of data analysis. We present design principles and a framework for data collection and knowledge generation in citizen science settings. We also describe *Common Sense Community*, a community tool for analyzing air quality data. Unlike prior systems, ours breaks analysis tasks into small, discrete applications designed to facilitate novice contributions. These applications use novices' interest in their own exposure to air quality as an entry point and provide gateways to encourage users to annotate, inspect, and validate their community's data. An evaluation we conducted with community members in an area with air quality concerns indicates that these applications help participants identify relevant phenomena and generate local knowledge contributions.

3. *Parallelizing analysis by using paid crowds to generate and assess explanations.*

Many datasets and analyses are too large to be managed by a single analyst or even a small team. While some data analysis problems have natural communities of interest who can help make sense of data, this is often not the case. Moreover, analysts often have no easy way of motivating large numbers of users to participate or ensuring that users make high-quality analytic contributions. Consequently, we also investigate methods for incorporating paid crowd workers into the process of data analysis. We propose a workflow in which an analyst selects a set of charts and asks paid crowd workers to explain specific visual features like outliers, trends, and patterns. We expose this information in an explanation-management interface that allows analysts to interactively filter and sort responses, select the most plausible explanations and decide which directions to explore further. Based on our experiences deploying this workflow, we outline strategies for increasing the quality of crowdsourced results, collecting explanation provenance, and handling redundancy and validate them through a series of experiments.

Finally, we propose additional extensions and future work not addressed in the thesis, including possible alternate models for crowdsourcing analysis, strategies for engaging domain experts, and supporting ad hoc analysis via social media. We also consider how future social data analysis tools might help users move beyond analysis to support data-driven presentations and storytelling.

Chapter 2

Related Work

This thesis builds on prior work in the analysis and information visualization communities, as well as more recent work on citizen sensing and crowdsourcing.

2.1 Sensemaking

At the core of data analysis is sensemaking—the iterative process by which we collect, parse, filter, organize, and manipulate data to solve problems, make decisions, and communicate results [83]. Prior research offers a number of lenses through which we can analyze the process of sensemaking and design to support it. Early work by Russel et al. [88] introduced a formal model that segmented the process of sensemaking into a four-stage “learning loop” in which users looking at a problem or dataset (1) search for possible representations, (2) collect data and use it to instantiate those representations, (3) shift, merge, and otherwise adjust representations, and (4) consume the results to process data or make decisions. The learning loop model provides a very high-level overview for describing a variety of problem solving tasks. It also highlights steps in the sensemaking process (data extraction, clustering results, etc.) where automated tools can make a difference for many kinds of learning tasks.

While Russel, et al. frame sensemaking broadly, Pirolli and Card [83] refine their model from a data analysis perspective. Based on cognitive task analyses conducted with intelligence analysts, Pirolli and Card introduce a more nuanced sensemaking model (Figure 2.1). This model breaks the loop into two sub-cycles—a foraging loop and a sensemaking loop, each with discrete sub-steps,

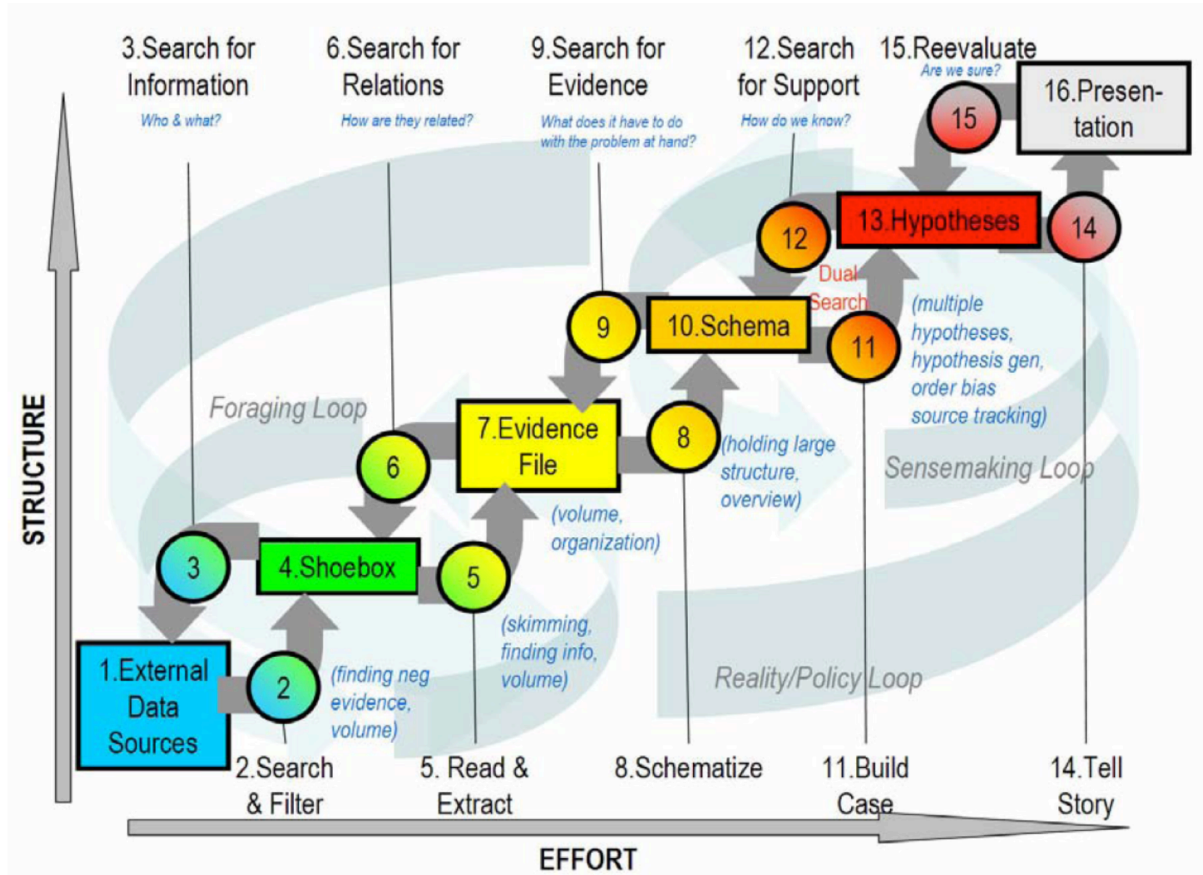


Figure 2.1: Pirolli and Card's sensemaking cycle for intelligence analysts (originally published in [83]). The model breaks the sensemaking process into two sub-cycles—a foraging loop and a sensemaking loop, each with discrete sub-steps.

and highlights specific data mining and analysis tasks like generating hypotheses and collecting evidence. Meanwhile, Card et al. [17] underscore the importance of visualization and interaction to the sensemaking process. They illustrate how data visualizations can bring human perception to bear to identify trends and outliers, filter and refine datasets to identify key elements of interest, and engage in iterative refinement to solve data-driven problems.

Together, this prior work [83, 17] emphasizes the iterative and multi-stage nature of sensemaking. This work also highlights how visualization and organization tools can serve a key role at specific stages in the sensemaking process. Historically, the sensemaking literature tended to consider sensemaking largely from a single-user perspective. However, because these models break the analysis process into discrete steps, they can help developers identify opportunities for collabora-

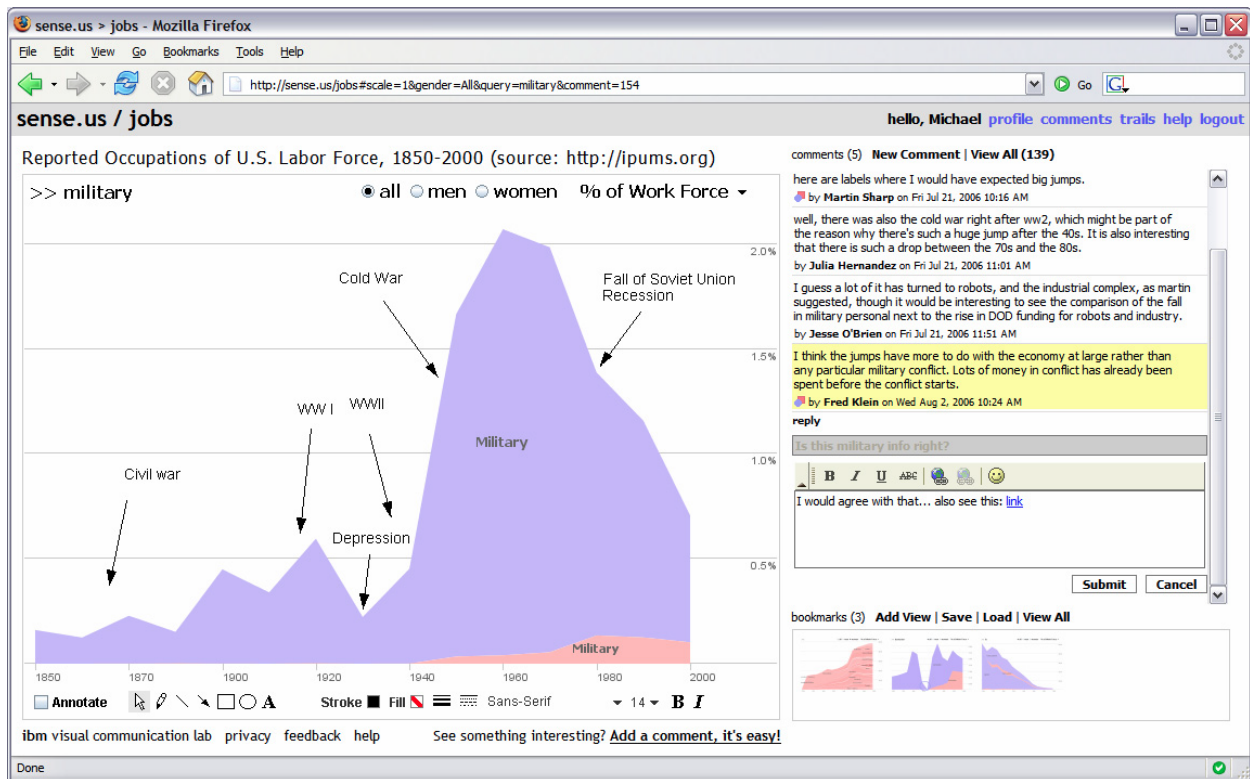


Figure 2.2: A screenshot of Heer et al.’s Sense.us system (originally published in [48]). Sense.us pairs an interactive visualization of the U.S. labor force (left) with discussion tools (right). The visualization is a stacked time-series showing the U.S. labor force, broken down by occupation and gender. The current view shows the percentage of the workforce employed in the military.

tion and parallelization within the analysis process. Analysis steps such as generating hypotheses, searching for evidence, and organizing content are all amenable to parallelization and can benefit from the diverse perspectives and expertise of multiple collaborators.

2.2 Social Data Analysis

The past half-decade has seen considerable research on tools for supporting sharing and collaboration using visualizations, both in academia and in the commercial sector. Heer and Agrawala [46] survey a wide range of asynchronous collaboration tools and discuss design considerations for these collaborative visualization tools. Our work is positioned squarely in this space.

The systems and strategies we explore build upon prior web-based tools for social data analysis including sense.us [48], Many Eyes [111, 27] that support forum-style discussion around interactive data visualizations. Both of these tools pair interactive data visualizations with integrated commenting and discussion tools. Sense.us (Figure 2.2) focuses on providing a set of highly-interactive visualizations of US census data paired with threaded comments that can be attached to individual views and graphical annotation tools that let users sketch and add highlights on top of the visualizations. Many Eyes, in contrast, focuses on providing a set of visualization templates that users on the web can populate with their own data. These visualizations appear on the Many Eyes site, but can also be embedded in outside web pages. The tool makes complex visualizations like interactive bubble charts and word diagrams available to a wide audience and, over the past five years, has allowed many thousands of users to generate and share visualizations across the web. Both Sense.us and Many Eyes demonstrate the power and potential of web-based data analysis with many participants, but they also highlight some of the shortcomings associated with these tools. Integrating commenting and discussion with visualization is difficult—Many Eyes places commenting below-the-fold where it is often missed, while comments in Sense.us are attached to individual views and can be difficult to find or organize. Incentivizing users and getting good analytic results from groups can also be difficult, since users may not have a good understanding of the analysis process and may not engage productively with one another.

Other recent tools and frameworks for systematizing the analysis process have also included a social component. Perer and Schneiderman’s Systematic-Yet-Flexible (SYF) framework [80] walks analysts step-by-step through a set of guided tasks in order to analyze a dataset. In the process, SYF users can create and share annotations that are associated with free text discussion. Eccles, et al’s GeoTime Stories [33] pairs a geospatial and timeline visualization tool and also adds support for annotation via a shared document editor. However, the social features in these tools have tended to be ancillary, rather than the core thrust of the system. Thus far little research has explored the tradeoffs associated with including social tools or encouraging collaboration during the analysis process.

At the same time, a number of commercial visualization tools allow users to comment on visualizations online, but do so in a limited way. Tableau Server [102] allows analysts to share workbooks that include annotation layers on top of visualizations, while SpotFire Decision Site Posters [105] support comments alongside visualization posters on the web. The past half-decade has also seen the rise of numerous “YouTubes for Data,” including Data360.org [28], Swivel.com [101], Verifiable.com [110], and more recently BuzzData.com [16] and Visual.ly [113]. These sites have focused

on allowing users to share and comment on simple visualizations and infographics, but many have failed to garner significant usage and a large number of these services have subsequently folded.

These efforts serve as a testament to the widespread enthusiasm for social data analysis, but also illustrate some of the problems it poses. While adding support for sharing or commenting to web-based visualizations makes collaboration possible, it does not ensure that collaboration will occur or that it will be fruitful. To produce knowledge from shared observation, we must understand how to organize users as well as the insights they produce. We need to identify points in the sense-making process (Figure 2.1) that can benefit from the efforts of multiple users. Moreover, we must learn how to scaffold both individual users and broader communities into the sensemaking process.

2.3 Designing for Collaboration and Analysis

Just as theories of perception guide the design of visualizations, we look to theories of social interaction to guide the design of the social analysis tools around them. Our thinking about collaboration and analysis behavior draws heavily on the sensemaking literature, but builds on other social theories as well. For example, Clark and Brennan’s [22] research on *common ground*—the shared understanding needed for successful communication—implies that collaborators are more effective when they can refer to a shared visual environment to ground each other’s actions and comments [39]. This observation has led designers of collaborative analysis systems to support synchronous view sharing [5] as well as asynchronous sharing and reference through bookmarking and graphical annotation of visualization states [33, 61].

In the context of asynchronous collaboration, work is often broken down into units that can be completed in parallel. In such situations, collaborators need mechanisms to maintain *awareness* [18, 31] of each other’s actions and to *synthesize* individual contributions [6]. In collaborative visual analysis, synthesis often means integrating comments and annotations associated with particular visualization states or data subsets. To reduce the cost of integration, recent systems have provided keyword search of collected comments and tagging of datasets with arbitrary keyword labels [48, 80, 102]. Others support the creation of “topic hubs” [111] for organizing analyses around topical themes. These systems simplify the process of finding commentary relevant to a topic of interest. To facilitate more consistent results, contributions may also be made more formal; tag vocabularies can be (partially) standardized to provide a shared lexicon for important features of the comments, e.g., to note the presence of a hypothesis or action item [30, 44].

A different approach is to use a shared editing (wiki) model rather than a discussion model. For example, Pathfinder [68] provides wiki-style semi-structured “Milestones” (akin to Wikipedia’s Template Messages [118]) to encourage collaboration. More recent extensions to Many Eyes have also included “wikified” service that enables visualizations to be embedded in wiki text [71]. Similarly, Eccles et al.’s GeoTime Stories uses a single text story that contains links to specific visualization states as a means to share analysis stories [33]. These systems integrate contributions via shared editing and the model remains largely informal: contributions can be arbitrary in nature and analysts perform the integration manually in the text.

Researchers have also explored highly formalized schemes for integrating analytic work. Argumentation systems [44, 14, 81] typically model hypotheses and evidence in a network structure but provide rigid constraints on the forms of input that analysts can make. These highly-formalized models can support computational aggregation and inference. However, formalized models can also increase “viscosity” of the system—making it more difficult to reorganize and manipulate—and place high cognitive demand on users[9], making it more difficult to contribute.

Some systems incorporate similar schemes in a more lightweight fashion: for example, the Analyst’s Sandbox [123] allows analysts to tag observations as evidence for or against a hypothesis using direct manipulation gestures. Tree Trellis and Table Trellis [21] support aggregation and comparison of linked free-text claims, but are intended largely for introspecting existing sets of claims rather than supporting ongoing analysis. Evidence matrices are a similar approach motivated by the theory of Alternative Competing Hypotheses [8]. Multiple hypotheses constitute the rows of the matrix, while collected evidence constitutes the columns. Similar to argumentation structures, the cells of the matrix are populated with scores representing the degree to which the evidence confirms or disputes the hypothesis. Such formal systems may lead to premature commitment since they can force analysts to think synthetically from the start rather than building on exploratory analysis. In contrast, we use a more lightweight model in which analysts can categorize and connect contributions in an ad hoc fashion, supporting both information foraging and synthesis [82].

2.4 Citizen Science and Environmental Monitoring

While early thinking on sensemaking and social data analysis originated largely in the context of business and intelligence analysis, a more grassroots emphasis on data collection and analysis has recently emerged under the banner of citizen science.

Community-based environmental monitoring efforts have a deep and varied history that has been well documented in the environmental justice literature, illustrated by numerous examples of “backpack studies” and volunteer monitoring programs [24]. These examples have demonstrated the effectiveness of community participation in the collection of environmental data. O’Rourke and Macey discuss the use of “bucket brigade” sampling in which a mix of participants in different roles coordinate to carry out observation, sampling, and analysis of refinery emissions [78]. Other work has documented the use of community air quality sensing to identify polluters and enforce standards for diesel bus emissions [72, 66]. This citizen-centric ethos has also begun to surface in government monitoring programs for water quality and waste [74].

Over the past few years, the intersection of the citizen sensing movement and mobile technology has produced an abundance of new tools for distributed, citizen-led collection of environmental data [58, 87, 13, 99]. However, these initiatives have primarily engaged citizens in the process of data collection, deferring data analysis scientists and domain experts.

Luther et al.’s Pathfinder [68] is unique in that it integrates both collaboration and visualization tools to support citizen science tasks. Pathfinder allows communities to share data and use a set of wiki-based collaboration tools to pose hypotheses and discuss findings. However, the system focuses primarily on providing tools for organizing wiki discussions and including visualizations, and does not attempt to scaffold novice users into the analysis process.

2.5 Crowdsourcing

The rise of online labor marketplaces such as Amazon’s Mechanical Turk (www.mturk.com) also has deep implications for data analysis. Human computation—the integration of humans into computational processes to solve problems that computers are not yet well suited for [86]—offers a new set of tools, interaction models, and incentive structures that help us to parallelize and systematize data-driven problem solving.

In the human-computer interaction literature, researchers have focused on the use of paid crowdsourcing to supplement purely computational approaches to problem solving and user testing [86, 59, 37]. In the context of visualization, recent work has used crowdsourced workers to perform graphical perception experiments on the effectiveness of charts and graphs [47, 62] and to annotate data sets for computer vision [95]. Other work has examined how to incorporate human computation into larger workflows. Soylent [7] uses paid workers to perform document editing tasks within a word processor, using a Find-Fix-Verify pattern to break editing tasks into smaller sub-tasks. We also take inspiration from human computation frameworks like Crowd-Forge [60], Jabberwocky [1], and TurKit [67] that provide general-purpose MapReduce-style programming models for leveraging crowds to perform complex tasks. We explore how these kinds of techniques can be applied directly to help break down complex data analysis operations into microtasks so that an analyst can enlist many workers to perform in parallel.

Thus far, the use of paid crowdsourcing for data visualization and analysis has been limited almost exclusively to graphical perception tests [47, 62], social experiments [53], and data collection [107, 94]. However, a number of online websites have recruited volunteer workers to take part in small, “Games With A Purpose”-style [2] visual analysis tasks. Sites like NASA Clickworkers (discussed in [6]), GalaxyZoo [38], and Stardust@Home [115] use redundant workers to annotate visual plots of crater locations while Planet Hunters [38] and DataMarket’s “Hot or Not, for Data” [29] allow volunteers to highlight features in time series data or mark apply a binary “Interesting/Not Interesting” label to a dataset. However, none of these sites allow workers to explain the features they have marked or engage them in deeper analysis.

More recently, researchers have begun to explore the application of distributed collaborators to sensemaking and analysis tasks. Fisher, et al. [36] examine how distributed collaborators iteratively build an understanding of information by organizing their work in shared knowledge maps. However, they focus on small groups of collaborators performing open-ended tasks that are less data-oriented. Other researchers have also explored “instrumenting the crowd” [89, 93] by augmenting crowd-based tasks to track workers’ behavior and automatically assess the quality of their work.

Chapter 3

Structured Support for Collaborative Visual Analysis

As we design tools for social data analysis, we must consider how the systems for discussing and communicating questions, observations, and findings support the analysis process. Prior web-based collaborative visual analysis systems—including sense.us [48], Many Eyes [111], and DecisionSite Posters [105]—facilitate such collaboration by allowing analysts to link freeform text comments and graphic annotations to specific views or states of an interactive visualization. However, these systems have primarily focused on using comments to share questions and observations in exploratory analysis, while ignoring more complex analytical tasks such as gathering evidence, organizing findings, weighing alternatives, and synthesizing results. They provide only basic tools for navigating and organizing the comments, either via bookmark trails [48] or general-purpose tags/topic hubs [102, 111]. As the number of comments grow, making sense of them can become a daunting task. Late-joining collaborators must read through lengthy discussion streams and manually synthesize results.

We explore how visualization tools can provide stronger support for multi-user analysis via the design of CommentSpace (Figure 3.1), a collaborative visual analysis system that enables analysts to annotate visualizations and apply additional kinds of structure. In CommentSpace, analysts

Portions of this chapter previously published by the author, Jeffrey Heer, Joseph M. Hellerstein, and Maneesh Agrawala in [121].

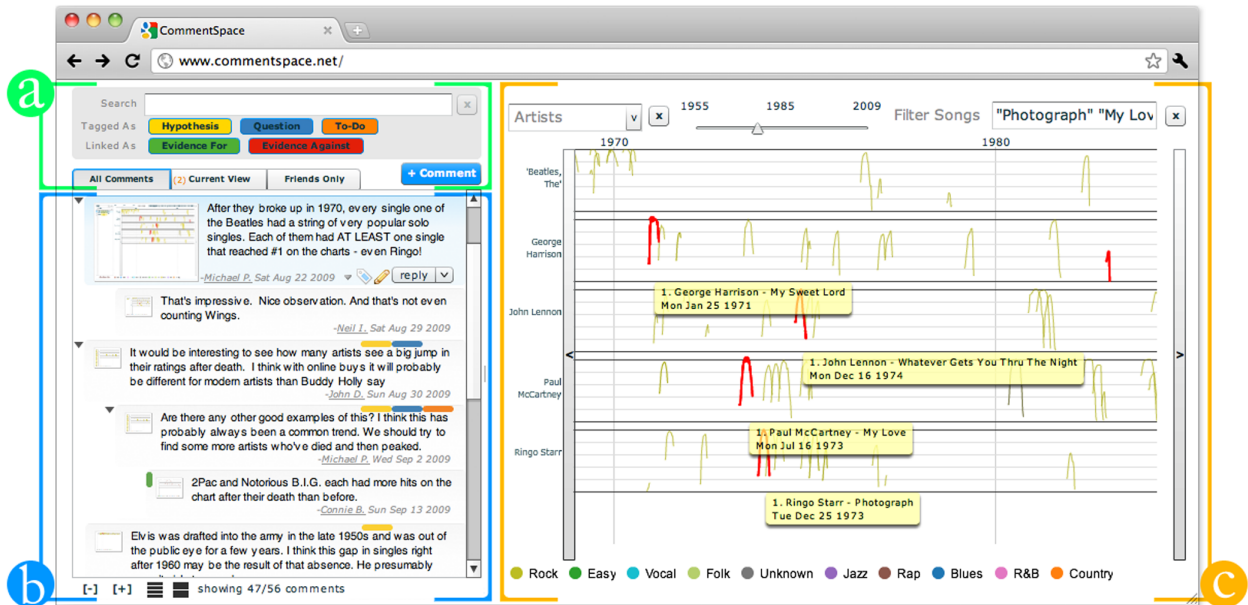


Figure 3.1: CommentSpace provides a threaded discussion area with search and filtering controls (a, b) alongside an interactive visualization (c). This visualization shows data from the Billboard Hot 100 chart—the current view shows the rise and fall of all top 100 hits between 1964 and 1980 by members of the Beatles. Color-coded bars on comments indicate tags and links (e.g. *hypothesis*, *evidence-for*, etc.).

can organize discussions by adding (1) *tags* that consist of descriptive text attached to comments or views; and (2) *links* that denote relationships between two comments or between a comment and a specific visualization state or view. The resulting structure can help analysts navigate, organize, and synthesize the comments, and move beyond exploration to more complex analytical tasks.

In particular, we focus on tags and links that support hypothesis generation and evidence gathering—helping analysts collect and organize new evidence, identify important findings made by others, and synthesize their collective insights. For example, an analyst may tag a comment as a *question* or a *to-do*, indicating a point of interest or contention. Another analyst might then respond by posting a *hypothesis*, to which other analysts might link additional comments or views, specifying *evidence-for* or *evidence-against* relationships. Visualizing hypotheses and support within threaded discussions (Figure 3.1b) can help analysts identify related comments and views and then connect them into coherent arguments and narratives. Tags and links also make it easier to locate comments that are relevant to particular analysis tasks. For instance, a new analyst might filter the comments by the *question* tag to see a list of unanswered questions and check if she can contribute

answers based on her own expertise. Analysts can also use tags and links to organize existing comments and gather scattered evidence for or against a hypothesis in one location. Such structured organization can help them weigh competing evidence and synthesize related comments.

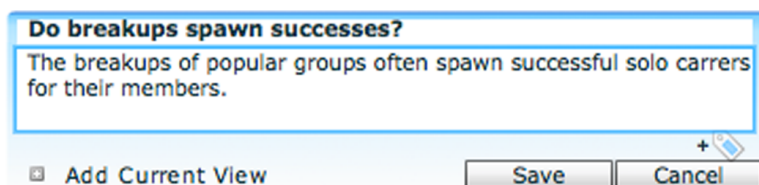
We designed CommentSpace as a modular software component for authoring, structuring, and navigating text comments. CommentSpace can run in conjunction with any interactive visualization system or website that treats each view of the data as a discrete state. The system must produce a vector of state parameters for each view it generates and be able to render a view from a given state vector. Thus, the state vector serves as a bookmark for returning to a view and for linking a view to comments. Using this mechanism, CommentSpace supports discussions that span a variety of websites and visualization systems.

3.1 CommentSpace

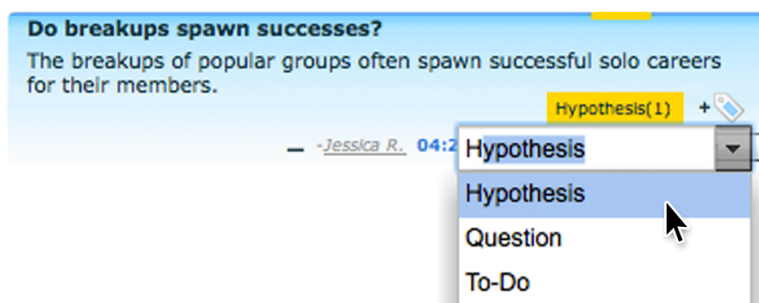
CommentSpace (Figure 3.1) consists of a threaded, forum-like list of comments (a) along with search and filtering tools (b) paired with an interactive visualization (c). The visualization pictured in Figure 3.1 shows data from the Billboard Hot 100 music chart and is based on a design from the New York Times [10]. It depicts the chart rankings of songs by various artists over time. Viewers can observe the rise and fall of individual songs as well as long-term trends in the ranking of artists and genres. They can interactively browse the visualization, hiding and showing artists and filtering to highlight individual songs.

To illustrate the use of CommentSpace, we consider a scenario in which a group of analysts are carrying out an analysis task using this visualization.

While reading through existing comments, Jessica wonders if the breakup of popular groups often spawns successful solo careers for their members. She clicks the *+ comment* button to open a new comment and posts her hypothesis.

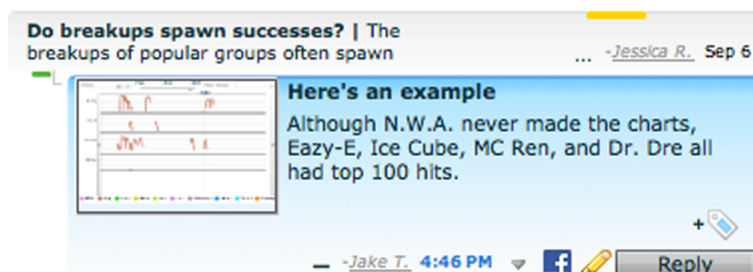


She then tags the comment as a *hypothesis* by clicking the blue tagging menu icon on the comment.



Each tag in our vocabulary is associated with a unique color. A yellow tag marker helps analysts visually identify Jessica's hypothesis as they browse and indicates that the comment is a candidate for further evidence or argument. A tally next to the marker (in this case (1)) indicates the number of analysts who have applied the same tag to this comment.

CommentSpace also supports links that indicate relationships between pairs of comments and between comments and views. Later, a second analyst, Jake, spots Jessica's hypothesis and, intrigued, begins to hunt for supporting evidence. He browses the visualization and builds a view showing the chart success of the former members of California hip-hop group N.W.A. that supports Jessica's claim. He then replies to the original hypothesis, specifying an *evidence-for* relationship, and describes this new view with a comment.

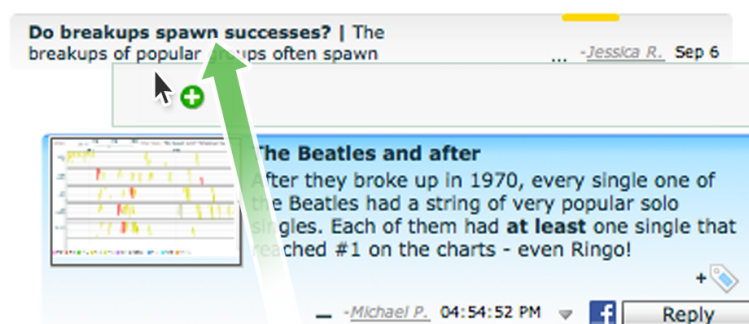


His new observation is threaded into the discussion. It appears below the original hypothesis and is labeled with a small green *evidence-for* link marker on its left side. Jake adds the current view, so a thumbnail of the current visualization state appears next to the comment. Clicking on this thumbnail loads the view into the visualization panel, allowing users to quickly return to it.

Later, Jessica searches for additional evidence relevant to her hypothesis. Using the search controls at the top of the comment panel (Figure 3.1a), she filters to show only those comments containing the words “broke up”. By clicking the legend below the search box, she can refine her search further to show, for example, only comments that are flagged as *hypotheses* or *evidence-for*.



Her search uncovers another observation that shows a long string of hits by John, Paul, George, and Ringo after the breakup of the Beatles (also shown in Figure 3.1). Jessica then drags this observation to her initial comment and links it as *evidence-for* her original hypothesis.



