A Cover Sheet

Please replace this page with the cover sheet.

Title: Data Intensive Computing: Scalable, Social Data Analysis

PI: Maneesh Agrawala, Associate Professor
Phone: 510-643-8220
E-Mail: maneesh@eecs.berkeley.edu

Co-PI: Jeffrey Heer, Assistant Professor
E-Mail: jheer@cs.stanford.edu

Co-PI: Joseph M. Hellerstein, Professor
E-Mail: hellerstein@cs.berkeley.edu

Contents

List of Figures

C Project Summary

Data Intensive Computing: Scalable, Social Data Analysis

Maneesh Agrawala*, Jeffrey Heer†, Joseph M. Hellerstein*
*University of California, Berkeley and †Stanford University

Analysts in all areas of human knowledge, from science and engineering to economics, social science and journalism are drowning in data. As a result we must rethink how we design the tools and techniques for exploring, analyzing and communicating data in a manner that scales as both the data and the organizations analyzing it grow in size. Throughout the data lifecycle, sensemaking is often a collaborative process. As different analysts each contribute to data acquisition, cleaning, analysis, and interpretation they contribute contextual knowledge that deepens understanding. At times may disagree on how to interpret data, but then work together to reach consensus. Many data sets are so large that thorough exploration by a single person is unlikely. In short, social cognition plays a critical role in the process of scalable data analysis. We believe that new analysis tools that address human cognitive characteristics, social interaction and data analytics in an integrated fashion can improve our ability to turn data into knowledge.

Our hypothesis is that scalable data analysis requires social interaction and therefore social context must be embedded in data analysis tools. The goals of this research are (1) to understand how social interaction and an understanding of social context can facilitate successful data analysis, (2) to develop models and tools for representing and annotating data transformations, visualizations, and social activity (e.g., textual and graphical annotations, discussions, links, tags), and (3) to design and test visual interfaces that leverage our tools to support collaborative analysis practices, including data entry, transformation, visualization, and interpretation. Central concerns of our research include (a) a focus on enabling social interaction throughout the data life-cycle and (b) the use of scalable data transformation routines that can return results in a time frame concordant with interactive, exploratory data transformation and analysis.

Intellectual Merit: This research will improve our understanding of the effects of social interaction and social context in group sensemaking and inform the development of new data analysis tools. Our tools will serve as a petri dish for exploring social collaboration processes throughout the life-cycle of massive data sets, enabling us to identify effective processes and interfaces for sensemaking and to gauge the effects of integrating contextual social metadata into analysis tools. We will evaluate the effectiveness of our systems through a combination of system benchmarks and both quantitative and qualitative user studies. By synthesizing methods from the database, visualization, and human-computer interaction research communities, our work aims to promote cross-fertilization across sub-disciplines of computer science to address shared research problems.

Broader Impacts: This project should inform and improve data analysis practices in many fields of critical importance to society, including business, intelligence, science, and public policy. Our results will provide new models, scalable analysis methods, and visual interfaces for analysts in these areas. We believe that expanding analysis tools to incorporate social interaction will improve both the scope and quality of analytic results, while also providing useful tools for analyzing a variety of social communication networks.

Keywords: visualization, scalable-analytics, social computing, social networks, online aggregation, provenance

D Project Description

1 Introduction

> "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, ... [b]ecause now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it..."
>
> *Hal Varian, Google's Chief Economist [65]*

Analysts in all areas of human knowledge, from science and engineering to economics, social science and journalism are drowning in data. New technologies for sensing, simulation, and communication are helping people to both collect and produce data at exponential rates [29,30,49]. As a result we must rethink how we design the tools and techniques for exploring, analyzing and communicating this abundance of data.

In order to produce real value from data, we must make sense of it. Such *sensemaking* – turning data sets into knowledge – is a basic motivation for database and data mining research. In addition to the systems, algorithms and statistics, sensemaking is a fundamental challenge in human-computer interaction. It requires integrating large-scale data storage, access, and analysis tools with contextualized human judgments about the meaning and significance of patterns in the data.