CommentSpace also provides a copy-paste mechanism for linking comments that are distant from one another or visible under different filtering conditions.

The linked comment now appears below her hypothesis in the threaded discussion. Unlike in standard threaded discussions, such linked comments can appear in multiple places in the comment tree, as the linking makes them part of multiple threads. Thus, the original hypothesis serves as a hub for multiple discussions and observations. Other analysts may reply to it or link in additional comments and views from elsewhere. As the set of comments grows over time, Jessica can

quickly return to her original hypothesis comment and filter to see the evidence for and against it. Later, when the analysts begin to organize their findings and synthesize results, they can use tags and links to organize its children into separate chains that contain only the comments that are relevant to their result.

3.2 Tags and Links

CommentSpace introduces a general model in which analysts can tag comments and create links between comments, between visualizations, *and* between comments and visualizations. Analysts can link comments to multiple visualization states and situate them in not just one, but many threaded discussions. For example, the same comment can appear in an ongoing discussion as well as in a collection of evidence for a particular claim. When multiple analysts apply the same tag or link to a comment, the tag's tally increases—indicating agreement on that classification or relationship.

We focus on exploring the impact of a small, fixed vocabulary of tags and links identified through content analyses in prior collaborative visualization systems [48, 112]. Using a breakdown of the comment types generated in sense.us and Many Eyes as a guide, we selected a minimal set of tags that were common, descriptive, and actionable. The set we selected is tailored towards hypothesis generation and evidence gathering tasks and includes tags for identifying *questions* and *hypotheses* as well as links for indicating *evidence-for* and *evidence-against* a hypothesis. We also include a *to-do* tag for indicating unfinished work. Implicit *reply-to* links are used to maintain the threaded conversation structure and *created-on* relationships are generated between comments and the views they are attached to. We used this small, fixed vocabulary because more flexible free tagging vocabularies can take time to evolve and establish tag meanings [20, 40]. A fixed, task-specific vocabulary also limits analysts' ability to apply tags or links whose meaning is ambiguous or generic and forces them to articulate consistent kinds of structure. Using a fixed vocabulary allowed us to explore the impact of tags and links on particular analysis behaviors without the added complexity of an evolving, community-specific vocabulary.

CommentSpace supports “doubly-linked discussion” [48], whereby authors can follow links between comments and views and only the comments associated with the current view are visible. Doubly-linked discussion can facilitate serendipitous discovery of new comments as users interact

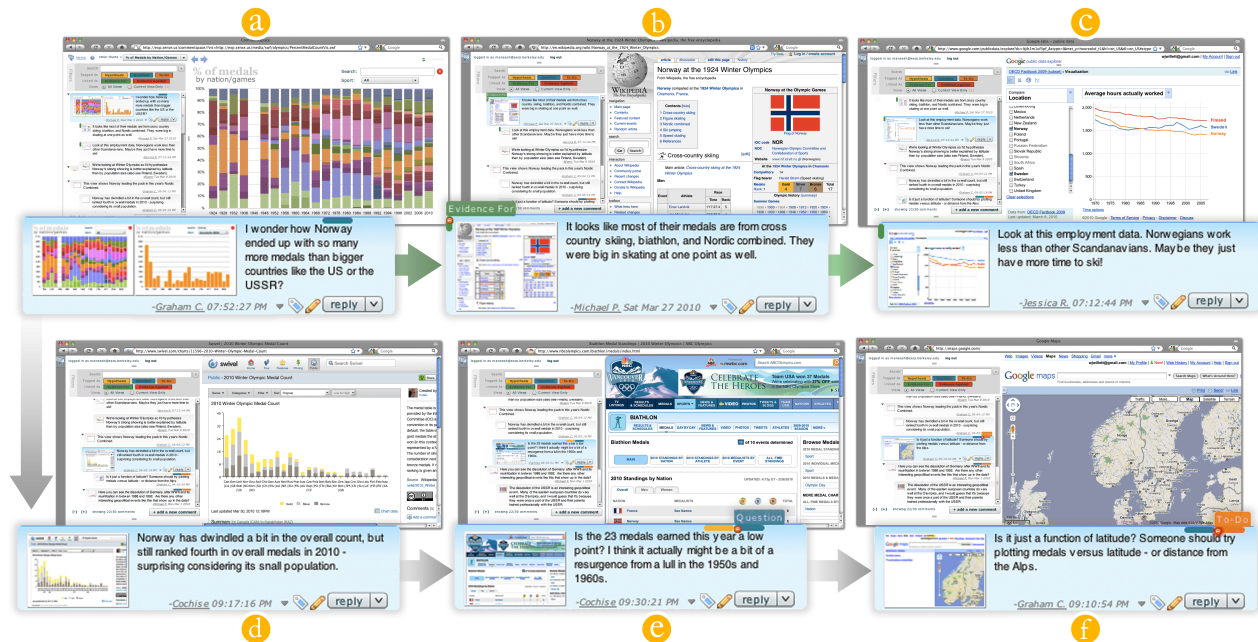


Figure 3.2: Using the Firefox extension, CommentSpace can facilitate discussion across the web. Here, a discussion begins on (a) a custom Flash visualization of medal counts from the Winter Olympics and incorporates information from (b) Wikipedia, (c) a specific view from Google Public Data Explorer, (d) a chart from swivel.com, (e) an official Olympics webpage, and (f) a view from Google Maps. Replies are shown as grey arrows (a→d,d→e,e→f) and *evidence-for* links are illustrated as green arrows (a→b,b→c).

with the visualization, but makes it more difficult for discussions to span multiple views. To address this limitation, CommentSpace allows analysts to toggle between a doubly-linked comment panel that shows only comments for the current view and a version that shows all comments. Unlike in sense.us, this master comment list is visible alongside the visualization and users can toggle between the two comment panels using tabs directly above the panel (Figure 3.1b). This approach encourages discussions that span multiple views and makes it easier to investigate other views without losing track of the current thread.

3.3 Design Details

CommentSpace is implemented as an Adobe Flash application that can be embedded in web pages containing interactive visualizations or run as an extension for the Firefox web browser. When embedded with a set of visualizations on a site, CommentSpace provides a browser-independent

commenting environment that can be tightly coupled with those particular visualizations. When used as a Firefox extension, the commenting panel is accessible via a browser sidebar rather than embedded within the page. This version supports linking to and commenting on visualizations as well as *any* view of a web page with a unique URL. Thus, it enables social discussion and evidence gathering across the web and allows collaborators to incorporate information from outside sources in their analyses, as seen in Figure 3.2.

3.3.1 State Saving and Visualization Support

CommentSpace can be paired with any visualization that implements a simple interface for setting and getting visualization state. The visualization must be able to produce a vector of state parameters for each view it generates, and also render a view from any state vector it produced. These state vectors serve as bookmarks for returning to views or for linking views to comments. Whenever a state change occurs, the visualization must dispatch an event, notifying CommentSpace of the change. Whenever a tag is applied to a comment or a comment is linked to a view, CommentSpace serializes and saves a copy of the state in JavaScript Object Notation (JSON). The CommentSpace web service stores and indexes these state vectors and passes them back to the visualization whenever a state needs to be reloaded.

The browser extension treats URLs as the state vector and thereby makes it possible to link comments to any web page. The extension listens for changes to the current URL (including changes after the fragment identifier, “#”) and generates a state vector incorporating the URL. This approach is well suited for rich Internet applications like Google Public Data Explorer [43] that provide unique URLs at every visualization state, and makes a compelling argument for designers to build visualizations that provide stateful URLs which update dynamically when the view changes [46]. However, we also include site-specific code to extract state vectors from some useful sites like Google Maps that can generate stateful URLs but don’t automatically update the address bar.

3.3.2 Social Sharing and Filtering

As Viégas et al. [112] observed, discussions and continued interactions around visualizations on the web are often more fruitful when they occur within existing communities. To support and encourage analysis within existing groups, CommentSpace also provides several social sharing and filtering tools. Users who log into CommentSpace using a Facebook account can share individual

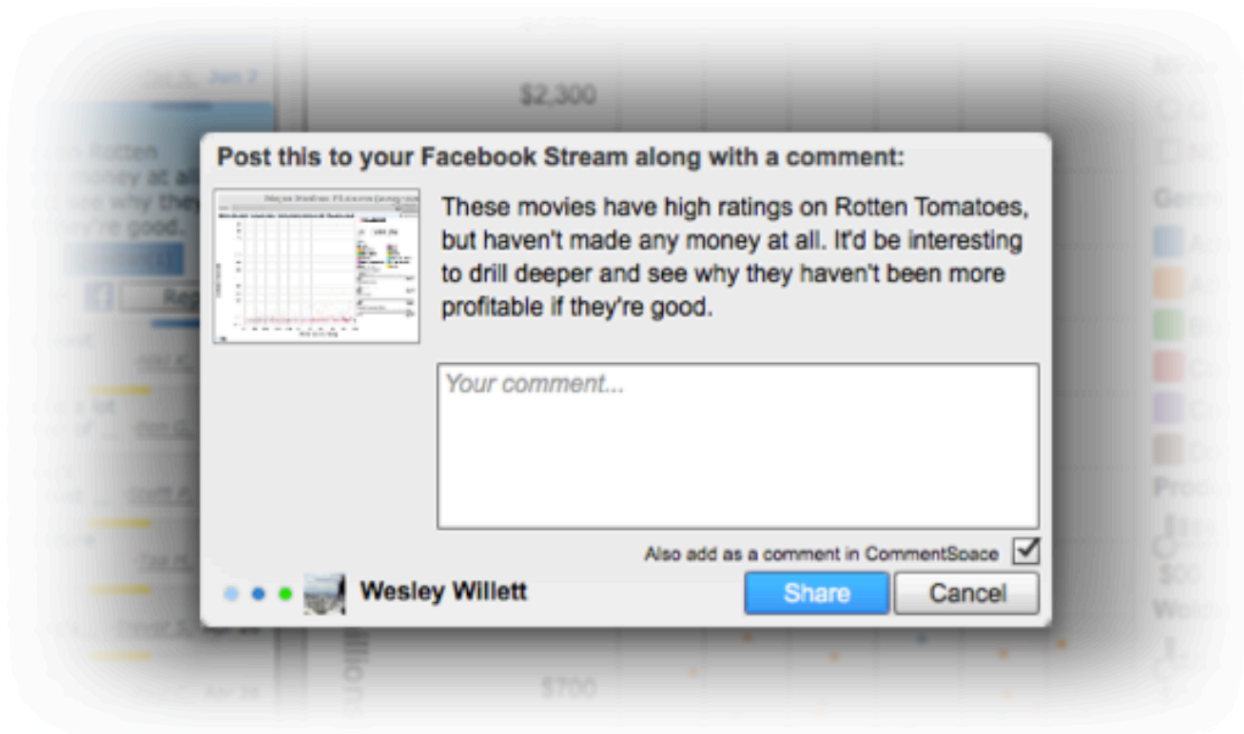


Figure 3.3: Share dialog from a version of CommentSpace with Facebook integration. Copies of comments posted to Facebook via this share dialog are also retained in the CommentSpace comment stream for later analysis.

comments and visualization views via their Facebook stream (Figure 3.3) and can generate unique URLs to share views by email or IM. They can also filter the comment graph using their Facebook contacts, showing only comments generated by neighbors in their social network.

3.4 Evaluation

We conducted two controlled studies and a live deployment to test whether tags and links would improve users' performance on common analysis tasks. In the first study, we tested the impact of tags and links on two specific analysis subtasks: (A) classifying comments left by others and (B) gathering evidence using comments. We also examined usage in a live deployment to assess commenting behavior during exploratory analysis. Finally, we conducted a smaller, qualitative study in which analysts used CommentSpace to perform a complex, multi-stage analysis with exploration, organization, and synthesis phases.

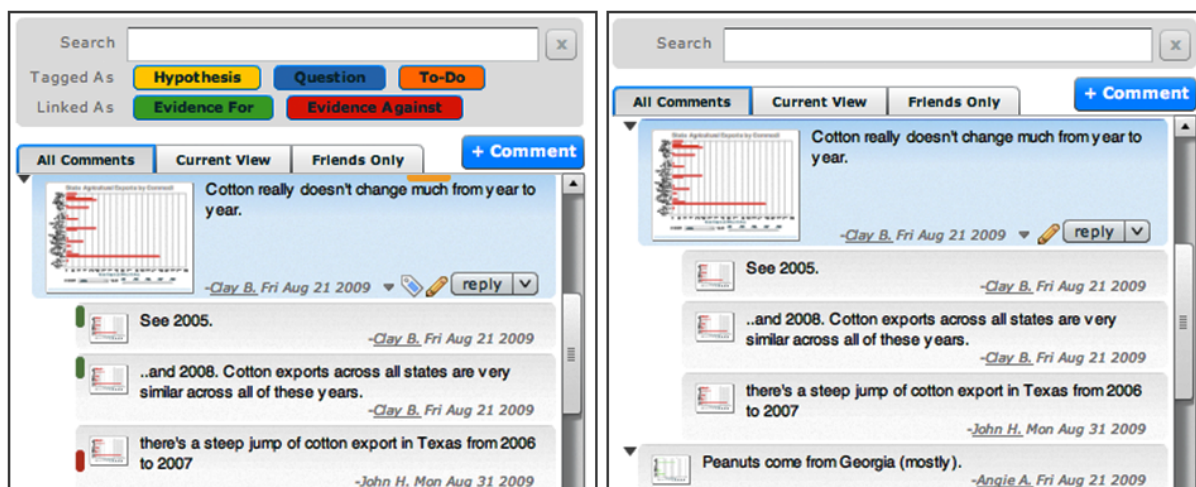


Figure 3.4: Versions of the interface seen in the *tag* (left) and *no-tag* (right) conditions. Users in the *tag* condition gain tag filtering controls and see colored tag and link markers on comments.

In both studies we compared a version of CommentSpace with tags and links (the *tag* condition) to a version similar to sense.us [48] that provided little support for structuring discussion (the *no-tag* condition). In the *no-tag* condition participants could author new comment threads, reply to existing comments and perform text searches but could not author or view tags and links. In the *tag* condition, participants could add *hypothesis*, *question*, and *to-do* tags along with *evidence-for* and *evidence-against* links. Additionally, *tag* participants could search and filter the comments by their tags and links. Figure 3.4 shows the commenting interfaces for the two conditions.

3.4.1 Study 1: Tagging and Linking in Analysis Subtasks

We first explored the effect of tags and links on two evidence gathering subtasks: (A) classifying comments made by others and (B) authoring comments when gathering evidence.

Methods

We recruited 24 paid participants (15 female, 9 male) via mailing lists and a research participation pool. Subjects were university students from a variety of majors. We conducted a between-subjects study in which 12 participants used the *no-tag* interface, while the other 12 used the *tag* interface.

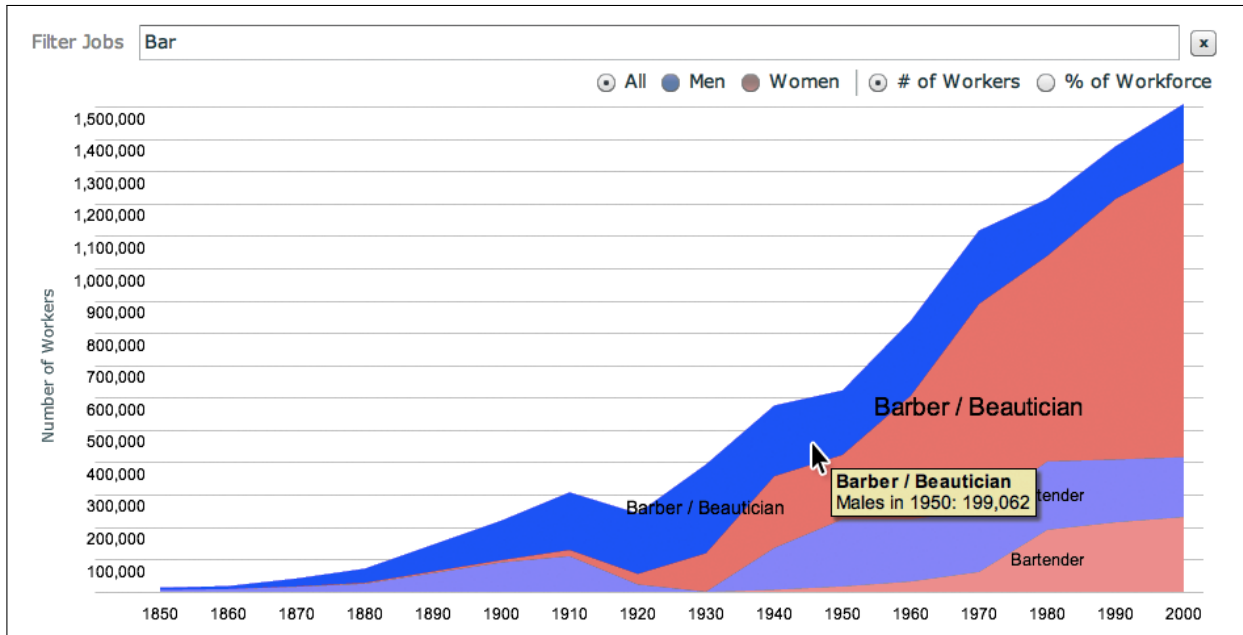


Figure 3.5: Interactive visualization of occupation data used in tasks A and B. This stack graph shows the size of the U.S. workforce since 1850, broken down by occupation and gender.

Task A: Identifying and Classifying Comments. Our first task examined how late-joining analysts navigate existing discussions to find comments relevant to a given hypothesis. It also tested whether the presence of tags and links helps users classify those comments more accurately. We anticipated that tags would provide common ground, leading to more consistent categorization of comments, and would make filtering and search more productive. Specifically, we hypothesized that:

- (1) Users will identify evidence relevant to a particular claim with greater accuracy when tags and links are present.
- (2) Users of a tag-enabled system will use filtering and search tools more extensively to identify relevant evidence.

We gave participants a visualization of U.S. occupation data similar to the one used in sense.us (Figure 3.5) and a corpus of 181 tagged seed comments drawn from that system [48]. The author tagged all *hypotheses*, *questions*, or *to-dos* in this set and added links between each hypothesis and every comment that provided *evidence-for* or *evidence-against* it. During the study, we asked participants

to identify as many comments as possible that provided evidence for or against one specific hypothesis: *Stereotypically male jobs have remained almost entirely male even as women have joined the work force*. The version of the seed corpus shown to participants contained 10 comments linked as *evidence-for* or *evidence-against* this hypothesis. It also included another 12 comments linked to other hypotheses.

We gave participants 15 minutes to examine and categorize comments that provided evidence for, provided evidence against, or were otherwise related to the claim. Since participants in the *no-tag* condition could not mark comments by tagging them, we asked all participants to write the three-digit identification number of each comment in the appropriate column of a paper worksheet. Subjects were not allowed to add comments, tags, or links during this task. The total number of comments was large enough that reading every comment individually in the allotted time was difficult.

As a baseline, three experts (the author and two research collaborators) also independently coded the comments using the same guidelines as the participants, but with no time limit. Out of 181 comments, the experts identified 9 comments as evidence for the claim, 24 comments as evidence against it, and 19 comments as related but not evidence.

Task B: Gathering Evidence as Comments. We designed the second task in Study 1 to explore comment authoring in an evidence-gathering task. We instructed participants to spend 20 minutes locating views and generating comments that provided evidence for or against the claim they investigated in Task A. We told subjects that subsequent users would see their comments when attempting to carry out Task A, and encouraged them to organize their comments so that later users could easily find the relevant ones. All participants began the task with the same set of seed comments they had seen in Task A.

We hypothesized that tags would help users identify unanswered questions and other relevant comments more easily, and that they would encourage users to organize their discussions around those comments. Specifically, users in the *tag* condition would be more likely to reply to existing threads and—in particular—more likely to reply to comments identified as hypotheses or questions.

Within Group Agreement					
Group	Evidence For	Evidence Against	Related	Unrelated	Average Kappa
(E)xpert	0.572	0.553	0.400	0.839	0.590
(T)ag	0.273	0.417	0.113	0.405	0.302
(N)o-tag	0.264	0.285	0.136	0.363	0.262

Between Group Agreement					
Pair	Evidence For	Evidence Against	Related	Unrelated	Average Kappa
E-T	0.335	0.425	0.151	0.444	0.339
E-N	0.314	0.302	0.183	0.412	0.303
T-N	0.276	0.338	0.105	0.384	0.276

Table 3.1: Average Fleiss’s kappa values showing within- and between-group agreement for expert, tag, and no tag groups. A kappa of 0 indicates no agreement, while a kappa of 1 indicates perfect agreement. Color redundantly encodes kappa values—darker colors correspond to higher agreement.

Results

Classifying Comments. To test our first hypothesis, we compared the lists of comments classified by each participant in Task A. Because the data are not normally distributed, we report median and median absolute deviation (MAD) and we use the non-parametric Mann-Whitney U-test for significance. Participants classified a similar number of comments in both conditions, (*Median* = 26.5, *MAD* = 4.5) in the *tag* condition and (*Median* = 25, *MAD* = 5) in *no-tag* and there was no significant difference. However, participants in the *tag* condition categorized significantly more ($U = 32.5, p < 0.024$) comments as *evidence-against* (*Median* = 15, *MAD* = 3) than those in *no-tag* (*Median* = 10, *MAD* = 3), showing that tags and links impacted categorization.

To assess the accuracy of users’ categorizations, we compared the level of agreement between comment categorizations made by our subjects and those made by the experts. We measured *consistency* (agreement with others in the same condition) and *accuracy* (agreement with the experts) by computing average within- and between-group Fleiss’s kappa values based on subjects’ and experts’ categorizations (Table 3.1). In general, the experts were the most consistent, followed by subjects in the *tag* and then *no-tag* conditions. More importantly, the *tag* group was more accu-

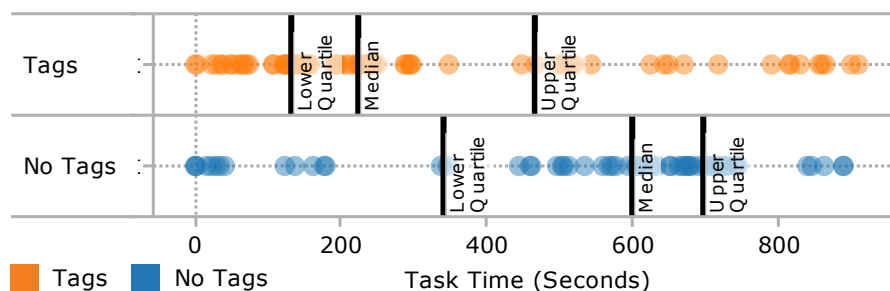


Figure 3.6: Timing of search and filtering operations in Task 1 (in seconds since the beginning of the task).

rate—agreeing with the experts more than the *no-tag* group across each of the categories, with the level of agreement on *evidence-against* being the most pronounced. This improvement indicates that tags and links encourage consistent labeling and improve shared understanding of comments for late-joining participants.

Filtering and Search Because they had access to additional tag and link metadata relevant to their task, we had hypothesized that participants in the *tag* condition would filter and search more extensively.

The activity logs for Task A show more total search and filtering operations by participants in the *tag* condition (*Median* = 10, *MAD* = 6) than the *no-tag* condition (*Median* = 4, *MAD* = 2), but this difference was not significant ($U = 46.5, p = 0.0749$). However, participants in the *tag* condition were far more likely to search and filter early in the task. On average, more than half the search and filtering operations in the *tag* condition came in the first four minutes of the task, while participants in the *no-tag* condition took until almost the ten minute mark to complete half of their filtering and search operations (Figure 3.6). Participants using *tags* searched and filtered significantly earlier than participants in the *no-tag* condition ($U = 2937, p < 0.0005$).

This data provides a possible hypothesis for the increased level of consistency and accuracy in the *tag* condition. Because subjects in the *tag* condition filtered and searched earlier, they were more likely to find clearly marked pieces of evidence early on. This evidence may have helped calibrate their categorization, making them more likely to mark pieces of evidence for and against the prompt consistently and accurately. Meanwhile, our observations of activity traces indicate that *no-tag* subjects were more likely to scroll sequentially through the list of comments, marking comments as *evidence-for* even if they were only marginally related.

Gathering Evidence In Task B, we had hypothesized that users in the *tag* condition would be more likely to respond to existing threads, especially those containing hypotheses or questions. Our results showed that participants generated similar numbers of comments in both the *tag* (*Median* = 12, *MAD* = 4) and *no-tag* (*Median* = 12.5, *MAD* = 4) conditions, but those in the *tag* condition generated significantly more replies (*Median* = 7, *MAD* = 3.5) than those in *no-tag* (*Median* = 2, *MAD* = 1.5) ($U = 32$, $p = 0.0226$). Moreover, a chi-square test shows that participants were significantly more likely to reply to existing discussions when tags were present ($\chi^2(1, 308) = 27.45$, $p < 0.001$), confirming our hypothesis. These results suggest that tags and links helped *tag* participants identify and build upon interesting observations and encouraged them to organize their findings.

3.4.2 Live Deployments and Exploratory Analysis

We also conducted two, one-month live deployments of CommentSpace to test its social sharing and filtering features. During these deployments, we paired CommentSpace with ten different interactive Flash visualizations (including those shown in Figures 3.1, 3.2, 3.5, and 3.7) and made them publicly available at www.commentspace.net. While tagging and linking were available during most of the deployment and were explained on a help page, we did not specifically instruct users to apply tags and links during their analysis.

Over the course of deployment, the site received about 6,000 page views from over 850 unique visitors. Of those visitors, 180 created an account on the site or logged in using a Facebook ID; 32 of those users left a total of 123 comments. While the number of registered users and comments is relatively small, the ratio of comments per user (0.68) is higher than for Many Eyes (0.31), the only comparable social data analysis site for which statistics covering a similar time period after launch were readily available [111].

Most of the analytic behavior reflected in these comments was exploratory. Users authored questions and made observations, but few posited hypotheses or responded to prior comments with pieces of related evidence. The lack of evidence gathering behavior was accompanied by a low level of tagging and linking. During our deployments, users with access to tagging and linking tools authored only 5 tags and a single link.

Based on these experiences in the live deployment as well as earlier pilot studies, we suspect that participants in our open-ended exploratory tasks did not have enough incentive to tag or link comments. Because participants in such tasks have no specific reason to revisit their own comments or

those of others, they have little motivation to organize or label comments during exploration. The superficial nature of users' comments suggests that more specific tasks and incentives are required to facilitate the transition from exploration to more complex modes of analysis. We revisit this observation in Chapter 5, where we demonstrate how analysts can use small monetary incentives to encourage crowd workers to generate explanations and candidate hypotheses en masse.

3.4.3 Study 2: Exploration, Organization and Synthesis

Neither Study 1 nor the live deployment examined how analysts might use tags and links to synthesize new findings and make decisions. In addition we found that users do not have strong incentives to author tags and links during open-ended exploratory analysis. Heer and Agrawala [46] suggest that managing the division of work and providing appropriate incentives are important considerations in designing collaborative visual analysis systems. We designed a second study to investigate these issues.

In Study 2, teams of participants completed a complex three-phase analysis task, consisting of a directed exploration phase, an explicit organization phase in which participants were encouraged to tag and link their comments as evidence for or against specific hypotheses, and a synthesis phase in which they used the organized comments to make decisions and explain them in writing. We managed each phase more explicitly and gave participants greater incentives than in Study 1 or the live deployments. In particular, we gave participants smaller more specific tasks, especially in the organization phase. As a form of social-psychological incentive, we explained how team members would benefit from one another's work and told participants that the best-written synthesis results would receive an extra monetary reward.

Methods

We recruited 12 paid participants via campus mailing lists. We divided participants into two six-person teams; one team worked together using the full, *tag* version of CommentSpace while the other team used the *no-tag* version. We asked teams to carry out a series of exploration, organization and synthesis tasks using an interactive visualization (Figure 3.7) of estimated return on investment for US college students [116]. Each team shared a comment workspace populated with 70 seed comments drawn from earlier pilot studies.

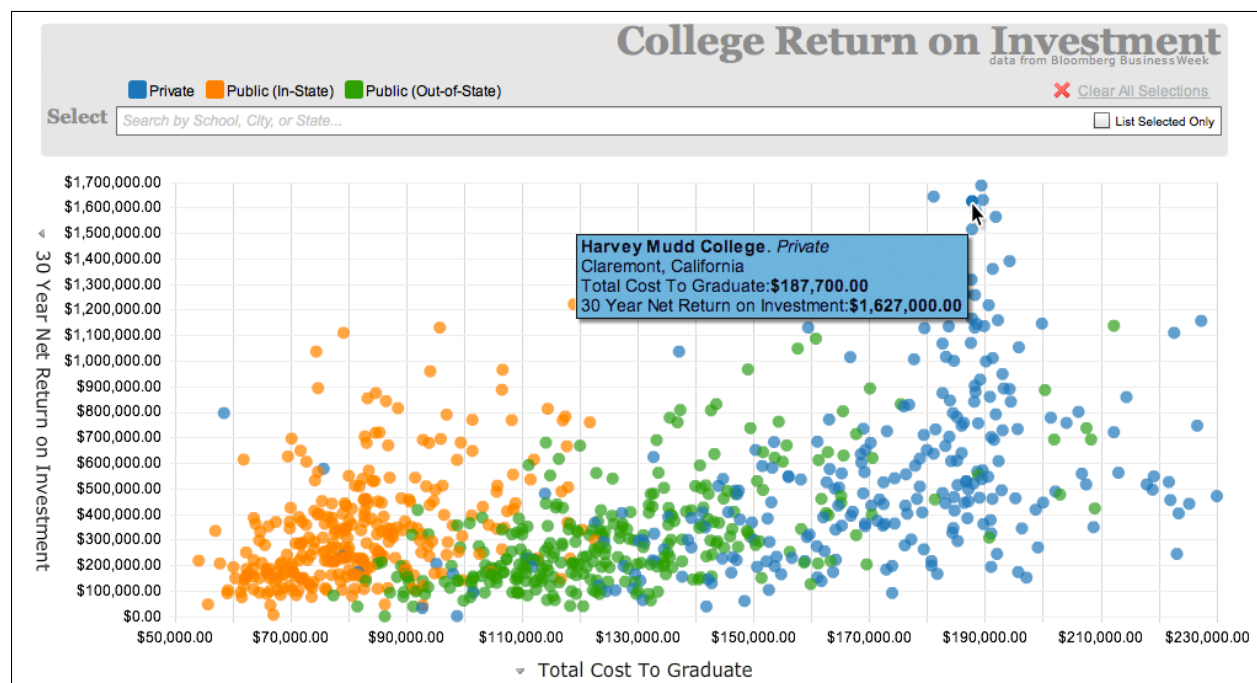


Figure 3.7: Interactive visualization of college return on investment data used in Study 2. This view plots universities according to their graduation rate and annualized return on investment. Color indicates public (in-state or out-of-state) and private universities.

We paid all participants an initial fixed amount (\$20) for participating in the study. In order to encourage participants to actively engage in the tasks, we also promised an additional, larger monetary reward (\$50) to the two participants who produced the best-written results (as scored judged by a team of experts).

In the exploration phase, we instructed participants to explore the visualization and the existing discussion, then leave comments documenting their findings. We encouraged participants to focus on two general areas of inquiry: “*The relationship between graduation rate, the total cost of attendance, and return on investment*” and “*The distribution of schools from each of the university systems in California.*” We gave participants 36 hours to complete the task, and we instructed each participant to leave at least 10 comments.

In the organization phase, we instructed participants in the *tag* condition to organize their team’s comments. We asked subjects to organize comments by topic, tag them, and link evidence to related hypotheses. To focus the task, we provided two hypotheses as prompts: “*There is a clear corre-*

lation between graduation rate, the total cost of attendance, and return on investment” and “There are consistent differences in the graduation rates, tuition, and return on investment between the University of California schools, California State schools, and private universities in California.” We instructed the *tag* participants to add links and tags until they were satisfied with the overall organization of the workspace. Because it was not possible to organize content in the *no-tag* condition, we instead asked *no-tag* participants to spend time reviewing the comments left by their team members. Members of each team carried out the task asynchronously over a 24-hour period. During that time they were free to iterate and build upon one another’s work.

Finally, in the synthesis phase, we asked all participants to complete a decision-making task using the visualization and the comments generated by their team. We posed two decision making tasks based on the earlier prompts. In the first, we asked each subject to “Produce a ranking of the top schools based on the relationship between graduation rate, the total cost of attendance, and return on investment.” In the second, we asked students to “Distribute a pool of imaginary funds amongst the public, in-state, and out-of-state schools in California.” We chose these questions to force participants to think critically and construct an argument that built on the exploratory analysis and organization they had completed. We asked participants to provide a short (1-2 paragraph) response to each prompt and to cite the ID numbers of each of the comments that informed their decision. Participants authored their synthesized responses in a web form, rather than in CommentSpace itself. During this task, participants used CommentSpace to revisit comments and views. They could also copy and paste references to comments directly into their responses. These citations, along with post-study surveys and interviews with select participants, allowed us to connect the synthesis behavior in this phase to the exploration and organization in the earlier phases.

Results

All 12 of our recruits completed the exploration and organization tasks. Of these, ten (6 *tag*, 4 *no-tag*) completed the synthesis task. The two remaining participants dropped out due to scheduling conflicts. We examined all comments generated by the participants and scored them to assess their length, quality, and relevance to topic. We removed one participant in the *tag* condition who produced short, incomplete comments after the task deadlines had expired.

Because of the scope and duration of Study 2, we used a smaller number of participants than in Study 1. Due to the small sample size, most numerical results of this study do not achieve statisti-

cal significance. Nevertheless, we believe the qualitative results and feedback from interviews are indicative of real-world usage by teams of analysts.

Exploration. During exploratory analysis, participants in both conditions authored roughly the minimum number of comments ($Median = 10, MAD = 0$). Three *tag* subjects applied at least one tag, but no participants tagged heavily, and none authored links. This mirrors the results from our live deployment and suggests that organization requires additional motivation. However, our current study does not rule out the possibility that these low numbers could be the result of usability issues or a cognitive mismatch between the task and the tool.

Organization. In the organization task, the five *tag* participants applied 84 tags and 15 links across 60 of the 138 comments in the workspace. *Tag* participants added the majority of their tags (83%) to comments authored by other users, indicating that they actively considered comments other than their own. There was also very little disagreement when tagging. Two or more users added identical tags to 14 comments, but no two users ever added competing tags or links to the same comment. This result suggests that, even without explicit coordination, users can author tags and links that organize the content without conflicting with one another.

While we also asked participants in the *no-tag* condition to review the comments left by other participants during the second phase, our logs show that *no-tag* participants spent less time in this phase ($Median = 12$ minutes, $MAD = 6$ minutes) than *tag* participants ($Median = 23$ minutes, $MAD = 13$ minutes) and examined fewer comments.

Synthesis. We found that *tag* participants produced longer responses in the synthesis task ($Median = 3082$ total characters, $MAD = 574$) than those in the *no-tag* condition ($Median = 1480$ total characters, $MAD = 487$). To compare the quality of the responses, three independent expert evaluators (one of whom was an author) rank-ordered the anonymized responses from best (1) to worst (9) based on their clarity, consistency, and use of comment citations. The average Spearman's rank correlation coefficient between the evaluators was 0.70, indicating good inter-rater reliability. For each response, we averaged the rankings from all three evaluators to compute an average rank. Comparing the average ranks of all responses, we found that *tag* participants ranked significantly better ($Median = 3.83, MAD = 0.5$) than those in the *no tag* group ($Median = 6.17, MAD = 1$) using a Mann-Whitney U test ($U = 5.5, p < 0.0013$). *Tag* participants also cited more comments in

their responses (*Median* = 10, *MAD* = 3) than the *no-tag* participants (*Median* = 6, *MAD* = 1). In addition, 79% of the comments cited by *tag* participants had been tagged or linked in the organization step and comments that had been tagged or linked were nearly three times more likely to be cited than those that had not. These results mirror our post-study interviews, which suggest that the organization task helped *tag* participants gain a better understanding of the findings, which they carried over to the synthesis task.

The stronger synthesis responses authored by *tag* participants reflect both their use of tags and link structures during synthesis and the increased awareness of the comments they gained in the organization task. *Tag* participants spent more time in the organization task than their *no-tag* counterparts and visited more comments and views while doing so. However, *tag* participants also cited comments that had been linked together during organization, but had not previously been adjacent to one another, suggesting that they used the tag and link structure directly when generating their result.

3.5 Discussion

Our studies demonstrate that tags and links can help participants identify and organize information in a collaborative visual analysis tool. We offer a few concrete takeaways regarding the use of tags and links for collaborative evidence gathering and synthesis tasks:

1. *Analysts using tags and links were more consistent and more accurate when classifying comments.* This result suggests that tags and links are useful when establishing common ground and can help late-joining participants get up to speed in ongoing discussions. We note however, that consensus among analysts is not always desirable and may be symptomatic of groupthink. Competing and divergent interpretations are often desired, in which case tag vocabularies need to be designed to encourage this.
2. *Analysts using tags and links searched and filtered significantly earlier and classified content more accurately than no-tag participants.* Tags and links affect how analysts explore and help them calibrate the way they categorize findings. Developers should be careful to select tags and links that encourage desired types of contributions.

3. *Analysts were significantly more likely to reply to existing discussions when tags were present.* This result shows that tags and links encourage contribution and continued discussion and can be used in collaborative visual analysis systems to promote more focused dialog.
4. *In our live deployments and pilots studies, analysts did not have enough incentive to tag or link comments during open-ended exploration.* Because analysts in such tasks often have no immediate reason to revisit their comments, they have little motivation to author additional structure, even if that structure may be useful later. Developers and managers need to guide participation using explicit tasks and incentives in order to facilitate the shift from exploratory analysis to deeper analytical tasks like organization and synthesis. We consider one approach to incentivizing participation in Chapter 5, in which we pay paid crowd workers to perform highly-structured hypothesis-generation and organization tasks.
5. *Tagging and linking resulted in better synthesis when conducted as part of an explicit organization task than when conducted during emergent exploratory analysis.* This result suggests a staged approach to collaborative analysis, wherein users first explore a data set, identifying interesting patterns and outliers, then organize those observations to facilitate deeper analysis. Such behaviors have precedent in Wikipedia, where an entire class of contributors categorize articles written by other editors [114]. The lightweight structure provided by tags and links makes this staging possible.

The stronger results produced by *tag* participants likely reflect both their use of tags and link structures during synthesis and the increased awareness of the comments they gained in the organization task. In Study 2, *Tag* participants spent more time in both the organization and synthesis task and visited more comments and views while doing so. In several cases, *tag* participants cited comments that had been linked together in the organization step, but had not previously been adjacent to one another, suggesting that they used the tag and link structure directly when generating their result. One direction for future work is comparing the importance of *authoring* the structure versus *referencing* it. For example, would participants perform as well if the task of organizing the data was performed by a single moderator, rather than distributed amongst the entire team?

Finally, while we have considered a small set of tags and links tailored to hypothesis generation and evidence gathering, other tasks may be better served by free tagging or by other custom vocabularies. Tasks like clustering, for example, might benefit from tags like *interesting* and links like

related-to that serve as flags or bookmarks and allow collaborators to quickly organize ideas but imply less about the comments' content or relationships between them. In very large workspaces, tags and links like *irrelevant* and *unrelated* could allow analysts to dismiss comments and prune unwanted structure. These dismissal tags could also help combat groupthink and errant tagging by providing analysts asked to curate an entire workspace with an alternate meaningful action when no tags apply.

Chapter 4

Scaffolding Mobile Sensing and Analysis for Novices

The addition of social tools to data analysis environments allows analysts to exchange ideas and pool their analytic effort. However, as tools for collecting and distributing data grow more widespread, new social data analysis applications seem likely to emerge outside of traditional analysis environments and with non-traditional, potentially novice users. Due to the increased availability of sensing technologies, citizens and novice users have new opportunities to pursue the kinds of data collection and analysis that were once handled almost exclusively by professional scientists and analysts [26]. Leveraging this citizen engagement effectively, however, requires not only tools for data collection but also mechanisms for understanding and utilizing citizens’ “local knowledge”—the experiential and cultural context, insights, and expertise unearthed through collaboration between locals and experts [24]. For example, while sensing systems may be able to detect the presence of a pollution source, local insight may be required to identify the source or reveal populations affected by it.

However, in the domain of air quality monitoring, most mobile monitoring systems [34, 57, 79] have tended to emphasize improving environmental awareness, or have taken creative approaches

Previously published by the author, Paul Aoki, Neil Kumar, Sushmita Subramanian, and Allison Woodruff in [119].



Figure 4.1: A personal air quality sensor (left). Community members with sensors (right).

to presenting and collecting this data through artful visual presentation [84], provocative platforms [25], and gameplay [76]. These systems have not focused on enabling direct citizen engagement in the data analysis process.

Meanwhile, most tools for viewing and analyzing sensed data do not explicitly support collaboration and are not designed to elicit or compile these kinds of local insights. Analysis tools are generally not accessible to novice users, since they tend to assume a high level of technical and scientific literacy. We seek to understand how interactive systems for supporting citizen science can facilitate input from novice users and provide scaffolding that allows them to make greater local knowledge contributions.

This research was conducted as part of the broader Common Sense project [4, 32], a mobile sensing program that aimed to deploy distributed air quality sensors in the service of practical action. The Common Sense project served as a research testbed to explore participatory sensing, examining issues such as the relative accuracy and resolution of community-sensed data versus data collected in professional fixed installations. The project also focused on developing models for facilitating engagement and cooperation between community members, citizen scientists, activists, and other stakeholders in the air quality ecosystem.

Whereas traditional air quality monitoring organizations utilize coarse, representative measurements from a relatively small network of fixed sensors, we focus instead on a mobile participatory sensing [15] approach in which large numbers of personal, mobile sensors are deployed within communities. This approach allows the community members impacted by poor air quality to en-

gage in the process of locating pollution sources and exploring local variations in air quality. It leverages citizens' desire to understand personal exposure and knowledge of their communities to help effect change.

4.1 Motivating Fieldwork

Before deploying our mobile sensing platform with community members, we wanted to understand how those members factor into discussions about air quality and what roles they could play in data collection, analysis, and outreach. To gauge this, we conducted a concentrated investigation of the communities we hoped to engage with.

4.1.1 Methods

Over the course of several months, we interviewed novice community members as well as scientists, remediation consultants, government representatives and other stakeholders in order to understand their perspectives on air quality and assess the role that technological interventions could play in their environmental decision-making processes [4]. This included 14 formal, in-person interviews and approximately 30 informal interviews conducted either in person, by phone, or at community meetings. In these interviews, we discussed existing practices and used prototype sensors and interface mockups to explore people's reactions to potential mobile sensing tools. We recorded the formal interviews and took detailed field notes describing all of our interactions. Using these, we performed affinity clustering to identify a general set of emergent themes and design principles. We also performed more targeted clustering to identify common user needs, tasks, and motivations for community participation and engagement with environmental data.

4.1.2 Personas

Based on this fieldwork, we developed a set of personas to characterize the relevant stakeholders and identified a set of common tasks and questions associated with each. Because the system presented here is targeted primarily at community members and novice users, we will limit our discussion to the three most relevant personas: an *activist* or *community organizer* responsible for orchestrating actions and publicizing environmental issues, a *browser* who has an interest in environmental quality but is not directly involved with sensing, and a novice community member who might

act as a *data collector* (Table 4.1). While we focus here on tools for these community members and novice users, we believe it is also valuable to provide tools for (and promote dialog with) expert stakeholders with different needs, including *scientists* and *government regulators*.

	Activist/Organizer	Browser	Data Collector
Motivation	Specific concerns about the community with an emphasis on political change.	Likely to be interested in environmental and/or societal issues. Possibly concerned with political change.	Likely to have personal health issues.
Goals	Prove there is a problem. Determine neighborhood exposure. Pursue political change.	Understand broader environmental and societal impacts. See trends.	See personal, immediate data. Modify personal behavior. Pursue political change.
Desired Tools	Tool for community understanding and presentation.	Summaries, Interactive tools for exploring data.	Glanceable summaries, Alarms, Forecasting.

Table 4.1: Some of the key personas derived from our initial fieldwork.

4.1.3 Design Principles

Based on our fieldwork, we also extracted a set of design principles for developing tools to support visual analysis of sensed data. Some of the key issues are:

Support specific, goal-directed tasks. Participants were highly goal-oriented and motivated by specific issues such as “What is my personal exposure throughout the day?” or “What are hotspots in this area?”. “General” exploration did not engage them. As one interviewee put it, “You don’t want to look at the interface and say, ‘What is this supposed to tell me?’”

Show local and personally relevant data. Participants were most interested in data close to their homes and other locations they frequented, rather than the aggregate regional data typically provided by current air quality monitoring solutions. The interviews further suggest that many users may not engage unless they are driven by health concerns or some other issue that personally connects them

to the data. As one participant said, “Make the data as local as possible. People want to see their house, their block, not a general neighborhood, not a general area.”

Elicit latent explanations and expectations. Community members have local knowledge and expertise, such as beliefs about sources of pollution in their neighborhood. However, our interviews suggest that it is often difficult for them to translate this knowledge into specific queries. While community members were good at generating high-level or vague questions (e.g. “How does the freeway impact air quality?”), they had fewer immediate instincts about how to break these questions down. Therefore, it is important to provide tools that help community members draw on their personal knowledge, for example by proactively prompting users with possible queries or by walking them step-by-step through an exploration of the data.

Prompt realizations. As mentioned above, community members have significant local knowledge that could be helpful in interpreting local environmental data. Accordingly, it is valuable to present views of the data that are perceptually suggestive of various possible patterns, and therefore prompt spontaneous realizations that draw on the users’ local knowledge. For example, a view that aligns readings from multiple days may prompt a user to realize that repeated spikes at a site are the result of a recurring event—for example, a delivery truck unloading.

Beware of “language” barriers. Current tools to which community members have access, such as the EPA EnviroMapper [35], are technically complex and require a moderate level of scientific knowledge (for example an understanding of pollutant concentrations in parts per million). Novice users may benefit from scaffolding to introduce scientific language, and tools that target novice users should not require an understanding of such language.

Avoid inundating users. Understandably, participants did not want to be overwhelmed with unnecessary information and complexity (particularly if the information was somewhat new to them or was beyond their level of expertise). Therefore, staged or gradual presentation of information is desirable.

4.1.4 A Framework for Knowledge Generation in Citizen Science

Drawing on our personas and design principles, we derived a framework (Figure 4.2/Table 4.2) for describing data collection and local knowledge generation in a citizen science setting. This framework does not just describe the existing ecosystem or citizen science applications. Rather,

it builds on the key findings and user needs we identified in our fieldwork and describes operations an ideal citizen science solution might address. As such, the framework serves as a potential blueprint for designing new citizen science tools and for assessing existing ones.

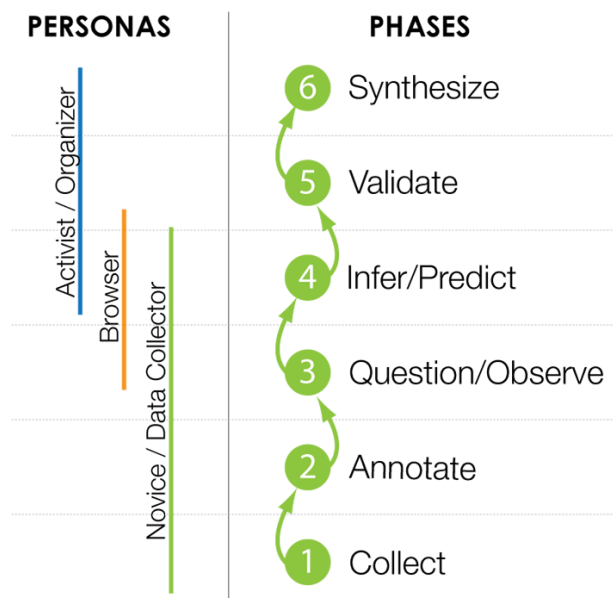


Figure 4.2: Our framework for knowledge generation in citizen science (right). Personas (left) are shown in their intended phases.

In this framework, we divide the process of collecting, analyzing, and synthesizing environmental data and local insights into six phases: *collect*, *annotate*, *question/observe*, *predict/infer*, *validate*, and *synthesize*. While these phases can build on one another, they are not necessarily linear and individual participants do not necessarily participate in all of them. Rather, each involved stakeholder may engage in the process at a few phases and the various members of the community together carry out activities at all phases. The various phases each serve different functions and can build on one another but do not always do so. These phases may also be iterative—for example, answering questions and validating predictions may require additional data collection.

The phases detailed here dovetail with formulations of the scientific method, and some steps (*question*, *predict*, and *validate*) echo the question-hypothesize-test formulations seen in the science education literature. The collection, inference, validation, and synthesis stages in the framework also have analogues in the sensemaking cycle (Section 2.1). Rather than describe the process of sensemaking, however, our framework outlines stages through which novice users may progress as they engage with the process of citizen science. We developed the framework to help developers and organizers envision the stages at which various stakeholders participate in citizen science and identify leverage points for scaffolding novices into the analysis process. As a result, the framework's stages are more general than those in the scientific method or sensemaking cycle, and describe activities that need not necessarily be formulated in the language of scientific discourse. Questions,

6. Synthesize.	Participants, with professional analysts, domain experts, and regulators, integrate data and knowledge generated in prior phases to produce summary documentation that can support activism, inform policy decisions, and impact regulations.
5. Validate.	Greater overlap between participants collecting data and other stakeholders. Participants may look for additional data to corroborate their own findings and organizers may also make requests for additional data or enlist the help of outside entities including domain experts and professional analysts to help verify insights and predictions.
4. Infer/Predict.	Building on questions and observations, participants make predictions and inferences about the observed phenomena (“I think values get worse around rush hour.”, “Higher counts here seem to indicate a nesting site.”). These may be less clearly articulated than in a formal analysis, but can contain local insights. In these predictions, regardless of their precise formulation, lie some of the most important pieces of local knowledge that community members can contribute.
3. Question/Observe.	Using their own data and data collected by other participants, participants can begin to ask basic questions and identify trends. These can be introspective (“What is my personal exposure to pollutants?”), “Is there graffiti near my home?”) or generally inquisitive (“Are there parcel by parcel trends in the appearance of a particular bird species?”).
2. Annotate.	Participants add additional insights to contextualize and supplement data (e.g. when, where, and under what conditions was the data collected) and provide indicators of data quality.
1. Collect.	Participants use sensors to record raw data or observe phenomena and make manual observations. Most existing citizen science places a strong emphasis on this collection phase.

Table 4.2: Framework phases in detail.

predictions, and inferences generated by community members can contribute valuable insights that inform a more formal and rigorous process of scientific analysis without necessarily being framed as such.

Finally, while we frame this process in terms of air quality monitoring for the sake of this discussion, the framework itself is applicable to a broad range of citizen science projects including other environmental and health monitoring efforts.

Collect

In this phase, data collectors engage in various data collection activities, including using sensors to record raw data or observing phenomena and manually recording observations (as in traditional citizen science activities like the Christmas Bird Count [50]). Most existing citizen science places a strong emphasis on this collect phase.

Annotate

After data has been recorded, data collectors provide additional insights that contextualize and supplement it. Collectors can include additional information that helps explain the data; for example, if a peak in the data corresponds to an event they observed during collection. They can also include information about the data gathering process (when, where, and under what conditions was the data collected) or comments about data quality.

Question/Observe

Using their own data and data collected by other participants, data collectors (as well as browsers and activists) can begin to ask basic questions and identify trends. These questions can be introspective (“What is my personal exposure to pollutants?”, “Is air quality bad at my home?”) or generally inquisitive (“Where is air quality good and bad?”, “Are there block-by-block trends in air quality?”). Some of these questions, including those dealing with personal exposure, can often be answered directly using the collected data, while others are more abstract. These questions can be implicit or explicit and may be driven by the data or by existing assumptions and expectations. Users may also observe and note apparent trends (for example, higher levels of a pollutant at different times of day) or other phenomena of interest (high levels at an unexpected intersection).

Infer/Predict

Building on these questions and observations, data collectors, browsers, and activists can begin to make predictions and inferences about the observed phenomena (“I think values will get worse towards this intersection.”, “Higher readings here seem to indicate a source.”). The observations and inferences made by community members may be less clearly articulated than in a formal analysis, but can contain local insights. While this phase often resembles the “hypothesize” stage seen in formulations of the scientific method, participants’ predictions and insights may not necessarily be framed as clearly testable hypotheses. They may only suggest the existence of a trend or its repeatability rather than proposing a mechanism for it. In these predictions, regardless of their precise formulation, lie some of the most important pieces of local knowledge that community members can contribute.

Validate

At this phase, contributions from data collectors are more likely to overlap with those of activists and organizers. Here, data collectors, browsers, and organizers may look for additional data to corroborate their own findings and organizers may also make requests for additional data. Additionally, organizers may enlist the help of outside entities including domain experts and professional analysts to help verify insights and predictions generated by collectors and browsers.

Synthesize

At the highest level, activists and organizers must integrate the data and knowledge generated in prior phases to produce documentation, reports and other deliverables. Again, organizers may involve domain experts and professional analysts, along with administrators and regulators, in order to generate summary documentation that can be used to support activism, inform policy decisions, and enforce regulations.

This framework (and particularly the *annotate*, *question/observe*, and *infer/predict* phases) provides a blueprint for scaffolding novice users’ progression from initial elicitation through more involved and integrated questions and contributions. In this chapter, we focus on applications that engage novice users and guide them through these initial phases. We defer discussion of validation and synthesis, which tend to utilize more specialized sets of tools for more expert users.

4.2 The Common Sense Community Site

Building on the framework and our design principles, we designed and built the Common Sense Community site, a suite of task-oriented applications that allow community members to participate in the collaborative analysis of local air quality data. While the site is targeted primarily at novice data collectors in a low-income urban area, it is also designed to be accessible to more specialized participants (browsers, organizers, scientists, administrators, and regulators) who may engage in the analytic process at different phases.

The set of visualizations is designed specifically to facilitate the incremental progression of novice community members through multiple phases of analysis. A person may begin by collecting data or asking questions about data collected by other community members and progress through structured phases, triggering new kinds of insights. Over time, engaging in this process can allow novices to become more adept contributors.

Providing a suite of simple task-oriented applications rather than a more general analysis tool has several benefits. First, it lowers barriers to entry. Participants do not need to learn a complicated tool in order to contribute. In turn, engaging in this process encourages legitimate peripheral participation [64] and allows novice users and participants with little computing experience to take part. Whereas more general analysis tools such as Excel, Tableau [102], or Matlab require greater familiarity with formal analysis processes, these individual applications allow users to answer specific questions and can guide them towards particular kinds of insights. Figure 4.3 shows approximate mapping between our applications and the framework discussed previously.

4.2.1 Collecting Data

Users collect air quality data using mobile sensors designed as part of the broader Common Sense project [32]. These sensors (Figure 4.1) are designed to be self-contained and unobtrusive monitoring devices that can be clipped to a bag or carried as an accessory. The units feature a custom board design and embedded software that can be deployed with commercial carbon monoxide, nitrogen oxides, and ozone gas sensors. As users carry these sensors with them throughout the day, the units transmit live sensor reading and GPS data to a database server over a GSM data network connection. Users can also upload data from offline air quality sensors.

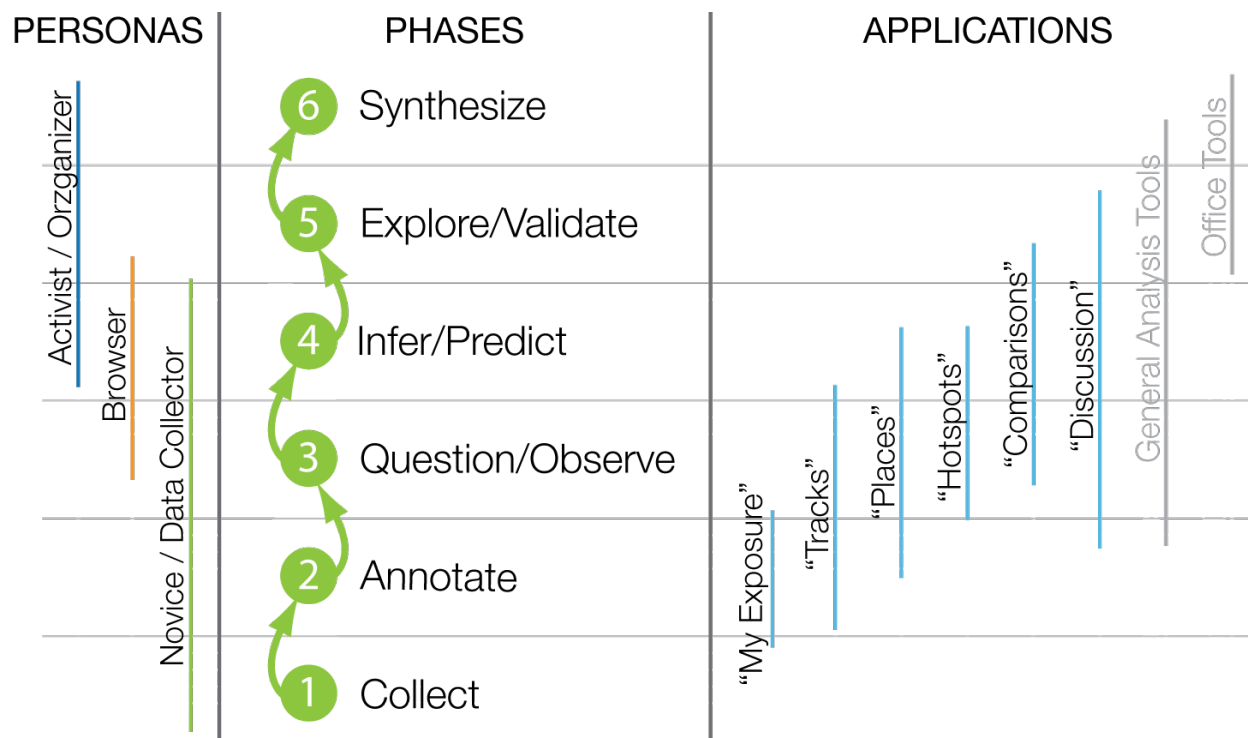


Figure 4.3: Our framework for knowledge generation in citizen science with our applications (right) shown in their intended phases.

4.2.2 Applications

To display this data, we built simple visual analysis applications (Figure 4.4) that target common, representative tasks and questions that we identified through our fieldwork. These applications included: monitoring personal exposure, inspecting recorded tracks, identifying locations with poor air quality, and eliciting possible sources. These targeted applications exemplify our approach to designing for citizen science—modular, accessible applications that serve specific needs and which together scaffold the process of local knowledge production. Users begin by selecting an application that serves a particular need (e.g. “see my personal exposure”) from a portal site. They then move between applications via a tabbed interface. We also provide gateways designed to allow participants to build familiarity with simpler, more targeted tools and then transition in a natural way to more complex tools designed to elicit different types of insights. This facilitates the transitions between *annotation* and *questioning* or *questioning* and *inference* we described in our framework.

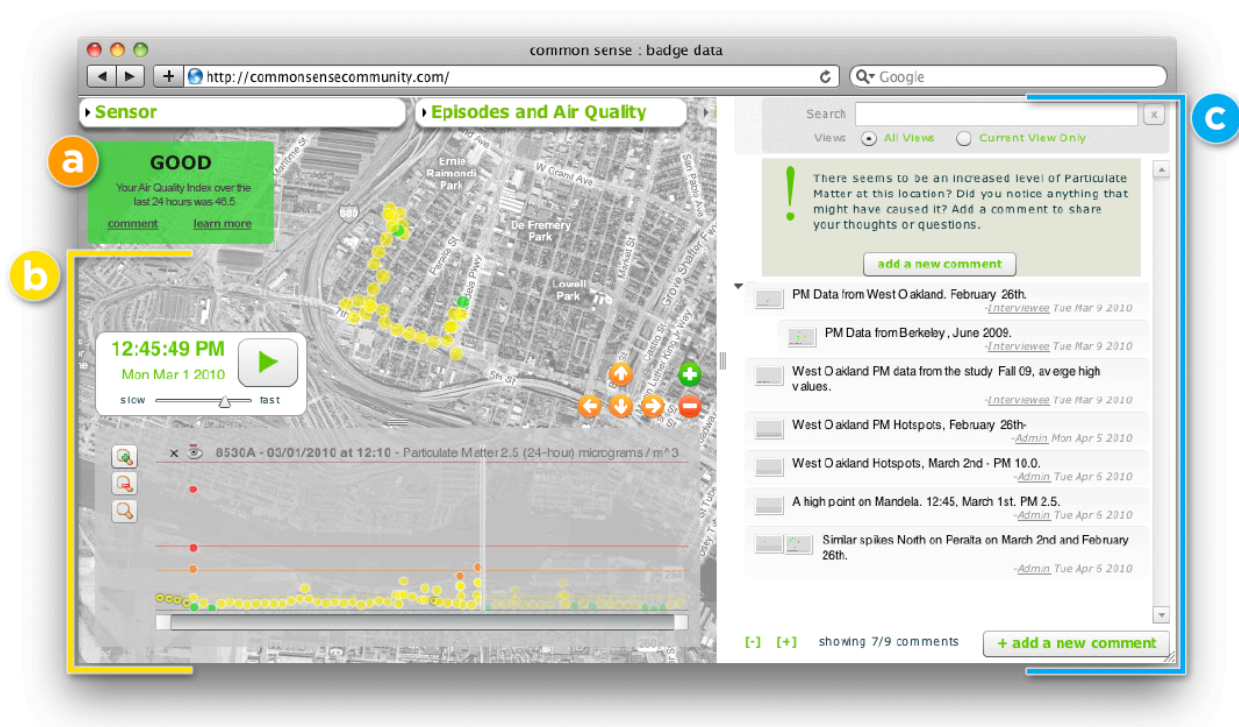


Figure 4.4: The Common Sense Community Site showing data collected by a single user. The My Exposure widget (a) and Tracks visualization (b) are visible along with the commenting panel (c).

In each of these applications, users can record their questions and insights by leaving comments attached to individual views of data. Each application includes an embedded version of the CommentSpace commenting panel (Figure 4.4c) that participants can use to annotate and discuss their findings. To encourage deeper participation we augmented the default CommentSpace panel with dynamic prompts designed to elicit questions and observations, along with educational prompts designed to help scaffold novice users' understanding of the domain. We describe several applications in detail below.