Sensemaking is a process that needs to scale across both data and organizations. It is typically a collaborative process that occurs throughout the data lifecycle, as different people contribute to data acquisition, cleaning, analysis, and interpretation. Studies of information workers [56,58] have demonstrated that sensemaking is highly iterative, as an insight gained from a visualization may suggest the need for additional corroborating data or highlight a data cleaning error. Analysts may disagree on how to interpret data, and then work together to reach consensus. Many data sets are large and multifaceted, and thorough exploration by a single person or a single algorithm is unlikely. As a result, recent attention has focused on the critical role of social cognition in the process of data analysis [63], investigating tools for sharing, annotation, and group deliberation of visualized data [7,9,35,44,61,70].

Moreover, social interaction depends on an understanding of social context, including the skills, inclinations, past actions, and relationships among collaborators [14,18,26]. The distributed and often asynchronous nature of computer-mediated collaboration, as well as the sheer volume of data and people at hand, complicates the process of making sense of the social environment. We believe that new tools that address human cognitive capabilities, social interaction and data analytics in an integrated fashion are crucial for improving our ability to turn data into knowledge.

Our hypothesis is that scalable data analysis requires social interaction, and therefore social context must be embedded in data analysis tools. The goals of this research are (1) to understand how social interaction and an understanding of social context can facilitate successful data analysis, (2) to develop models and tools for representing and annotating data transformations, visualizations, and social activity (e.g., textual discussion, graphical annotation, links, tags), and (3) to design and test visual interfaces that leverage our tools to support collaborative analysis practices, including data cleaning, transformation, visualization, and interpretation. Central concerns of our research include (a) a focus on enabling social interaction throughout the data lifecycle and (b) the use of scalable analysis routines that can return results in a time frame concordant with interactive, exploratory data transformation and analysis.

2 Motivation and Objectives

Our goal of scalable, social data analysis tools is motivated by our prior research on collaborative analysis environments, social network visualization, and interactive analysis of large data sets.

2.1 Social Data Analysis

To explore the potential of incorporating social interaction with visual analysis, we built Sense.us, a web application for collaborative sensemaking of 150 years of United States census data [35]. Sense.us integrates visualizations of demographic data with features for collective analysis (Figure 1). Users can attach commentary and annotations to views, share collections of views, and engage in discussion. Novel bookmarking and indexing features facilitate view sharing and reduce cross-talk between related visualization states.

We studied usage of the system through a live deployment and a series of laboratory studies, and conducted a *content analysis* [46] of recorded usage. We found that users often combined their knowledge in cycles of observation and hypothesis to make sense of trends in the data. For example, one observer noted a decline in the number of dentists in the labor force. Other people then hypothesized possible explanations, including fluoridation of the water supply and an increasing stratification between dentists and hygienists over the last century. In other cases, users explored topics such as changing gender roles, the impact of technology on the job market, and correlations among the wax and wane of occupations (Figure 1). We observed that social features helped mobilize users in the process of identifying interesting trends and generating hypotheses, and that exposing social activity regularly catalyzed new explorations by collaborators.

Based on these observations we have designed mechanisms to help analysts effectively allocate their attention. Scented widgets [75] are user interface controls with embedded visualizations that depict the visitation and comment counts for visualization views reachable from the current application state. In a controlled experiment we found that such cues can simultaneously promote visits to popular or controversial views and, by revealing under-visited regions of the data, increase the number of unique discoveries made by users.

Our research presaged a flowering of data sharing and visualization sites on the web, including IBM's Many-Eyes.com [70], Swivel.com, Data360.org, Google Fusion Tables and commercial products such as Tableau Server. These services enable users to upload data sets, visualize them using a palette of visualization types, and attach comments to data sets and visualization states.

Though initial experiences suggest the potential of social data analysis, Sense.us and subsequent tools lack many features that we believe are essential for fully realizing the value of social sensemaking. Within Sense.us and Many-Eyes.com, comments are associated with a single visualization view, hampering the ability to discuss multiple views and datasets. Yet, real-world analysis often requires bringing together in-