My Exposure

Our first application is a widget that helps users answer one of the most common questions we observed in our fieldwork: "What is *my exposure* to a pollutant?" Many of the community members we interviewed suffered from allergies or respiratory disease exacerbated by the poor air quality in their neighborhood, and expressed a desire for tools that would help them gauge and mitigate

their exposure. To meet this need, we developed the My Exposure widget (Figure 4.5, Figure 4.4a). My Exposure shows a single aggregated measure of the pollutants measured by a participant's sensor, normalized over time to the EPA's Air Quality Index (AQI) [73]. Because the majority of people are not familiar with raw pollutant concentrations, all of the visualizations on the site also use the AQI color encodings and category descriptors—"Good" (green ■), "Moderate" (yellow ■), "Unhealthy for Sensitive Groups" (orange ■), "Unhealthy" (red ■), "Very Unhealthy" (purple ■), and "Hazardous" (maroon ■)—in addition to providing actual values.



Figure 4.5: Two views of the My Exposure application.

For community members carrying our air quality sensors, this application acts as an entry point to the site and serves an ongoing need that is likely to garner repeat visits. To encourage participants who are initially only curious about their exposure to further explore their data, we placed the *My Exposure* view adjacent to the *Tracks* application (discussed momentarily).

Tracks

The *Tracks* application (Figure 4.4b) provides a simple way for novice users to observe and ask questions about pollution data from their own sensor. In this visualization, pollution measurements are plotted on a map and also appear in a timeline below the map view. The application behaves like a media player and provides a play/pause button, a play-back speed control, and a draggable thumb on the timeline that can be used to scrub back and forth in the dataset.

As mentioned above, in each of our applications, participants use the commenting panel (Figure 4.4c) to annotate and discuss their findings. This panel is collapsed by default to avoid overwhelming the user, but expands to display intelligent prompts designed to elicit questions and observations. For example, when a participant plays back data from their own sensor in the *Tracks* application, the interface pauses briefly whenever a dramatic spike occurs in the data and actively prompts the user to document the change. The user can choose to either enter a comment or continue playback. If no action is taken, playback resumes after a brief interval. Users can also pause playback at any point to enter comments or questions.

Places

Our fieldwork indicated that users' initial inquiries about air quality are often location-centric ("What is air quality like in my neighborhood?", "Are we protecting our 'treasures', our schools, hospitals, libraries, parks, etc.?"). To help facilitate questions and observations of this type, we provide a location-centric *Places* visualization. When a user starts the visualization, they are prompted to enter an address and a time range. The application then produces an interactive map showing all data collected by any sensor near the specified address during those times. Whereas the *Tracks* application is designed to mimic the functionality of a media player, *Places* is designed to feel similar to online mapping tools like Google Maps [41]. The map can be panned and zoomed and the data points plotted on it can be played back chronologically.

We include gateways that allow users to enter the *Places* view from within other applications. When using another application, a user can click a "see more for this location" button to transition to the *Places* view, centered on the location visible in their current application.

Hotspots

The *Hotspots* visualization (Figure 4.6) helps users identify regions with the best and worst air quality over a period of time. The application is intended to help users answer questions about where and when levels are high and low. It draws on the notion, frequently seen in our initial interviews, that "worse things are exciting" and uses this to provoke insights regarding new locations and unexpected sources.

Using a range slider, users select whether to show regions with high or low pollution levels. Readings that match the specified thresholds are then plotted on a map similar to the one used in the *Places* view. Users can also transition to this visualization by clicking the "see other places with readings this high/low" gateway from within the *Tracks* or *Places* applications.

Comparisons

The *Comparisons* visualization (Figure 4.7) is designed to support inference and help users identify repeated sources and relationships between them. The *Comparisons* visualization presents users with a set of discrete "episodes"—short windows of time in which some notable event occurred in the recorded air quality data. These episodes can be the largest spikes seen in an area over the course of a period of time, or the periods of time with the highest variance.

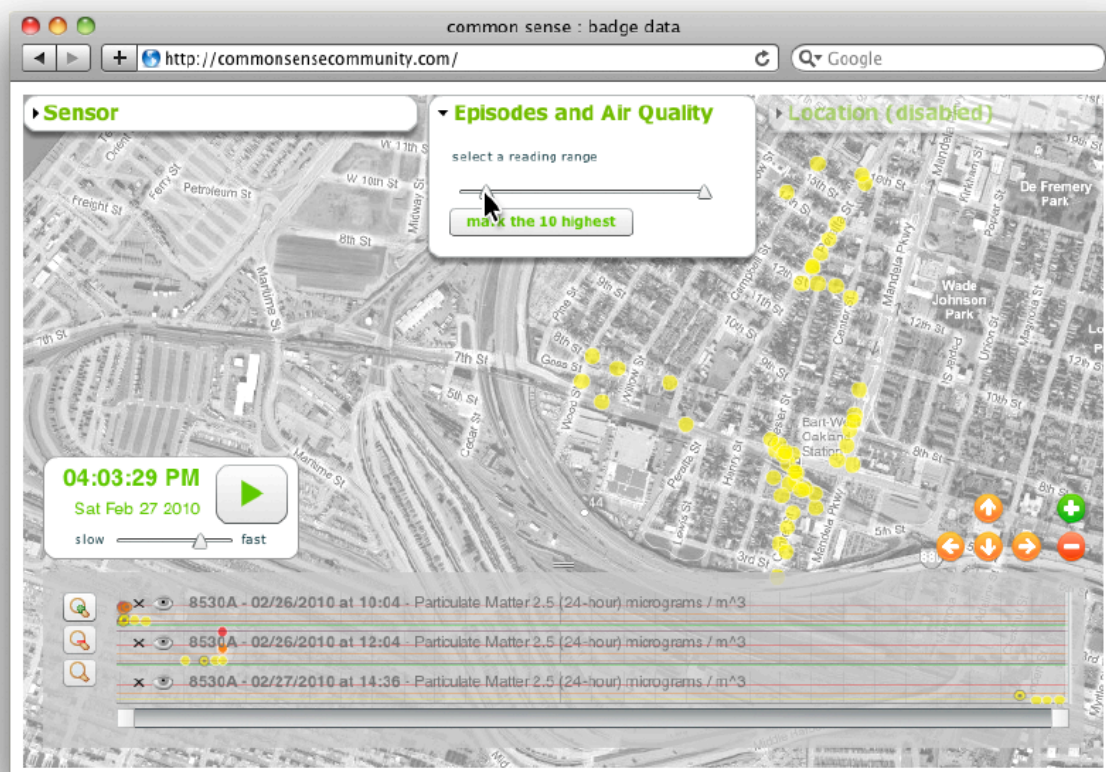


Figure 4.6: The Common Sense Community Hotspots visualization. Users can adjust a range slider to find areas where readings from multiple different sensors are regularly high or low.

Two observations from our fieldwork led us to focus on identifying spikes. First, we noted that people often wanted to “examine an event, not a timeline,” seeing detailed data at the scale where the event was apparent, rather than at the level of the entire dataset. Second, we anticipated that by grouping together sets of episodes that would otherwise appear separately, this view would prompt noticings and inferences that might not emerge otherwise. In the *Comparisons* view, these episodes are displayed as a set of small multiples [108] alongside a map that also plots that same data. The small multiples are linked to the map so that brushing a plot focuses that event in both views. This allows users to compare the events spatially as well as temporally.

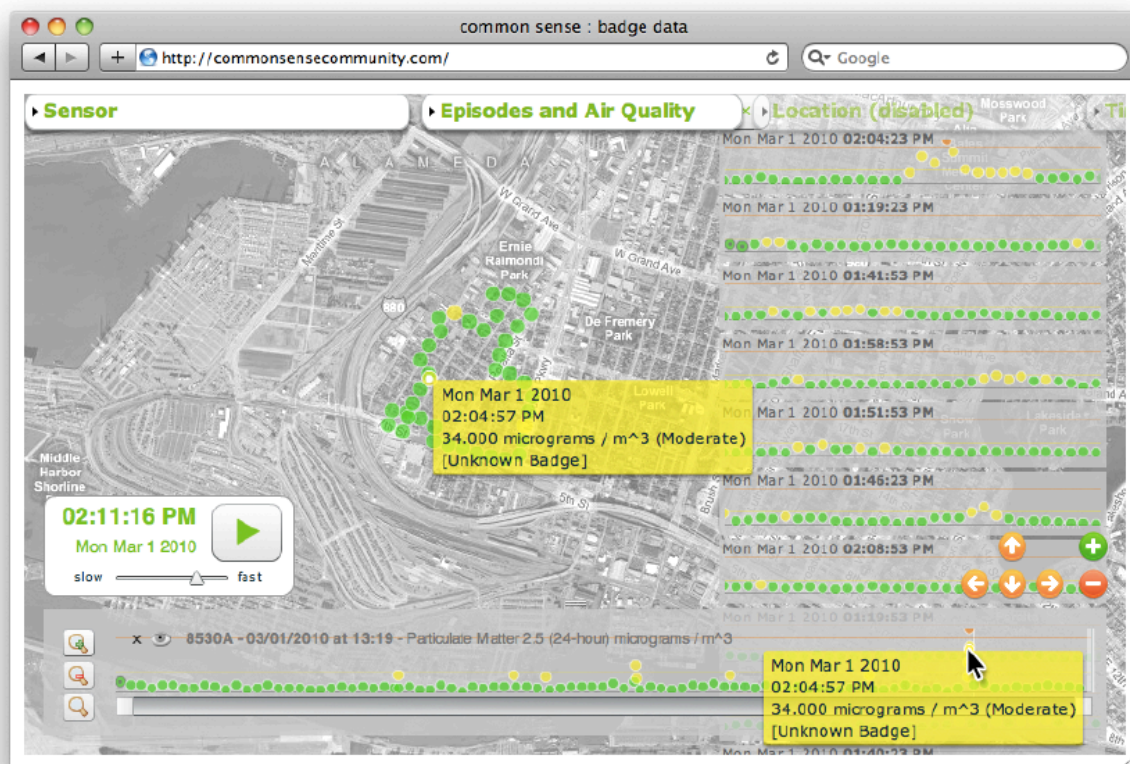


Figure 4.7: The Common Sense Community Comparisons visualization. Small multiples of the timeline (right) showing the 10 highest episodes recorded during the past day. Hovering a timeline jumps to the portion of the map showing those readings. The commenting panel is hidden at right.

Discussions

In addition to the collapsible commenting pane that accompanies each one of the visualizations, the site features a *Discussions* view—a separate instance of CommentSpace that serves as a central location for viewing all comments and provides a forum-like interface for further discussion. All comments and annotations left by users in the other applications are visible here as separate threads and users can compare and build on observations and insights from multiple applications.

4.3 Evaluation

We deployed an early version of the site with community members in a low-income urban neighborhood with poor air quality. There, we carried out interviews and think-aloud assessments to help characterize participants' use of the tools. We wanted to understand which visualizations were perceived to be useful and approachable and assess whether this set of tools facilitated activities at the various phases identified in our framework, such as emergent prediction and observation.

4.3.1 Methods

During our assessment we carried out seven interviews with nine community members. We recruited participants through a local non-profit organization that focuses on environmental monitoring and awareness. Five of the participants were affiliated with the non-profit and had participated in air quality monitoring activities through the organization. Most of the participants we surveyed were members of a small and relatively tightly knit community and the majority knew one another in some capacity. Participant ranged in age from the mid-teens to late 40's and had a variety of education levels, including four middle- and high-school students and some participants without high school degrees.

We conducted all of the interviews at the office of the non-profit. We started each session with a brief interview designed to assess participants' knowledge of air quality issues and the impact of air quality on their community. In our discussions, we emphasized the impacts of particulate matter and described its sources. We then gave the participants a particulate matter sensor and asked them to take samples in a several block radius around the office. We asked participants to choose a route that they thought would maximize the amount of particulate matter detected. During the sampling process, the interviewer walked with the participants and asked them to describe their route choice and identify potential sources in the area. We used a commercial particulate matter sensor rather than our custom hardware since particle pollution is of particular interest in the target neighborhood.

Once they returned to the non-profit, participants used an early version of the Common Sense Community site to examine their data as well as data gathered by other participants. We conducted a one-hour think-aloud evaluation with each participant in which they were instructed to interact with the site and verbally relate their thought processes and any questions or insights that occurred to them. Participants used a version of the site that included the *Tracks*, *Places*, and

Comparisons visualizations detailed above. In the *Places* and *Comparisons* views, each participant had access to his or her own measurements as well as measurements taken by all of the previous participants. Because users only had access to data collected by a small group of participants in short windows over the course of a few days, we were unable to test the *Hotspots* visualization, which was designed to leverage larger datasets.

We recorded each of these interviews and coded participants' interactions with the site to assess whether or not they fit within our framework. We also performed clustering to extract key findings that emerged.

4.3.2 Scaffolding and Navigation Strategies

Most participants were able to explore the visualizations and inspect the data that they had collected without much confusion. The majority began by identifying their current location on the map and followed the track they had recorded, looking for peaks either on the map or in the timeline. Most voiced questions and observations about the data and a few made additional inferences or predictions. We report key observations that correspond to each of the phases in our framework.

Collect. During their interviews, almost all of the users identified a nearby freeway and trucking lots as the most likely sources of pollution and most chose routes that took them along a nearby frontage road. The students we interviewed all observed the readouts on their sensors attentively as they walked, looking for spikes and actively seeking out areas with higher readings. All other participants used the sensor more passively and traversed areas that they predicted would be more polluted without actively noting the levels there.

Annotate. Using the *Tracks* view, several participants observed distinct peaks and verbally ascribed them to events that occurred or features that they passed while they were walking ("All the trucks [get on the highway there].", "That's the new construction there."). Participants also tended to note readings taken adjacent to locations that interested them ("At least we don't have any red marks near the park..."). In two cases, participants had observed increased particulate matter levels on the sensor as they walked and directly attributed a peak to a particular source.

Question/Observe. Most participants asked questions and made remarks about locations (“Where was that again?”), data (“Was [that spike] at an intersection?”), and other participants (“Where did she go?”, “Which person did that come from?”). Participants also asked broader questions about day-to-day and month-to-month trends. For example, one wondered whether pollution levels would change during the rainy season and another asked “Would it be different if there was wind?” A few participants also noted locations on the map without data and contributed additional anecdotes and pieces of information about them.

Infer/Predict. Based on the data and their initial questions and observations, several participants made inferences about the behavior of phenomena they observed. For example, one participant compared her readings with those from a participant earlier in the day and noticed that her own were higher. She inferred that the level of particulate matter might be impacted by the change in temperature.

Another participant investigated the data he had collected and extrapolated from it to predict air quality readings further along the frontage road saying, “I wouldn’t doubt that it gets worse around the bend.” Talking about a several-block radius, he also made a prediction about the health impacts of pollutants in the area. He noted, “Just in this radius I can honestly say [...] at least half the kids have asthma. At *least* half.” He supplemented this prediction with a quick calculation, “Fifteen residences per area so ... that’s probably about a good 500 kids.”

Validate and Synthesize. This set of interviews involved only novice community members and incorporated only data collected during those participants’ sessions. As such, we did not emphasize the *validate* and *synthesize* phases in this study.

4.3.3 Usability

Based on our fieldwork, we were mindful in our design process of the computer literacy of the target population. As one participant in our initial interviews noted, “There’s still that big digital divide in [our city] and all poor neighborhoods.” Therefore, we were pleased that the system was generally usable by all participants. The study did reveal a few straightforward usability issues, which we are addressing, such as the need to make the playback controls more visible. These issues did not appear to impact the results discussed below.

4.4 Discussion

Here we discuss trends and activities we observed across all of our interviews.

4.4.1 Health and Personal Safety

As expected, displays tailored to personal use seemed to be an effective tool for engaging users in the process of citizen science. The most interested and receptive participants each had a personal or family health concern (asthma, allergies, or some other reaction) that they attributed to air quality. One asthmatic participant who bicycles and does not own a car expressed a desire to use the data to vet safe cycling routes, stating, “This has brought to mind—you’re gonna get exercise, but what are you breathing in?” Participants with small children also expressed a strong desire to use the tool on a regular basis to help minimize exposure.

4.4.2 Socializing

Although we conducted interviews separately and the sequential nature of the interviews did not facilitate conversations or dialogues using the commenting tools, we did see social interactions between participants when they viewed one another’s data. Several participants asked questions like, “Which person did these come from?” and “Whose was whose?” and were eager to compare their tracks against those recorded by previous participants. In particular, those from the same social circle were interested in knowing which of their friends had collected data, where they had walked, and how “well” they had done. For example, one participant located a friend’s track and followed it for the entire length, noting each location she had visited and commenting, “She was pretty good, [she found a few orange ones.]” Comparing tracks in a competitive way was also common, particularly among the students we interviewed. One group of younger students, for example, was excited to discover that their readings were higher than those of other participants. This excitement suggests a competitive impulse that we might also leverage to encourage participation—possibly by introducing game-like elements to the collection and annotation processes.

During the interviews, several participants attributed their continued awareness and investment in air quality to a particular community organizer. One participant observed, “You could say *she’s* our resource when things are happening. If she feels we need to know, then it’s up to us to get involved.” These comments suggest that, at least within this community, maintaining long-term interest and investment depends, in part, on leveraging these kinds of key community members.

While we observed users' reactions to one another's data, the linear nature of our interviews did not allow us to observe exchanges or evolving social use of the system. A longitudinal study with more users is needed to understand these social aspects of the system and to gauge the impact of larger amounts of data and discussion on the analyses that participants undertake.

4.4.3 Exposing Preconceived Notions

A number of our participants approached the data not from an inquisitive standpoint, but rather expecting to find validation of their expectations about air quality. We noted comments from a number of participants that suggested implicit assumptions about areas ("On Fourth Street, that makes sense."—referring to an area adjacent to a major freeway) and expectations about how bad pollution levels would be ("[If you sampled this area] you'd see lots of red"). One participant, in particular, was surprised that the level of particulate matter she recorded was low, stating, "I feel like it should be a little stronger with picking up certain particulates and fumes. I *know* there should be a lot more out there because there are a lot of businesses and industrial stuff." To test their assumption, the participant requested to take the sensor out again and collected additional data.

In some cases these kinds of assumptions may function as implied hypotheses and predictions that participants can immediately begin to validate and build on. However, as in the case of the latter participant, preconceptions can sometimes generate mistrust in sensors and tools that do not reinforce these existing notions. Understanding how to circumvent these preconceptions and help novices build an informed understanding of these tools remains an important area for future work.

4.4.4 Visualizations as a Catalyst for Discussion

We also observed several participants who used the map extensively as a catalyst for discussion. These users would point and navigate to areas with strong personal relevance including their homes, schools, and public areas, even when no air quality data for that particular region was present. One interviewee, in particular, used the map to discuss pollution sources outside the zone in which he had collected data and to make predictions about sources and impacts there. He first predicted that there might be "really high values" in main intersections adjacent to a nearby port and shipping terminal, stating, "I can only imagine [it gets worse toward the intersections.]" He then contributed a number of anecdotes about locations in and around the port including spots where

diesel trucks idle, areas where water quality has been impacted by dredging, and an isolated residential building in the industrial zone. These anecdotes were often very specific and drew on his experience as a port worker and volunteer air monitor—for example:

“Here—definitely this intersection—we did some of the survey in this area last year. Here, right here—this is a fuel station. It’s a truck fuel station. This is where all the trucks get on the freeway. All the trucks are always right here—along [Street 1] and [Street 2] and um, [Street 3] and [Street 2]. I know for sure, these monitors are not going to catch moderate here. Lucky enough, nobody lives on these blocks. All business, all industry.”

These kinds of observations are key examples of the types of local insights community members may bring to the table and which we hope to elicit.

4.5 Additional Design Considerations

While we have explored community analysis and sensemaking in the context of air quality monitoring, we believe our framework is applicable across a wider range of citizen science domains. Depending on the nature and limitations of a particular community and domain, several additional considerations may impact the application of our framework.

4.5.1 Qualitative vs. Quantitative data collection

In cases where the data collection occurs manually or where it is qualitative rather than quantitative, participants can annotate the data as they collect it. As a result, the collect and annotate steps may overlap. In all cases, allowing participants to annotate data with additional contextual information (notes, photos, or other metadata) at the time of collection can provide additional insights that may be lost if data is annotated post-hoc.

4.5.2 Privacy and Security

Privacy and data security are serious system-level concerns in citizen sensing tasks [24] and also affect how users access and explore community data. In some cases, the personal nature of the data collected may make it problematic to share data among participants, making it more difficult to

transition from observing one's own data to asking questions and making inferences about broader trends. For example, epidemiological monitoring in which information about participants' medical histories are collected would require further levels of anonymization or authorization. Any activity in which participants' location or activities are either explicitly or implicitly tracked may also require anonymization—which may limit participants' ability to ask questions and make predictions—or require participants to agree to make their data open to the community. Similarly, if citizen-sensed data is sensitive in nature, it may be difficult to share openly, even within the community. For example, the precise locations of vulnerable archaeological sites or endangered species may need to be protected in order to ensure that the cultural or environmental resources being tracked are not further disturbed.

4.5.3 Stakeholder Goals and Competing Interests

Multiple communities, and even members within the same community may have goals that are not compatible. For example, both hunters and recreational bird watchers may be interested in tracking and understanding the ranges, movements, and condition of a bird species, but the communities may be ideologically incompatible. If a clear, unambiguous community goal can be articulated, the validation and synthesis phases can be designed explicitly to support it. Otherwise it may be important to ensure the verifiability of collected data— particularly in the validation phase—in light of competing interests.

4.5.4 Importance of Discussion Tools

The presence of tightly integrated discussion features in community-oriented tools is critical in order for constructive community-driven knowledge generation to proceed. Persistent commenting tools enable participants to transition between the annotate, question/observe, and predict/infer phases while still retaining access to all prior discussions and contributions. By coupling discussion to the visualizations and analysis tools (and ultimately to the raw data), a system can promote fluid discussion that is grounded in the underlying data, even as the community's analysis becomes increasingly abstract.

Chapter 5

Crowdsourcing Social Data Analysis

Many datasets and analyses are simply too large to be managed easily by a single analyst or even a small team. Automated data mining tools can find recurring patterns, outliers and other anomalies in the data, and help analysts find potential points of interest in big datasets. However, only people currently can provide the explanations, hypotheses, and insights which are necessary to understand them [83, 88]. While tools like Sense.us [48], Many Eyes [111], and CommentSpace (Chapter 3) are designed to support large-scale analysis involving many participants, but such analyses do not typically occur in the wild.

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. To illustrate the lack of emergent analysis, we conducted a survey in April 2012 examining the commenting behavior on visualizations in the Many Eyes site. We found that from 2006 to 2012, Many Eyes users published 294,646 data sets but generated only 128,478 visualizations and left only 17,340 comments. We then randomly sampled 100 of the visualizations containing comments and found that just 11% of the discussions provided a plausible hypothesis or explanation for the data in the chart. This low rate of commenting may represent a shortage of viewers or may

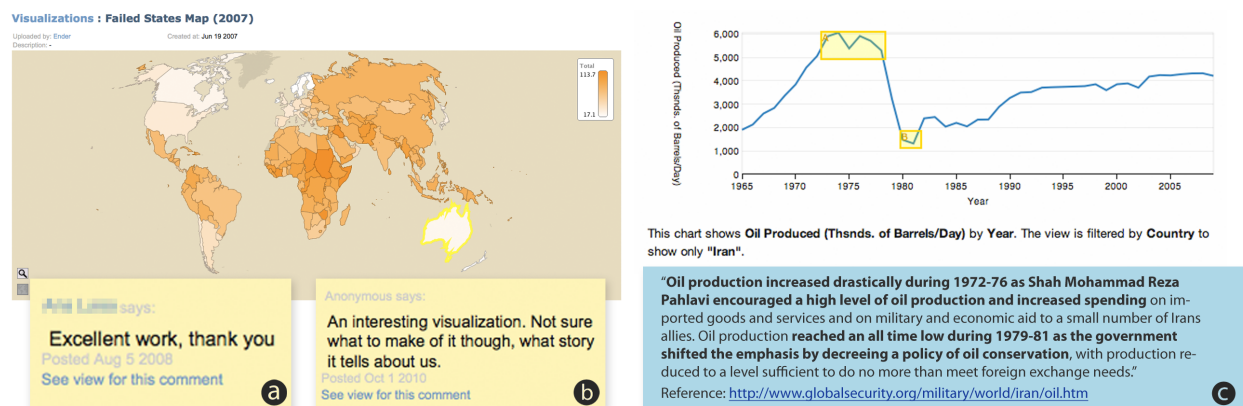


Figure 5.1: Comments on social data analysis on sites like Many Eyes (a,b) often add little value for analysts. We show that crowd workers can reliably produce high-quality explanations (c) that analysts can build upon as part of their broader analyses.

be due to *lurking*—a common web phenomenon in which visitors explore and read discussions, but do not contribute to them [117, 75]. When comments do appear, they are often superficial or descriptive rather than explanatory (Figures 5.1a, 5.1b). Higher-quality analyses sometimes take place off-site [27] but tend to occur around limited (often single-image) views of the data curated by a single author.

Controlled studies of social data analysis systems like sense.us [48] and CommentSpace (Chapter 3) have shown that analysts and enthusiasts in more structured environments can share the process of exploring datasets, proposing hypotheses, and seeking out new insights. However, in these cases, eliciting high-quality explanations of a visualization required seeding the discussion with prompts, examples, and other starting points designed to encourage high-quality contribution. Moreover, depending on ad-hoc exploration by loosely-coupled cadre of users can give poor coverage of a dataset. Users may miss important views if they only flock to the most popular or easily accessible views of the data. Ultimately, marshaling the analytic potential of crowds calls for a more systematic approach to social data analysis—one that explicitly encourages users to generate high-quality hypotheses and explanations.

In this chapter we show how key sensemaking tasks like generating explanations can be broken down and performed systematically by paid crowd workers. We develop an analysis workflow (Figure 5.2) 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 5.1c, 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.

Although the simplest form of the analysis microtask asks crowd workers to “*Explain why a chart is interesting.*” prompting users this way can result in irrelevant, unclear or speculative explanations of variable quality. The explanation may be irrelevant to the analyst—charts often contain many interesting features (e.g. peaks, valleys, steep slopes, flat regions, overall trends) that a worker could explain, but the analyst often cares about one, specific feature. The worker may attempt to scam the task or may not attend to the relevant visual features of the chart. The worker may not know what views of the data look like or what is required of a high-quality explanation. The explanation may also be based on speculation or assumptions that are not supported by outside sources.

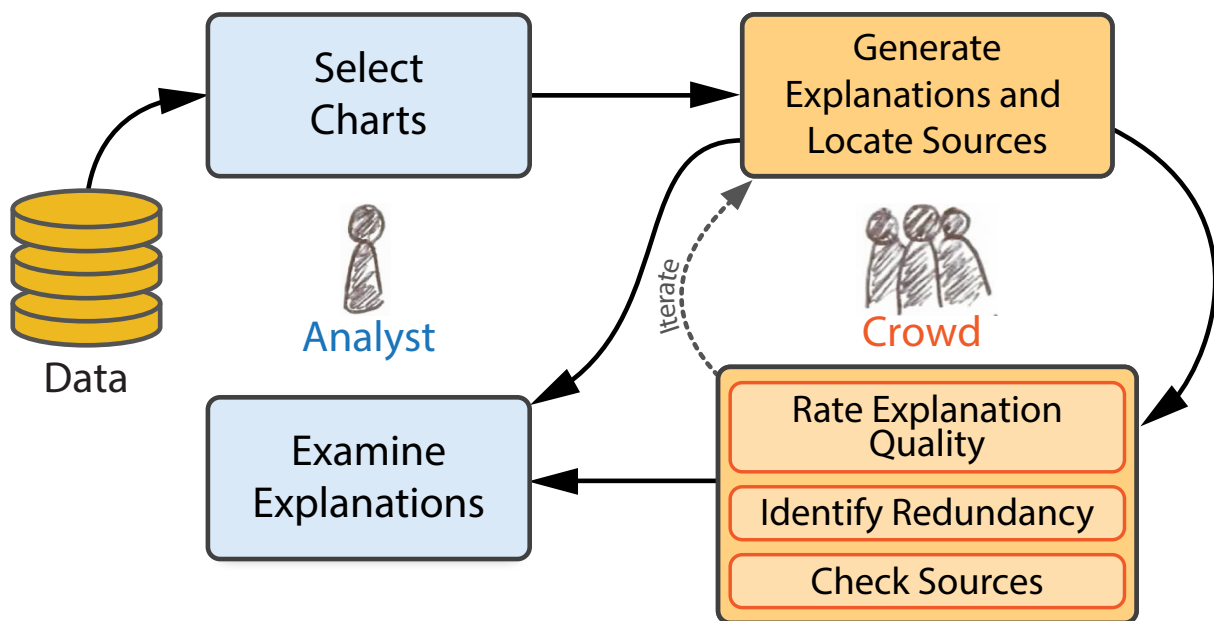


Figure 5.2: Our workflow for crowdsourcing data analysis. In our workflow an analyst first selects charts, then uses crowd workers to generate possible explanations for outliers, trends and other features in the data. Other workers then rate the explanations, check their sources, and identify redundancies, before returning results to the analyst.

To address these concerns, we propose a set of seven strategies that address problems encountered when eliciting responses and improve the quality of the worker-generated explanations of data.

We also focus on helping analysts make sense of the large sets of explanations generated by crowd workers. Workers operating in parallel often produce many redundant responses that give the same general explanation for a trend or outlier. Analysts must spend time filtering and condensing these redundant responses to identify unique explanations and determine if redundant explanations corroborate one another. Because individual workers have different competencies and domain knowledge, some of the explanations they produce are more plausible—more likely to be true—than others. Determining which explanations are plausible and which are not is often difficult, in part, because workers' explanations often lack detailed provenance—information about the sources used to produce the explanation. In these cases, analysts cannot determine whether a worker's explanation is derived from a reputable source or is merely the worker's own speculation. We explore a range of criteria that analysts may use to filter and organize explanations and decide whether or not they are plausible. We then demonstrate two sets of techniques to help analysts manage crowdsourced explanations:

(1) We explore two strategies (*distributed comparison* and *manual clustering*) that use crowd workers to detect redundant explanations. Using our manual clustering approach—in which multiple workers cluster explanations and we select the *most-representative* clustering—we can reliably generate clusterings that are as good as those produced by experts.

(2) We help analysts gauge the plausibility of explanations by exposing more detailed explanation provenance. We record workers' browsing behavior in an embedded web browser. We also introduce highlighting tasks that allow workers to make finer-grained citations by marking paragraphs and sentences on the web pages they cite. Additionally, we show how workers can help verify the provenance of others' explanations via source-checking tasks.

Finally, we provide an explanation-management interface that allows analysts to interactively explore clustered explanations and examine their provenance. Using this interface, analysts can quickly group and filter responses, in order to determine which explanations should be further considered.

5.1 A Workflow for Crowdsourcing Data Analysis

Hypothesis (or explanation) generation is a key step of the sensemaking model (Chapter 2) and requires human judgment. Developing good hypotheses often involves generating a diverse set of candidate explanations based on understanding many different views of the data. When analyzing a dataset, an analyst may need to explore many different views of the data and build an understanding of them. Developing such understanding usually requires generating a diverse set of candidate explanations and hypotheses. Our techniques allow the analyst to parallelize the work of generating explanations by dividing it into smaller *microtasks* and efficiently distributing these microtasks across a large pool of workers.

We propose a four-stage workflow (Figure 5.2) for crowdsourcing data analysis. In our workflow, 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 first selects a set of charts for analysis. The analyst may interactively peruse the data using a visual tool like Tableau [102] to find charts that raise questions or warrant further explanation. Alternatively, the analyst may apply data mining techniques (e.g., [54, 65, 122]) 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 data mining techniques to find these three kinds of features in time-series data. Given a set of time-series charts these scripts 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 personally or post the charts directly to crowd workers to begin eliciting explanations.

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 5.3). The analyst can present each microtask to more than one worker to generate a more diverse set of responses. Presenting microtasks to multiple workers costs more and may also take more time.

5.1.1 Rating, Clustering, and Checking 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, unique explanations, or explanations based on the most reliable sources. In the third stage, the analyst enlists crowd workers to aid in this sorting and filtering process.

If the quality of the explanations generated by crowd workers is inconsistent, an analyst can have a second group of workers complete *rating microtasks* (Figure 5.4) in which they score the explanations. Each rating microtask 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 and readability of each response on a numerical (1-5) scale. We combine these ratings into a numerical *quality* score (0-5) that measures how well a worker's response explains the feature they were asked to focus on.

$$quality = (clarity \times relevance)$$

Multiplying by the binary relevance score gives irrelevant responses a quality score of 0, while all relevant responses receive a 1-5 quality score based on their clarity.

Multiple workers operating in parallel often generate duplicate or overlapping explanations and can create additional work for the analyst. We include redundancy microtasks in which we ask workers to identify and consolidate these redundant explanations. Analysts may also need to determine whether or not each explanation came from a reliable source. We include source-checking microtasks in which crowd workers check the explanations and sources generated by others and identify direct citations.

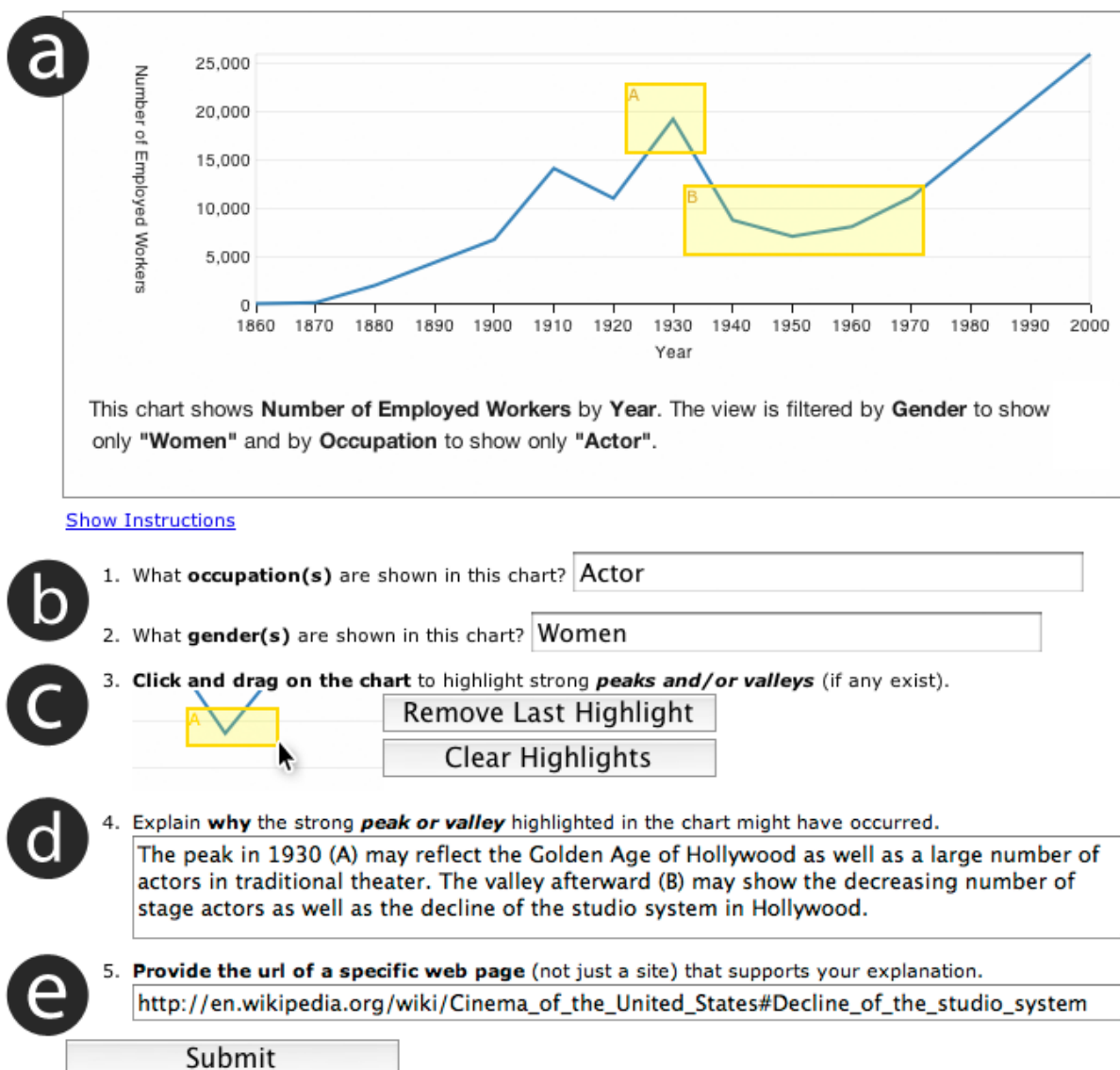
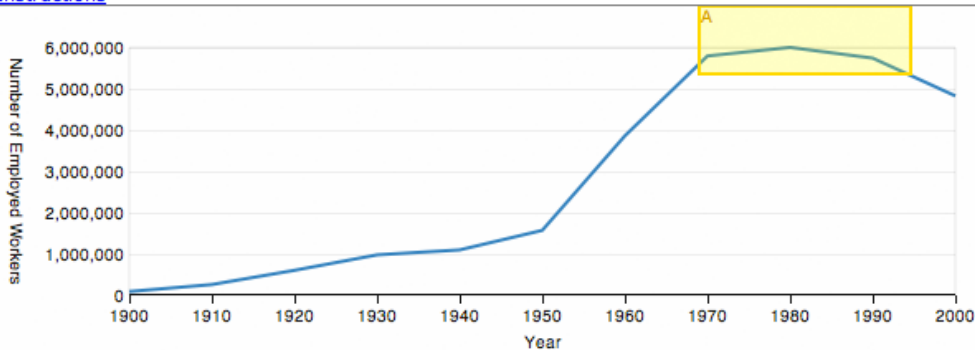


Figure 5.3: An example **analysis microtask** shows a single chart (a) along with chart-reading subtasks (b) an annotation subtask (c) and a feature-oriented explanation prompt designed to encourage workers to focus on the chart (d). A request for outside URLs (e), encourages workers to check their facts and consider outside information.

[Show Instructions](#)

a



This chart shows **Number of Employed Workers** by **Year**. The view is filtered by **Gender** to show only "Women" and by **Occupation** to show only "Secretary".

b

Prompt: Explain **why** any strong **peaks and/or valleys** in the chart might have occurred.

Response R1: "The peak is is the 1980s to 1990s and then begins to decline. This is the time period when women were becoming more prominent in the workforce, but the decline is due to the time period where women began to move up in the work world and secretaries was no longer the typical job."

c

1. What **occupation(s)** are shown in this chart?

2. Does this response provide an explanation for **why** the highlighted peaks and valleys in the chart might have occurred? ☒ Yes ☐ No ☐ None Present

d

3. How **clear** and **specific** is the response (Not Clear/Specific) ← ☐ 1 ☐ 2 ☒ 3 ☐ 4 ☐ 5 → (Very Clear/Specific)

4. How **plausible** is the response? (Not Plausible) ← ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 → (Very Plausible)

e

- For each response, identify if any of the other responses give better explanations for the peaks and valleys.

Response R1: With the advent of computers and advanced automation of offices, jobs for secretaries have been in decline since the 1980s. (Reference: http://www.allthingspolitical.org/careers/secretary_job_outlook.htm)

Does one of the other responses give a better version of this same explanation:

☐ R1 ☐ R2 ☐ R3 ☐ R4 ☐ R5 ☒ R6 | ☐ No

f

In a few short sentences, explain **specifically** why you ranked the responses in this order. (Responses that do not provide a clear reason for their ranking will be rejected.)

R3 most clearly attributed the dip in the 1980s and beyond to the diversification of career categorizations. R1 hints at a similar point but is less clear.

Figure 5.4: 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).

5.1.2 Examining and Managing Explanations

Finally, once workers have generated explanations, the analyst can view the responses in an explanation-management interface. Using the tools provided by our management environment, analysts can filter, sort, and organize the crowd commentary and decide which explanations and areas of the dataset to pursue further. An analyst may also choose to have workers iterate on a task, generating either additional unique explanations or explanations that improve on the best responses from a prior round.

5.2 Strategies for Eliciting Good Explanations

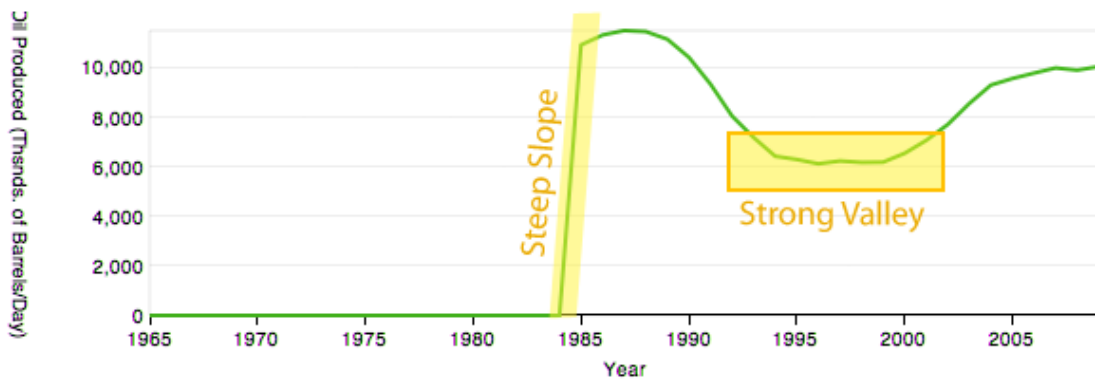
Simply asking workers to look at a chart and explain why it is interesting may not produce good results. We consider five types of problems that can reduce the quality of these explanations and discuss strategies (S1-S7) designed to mitigate these problems.

Note: For illustration we focus our discussion of strategies around two time series datasets (Figure 5.5); historical data on world oil production by nation from 1965-2010, and US census counts of workers by profession from 1850-2000. We consider more datasets later in Section 5.6.

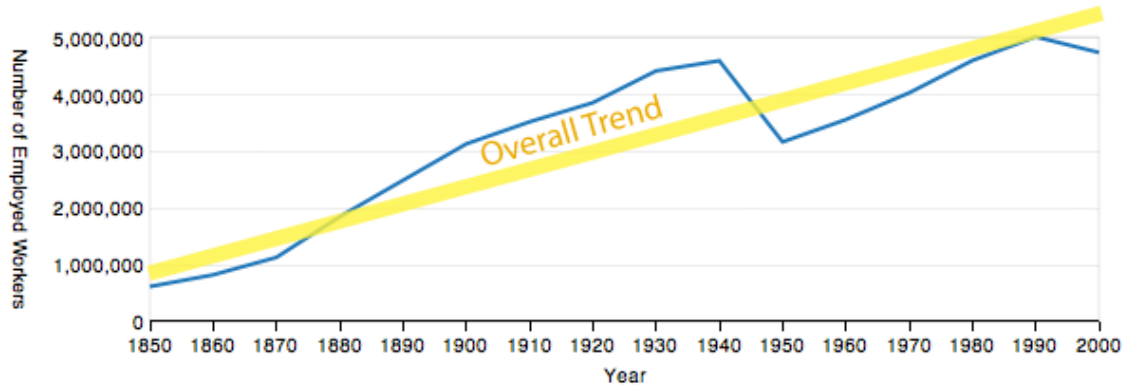
5.2.1 Problem 1: Irrelevant Explanations

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 aspects of the chart that are 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 that interest the analyst increases the likelihood that workers will provide relevant explanations. Consider the line charts in Figure 5.5. 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 multiple dimensions.



This chart shows **Oil Produced (Thousands of Barrels/Day)** by **Year**. The view is filtered by **Country** to show only "Russian Federation".



This chart shows **Number of Employed Workers** by **Year**. The view is filtered by **Gender** to show only "Men" and by **Occupation** to show only "Laborer".

Figure 5.5: Sample charts from the oil production and US census datasets used in our examples and experiments. Depending on their interests analysts may wish to focus workers' attention on a variety of different features of a chart, including *slopes*, *valleys*, and *overall trends*.

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.

5.2.2 Problem 2: Unclear Expectations

Workers may not understand 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 users to a dataset or feature type before they begin, 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 [12]. 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 datasets analyzed earlier. However, the data, chart type, and desired features should be similar to the new dataset. More interactive training, in which workers complete the example tasks themselves and then compare their responses against the example responses provided by an analyst, could also be used to more strongly communicate the style and content of a desirable response.

5.2.3 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 5.3e). Requiring such links may encourage workers to fact-check more speculative answers, uncover useful resources that the analyst can use later in the analysis process. It also provides information about the provenance of the explanation that analysts can use when assessing the explanation’s quality. However, asking workers to gather outside references may also increase the time and effort associated with a microtask, and may increase worker attrition. This strategy is best-suited to domains with public data and broad accessibility such as demographics, economics, and campaign finance, or where clear citations are important to the analyst. However, finding references may be more difficult in niche domains where web resources are limited.

5.2.4 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 5.3b) 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 baseline. Additionally, because “gold standard” answers to such chart reading questions are known a priori, we can automatically check workers’ answers and eliminate incorrect responses that indicate spam 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 [77,95]. In our case these questions also help workers pay 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 that workers first annotate relevant features of a chart and then explain those features (Figure 5.3c). Such annotations encourage attention to details and support deixis [49], 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 [46].

S6. Use pre-annotated charts . Alternatively, the analyst can pre-annotate visual features in the chart (Figure 5.5) 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.

5.2.5 Problem 5: Lack of Diversity

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

S7. Elicit explanations iteratively. Analysis microtasks can be run in multiple, sequential stages, in which workers see a chart along with the best explanations generated in prior stages. The analyst may elicit more diverse explanations by asking workers to generate explanations that are different from the earlier explanations. Alternately, the analyst can increase explanation quality by asking workers to expand and improve upon the earlier explanations.

5.3 Assessing Explanation Plausibility

Even if workers produce clear, relevant, and well-grounded explanations, it is still up to the analyst to examine each one to determine if it is plausible and if she should explore it further. Because workers can generate dozens or even hundreds of candidate explanations, identifying the most promising ones can become a time-consuming process that requires considerable effort from the analyst.

However, we can leverage the fact that workers' explanations are often redundant and are usually supported by known sources on the web. Information about explanation redundancy and provenance can help an analyst prioritize sets of redundant answers and quickly assess the plausibility of the possible explanations for the same phenomenon. We present a set of additional crowdsourcing techniques for identifying redundant explanations and providing provenance information that can help analysts evaluate candidate explanations.

When considering explanations for trends or outliers, an analyst's key task is to determine if each explanation is likely to be true and decide whether it should be discarded, retained, or explored

further. Analysts use a number of criteria to assess how plausible a candidate explanation is. We enumerate several key criteria:

C1: Text Clarity and Specificity. Some fraction of crowd workers typically satisfice—they perform the minimum amount of work to complete the task—and may generate poorly-constructed, unspecific, or logically implausible results. By comparison, well-written explanations that appear internally consistent can instill greater confidence in the explanations’ veracity.

C2: Explanation Frequency. If an explanation is proposed multiple times by different workers, it may indicate that the explanation is more likely to be a good one [100]. Conversely, a lack of redundant explanations may signal that there are many likely answers, and the odds that the workers have found the most plausible one are lower [107]. Clustering redundant explanations and indicating the frequency with which each explanation occurs can help analysts make these assessments.

C3: Explanation Provenance. An analyst can also use information about the source from which an explanation was taken, in order to help determine if it is plausible. To make this judgment, the analyst needs to understand both where the explanation originated and how it was collected or generated by the worker. An analyst may use provenance data to answer a number of specific questions about an explanation:

C3.1: Does the explanation cite a reputable source? If an explanation draws from a source the analyst is familiar with, the analyst can also leverage his or her knowledge of the source to help decide if an explanation is plausible. Citing a source that an analyst recognizes and trusts (for example a known news organization or reference) may bolster the explanation’s credibility, while citing an unknown or disreputable source may diminish it. Similarly, surfacing details about the cited source and other resources used by the worker as they derived the explanation can help analysts make this assessment.

C3.2: Does the content of the explanation come from the source or the worker? In our experience, workers who are not domain experts (including most workers on crowd marketplaces like Mechanical Turk) are more adept at extracting good explanations from sources than they are at producing explanations on their own. As a result, explanations that repeat or paraphrase facts and inferences from a good source tend to be more credible than explanations based on facts or inferences produced entirely by the worker. As such, an indication of whether or not the content is copied or paraphrased directly from the source can help analysts assess plausibility more easily.

C3.3: Is the explanation corroborated by multiple sources? If multiple versions of the explanation cite the same source, it indicates a reliance on that source. If the source is known and trusted, this reliance can increase confidence in the explanation. Alternatively, if multiple explanations cite an unknown source, it can suggest that the source is one that the analyst may wish to consider directly. Finally, multiple versions of an explanation that cite *different* reputable sources may increase confidence even further, since sources can corroborate one another [124].

5.4 Identifying Redundancy via Crowdsourcing

Grouping redundant explanations together can keep analysts from spending time considering duplicate explanations and can help analysts see which explanations are frequent or corroborated by multiple sources. However, determining whether multiple explanations are redundant is a difficult and somewhat subjective task.

The research community has produced numerous automated text similarity and topical clustering methods [69]. However, automated approaches tend to rely on the assumption that similar explanations will use similar language. These measures of explanation similarity can fail when explanations use different terms to describe the same phenomenon (e.g., “layoffs” instead of “downsizing”) or when the connection between two comments requires outside knowledge (e.g., the notion that widespread “layoffs” may be related to an “economic downturn”). Moreover workers’ explanations are typically short and the total number of explanations for a single feature can be small (sometimes less than 10 in our examples). Small text corpuses like these present a challenge for text similarity algorithms, since word co-occurrences tend to be very sparse, making it difficult to produce reliable clusters [96].

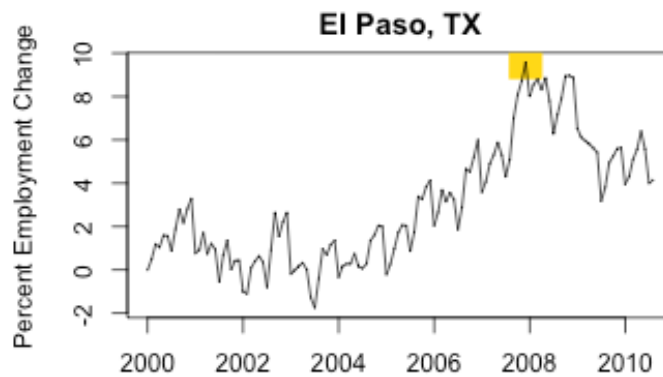
In contrast to automated approaches, human workers can leverage semantic information and outside knowledge to cluster sets of textual explanations. However, the limits of human attention ensure that individual workers can only examine a limited number of explanations at one time. Workers may also cluster explanations differently from one another, making it challenging to integrate clusterings obtained from multiple workers. As a result, crowd-based clustering approaches must provide a means of distributing the clustering tasks across workers and combining their efforts.

We explore two different crowdsourced approaches for clustering explanations: (1) In our *distributed comparison* approach, workers compare responses two at a time and mark any pairs of responses that give the same explanation. Our system then aggregates results from multiple workers to cluster the complete set of explanations. (2) In our *manual clustering* approach, multiple workers consider all of the responses (around 10 in our examples) and organize them into clusters. Our system automatically selects the best clustering from amongst those produced by the workers.

5.4.1 Distributed Comparison

In the *distributed comparison* approach (Figure 5.6), we ask crowd workers to examine pairs of explanations and indicate whether or not they are redundant. Using multiple workers, we collect at least 5 judgments for every pair of explanations, then average the binary similarity judgments to produce an average similarity score for the pair. To limit the impact of workers who attempt to game the task, we include pairs of gold standard explanations with known similarity, and remove results from workers who fail to mark them correctly. We then use these similarity scores to group the explanations in to a fixed number of clusters using k -means clustering.

One challenge when using k -means is picking the number of target clusters, k . We chose a heuristic for selecting k based on our own experiences clustering sets of explanations generated by workers. We found that the median number of clusters in the sets of explanations we considered was $k = 0.7 * n$, where n is the number of total explanations in the set. Because the proportion of redundant explanations can vary from set to set and is dependent on the semantics of the data, rules of this form are an imperfect solution. However, in our experience, our heuristic typically produces a value closer to the actual number of redundant clusters we observed than other common heuristics do (Tibshirani et al. [106] provide an overview of a number of methods for selecting k). We use our method to set k in all of our experiments that use k -means.



Prompt: Explain **why** the strong **peak or valley** highlighted in the chart might have occurred.

Response: " During that time the El paso government had a lot of money going in to new projects, fort bragg was becoming the home of several new troops and their families, 1.3 billion dollars went to improving their roadways and school systems. "(Reference: newspapertree.com/opinion/3561)

Response: " The University of Texas at El Paso (UTEP) started construction in that time period. The military base outside El Paso continued to hire contractors to support the base for the Iraq and Afghanistan support. And the Department of Transportation was expanding in the area at that time. "(Reference: <http://newspapertree.com/opinion/3561>)

Do these two responses give the same general explanation for the peaks and valleys in the chart?

- ☐ **Yes. Both responses give the same general explanation.**
- ☐ **No. The responses do not give the same explanation.**

Figure 5.6: In the *distributed comparison* approach, we show workers pairs of explanations for a phenomenon and ask them to decide whether or not the two explanations are redundant.

While this approach decomposes clustering into small tasks that are easy for workers to perform, it scales poorly as the number of explanations increases. Assessing redundancy for all pairs requires $\binom{n}{2}$ operations and the number of comparisons grows quadratically as the number of explanations grows. Framing these tasks as triplet-based comparison tasks and sampling to build a partial similarity matrix (as in Tamuz et al.'s "crowd kernel" [103]) can reduce the number of

total comparisons. Another solution may be to use a matrix completion approach similar to the one proposed by Yi et al. [125] to build similarity matrices without asking workers to compare all pairs of items. However, both of these approaches create approximations of the complete worker-generated similarity matrix, and may produce similarity scores for some pairs that were not intended by workers. As a result, we opt to build the complete similarity matrix by eliciting multiple worker comparisons for every pair of explanations. In the future, we hope to evaluate and employ approaches like these to reduce the number of worker comparisons necessary to build the similarity matrix.

Additionally, because workers never see all of the explanations at once, they may miss redundancies that require context from other explanations in the set. For example, four responses attributing employment growth in El Paso to (A) *“a new medical complex”*, (B) *“a new medical center at UTEP”*, (C) *“construction on the university campus”*, and (D) *“constructions of new building on campus”* might be split into two separate clusters if considered in isolation. If presented as a series of binary comparisons, workers are likely to group A and B together because they both mention the medical complex, and are likely to group C and D because they discuss university construction. However, seeing the larger set of explanations together could give a worker the opportunity to realize that all four explanations are actually attributing growth to the same hospital construction project.

5.4.2 Manual Clustering

Due to the many issues of distributed comparison, we developed a second clustering approach in which workers examine all of the explanations for a chart and group them manually. Displaying the full set of explanations gives workers the opportunity to identify clusters (like the one described above) that may not be obvious without additional context.

To simplify the task of specifying clusters, we created a system where workers group comments by color-coding them. In each manual clustering task (Figure 5.7), workers see the full set of explanations for a chart and can color code each explanation by clicking in the palette attached to it. When a worker assigns the same color to multiple responses, the system moves the responses next to one another, creating visually distinct clusters. These clusters allow workers to see their clusters as they create them and compare similar comments side by side without having to rely as strongly on their working memory.

Prompt: Assign each response a color by clicking the color bar below it.

If multiple responses give the same general explanation, assign the same color to each of them. Items with the same color will be moved together to help you compare them.

Response R2:Due to BRAC (Base Realignment and Closure), Fort Bliss will grow by 11,500 troops and is estimated to boost the El Paso economy by over \$4 billion annually.

(Reference: realitytimes.com/rtpages/20090206_hotmarket.htm)



Response R7:Expansion of Fort Bliss.

(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus)



Response R5:The Medical Center of the Americas opened the first new medical school.

(Reference: newspapertree.com/opinion/3561-the-el-paso-stimulus)



Response R3:It is possible that employment rate grew a little due to Rick Perrys strategies of creating jobs especially government jobs.

(Reference: www.usnews.com/...perry-created-jobs-in-texas)



Figure 5.7: In the manual approach, we show workers all explanations for a chart and ask them to create clusters by marking redundant explanations with the same color. Similarly-colored explanations are grouped together on-screen, allowing workers to see their clusters in context.

Clustering explanations is a subjective task and the boundaries between clusters can vary depending on subtle interpretations of the explanation text. As a result, multiple workers—even well-intentioned and well-informed ones—may produce different clusterings. Because many different clusterings may be valid, it is difficult to identify one clustering as the most correct or to combine the clusterings produced by multiple workers into a single clustering.

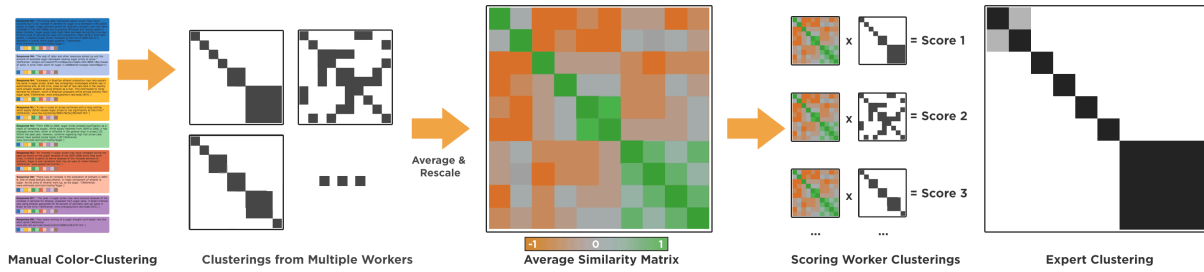


Figure 5.8: Illustration of our algorithm to select good worker clusterings from a larger set of possible clusterings. Workers’ manual clusterings (left) are transformed into similarity matrices (center-left) and averaged to create an average similarity matrix (center). Individual clusterings are then compared against the average to choose the most representative (center-right), which typically strongly resembles clusterings generated by experts (right).

To design an algorithm for selecting the best clustering from a set, we built on several observations:

1. If multiple workers agree that a particular group of explanations should be clustered together, there is a high likelihood that that grouping indeed reflects similarities in the explanations’ content [100]. As a result, we assume that the clusterings that are the most dissimilar from all other clusterings for a given chart are likely to be bad, while the clusterings that are the most similar to all the others are likely to be good.
2. Most systematic errors (e.g., a worker satisficing by lumping all explanations into a single cluster) can be caught by including gold-standard tests and by eliminating workers who complete the task in less time than it would take for a fast reader to parse all of the explanations. Other errors tend to be noisy (e.g., a worker satisficing by randomly clusterings explanations) and are not usually duplicated by multiple workers.
3. A single worker’s clustering is more likely to be internally consistent and understandable to the analyst, because it reflects a single set of judgments made in-context with one another. Therefore, choosing a single worker’s clustering is preferable to combining results from multiple workers.

Based on these insights, we designed a procedure for extracting the *most-representative* clusterings from a set of clusterings generated by multiple workers (Figure 5.8). The rating scheme we use is

based purely on the correspondences between workers' clusterings, without any a priori knowledge of the content or quality of the explanations.

First, we ask a multiple workers to cluster the explanations for a chart using the manual clustering interface (Figure 5.8 left). We then construct a separate *cluster similarity matrix* for each worker's clustering (Figure 5.8 center-left). Each row and column in this matrix corresponds to one of the explanations in the set. We initialize all elements in this matrix to 0, then assign a 1 to each element where the worker placed the explanation on the corresponding row and the explanation on the corresponding column into the same cluster.

Next, we average together the matrices from all of the workers who clustered the set. This produces a single *average similarity matrix*, which we normalize to a range of -1 to 1 (Figure 5.8 center). Positive values in this matrix correspond to pairs of explanations that were clustered together by the majority of workers, while negative values correspond to pairs that the majority of workers did not put in the same cluster. This matrix gives a sense which pairs of explanations are highly likely to belong in the same cluster and which are unlikely to belong together.

Finally, we select the *most-representative* clustering—the clustering from a single worker that most closely matches the average similarity matrix. We treat the positive and negative values in the average similarity matrix as rewards and penalties for a single worker's clustering as follows. We individually multiply each worker's binary similarity matrix with the average similarity matrix element-wise, and sum the values of all the elements in the product to obtain a final score (Figure 5.8 center-right). We retain only the clustering which produced the highest total score. The resulting clustering groups together the most pairs of explanations in a similar way to the majority of workers, and is thus the most likely to be correct. We surface only this most representative clustering to the analyst.

5.5 Explanation Provenance

To make judgements based on source reputability, analysts need information about the websites workers use to produce their explanations. We consider a set of techniques to help analysts make these assessments.

Examine a line chart showing employment change in a US city and briefly explain it.
 Requester: visualizationlab.ucb
 Qualifications Required: Location is US
 Reward: \$0.40 per HIT
 HITs Available: 10
 Duration: 30 minutes

a. explanation-generation task

Each of the charts in this HIT shows the percent change in the number of workers employed in a single US city since January 2000. (For example, if a city has a score of 5 during a month, it means that the number of people employed in that city was 5% larger during that month than in January 2000).

b. embedded web browser

en.wikipedia.org/wiki/Fort_Bliss

Fort Bliss
 From Wikipedia, the free encyclopedia
 Coordinates: 31.801847°N 106.424608°W

It has been suggested that *Fort Bliss shooting* be merged into this article or section.
 (Discuss) Proposed since June 2011.

Fort Bliss
 Part of Army Forces Command (FORSCOM)
 El Paso County, Texas and Doña Ana / Otero counties,
 New Mexico, Southwestern United States

Contents [hide]

- History
 - The Pershing Expedition
 - World War I and World War II
 - The Cold War
 - Base Realignment and Closure
 - The War on Terror

Explanation 1
 The expansion of Fort Bliss and base realignments added 14,000 jobs in the region in between 2005 and 2008.

Source:
 Use the embedded browser on the right to find evidence for your explanation. Select text that supports your answer and click to include it here.

c. explanation and source fields

d. highlighting tools

Finished with this HIT? Submit HIT Return HIT

Figure 5.9: An analysis microtask (A) is paired with an proxied web browser embedded inside the task (B). The explanation prompts in the interface (C) are linked to highlighting tools (D) that let workers cite specific sections of source documents.

5.5.1 Logging Activity and Sources

In addition to the basic analysis microtasks introduced previously (Figure 5.3), we also developed a version with an embedded web browser (Figure 5.9) that provides a record of workers' browsing activity during each task.

Recording the sites workers visit as they perform microtasks is difficult to implement in practice because the same-origin policy [90] implemented by modern web browsers prevents code from one internet domain from accessing web pages loaded from other domains. As a result, our microtasks cannot monitor activity that occurs in browser windows or tabs that do not originate from our site. This restriction would normally make it impossible to capture workers' web browsing and search activity as they complete the task.

We circumvent the restriction by having workers browse and search for sources using a custom web browser embedded within the analysis microtask (Figure 5.9B). This custom browser consists of a set of browser controls and an IFrame that loads web pages via our own custom proxy server. Requesting and then serving pages via our server (Figure 5.10) allows us to log each page workers visit and track any web searches they make as they forage for sources and candidate explanations.

For technical and security reasons, we do not proxy content served using protocols other than HTTP and do not handle third-party cookies. As a result, we cannot load content from sites that require users to authenticate or log in. Additionally, we cannot guarantee that workers perform all of their browsing within our proxied interface rather than in another browser window. However, our analysis of log data suggests that most of the sites workers visit are rendered appropriately via the proxy and that workers are active within our browser window for the majority of the time they spend on the task.

5.5.2 Supporting Fine-Grained Citations

Typically, when a worker cites a web page to support an explanation, only a small portion of the page (a paragraph or even a few sentences) is directly relevant to their explanation. Page-level citations can make it difficult for analysts or workers in rating microtasks to assess as source, since they may need to examine the entire web page to find the relevant text. We support finer-grained source citations by allowing workers to highlight specific blocks of text within pages as sources.

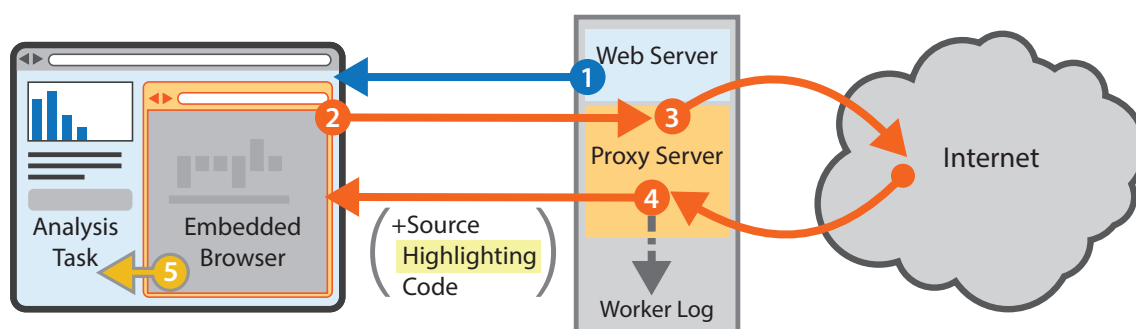


Figure 5.10: In our instrumented tasks, analysis microtasks are loaded from our web server (1). When workers look for evidence using the embedded web browser inside the task, page requests are redirected via our proxy server (2). The proxy server requests pages from their source (3), then logs them and injects custom highlighting code (4). Workers can then highlight text in embedded browser to have it included directly in their explanations (5).

We add highlighting controls to existing web pages by injecting custom code into each page as it is delivered by our proxy server. When a worker identifies a block of text on a proxied page that provides or supports their explanation, they can highlight the text and then click on an overlay (Figure 5.9D) to mark it as a source. We save the selected text and the URL of the page along with the explanation.

5.5.3 Detecting Copying and Paraphrasing

Understanding whether an explanation came directly from the source or the worker can be important when assessing the plausibility of a response. In general, we know relatively little about the domain expertise of workers recruited in a marketplace like Mechanical Turk. Therefore, our default assumption is that explanations that directly paraphrase a reputable source are likely to be plausible and are more desirable for the analyst. When workers add their own ideas and inferences to an explanation, we assume the explanation is less likely to be plausible, and the analyst may wish to either disregard the explanation or check the source themselves.

While people can generally identify whether or not an explanation is derived or paraphrased from a source, paraphrasing is difficult to detect automatically. However, these source-checking tasks are readily amenable to crowdsourcing.

In our workflow, we use *source-checking microtasks* to determine whether or not explanations are drawn directly from a source. In these microtasks, workers examine an explanation generated by another worker, along with the source document from which they derived it, and indicate whether the explanation “is copied or paraphrased from the cited source”.

5.6 Deployment

We have deployed our crowdsourced data analysis workflow on Amazon’s Mechanical Turk and used workers to generate 850 explanations for 60 different charts drawn from 15 different datasets.

Our deployment included the jobs and oil production 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

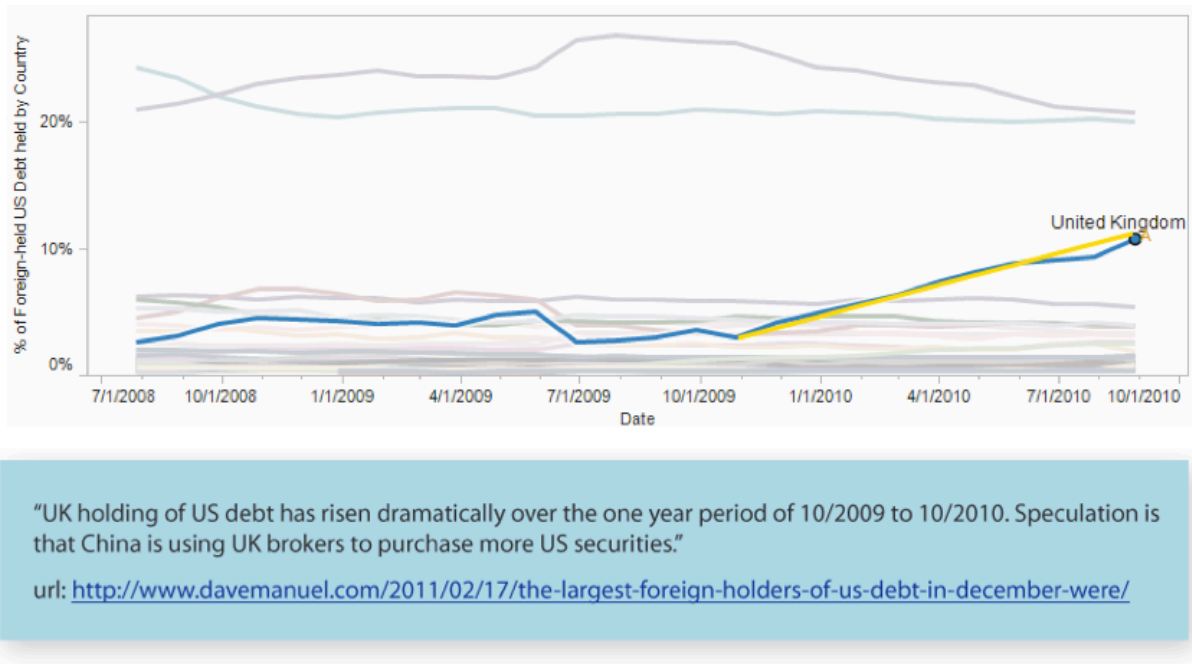


Figure 5.11: An example of an high-quality explanation generated by a crowd worker for a chart showing changes in foreign holdings of US sovereign debt since 2006. Here, a worker provided a novel hypothesis for the dramatic increase in UK holdings of US debt and cited a web site that provided further detail.

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). These datasets are examples of a rich class of public-interest datasets that contain valuable insights but do not require extensive domain knowledge. As a proof-of-concept, we generated a set of 4 or 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.

Three experts (the author, along with two other researchers with analysis experience) sampled 332 of the responses generated by workers and scored their relevance and clarity. We then generated quality scores for each explanation using the quality metric described in the Section 5.1.1. We assigned *quality* ≥ 3.5 to 220 responses (66%), 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.

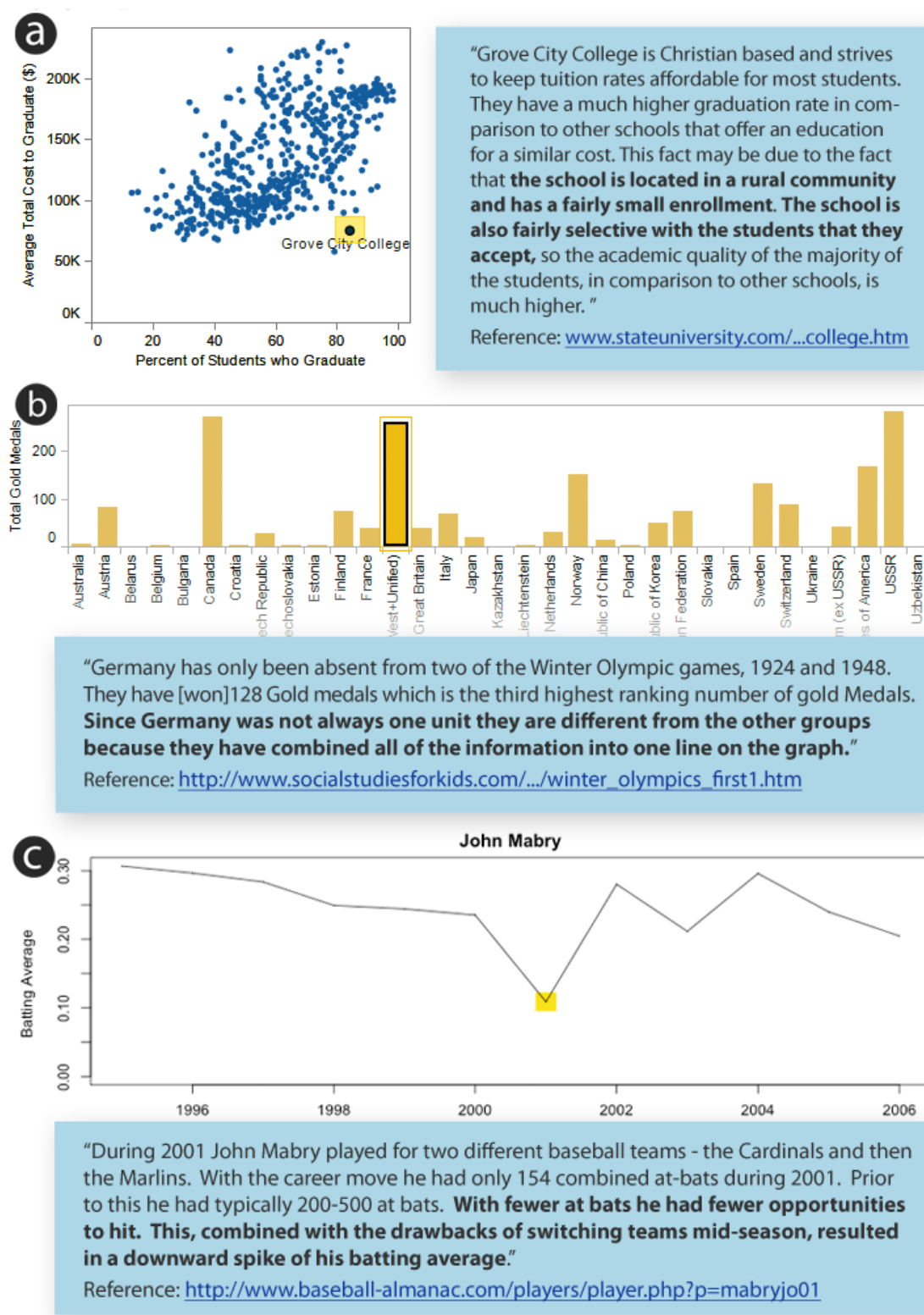


Figure 5.12: Sample explanations generated for charts showing university tuition and graduation rates (a), olympic medal counts by country (b), and historical batting averages (c). In each case we asked workers to provide explanations for a single outlier on a pre-annotated chart.

For example, one worker examining the US debt dataset suggested that a large spike in British purchases of US debt might be due to Chinese purchases through British brokers (Figure 5.11). Other novel insights are shown in Figure 5.12. Figure 5.12a shows one of a number of outliers in a visualization comparing cost to graduate versus graduation rate for major US universities. Workers posed common hypotheses for several low-cost, high graduation rate universities (including the one shown in the figure)—noting that they had religious ties and were often subsidized and selective. Figure 5.12b, shows how a worker identified an anomaly in the dataset behind a visualization of Winter Olympics medal counts by nation. The dataset combined counts from East and West Germany, which competed separately for much of the 20th century. Figure 5.12c, shows one worker’s explanation for a valley in baseball player John Mabry’s batting average in 2001. Five different workers all independently attributed this prominent valley to a midseason trade in 2001 that reduced his at-bats and impacted his performance. While these kinds of detailed, analytic responses and hypotheses are extremely rare in systems like Many Eyes, our approach is able to reliably elicit them for a wide range of datasets.

5.7 Evaluation

We conducted a series of experiments to test our strategies for eliciting good explanations, and evaluate our approaches for detecting redundancy and gathering explanation.

A full factorial experiment to evaluate all seven strategies for eliciting explanations 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 more subtle effects of three strategies—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). We also evaluated workers’ performance on rating and redundancy-detection microtasks, and tested their ability to detect copying and paraphrasing to verify explanation provenance.

5.7.1 Experiment 1: Strategies S1-S5 in Two Worker Pools

To evaluate the cumulative impact of the first five strategies (S1-S5) we asked one pool of workers to 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 [55] and can often provide results more quickly and cheaply than US-based workers, but studies of Mechanical Turk have shown that workers from outside the United States exhibit poorer performance on content analysis [92] and labeling tasks [19]. We designed this 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 first experiment, we ran 200 analysis microtasks using Mechanical Turk. We divided these microtasks into 8 experimental conditions:

$$2 \text{ strategy variants} \times 2 \text{ worker pools} \times 2 \text{ datasets} = 8$$

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 5.3) included a **feature-oriented prompt (S1)** asking workers to “explain why any strong peaks and/or valleys in the chart might have occurred”. 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 included an **annotation subtask (S5)** that instructed workers to highlight the peaks and valleys they explained. 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 [47, 70] 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 performed the tasks, three experts (including the author) scored each response and assigned it a quality score (as described in Section 5.1.1). The experts 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. Finally, the experts labeled whether or not each response referred to “*peaks or valleys*”, “*steep slopes or flat regions*”, or an “*overall trend*” in the data.

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 scores assigned by the experts were higher for US workers ($\mu = 2.23$, $\sigma = 1.79$) than for non-US workers ($\mu = 1.37$, $\sigma = 1.87$) (Figure 5.13), confirming our first hypothesis. Quality scores for US workers were higher, in part, because 83% of US responses contained relevant explanations, while only 42% of responses from non-US workers did so (per Section 5.1.1, irrelevant explanations receive a quality score of 0). Non-US workers frequently described the chart (34% of responses) rather than explaining it, or produced responses that were so poorly written we could not classify them, and (24% of responses). 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.

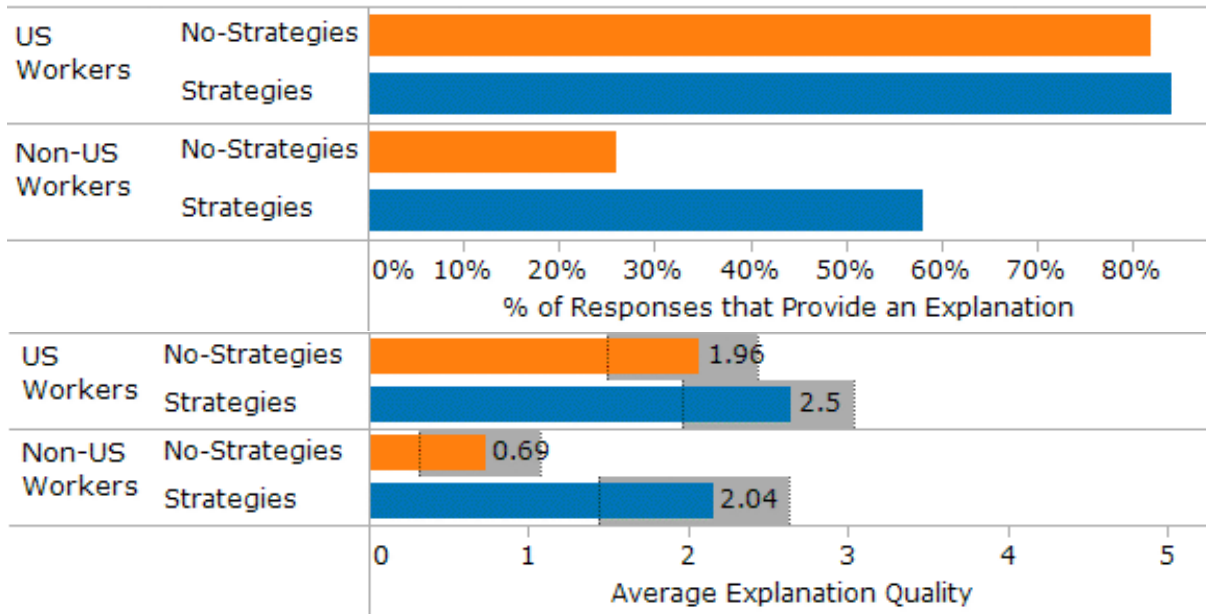


Figure 5.13: Percent of responses containing an explanation(top) and average explanation quality(bottom), by worker group (*US* / *non-US* workers) and strategy condition (*strategies* / *no-strategies*) in Experiment 1. Error bars give 95% confidence intervals.

We also found that across both the 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 ($\mu = 1.33$, $\sigma = 1.62$) ($F_{1,198} = 14.5$, $p < 0.01$), confirming our second hypothesis. 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.

Referencing Chart Features. 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 66% for workers in the *strategies* condition. Attrition was less than 24% 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.

5.7.2 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 effect of these strategies, we conducted a factorial experiment that varied each independently. We hypothesized that:

- (1) Feature-oriented explanation 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 refer to specific points in the chart and improve quality by increasing the number of responses that explained prompted features.

Methods

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

$$4 \text{ prompts} \times 2 \text{ examples variants} \times 2 \text{ annotation variants} = 16$$

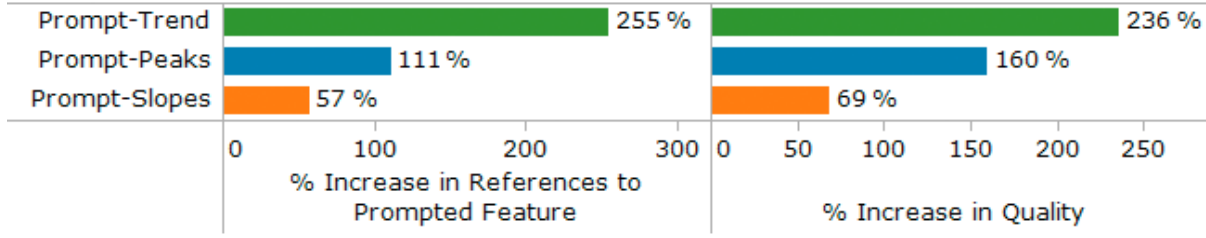


Figure 5.14: Percent increase in the number of references to the prompted feature (left) and the average explanation quality score (right) for each feature-oriented prompt (S1) condition in Experiment 2 over the control condition.

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 5.14). 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 the *prompt-control*. This increase was significant for workers in *prompt-trend* ($U = 372.0$, $p < 0.001$) and *prompt-peaks* ($U = 564.5$, $p = 0.008$), confirming our hypothesis for those two conditions. The increase in *prompt-slopes* ($U = 624.5$, $p = .064$)

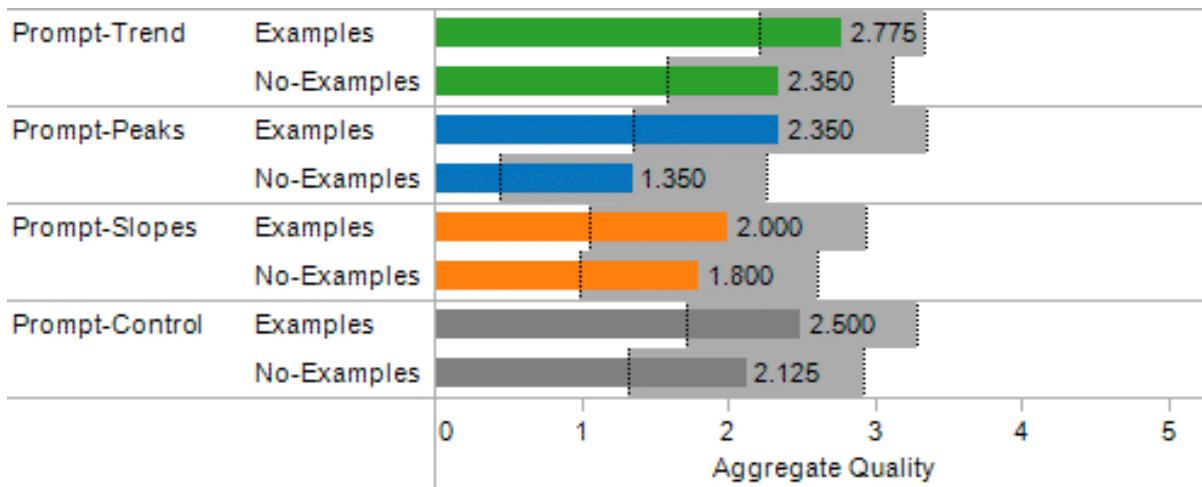


Figure 5.15: Average response quality by prompts (*prompt-trend*, *prompt-peaks*, *prompt-slopes*, or *prompt-control*) and examples (*examples*, *no-examples*). Error bars show 95% confidence intervals.

was not quite significant, probably because *prompt-control* workers were already more likely to explain slopes.

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 5.15), but the difference in quality was not significant ($U = 2717.5, p = 0.09$). Anecdotally, we observed that providing examples improved the consistency with which workers marked and annotated charts. Workers in the *worker-annotation* condition who saw examples of high-quality responses with annotated features, emulated those examples (Figures 5.1c and 5.4), usually marking a few clear peaks, slopes, or trends. Workers who did not see such examples created annotations that were more difficult to interpret and often annotated a larger number of 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. Most workers added either one or two annotations to the chart as they completed the microtask, but a few added as many as eight. 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 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 ($U = 3018, p = 0.08$).

5.7.3 Experiment 3: Reference Gathering

Based on results from Experiments 1 and 2, we hypothesized that including **reference gathering (S3)** would increase response quality. However we also hypothesized that the additional effort required to complete reference gathering tasks would contribute to high attrition. To test these hypotheses, we ran an additional 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 ($\mu = 2.22, \sigma = 1.94$) ($U = 211.5, p = 0.046$). The attrition rate was lower (46%) in the *no-gathering* than in the *gathering* condition (64%) but the difference was not significant ($\chi^2 = 2.209, p = 0.137$). Finally, we observed that 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 ($U = 175.5, p = 0.008$). Together, these results suggest that while reference gathering tasks produce useful references, they do so at the cost of speed and quality. As a result, more passive techniques for assessing provenance like those discussed in Section 5.5, may be preferable.

5.7.4 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 *worker-annotation* and *pre-annotation* conditions. However the number of responses 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 that if the features of interest are known a priori, both strategies perform well.

5.7.5 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

Participants in the *initial* condition generated 36 explanations, while those in the *iteration* condition generated 35 (many responses contained more than one explanation). 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.

5.7.6 Experiment 6: Rating

For rating microtasks to provide an effective means for sorting explanations, workers must be able to generate consistent ratings. To test consistency, 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 quality ratings assigned by workers would be similar to our own quality ratings.

Methods

We asked 243 Mechanical Turk workers to rate 192 different explanations across 37 charts. Using the interface shown in Figure 5.4, workers rated each response according to the criteria (relevance, clarity, and plausibility) described in Section 5.1.1. We compared these ratings against our expert quality ratings for the same results.

Results

In total, the workers produced 1,334 individual ratings for 192 different explanations.

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' quality scores and the experts' scores for each explanation were moderately correlated (average $\rho = 0.415$).

However, we found that we could produce results that were more strongly correlated with our own by instead using the mean score from multiple raters. We estimated the number of raters necessary to obtain a robust overall quality score by sampling from one to ten worker quality scores for

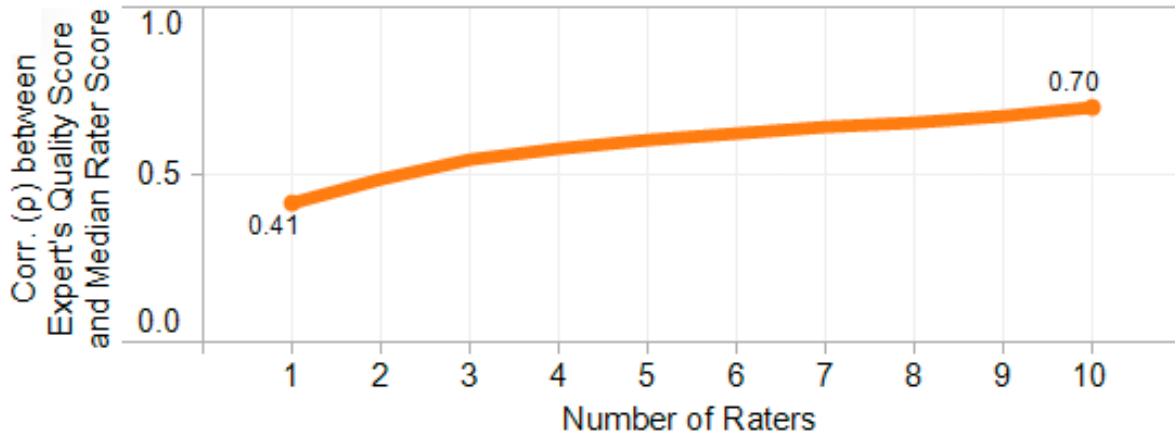


Figure 5.16: Agreement between workers' ratings and our own increases if we use the mean or median quality score from multiple workers. Using the mean from 5 or more workers gives strong ($\rho > 0.7$) agreement.

each response. For each number of workers, we averaged all of the selected workers' quality scores for each response, then computed the correlation between the mean quality scores and the expert scores (Figure 5.16). We found that averaging results from five raters produced quality scores that were strongly correlated with the expert scores (average $\rho = 0.726$), but adding additional workers gave diminishing returns.

5.7.7 Experiment 7: Redundancy

We also conducted an experiment to evaluate our two approaches for detecting redundant explanations. To compare our *distributed comparison* (Section 5.4.1) and *manual clustering* (Section 5.4.2) techniques we used both methods to cluster explanations for 12 different charts (each with between 10 and 20 explanations). We then compared the workers' clusterings against clusterings produced by the same three experts who scored explanation quality (Section 5.7.1).

In the *distributed comparison* condition, we created a comparison task for each pair of explanations given for each the 12 charts. This produced a total of 1,064 comparison tasks. We grouped tasks into batches of 20 and asked five unique workers to complete each batch. We paid workers \$0.20 for each batch. A total of 96 workers produced 5,032 comparisons. We then averaged all five workers' scores for each comparison and used k-means clustering to produce a final set of clusters (as described in Section 5.4.1).

In the *manual clustering* condition, we asked ten different workers to cluster the complete set of explanations for each of the 12 charts. Again, we paid workers \$0.20 for each task. A total of 91 workers participated, producing 120 total clusterings. We then extracted the single most-representative clustering for each chart.

To prevent workers from gaming the task, we included gold standard explanations in both conditions. In each task we added two stock explanations that we knew to be redundant and a third which we knew to be unique. We eliminated workers who failed to group the known redundant explanations together or who grouped the unique pair.

As a baseline, we also included an *unclustered explanations* condition, in which we kept the complete set of explanations for each chart without any clustering. We also compared our strategies against an *automated* condition in which we calculated the similarity between explanations based on the word overlap between them (using cosine similarity [96]), then clustered the explanations using k-means.

Results

Because clustering is subjective and no objective “best” clustering exists, we compared the results against manual clusterings generated by the expert raters. We hypothesized that the manual method would produce the clusterings that were the closest to the experts. We based our hypothesis on the observation that workers in the manual clustering conditions could see the complete sets of explanations at once and make clustering decisions with more complete context. We also expected results from the manual clustering method to be more similar to the experts because they are produced by a single worker, and are likely to be more internally consistent than results produced by aggregating multiple workers’ comparisons.

We compare clusterings against the expert clusterings using the F-measure, a symmetric similarity metric that is tolerant to small errors on large clusters, but intolerant to bi-directional impurities [3]. The F measure of a single cluster is the maximal harmonic average of the precision and the recall, and the F measure of an entire clustering is the weighted average of the F measures of all the clusters. Given two clusterings L and R , their F measure is:

$$F(L, R) = \sum_i \frac{|L_i|}{n} \cdot \max_j F(L_i, R_j)$$

where n is the total number of clustered elements, i ranges over the number of clusters in L and j ranges over the clusters in R , and L_i is the i ’th cluster in L and R_j is the j ’th cluster in R .

The function $F(L_i, R_j)$ is defined as:

$$F(L_i, R_j) = \frac{2 \cdot \text{Recall}(L_i, R_j) \cdot \text{Precision}(L_i, R_j)}{\text{Recall}(L_i, R_j) + \text{Precision}(L_i, R_j)}$$

where:

$$\text{Precision}(L_i, R_j) = \frac{|L_i \cap R_j|}{|L_i|}$$

$$\text{Recall}(L_i, R_j) = \frac{|L_i \cap R_j|}{|R_j|}$$

The F-measure similarity for two clusterings is reported on a range from 0 to 1, where 1 indicates that the clusterings are identical and 0 indicates that they are completely dissimilar. We scored each clustering by computing the F-measure between it and each of the three expert clusterings, then averaging the three results (Figure 5.17).

To calibrate our expectations, we compared the three *experts clusterings* against one another. On average, we found that their clusterings were quite consistent with one another ($F = 0.84$). Pairwise comparisons between the individual experts (E1-E2: $F = 0.84$, E1-E3: $F = 0.85$, E2-E3: $F = 0.83$) revealed that no one expert was an outlier.

An ANOVA showed a significant effect for clustering method on the average F-measure score ($F_{3,44} = 4.97$, $p < 0.01$). Pairwise t-tests also showed that selecting the *most-representative manual clustering* produced results that were significantly closer to the experts than the average *manual clustering* ($p < 0.01$). *Most-representative manual clustering* also produced clusters that were significantly closer to the experts than clusters produced in the *distributed comparison* ($p = 0.04$) and *automated* ($p < 0.01$) conditions or the results from the *unclustered* ($p < 0.01$) condition.

On average, the *unclustered* results were the least similar to the experts (average $F = 0.68$). This value is non-zero because even the clusters of explanations generated by experts often contain a number of singletons—explanations that do not cluster with any other. As a result, even an unclustered set gets the clustering right for these clusters of size one. Clusterings from the *automated* approach received a similarly low scores (average $F = 0.67$), confirming our intuition that text-based techniques are not well suited for clustering sparse, noisy data. Clusterings produced by the *distributed comparison* condition were somewhat more closely aligned with the experts' scores

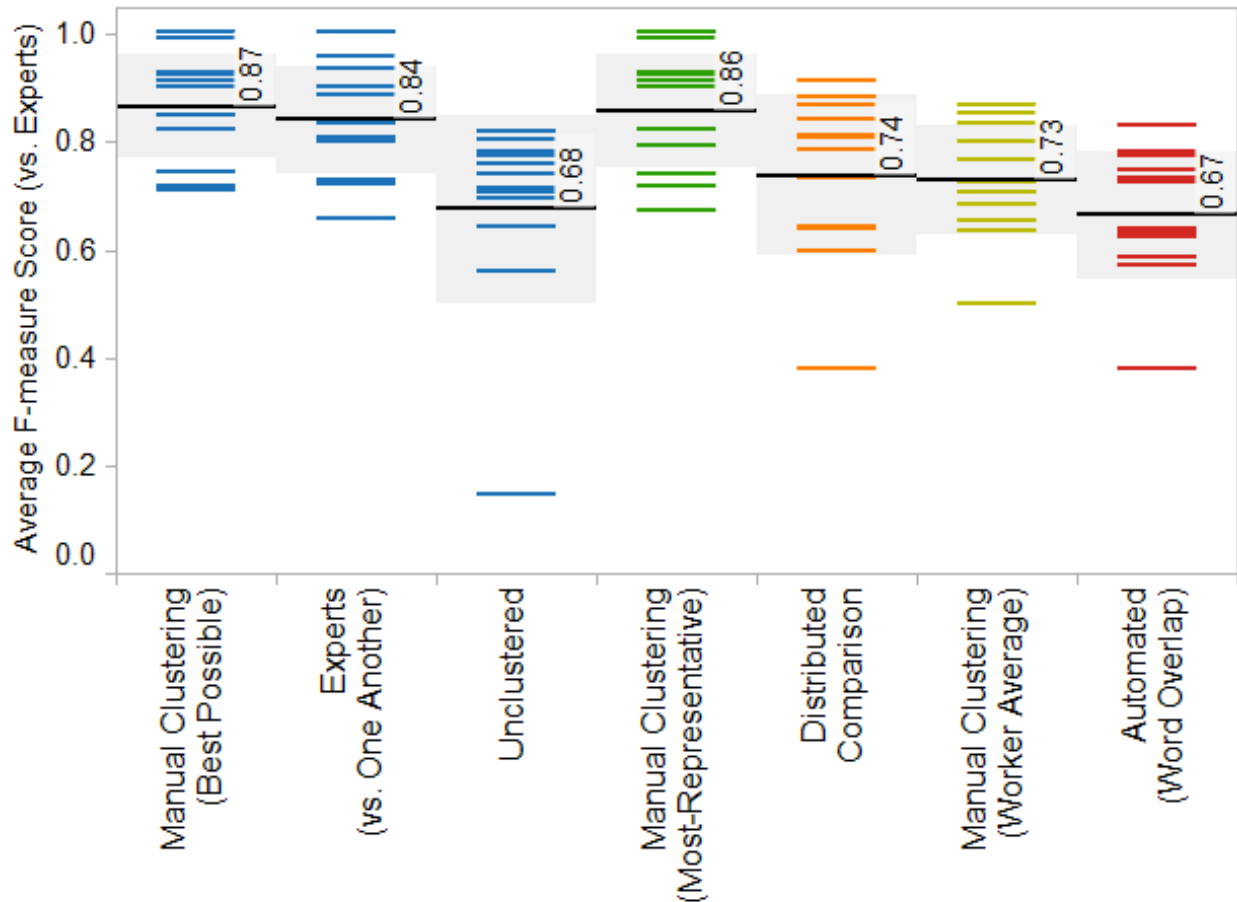


Figure 5.17: Results for each of our clustering selection methods. Each mark shows the average F-measure similarity between the experts' clusterings and the clusterings produced by the given clustering method. A separate mark is shown for each chart. The black line and grey bars give the average and standard deviation for each method.

(average $F = 0.74$) than the unclustered results. The clusterings produced by workers in the *manual clustering* condition were also a bit better (average $F = 0.73$). Choosing the *most-representative manual clustering* using the procedure in Section 5.4.2, however, produced better results across all 12 of our charts (average $F = 0.86$). For almost every chart, the most-representative selection algorithm chose the worker clustering that was the best possible match to the three experts. Moreover, the most-representative clustering was closer, on average, to all three of the experts than the three experts were to one another (average inter-expert $F = 0.84$). These findings suggest that choosing the most-representative clustering provides a reliable way of selecting high-quality clusterings.

5.7.8 Experiment 8: Copying and Paraphrasing

We also evaluated how well workers were able to identify paraphrasing from sources. To establish a baseline for how often workers' explanations are copied or paraphrased from the sources they cited, two of our three expert raters examined a sample containing 70 explanations from our deployment that included citations. The two experts individually examined each explanation and the source it cited and coded the explanation as either "copied or paraphrased from the cited source" or "not copied or paraphrased from the cited source". Afterward, the two experts worked together to resolve any differences, and produced a single gold standard. Of the 70 explanations, the experts marked 60% as copied or paraphrased from the source.

We then conducted an experiment to determine how reliably workers could detect paraphrasing. We randomly sampled 20 explanations of the explanations scored by the experts and presented each as a *source-checking microtask* to the crowd. Five crowd workers examined each explanation and source and voted whether the page was or was not "copied or paraphrased from the source". We then tallied these votes and assigned the winning label to each explanation.

The workers' final result matched the experts' for 75% of the explanations. All of the incorrect cases we observed were false negatives—workers indicated that results were not drawn from the source, while the experts deemed that they were paraphrased. The high number of false negatives suggests that workers as a whole used a more conservative definition of paraphrasing than the experts.

5.8 The Explanation Management Interface

Once workers have rated and clustered a set of explanations, we must surface that information in a way that allows the analyst to quickly browse the explanations and assess them. To this end, we developed an explanation-management interface (Figure 5.18 and 5.19) that provides a number of tools and visual cues intended to help analysts quickly find unique explanations and judge their

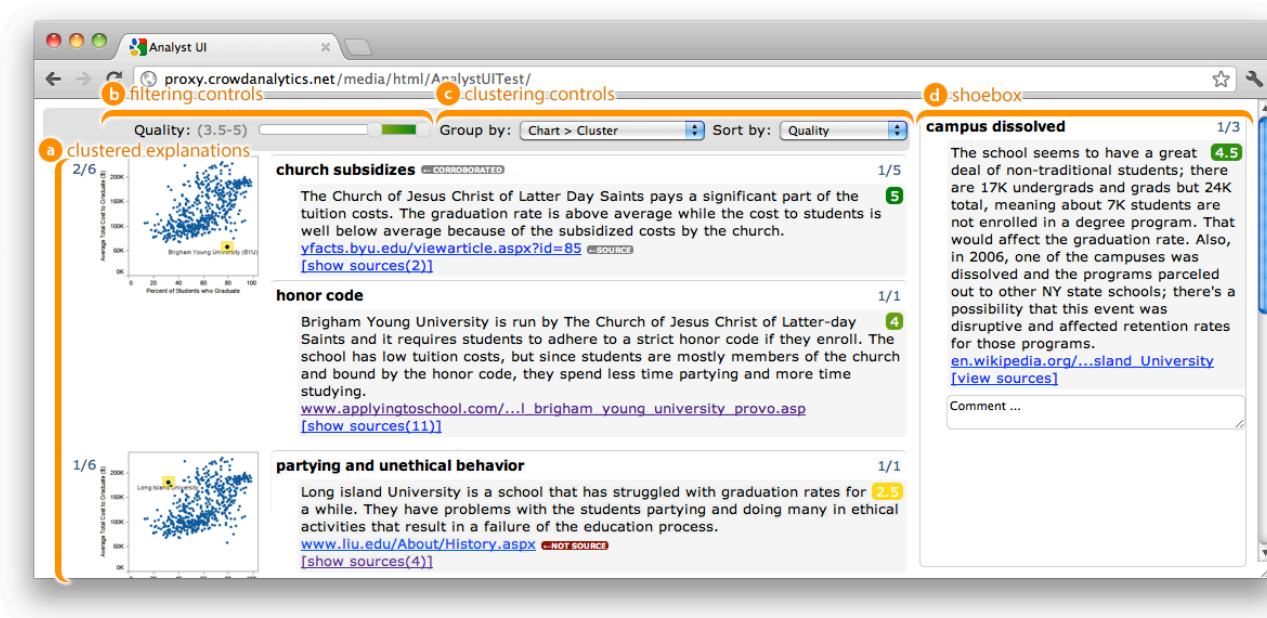


Figure 5.18: The explanation-management interface. Explanations (A) can be clustered and collapsed by chart, topic, and source. Filtering (B) and clustering (C) controls allow the analyst to hide low-scoring clusters and control how they are nested. Explanations, clusters, and charts, can be dragged to the shoebox (D) and annotated for later review. Figure 5.19 shows additional detail for a single cluster.

plausibility. We tailored the interface based on the criteria (C1 through C3) that we identified in Section 5.3:

C1 Text Clarity and Specificity

C2 Explanation Frequency

C3 Explanation Provenance

C3.1 Source Reputability

C3.2 Paraphrasing and Worker Additions

C3.3 Corroboration

Analysts can use this interface to browse, filter, and organize explanations generated by workers. Using the explanation-management tools, they no longer need to read through each and every explanation in order. Instead, they can explore clustered results, filter them by quality and frequency, and get a sense of their provenance. This section describes the various features of the interface in terms of the criteria they surface.

By default, the interface displays a list of explanations grouped first by chart view and then by cluster. Clusters are initially collapsed, so that only the explanation in the cluster with the highest quality score is visible. The clusters are also sorted based on their quality scores, so that the clusters containing the clearest, most plausible explanations are shown first. The analyst can expand clusters to inspect their individual members, and can filter the set of clusters based on a variety of attributes. In many cases, the analyst may wish to save interesting explanations to a “shoebox” [83] in order to revisit them later in the sensemaking process. Our interface allows analysts to save good explanations or groups by dragging them to a shoebox panel at the right of the screen (Figure 5.18D).

Each cluster in the interface includes a set of visual indicators designed to allow the analyst to quickly make judgements about the explanations in it, often without even reading them. These include explanation quality and frequency information (e.g., cluster size) as well as visual indicators that allow analysts to quickly determine explanation provenance.

5.8.1 Surfacing Explanation Clarity and Specificity

The interface displays the average quality scores generated by workers in *rating microtasks* (Section 5.1.1). We display the quality score in the upper right corner of each explanation (Figure 5.19G) and color the score using a red-yellow-green color scale. These quality indicators allow an analyst to quickly determine which explanations are more likely to be clear and specific (criteria C1). Analysts can also reduce the number of visible explanations by using the filtering controls at the top of the interface to hide explanations and clusters that do not contain explanations with high quality scores.

5.8.2 Surfacing Explanation Frequency

By default, the system collapses clusters of redundant explanations so that each cluster displays just the highest-quality version of the explanation. Each cluster also contains a count showing the total number of explanations in the cluster and how many are currently visible (Figure 5.19D). The highest-quality explanation serves as a summary of the cluster and reduces the amount of effort an analyst must expend to examine the explanation. An analyst can also use the cluster size to gauge the frequency and level of support for the explanation (criteria C2). If the analyst wants to inspect other versions of the explanation, they can expand a collapsed cluster by clicking on the cluster size indicator. Clicking on the indicator a second time re-collapses the cluster.

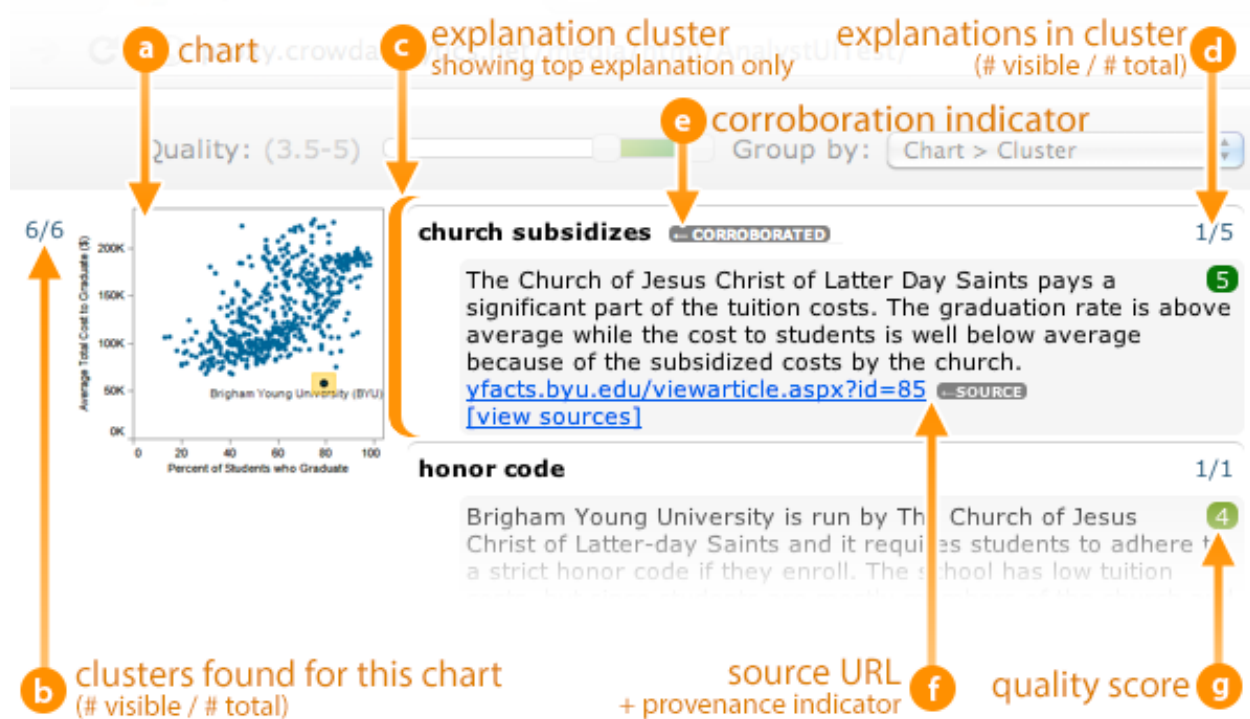


Figure 5.19: A closeup of the explanation-management interface introduced in Figure 5.18. Here we highlight a single chart (A) with two explanation clusters. Each chart includes an indicator (B), showing the number clusters of explanations for the chart. Each cluster (C) displays a count showing explanations it contains (D) and an indicator showing whether the explanation is corroborated by multiple sources (E). Each individual comment displays a source URL and provenance indicator (F) along with a color-coded quality score (G).

5.8.3 Surfacing Explanation Provenance

Each explanation also displays an abbreviated link to any web pages it cites (Figure 5.19F). These short links allow the analyst to quickly determine if the explanation is drawn from a source that they trust. The analyst can also click the link to view the source page along with any sections of the page highlighted by the worker (criteria C3.1).

If an explanation is of particular interest to the analyst, he or she can expose additional provenance information by clicking the “view sources” link on the comment. Clicking the link exposes the complete set of web pages the worker visited while generating the explanation along with detailed timing information. The analyst can use this list to locate and inspect other sources that

informed the explanation and help build an understanding of how a worker came to a conclusion (criteria C3.2).

If the analyst determines that a specific domain or web page is a good source, he or she may wish to directly explore other explanations that are drawn from that source. In our own experience, the sources which provide the best explanation for one chart may also provide good explanations for others (for example pages from the Bureau of Labor Statistics provide good explanations for changes in employment in many different US cities). Therefore, our interface also allows the analyst to group explanations based on the sources they cite to quickly find multiple explanations drawn from the high-quality sources.

5.8.4 Surfacing Paraphrasing and Worker Additions

In the explanation-management interface, we provide a provenance indicator next to the source URL (Figure 5.19F) of each explanation that more than 50% of *source-checking* workers identified as being either copied or paraphrased directly from the source. This indicator allows analysts to quickly identify explanations that are drawn directly from a source before reading them. Knowing an explanation was copied or paraphrased from a known source can allow an analyst to make confidence judgments based on that source's reputation (criteria C3.2). High-quality paraphrased explanations also serve as leads to help analysts identify good web resources that they may wish to utilize directly.

5.8.5 Surfacing Corroborating Explanations

An explanation that cites multiple reliable sources is more likely to be credible than one that sites only a single reliable source (criteria C3.3). Therefore an analyst may wish to know if multiple versions of an explanation in a cluster cite the same source or refer to multiple independent ones. In our interface, workers can assess this directly by expanding a cluster and grouping the responses within in by URL or domain. We also provide a "multiple sources" indicator in the heading of clusters that contain corroborating citations. Mousing over this indicator displays a list of sources along with the number of explanations in the group that cite them. This indicator serves as a shortcut for analysts, allowing them to quickly make confidence judgments based on corroborating sources without examining explanations or sources individually.

5.9 Discussion

Based on our experience collecting, clustering, and exploring crowdsourced explanations, we offer a few additional observations.

5.9.1 Explanation Segmentation

Our current implementation asks workers to separate distinct explanations into separate fields in the explanation microtask and allows them to select a different source text for each. However, in practice, many workers still give multiple candidate explanations as part of a single paragraph or sentence. Responses that contain multiple explanations can be difficult to group, since each one may include several distinct explanations that each belong in disparate clusters.

We addressed the issue of segmentation by creating the explanation-generation tasks that encouraged workers to enter each explanation in a separate text box. Providing separate text boxes and clear instructions reduced the number of responses that mixed multiple explanations. However, clean segmentation remains difficult to enforce, especially because explanations are often interrelated. Another possible approach might be to ask workers in intermediate *segmentation microtasks* to break apart compound responses into their constituent explanations. However, these tasks introduce the potential for information or intent to be lost as workers break apart or alter responses generated by other workers.

All of these issues are related to the broader issue of task granularity when crowdsourcing open-ended tasks. Breaking tasks into small, modular components makes it easier to compose tasks together and process results systematically. Small, straightforward tasks also reduce the potential for worker error, and make it easier to identify and discard poor results. However, small tasks may inhibit contributions from talented or knowledgeable workers, since they are not free to explore, author, or contribute outside the constraints of the task and cannot bring their expertise to bear on areas of the problem where it might be beneficial. As a result, balancing task simplicity and flexibility in a way that suits the expertise and trustworthiness of a worker pool remains a key challenge when designing new tasks.

5.9.2 Defining Redundancy

While we assume a particular definition of redundancy when clustering explanations, other types of clustering may be useful for analysts. We define redundant explanations as explanations that

give “the same general explanation for a trend or outlier”. This means that only explanations that make the same assertion about the trend and provide the same level of detail will be clustered. However, we observed that workers often produce explanations that are not strictly redundant, but are hierarchically related (for example, one explanation might attribute a drop in employment statistics to “an economic downturn” while another cites job losses in a specific industry). Clustering explanations hierarchically would allow analysts to consider high-level explanations and make confidence assessments about them before examining lower-level details.

5.9.3 Crowd Composition

Our approach assumes a crowd composed largely of non-expert workers whose responses may be of variable quality—for example, workers recruited in online task markets like Mechanical Turk. However, more complex analyses or datasets that require specific domain knowledge may call for the use of private crowds. We believe a similar workflow could be used to systematically collect and integrate findings from large crowds of trusted workers. In trusted crowds, some quality-control mechanisms could be relaxed, reducing the number of post-processing steps and giving workers more freedom to explore. For example, trusted workers could be given the freedom to manipulate the visualization and explore alternate views of the dataset that might inform their explanations. Trusted workers could also self-assess the quality of their explanations and sources, reducing the number of steps in workflow while still providing metadata that analysts can use to filter and re-organize their results.

5.9.4 Economics of Crowd Work

Because crowd markets remain a new phenomenon, many questions remain about the economic efficiency of crowd-based systems [52]. For example, it remains unclear whether it will be economical for analysts to employ large-numbers of novice workers on a short-term basis rather than cultivate a trained cadre of analysis specialists. Designers of social data analysis systems that employ crowd workers will also need to consider the ethical implications of using paid crowds and work to ensure that workers are compensated fairly and enjoy sufficient protections. This is especially important given that many proposed crowdsourcing platforms (including [63, 45]) employ workers in developing regions, where income levels are lower and fewer worker protections exist. A considerable body of recent research has focused on reducing the cost of crowd-based work, largely by minimizing the amount paid to workers [51]. However, future systems will need to strike a balance between cheaply and accurately performing analyses and ensuring that workers are treated fairly.

Chapter 6

Future Work

This dissertation has focused on several core research questions regarding social data analysis. However, many additional aspects of social data analysis remain to be explored. Future work beyond this thesis will support stakeholders not addressed in our examples and make the process of organizing, presenting, and sharing analysis more accessible.

6.1 Alternate Models for Crowdsourcing Analysis

In Chapter 5, we demonstrated that paid crowd work can be a viable tool for generating and rating hypotheses—one key component of the data analysis process. However, other steps in the sense-making cycle—including organizing content, comparing hypotheses, and searching for relations between observation—may also be amenable to crowdsourcing. Until now, we have considered parallel and iterative processes in which workers collaborate to produce and group evidence. However, competitive models of analysis—in which workers are offered incentives for producing better results than their peers or for disproving explanations and hypotheses created by others—also present a fruitful area for exploration. *What are the relative benefits and tradeoffs associated with competitive models of analysis?*

For example, we may be able to produce stronger explanations by providing incentives that encourage workers to evaluate, challenge, and validate one another's explanations and compete to generate the most likely or more diverse opinions. Techniques for hypothesis validation like Analysis of Competing Hypotheses [8] provide a systematic way of integrating many (sometimes com-

peting) contributions and controlling against bias. Moreover, they can be used to operationalize hypothesis testing into hypothesis generation, evidence collection, and cross-validation tasks that can be carried out by multiple users working in parallel (as in CACHE system [23]). We believe there is great potential for using similar techniques to perform hypothesis testing using online labor markets. For example, crowd workers could be instructed to carry out CACHE-style tasks in which they generate new hypotheses and search for evidence that invalidates the hypotheses generated by other workers. This approach could be paired with financial or social incentives that reward the workers who successfully disprove hypotheses and who produce hypotheses that survive elimination.

Kaggle [56] and other competition-based platforms for analysis represent another interesting point in the space. These tools allow individuals and teams of experts to compete to produce solutions to well-scoped data mining and prediction challenges—for example, developing the best algorithm to predict consumer shopping behavior on a website or identify celestial objects in high resolution telescope imagery. Typically these competitions are targeted at small groups of analysis experts and provide monetary rewards to the winners. Kaggle’s approach is less fine-grained than ours, with each team or individual completing the entire analysis in isolation and comparing only their results.

We suspect that fertile ground exists in the scales between our crowdsourcing work, which relies largely on small-scale novice labor in microtask markets, and these sorts of expert-level competitions. For example, it may be more productive to have teams compete for financial incentives on smaller pieces of the analysis, but periodically share findings and strategies. Future work should also focus on comparing rewards structures for competitive analysis and understanding which kinds of tasks are best suited to different competitive models.

6.2 Engaging Domain Experts

While our work has focused largely on analysts (Chapter 3), community members (Chapter 4), and crowd workers (Chapter 5), making sense of more complex, domain-specific datasets may require the input of domain experts. One promising thread of future research involves developing tools and strategies for identifying domain experts and incorporating their efforts. *How can we engage with outside domain experts and integrate their contributions into analyses?*

Domain experts provide valuable insights and expertise. Moreover, they may be able to answer questions about technical subjects that other workers and even analysts cannot. However, finding and engaging expert users on the web can be difficult. For example, during the development of CommentSpace, we deployed several versions of the tool live on the web—both on our own site and within news stories. Despite an effort to target these visualizations at particular communities and recruit experts to use them, very few visitors commented or contributed to the analysis. These efforts may have failed for a number of reasons. One possibility is that because the commenting was situated within our own site, rather than in the context of their existing community, expert users lack sufficient incentive to contribute. This observation is consistent with Danis et al.'s finding [27] that the most productive discussions in Many Eyes took place not in the context of the site, but offsite on blogs and forums where the visualizations were used as a “community component”. Users may also have been deterred by the relative complexity of the interface and the fact that the task (asking questions and generating hypotheses and evidence) was often ambiguous. These early deployments violated important design principles that emerged during our subsequent work with novice communities and crowd workers—they failed to provide clear, feature-oriented prompts.

One approach for eliciting input from outside experts is to extend a system like CommentSpace so that analysts can embed simple visualization views anywhere on the web, and collect responses and explanations in situ. On websites with a dedicated commenting mechanism, an analyst could embed only a visualization view and the system could collect responses by scraping the page and extracting comments. Where no commenting mechanism exists, analysts could elicit expert feedback by embedded question prompts similar to our analysis microtasks (Chapter 5) along with the visualizations. We hypothesize that embedding visualizations and questions directly in Q&A sites, forums, blogs, and other existing communities will allow analysts to engage domain experts and elicit feedback more easily. These new tools could lower the barrier to entry by incorporating strategies from the feature-oriented microtasks that proved successful in our crowd research.

For example, a simple embeddable CommentSpace web widget (Figure 6.1) could be used to elicit feedback from domain experts. Rather than directing experts to the CommentSpace site or embedding the entire CommentSpace interface in an outside site, an expert would export a single CommentSpace comment and its associated visualization views a self-contained widget. The widget could be embedded directly in forums, blog posts, Q&A sites, social media, and even interpersonal communications like email. Each widget would feature a clear prompt and would resemble the analysis microtasks used in our crowdsourcing work (Figure 5.3). On sites where no com-



Figure 6.1: Mockup of a possible design for an embeddable CommentSpace web widget designed to be embedded in Q&A sites, forums, and personal communications to elicit insights and feedback from domain experts.

menting mechanism exists, the widget would also include an answer field along with the prompt. Responses entered in the answer field could be submitted directly to CommentSpace as new comments or sent to crowd workers for rating and iteration. In Q&A sites or forums, a separate answer field would duplicate the existing functionality of the site. In these cases, we could embed only the visualization view(s) and the prompt, then extract experts' responses directly from the site.

This line of research still presents a number of challenges. Identifying domain experts qualified to explore a given dataset and an appropriate venue for eliciting explanations from them remains a difficult task. One possible approach for identifying experts may be to use data analysis and machine learning techniques to mine the content on Q&A sites in order to identify the people most likely to give a good response to an analysis question. Future research will also need to test a range of monetary and social-psychological incentives in order to better understand how to motivate experts to contribute.

6.3 Supporting Ad Hoc Social Data Analysis

While experimental tools like sense.us [48], Many Eyes [111], and CommentSpace make visualization tools more social by supporting embedding and commenting, they still build primarily on

proprietary visualization and commenting tools that make it difficult to share, combine, and build stories around data on the web.

However, a number of compelling recent examples of social data analysis have begun to occur outside these kinds of tools. One example which we have explored is the Google Books Ngram viewer [42], an interactive visualization of data from the Google Books corpus (Figure 6.2). After it was released in December 2010, the Ngram viewer elicited thousands of tweets, Facebook posts, and blog entries from users noting trends, commenting on data quality issues, and building narratives around their findings. This fountain of discourse was made possible by the designers' decision to provide unique, stateful URLs for every possible view of the visualization. These stateful URLs allowed users to share, save, and refer back to specific visualization views on Twitter, in blog posts, and elsewhere.

The volume of discussion and participation generated by tools like the Ngram viewer illustrates the value of designing visualization and analysis tools such that they are compatible with existing social media practices. Because views of the visualization were easy to produce and share, users not only shared them extensively, but also began to collect interesting or topical Ngrams views using lightweight blogging platforms like Tumblr [109]. Others used these as the starting point for more detailed blog posts and explorations of particular aspects of the data relevant to particular disciplines. This echoes one of the key findings from Many-Eyes [27]—that web visualizations often work best when they serve as a “community component” that communities can readily adopt, repurpose, and use within the context of their existing discourse. *How can we streamline the process of designing visualizations that are easy to annotate, share and embed, and which play nicely with existing social media?*

Developers can easily add stateful linking to simple visualizations, but adding them to visualizations that support complex navigation and filtering often requires considerable effort. Developers may have a hard time deciding which pieces of information about the visualization state are important to maintain and which are not. Moreover, information needed to reproduce the state of the visualization is often spread between the visualization definition, interface widgets, and event handlers, making it difficult to manage. There are also many other aspects of the chart beyond the state that users and analysts may wish to refer to—for example, the underlying data, selections, even individual datapoints. Providing ways of linking directly to selections, data points, and other chart elements supports deixis [49] and may enable deeper discussion, but implementing these linking mechanisms per-visualization requires considerable effort on the part of developers.

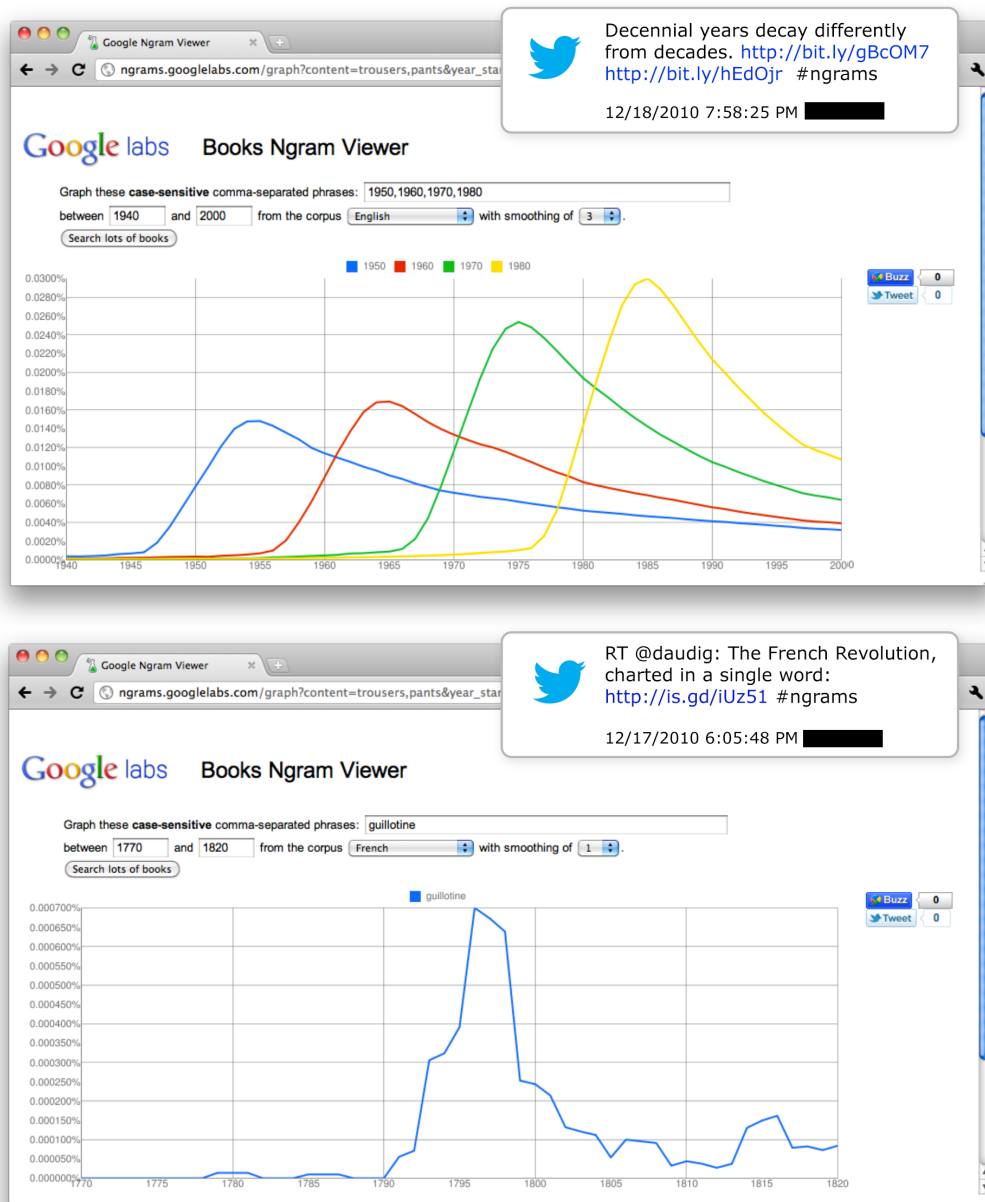


Figure 6.2: Several views of the Google Books Ngram viewer [42], an interactive visualization of the entire Google Books corpus, paired with Twitter messages discussing the views. The Ngram viewer displays the changing use of multi-word phrases in printed English over time. Using it, users noted and shared observations such as the sharp rise and gradual decay in references to decennial years (top) and linguistic artifacts like the pronounced usage of the word “Guillotine” during the period surrounding the French Revolution (bottom).

The lack of strong pointing and linking tools in most current web visualizations suggests a need for standards and best practices that make it easier for users to share, annotate, and build upon datasets and visualizations within the ecosystem of the web. Currently, D3 [11], Processing [85], and other frameworks provide easy platforms for developing visualizations. One promising approach may be to extend these frameworks to provide toolkit-level support for deep linking, annotation, sharing, and data provenance out-of-the-box. For example, it may be useful to provide libraries for serializing and deserializing visualization state and providing deep linking with less developer effort.

More research is also necessary to understand what sorts of pointing interactions have the most value during collaborative analysis and establish best practices and guidelines for supporting them. Visualization development environments that help designers and analysts construct new visualizations could also increase the consistency of sharing, pointing, and linking behaviors available across visualizations. Providing tools like these that make it easier for designers, journalists, and other end-users to add these capabilities to their visualizations could enable social discussions around data in a wider range of disciplines.

Another key challenge involves capturing the ad hoc discussion that occurs around visualizations on the web and extracting meaning from it. *How can we make it possible to find and collect comments about a visualization from the web and what can we do with these large sets of insights once we have them?*

One common strategy is to provide or suggest unique identifiers like hashtags, or shortened URLs that can be included in comments and social media posts. Comments with these identifiers can then be retrieved by searching the target networks and via platform APIs, where they exist. Marking and then searching for content this way works well for public services like Twitter. A related approach involves integrating tools for publishing, commenting, and sharing via social media directly into visualizations. Integrated sharing tools can make it easy for users to post visualization views to services like Facebook and Twitter in a standardized way, and can simultaneously log the views or comments that are published for later analysis. Figure 3.3 shows a custom sharing interface implemented within CommentSpace.

However, even if it is possible to capture large amounts of ad-hoc discussion and exploration of a dataset, it may be difficult to extract useful information from it. Ad hoc analyses like the one spawned by the Ngrams viewer can produce thousands of potential observations spread across as many different views of a dataset. Moreover, these observations may be much less organized than

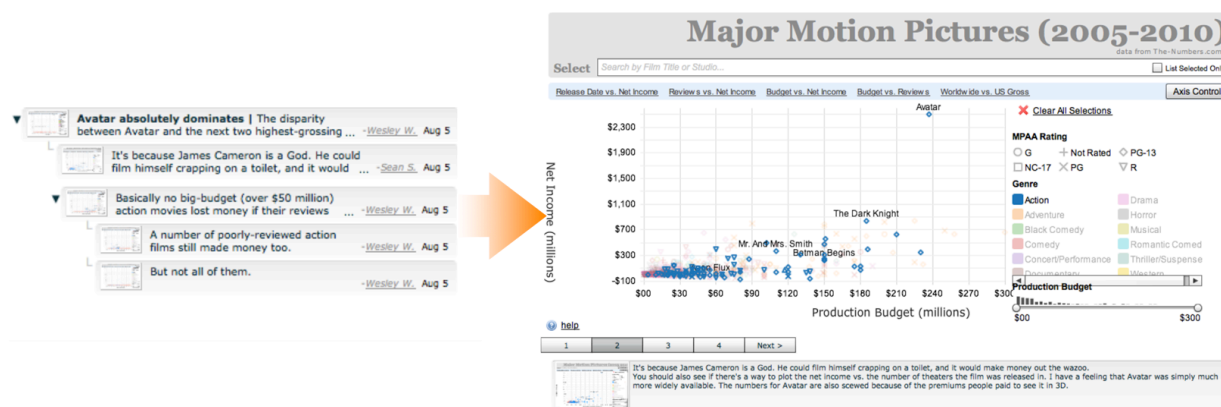


Figure 6.3: The CommentSpace slideshow extension allows authors to organize comments and visualization views to tell a story, then converts the sequence into an interactive slideshow. Here, a sequence of posts on a visualization comparing the cost of production against critic’s ratings for major motion pictures (left) becomes an interactive slideshow (right).

observations generated by our more systematic crowdsourcing framework. Such corpora call for new tools for mining discussions and performing meta-analysis to extract high-level concepts and identify the most useful and well-supported observations..

6.4 Visualization, Presentation, and Storytelling

Finally, social interactions around data, like most other human interactions, involve storytelling. In fact, most of the data visualizations presented on the web and in the media are constructed in the service of a particular story [91]. Storytelling serves a dual role in visual analysis. For the analyst, storytelling represents the final phase of sensemaking—where observations, hypotheses, and conclusions are synthesized into data-driven stories that communicate the results of the process. Simultaneously, storytelling provides a means for introducing new users to a dataset or topic and encouraging them to explore and engage with it. Yet few tools exist to support data-driven storytelling. *Can we provide systems that allow users to navigate, curate, and build stories out of the collective observations of many users? Can we make it easier to transform the products of social data analysis into reports, news stories, and presentations that allow them to communicate?*

We have begun to explore this problem in the context of CommentSpace by building an extension that provide more explicit support for storytelling. The extension provides tools that allow

users to organize sets of comments and visualization views and then convert them into interactive slideshows (Figure 6.3). Using the tool, authors can connect a sequence of views from one or more visualizations and pair each view with captions that explain and contextualize it. The tools also allows authors to control whether or not the interactive controls for each view should be enabled. By disabling interaction on early views that provide context and background, then enabling interaction on later views, authors can create “Martini Glass”-style narrative structures [91] that guide viewers gradually into a visualization.

However, this extension provides a very limited set of tools with which to author narratives based on analytic findings. Authors can connect views together, but have little control over the content or styling of views. Moreover, our tools don’t provide any unified mechanism for highlighting, emphasizing, and deemphasizing particular points in a visualization—selection, editing, and highlighting tools are limited to those provided in the visualizations. Because visualizations themselves cannot be edited or reformatted to fit the narrative, it can be difficult for authors to repurpose them to accentuate the important parts of the story or integrate them into a broader narrative. Future tools will need to offer more powerful publishing and sharing functionality that makes it easy for authors to create and present many different kinds of data-driven stories.

Social media reporting tools like Storify [97] and Storyful [98] allow users to create stories by collecting posts from Twitter and other social media streams and embedding them into news stories and blog posts. These tools take the process of gathering social media, selecting relevant content, and integrating it into a news story, and provide a streamlined interface that allows authors to locate content and quickly construct a story around it without writing code or dealing with layout. In doing so, these tools enable non-expert users to quickly build on pieces of social media content and them as core building blocks for news stories, reports, and presentations. Analogous tools for data visualization could allow analysts and journalists to gather comments, annotations, and other findings from data analysis and integrate them, along with tailored views of their analytic visualizations, into their stories and presentations.

Like social media reporting tools, data visualization reporting tools should allow authors to easily collect and organize multiple views of visualizations, social media posts and other content. Unlike social media content, however, visualizations designed for analysis tasks often are not well suited for public consumption. Instead, they tend to be visually complex and readers generally need considerable context and training to interpret them. As a result, a core challenge for data visualization reporting tools is to provide flexible end-user tools that allow authors and analysts

to tailor visualizations for public consumption. These tools must ease the process of annotating visualizations, selecting individual views, generating simplified and restyled representations of interactive graphics, and simplifying interaction to illustrate specific points.

Chapter 7

Conclusion

Human attention and domain knowledge are inherently finite. Therefore, we expect that single-user models of analysis will always limit analysts' ability to make sense of datasets that are large, complex, and span disciplinary boundaries. However—as we have noted throughout this thesis—the design of multi-user systems for data analysis is complex and nuanced. Designers and developers must tailor social data analysis tools to suit the interests and competencies of their various stakeholders, and each system often requires considerable tuning to produce the desired analytic results.

Over the course of this thesis we have explored several points in the design space of collaborative data analysis tools. By focusing on a few key user groups—small analysis teams, novice communities, and paid crowds—this work illustrates the range and diversity of useful approaches to social data analysis. Some analysis scenarios—for example, a journalist scouring a large public-interest dataset—may benefit greatly from the parallelization that crowdsourcing can bring to bear. Others—like analyses of small-scale environmental quality data—can benefit greatly from the local knowledge of community members, even if they lack analysis expertise. Teams of more expert analysts working in concert, meanwhile, can benefit from more robust techniques for organizing, discussing, and building on one another's findings.

The volume of data generated by governments, institutions, and individuals continues to grow unabated. As a result, the tools we use to explore data must continue to evolve. Advances in visualization, data mining, machine learning, and information retrieval will undoubtedly improve

the effectiveness of individual analysts. However, collaborative tools that multiply the impact of many stakeholders promise to compound these gains even further.

This dissertation and the three social data analysis tools presented herein point towards a future in which big data analysis tasks might engage not just one or two collaborators, but tens, hundreds, or millions. Moreover, the diversity of these systems suggest that future analysis tools will be anything but homogeneous. Rather, each new dataset or analytic problem brings with it new constraints and new stakeholders, but also new potential. This work offers just a few possible visions of these future tools, and suggests models for how we might pool our collective effort to tackle the next generation of big, data-driven problems.

Bibliography

- [1] S. Ahmad, A. Battle, Z. Malkani, and S. Kamvar. "The Jabberwocky Programming Environment for Structured Social Computing". In: *Proceedings of UIST*. 2011, pp. 53–64.
- [2] L. von Ahn and L. Dabbish. "Designing games with a purpose". In: *Communications of the ACM* 51.8 (2008), pp. 58–67.
- [3] E. Amigó, J. Gonzalo, J. Artilles, and F. Verdejo. "A comparison of extrinsic clustering evaluation metrics based on formal constraints". In: *Information Retrieval* 12.4 (July 2008), pp. 461–486.
- [4] P. M. Aoki, R. J. Honicky, A. Mainwaring, C. Myers, E. Paulos, S. Subramanian, and A. Woodruff. "A vehicle for research: using street sweepers to explore the landscape of environmental community action". In: *Proceedings of CHI*. ACM, 2009, pp. 375–384.
- [5] A. D. Balakrishnan, S. R. Fussell, and S. Kiesler. "Do Visualizations Improve Synchronous Remote Collaboration?" In: *Proceedings of CHI*. ACM, 2008, pp. 1227–1236.
- [6] Y Benkler. "Coase's Penguin, or, Linux and the Nature of the Firm." In: *Yale Law Journal* 112 (2002), pp. 369–456.
- [7] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. "Soylent: A Word Processor with a Crowd Inside". In: *Proceedings of UIST*. ACM, 2010, pp. 313–322.
- [8] D. Billman, G. Convertino, J. Shrager, J. Massar, and P. Pirolli. "Collaborative intelligence analysis with CACHE and its effects on information gathering and cognitive bias." In: *HCI Consortium Workshop* (Snow Mountain, CO,). 2008.
- [9] A. F. Blackwell et al. In: *Proceedings of the 4th International Conference on Cognitive Technology: Instruments of Mind*. London, UK: Springer-Verlag, 2001, pp. 325–341.

- [10] M. Bloch, S. Carter, J. Corum, A. Cox, and M. Ericson. *Jackson's Billboard Rankings Over Time - Interactive Timeline*. The New York Times (<http://www.nytimes.com/interactive/2009/06/25/arts/0625-jackson-graphic.html>). June 2009.
- [11] M. Bostock, V. Ogievetsky, and J. Heer. "D3: Data-Driven Documents". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2301–2309.
- [12] D. Boud. *Enhancing learning through self assessment*. London, UK: Kogan Page, 1995.
- [13] C. Brigham, E. Graham, S. Reddy, E. Yuen, and K. Mayoral. "Using smart phones and citizen scientists to map invasive species and track spread over time". In: *California Invasive Plant Council Symposium* (2009), pp. 28–30.
- [14] D. S. Buckingham Shum, A. Selvin, D. M. Sierhuis, D. J. Conklin, C. Haley, and P. B. Nuseibeh. "Hypermedia Support for Argumentation-Based Rationale: 15 Years on from gIBIS and QOC". In: *Rationale Management in Software Engineering*. Ed. by A. H. Dutoit, R. McCall, I. Mistrik, and B. Paech. Springer-Verlag, 2006, pp. 111–132.
- [15] J. A. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. "Participatory sensing". In: *Proceedings of Workshop on World-Sensor-Web: Mobile Device-Centric Sensor Networks and Applications*. 2006, pp. 117–134.
- [16] BuzzData. <http://buzzdata.com/>.
- [17] S. K. Card, J. D. Mackinlay, and B. Shneiderman, eds. *Readings in information visualization: using vision to think*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.
- [18] J. M. Carroll, M. B. Rosson, G. Convertino, and C. H. Ganoë. "Awareness and teamwork in computer-supported collaborations". In: *Interacting with Computers* 18.1 (2006), pp. 21–46.
- [19] D. Chandler and A. Kapelner. "Breaking monotony with meaning: Motivation in crowd-sourcing markets". In: *University of Chicago mimeo* (2010).
- [20] E. H. Chi and T. Mytkowicz. "Understanding the efficiency of social tagging systems using information theory". In: *Proceedings of Hypertext and Hypermedia*. ACM. 2008, pp. 81–88.
- [21] T. Chklovski, V. Ratnakar, and Y. Gil. "User interfaces with semi-formal representations: a study of designing argumentation structures". In: *Proceedings of IUI*. ACM. 2005, pp. 130–136.
- [22] H. H. Clark and S. E. Brennan. "Grounding in communication". In: *Perspectives on socially shared cognition* 13 (1991), pp. 127–149.

- [23] G. Convertino, D. Billman, P. Pirolli, J. Massar, and J. Shrager. "The CACHE Study: Group Effects in Computer-supported Collaborative Analysis". In: *Computer Supported Cooperative Work* 17.4 (2008), pp. 353–393.
- [24] J. Corburn. *Street Science: Community Knowledge and Environmental Health Justice*. Boston, MA: MIT Press, 2005.
- [25] B. da Costa. "Pigeonblog". In: *Net Works: Case Studies in Web Art and Design*. Ed. by X. Burrough. Routledge, 2012. Chap. 17, pp. 192–199.
- [26] D. Cuff, M. Hansen, and J. Kang. "Urban Sensing: Out of the Woods". In: *Communications of the ACM* 51.3 (2008), pp. 24–33.
- [27] C. M. Danis, F. B. Viégas, M. Wattenberg, and J. Kriss. "Your Place or Mine? Visualization as a Community Component". In: *Proceedings of CHI*. ACM. 2008, pp. 275–284.
- [28] Data360. <http://data360.org>.
- [29] DataMarket - "Hot or Not" for Data. <http://datamarket.com/featured/hot-or-not-data/>.
- [30] N. Diakopoulos, S. Goldenberg, and I. Essa. "Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists". In: *Proceedings of CHI*. ACM. 2009, pp. 799–808.
- [31] P. Dourish and V. Bellotti. "Awareness and coordination in shared workspaces". In: *Proceedings of CSCW*. ACM. 1992, pp. 107–114.
- [32] P. Dutta, P. M. Aoki, N. Kumar, A. Mainwaring, C. Myers, W. Willett, and A. Woodruff. "Common Sense: participatory urban sensing using a network of handheld air quality monitors". In: *Proceedings of SenSys*. ACM, 2009, pp. 349–350.
- [33] R. Eccles, T. Kapler, R. Harper, and W. Wright. "Stories in GeoTime". In: *Information Visualization* 7.1 (Dec. 2008), pp. 3–17.
- [34] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell. "BikeNet: A mobile sensing system for cyclist experience mapping". In: *Transactions on Sensor Networks* 6.1 (Dec. 2009), 6:1–6:39.
- [35] EPA EnviroMapper for Envirofacts. <http://www.epa.gov/emefdata/em4ef.home>.
- [36] K. Fisher, S. Counts, and A. Kittur. "Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users". In: *Proceedings of CHI*. ACM, 2012, pp. 247–256.

- [37] K. Fort, G. Adda, and K. B. Cohen. “Amazon Mechanical Turk: Gold Mine or Coal Mine?” In: *Computational Linguistics* 37.2 (2011), pp. 413–420.
- [38] L. Fortson, K. Masters, R. Nichol, K. Borne, E. Edmondson, C. Lintott, J. Raddick, K. Schawinski, and J. Wallin. “Galaxy Zoo: Morphological Classification and Citizen Science”. In: *Advances in Machine Learning and Data Mining for Astronomy*. Ed. by M. J. Way, K. M. Ali, and A. N. Srivastava. CRC Press, 2012. Chap. 11, pp. 213–236.
- [39] D. Gergle, R. E. Kraut, and S. R. Fussell. “Language efficiency and visual technology: Minimizing collaborative effort with visual information”. In: *Journal of Language and Social Psychology* 23.4 (Dec. 2004), pp. 491–517.
- [40] S. A. Golder and B. A. Huberman. “Usage patterns of collaborative tagging systems”. In: *Journal of Information Science* 32.2 (Apr. 2006), pp. 198–208.
- [41] Google Maps. <https://maps.google.com/>.
- [42] Google Ngram Viewer. <http://books.google.com/ngrams>.
- [43] Google Public Data Explorer. <http://www.google.com/publicdata/>.
- [44] T. F. Gordon and N. Karacapilidis. “The Zeno argumentation framework”. In: *Proceedings of the Conference on Artificial Intelligence and Law*. ACM, 1997, pp. 10–18.
- [45] A. Gupta, W. Thies, E. Cutrell, and R. Balakrishnan. “mClerk: enabling mobile crowdsourcing in developing regions”. In: *Proceedings of CHI*. ACM, 2012, pp. 1843–1852.
- [46] J. Heer and M. Agrawala. “Design Considerations for Collaborative Visual Analytics”. In: *Information Visualization* 7.1 (2008), pp. 49–62.
- [47] J. Heer and M. Bostock. “Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design”. In: *Proceedings of CHI*. ACM, 2010, pp. 203–212.
- [48] J. Heer, F. B. Viégas, and M. Wattenberg. “Voyagers and voyeurs: Supporting asynchronous collaborative visualization”. In: *Communications of the ACM* 52.1 (2009), pp. 87–97.
- [49] W. C. Hill and J. D. Hollan. “Deixis and the future of visualization excellence”. In: *Proceedings of Visualization*. IEEE, 1991, pp. 314–320.
- [50] History & Objectives Christmas Bird Count. <http://web4.audubon.org/bird/cbc/history.html>.
- [51] J. J. Horton and R. J. Zeckhauser. “Algorithmic Wage Negotiations: Applications to Paid Crowdsourcing”. In: *CrowdConf*. 2010.

- [52] J.J. Horton and L. B. Chilton. "The Labor Economics of Paid Crowdsourcing". In: *Electronic Commerce*. ACM, 2010, pp. 209–218.
- [53] J. Hullman, E. Adar, and P. Shah. "The Impact of Social Information on Visual Judgments". In: *Proceedings of CHI*. ACM, 2011, pp. 1461–1470.
- [54] C. B. Hurley and R. W. Oldford. "Pairwise Display of High-Dimensional Information via Eulerian Tours and Hamiltonian Decompositions". In: *Journal of Computational and Graphical Statistics* 19.4 (2010), pp. 861–886.
- [55] P. Iperiotis. "Demographics of mechanical turk". In: *New York University, Tech. Rep* (2010).
- [56] *Kaggle*. <http://www.kaggle.com>.
- [57] E. Kanjo, S. Benford, M. Paxton, A. Chamberlain, D. S. Fraser, D. Woodgate, D. Crellin, and A. Woolard. "MobGeoSen: facilitating personal geosensor data collection and visualization using mobile phones". In: *Personal Ubiquitous Computing* 12.8 (2008), pp. 599–607.
- [58] S. Kim, C. Robson, T. Zimmerman, J. Pierce, and E. M. Haber. "Creek watch: pairing usefulness and usability for successful citizen science". In: *Proceedings of CHI*. ACM, 2011, pp. 2125–2134.
- [59] A. Kittur, E. H. Chi, and B. Suh. "Crowdsourcing user studies with Mechanical Turk". In: *Proceedings of CHI*. ACM, 2008, pp. 453–456.
- [60] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut. "CrowdForge: crowdsourcing complex work". In: *Proceedings of UIST*. ACM, 2011, pp. 43–52.
- [61] N. Kong and M. Agrawala. "Perceptual interpretation of ink annotations on line charts". In: *Proceedings of UIST*. ACM, 2009, pp. 233–236.
- [62] N. Kong, J. Heer, and M. Agrawala. "Perceptual Guidelines for Creating Rectangular Treemaps". In: *Transactions on Visualization and Computer Graphics* 16 (2010), pp. 990–998.
- [63] A. Kulkarni, P. Gutheim, P. Narula, D. Rolnitzky, T. Parikh, and B. Hartmann. "MobileWorks: Designing for Quality in a Managed Crowdsourcing Architecture". In: *Internet Computing* 16.5 (2012).
- [64] J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, 1991.
- [65] E.-K. Lee, D. Cook, S. Klinke, and T. Lumley. "Projection pursuit for exploratory supervised classification". In: *Journal of Computational and Graphical Statistics* 14.4 (2005), pp. 831–846.

- [66] J. I. Levy, E. A. Houseman, J. D. Spengler, P. Loh, and L. Ryan. "Fine particulate matter and polycyclic aromatic hydrocarbon concentration patterns in Roxbury, Massachusetts: a community-based GIS analysis." In: *Environmental Health Perspectives* 109.4 (2001), pp. 341–347.
- [67] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. "Turkit: human computation algorithms on mechanical turk". In: *Proceedings of UIST*. ACM. 2010, pp. 57–66.
- [68] K. Luther, S. Counts, K. B. Stecher, A. Hoff, and P. Johns. "Pathfinder: An Online Collaboration Environment for Citizen Scientists". In: *Proceedings of CHI*. ACM. 2009, pp. 239–248.
- [69] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [70] W. Mason and D. J. Watts. "Financial incentives and the "performance of crowds"". In: *Proceedings of HComp*. AAAI. 2009, pp. 77–85.
- [71] M. McKeon. "Harnessing the Web Information Ecosystem with Wiki-based Visualization Dashboards". In: *Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1081–1088.
- [72] M. Minkler and N. Wallerstein, eds. *Community-Based Participatory Research for Health: From Process to Outcomes*. Jossey-Bass, 2011.
- [73] D. Mintz. *Technical Assistance Document for the Reporting of Daily Air Quality - the Air Quality Index (AQI)*. Tech. rep. US Environmental Protection Agency, Office of Air Quality Planning and Standards, 2009.
- [74] *Monitoring and Assessing Water Quality - Volunteer Monitoring*.
<http://water.epa.gov/type/rsl/monitoring>.
- [75] J. Nielsen. "Participation Inequality: Encouraging More Users to Contribute". In: *Jakob Nielsen's Alertbox* (Oct. 2006).
- [76] G. Niemeyer, A. Garcia, and R. Naima. "Black cloud: patterns towards da future". In: *Proceedings of Multimedia*. ACM, 2009, pp. 1073–1082.
- [77] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald. "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing". In: *Proceedings of HComp*. AAAI. 2011, pp. 43–48.
- [78] D. O'Rourke and G. P. Macey. "Community environmental policing: Assessing new strategies of public participation in environmental regulation". In: *Journal of Policy Analysis and Management* 22.3 (2003), pp. 384–414.

- [79] E. Paulos, R. J. Honicky, and B. Hooker. "Citizen Science: Enabling Participatory Urbanism". In: *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Ed. by M. Foth. IGI Global, 2009. Chap. 28, pp. 414–436.
- [80] A. Perer and B. Shneiderman. "Systematic yet flexible discovery: guiding domain experts through exploratory data analysis". In: *Proceedings of IUI*. ACM. 2008, pp. 109–118.
- [81] N. J. Pioch and J. O. Everett. "POLESTAR: collaborative knowledge management and sense-making tools for intelligence analysts". In: *Proceedings of CIKM*. ACM. 2006, pp. 513–521.
- [82] P. Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, USA, 2007.
- [83] P. Pirolli and S. Card. "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis". In: *Proceedings of International Conference on Intelligence Analysis 5* (2005).
- [84] Preemptive Media. *AIR – Area's Immediate Reading*. <http://www.pm-air.net/>.
- [85] *Processing.js*. <http://processingjs.org>.
- [86] A. J. Quinn and B. B. Bederson. "Human Computation: A Survey and Taxonomy of a Growing Field". In: *Proceedings of CHI*. ACM. 2011, pp. 1403–1412.
- [87] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin, and M. Srivastava. "Biketastic: sensing and mapping for better biking". In: *Proceedings of CHI*. ACM, 2010, pp. 1817–1820.
- [88] D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. "The cost structure of sensemaking". In: *Proceedings of CHI*. ACM. 1993, pp. 269–276.
- [89] J. M. Rzeszotarski and A. Kittur. "Instrumenting the crowd: using implicit behavioral measures to predict task performance". In: *Proceedings of UIST*. ACM. 2011, pp. 13–22.
- [90] *Same Origin Policy*. http://www.w3.org/Security/wiki/Same-Origin_Policy.
- [91] E. Segel and J. Heer. "Narrative Visualization: Telling Stories with Data". In: *Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 1139–1148.
- [92] A. D. Shaw, J. J. Horton, and D. L. Chen. "Designing Incentives for Inexpert Human Raters". In: *Proceedings of CSCW*. ACM. 2011, pp. 275–284.
- [93] V. S. Sheng, F. Provost, and P. G. Ipeirotis. "Get another label? improving data quality and data mining using multiple, noisy labelers". In: *Proceedings of KDD*. ACM, 2008, pp. 614–622.

- [94] *Smartsheet*. <http://www.smartsheet.com>.
- [95] A. Sorokin and D. Forsyth. "Utility data annotation with Amazon Mechanical Turk". In: *Computer Vision and Pattern Recognition Workshops*. IEEE. 2008, pp. 1–8.
- [96] B. Stone, S. Dennis, and P. J. Kwantes. "Comparing Methods for Single Paragraph Similarity Analysis". In: *Topics in Cognitive Science* 3.1 (2010), pp. 92–122.
- [97] *Storify*. <http://www.storify.com>.
- [98] *Storyful*. <http://www.storyful.com>.
- [99] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, and D Fink. "eBird: A citizen-based bird observation network in the biological sciences". In: *Biological Conservation* 142.10 (Oct. 2009), pp. 2282–2292.
- [100] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.
- [101] *Swivel*. <http://www.swivel.com>.
- [102] *Tableau*. <http://www.tableausoftware.com>.
- [103] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. "Adaptively Learning the Crowd Kernel". In: *arXiv.org* (2011).
- [104] J. J. Thomas and K. A. Cook. "Illuminating the path: The research and development agenda for visual analytics". In: *IEEE Computer Society* (2005).
- [105] *TIBCO Spotfire Decision Site*. <http://spotfire.tibco.com>.
- [106] R. Tibshirani, G. Walther, and T. Hastie. "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.
- [107] B. Trushkowsky, T. Kraska, M. J. Franklin, and P. Sarkar. "Getting It All from the Crowd". In: *arXiv.org* (2012).
- [108] E. Tufte. *Envisioning information*. Cheshire, CT, USA: Graphics Press, 1990.
- [109] *Tumblr*. <http://www.tumblr.com>.
- [110] *Verifiable.com*. <http://www.verifiable.com>.
- [111] F. B. Viégas, M. Wattenberg, F. v. Ham, J. Kriss, and M. McKeon. "Many Eyes: A Site for Visualization at Internet Scale ". In: *Trasactions on Visualization and Computer Graphics* 13.6 (2007), pp. 1121–1128.

- [112] F. B. Viégas, M. Wattenberg, M. McKeon, F. V. Ham, and J. Kriss. “Harry potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools”. In: *Proceedings of HICSS*. 2008.
- [113] *Visual.ly*. <http://visual.ly/>.
- [114] M. Wattenberg, F. Viégas, and K. Hollenbach. “Visualizing activity on wikipedia with chromograms”. In: *INTERACT* (2007), pp. 272–287.
- [115] A. J. Westphal, A. L. Butterworth, C. J. Snead, N. Craig, D. Anderson, S. M. Jones, D. E. Brownle, R. Fransworth, and M. E. Zolensky. “Stardust@home: A Massively Distributed Public Search for Interstellar Dust in the Stardust Interstellar Dust Collector”. In: *Lunar and Planetary Science* (2005).
- [116] *What’s Your College Degree Worth?* Bloomberg Businessweek interactive table. June 2010.
- [117] S. Whittaker, L. Terveen, W. Hill, and L. Cherny. “The dynamics of mass interaction”. In: *Proceedings of CSCW*. ACM. 1998, pp. 257–264.
- [118] *Wikipedia:Template messages - Wikipedia, the free encyclopedia*. http://en.wikipedia.org/wiki/Wikipedia:Template_messages.
- [119] W. Willett, P. Aoki, N. Kumar, S. Subramanian, and A. Woodruff. “Common Sense Community: Scaffolding Mobile Sensing and Analysis for Novice Users”. In: *Proceedings of Pervasive*. 2010, pp. 301–318.
- [120] W. Willett, J. Heer, and M. Agrawala. “Strategies for crowdsourcing social data analysis”. In: *Proceedings of CHI*. ACM, 2012, pp. 227–236.
- [121] W. Willett, J. Heer, J. M. Hellerstein, and M. Agrawala. “CommentSpace: Structured Support for Collaborative Visual Analysis”. In: *Proceedings of CHI*. ACM, 2011.
- [122] G. Wills and L. Wilkinson. “AutoVis: automatic visualization”. In: *Information Visualization* 9 (Mar. 2010), pp. 47–69.
- [123] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. “The Sandbox for analysis: concepts and methods”. In: *Proceedings of CHI*. ACM. 2006, pp. 801–810.
- [124] M. Wu and A. Marian. “Corroborating answers from multiple web sources”. In: *Proceedings of WebDB*. 2007.
- [125] J. Yi, R. Jin, A. K. Jain, and S. Jain. “Crowdclustering with Sparse Pairwise Labels: A Matrix Completion Approach”. In: *Workshops at the Conference on Artificial Intelligence*. AAAI. 2012.