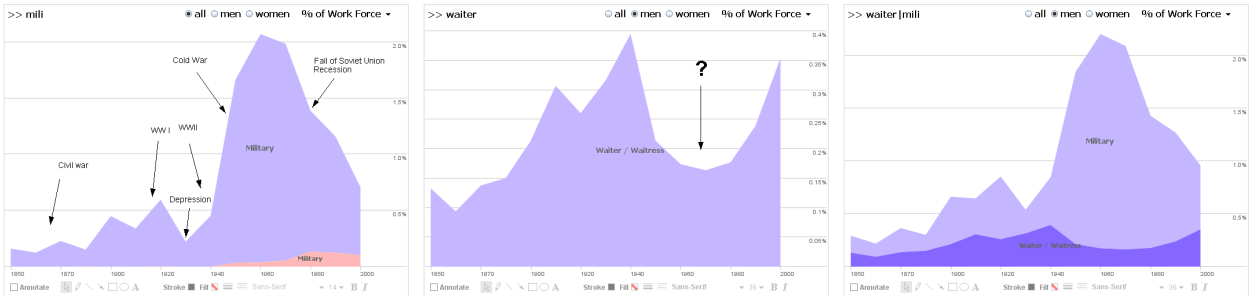

Figure 1: Collaborative sensemaking in Sense.us. (a) Multiple users debated the causes of military build-up while (b) another noted a drop and subsequent rise in male waiters in the late 1900's. (c) A different participant sought to connect the two, noting an inverse correlation between the occupations.

sights from multiple data sources or analysis methods [56, 63]. Similarly, annotations in Sense.us live in pixel space; one can not annotate the underlying data, limiting the scale of analysis by inhibiting computational re-use and analysis of annotation patterns. More *flexible data and interaction models for commenting and annotation are needed* to facilitate analyses spanning multiple data sets and visualizations.

Our studies of Sense.us found that roughly 16% of social commentary involved data integrity issues [35], as users posted comments to aid interpretation of missing data, problematic outliers, changing schemas, and data integration errors. Yet, existing social data analysis tools do not conform to the iterative nature of sensemaking. To correct an error in the data one must leave the system, correct the data manually, upload a new data set, and begin anew – often losing the connection between the prior analysis and the resulting analysis on transformed data. *Richer tools that enable interactive data transformation and record data provenance could empower analysts to collaboratively and opportunistically clean data as they go about their work*, keeping track of insights and changes across iterations.

Research Goal 1 *In this grant, we will develop improved tools and techniques that support social data analysis across the lifecycle of data, scaling over numerous collaborators and massive data sets.*

2.2 Interacting with Big Data

Most current social data analysis tools handle modest-sized data sets of at most a few thousand tuples. In areas such as business intelligence and scientific research, analysts need interactive analysis tools for multi-terabyte or even petabyte-scale data sets. One example we have worked with is a half-petabyte data warehouse at FOX Advertising Network comprising data on advertising, user behavior, and customer relationships. For a database of this scale, current social data analysis tools are limited to visualizing only the outputs of long chains of analysis [20]. But visualization and collaborative analysis can assist at every step of these analysis chains – especially those that are data-intensive. To this end, we believe that collaborative visualization and analysis tools need to handle *massive data sets at human interaction speeds*. In addition, there should be facilities to support *collaborative refinement of rich data processing pipelines* that span the data lifecycle from acquisition through transformation, cleaning, analysis and presentation.

Data analysis is fundamentally an iterative process in which analysts invoke tasks, receive responses, formulate the next tasks based on responses, and repeat. Interactive manipulation of the data with immediate feedback is essential for analysts to efficiently make sense of the data [40]. Ironically, as access to data has exploded in recent years, there has been a retreat to mainframe-style batch processing of Big Data that stymies iterative analysis. For example, Google's MapReduce framework (cloned in open source as Hadoop) produces no output for a job until all input data has been consumed. Long-running batch operations offer the most primitive possible user interaction, which frustrates the naturally iterative process of data analysis.

In earlier work we tackled the problem of providing interactive control and immediate feedback in several settings ranging from SQL queries to data cleaning tools to spreadsheet and map visualizations [36]. For example, our Online Aggregation system supports interactive data analysis through SQL queries. (Figure 2). Using continuous sampling methods, this system immediately computes a coarse statistical estimate in response to a query and visualizes the result as a dot plot with a confidence interval per dot. Over time, increasing sample size enables progressively refined estimates, which are animated in the visualization to show the refinement process. Initially the red dots oscillate in large steps as the estimates are rough, but as the query runs, the oscillations become tighter, visually indicating the rate of reduction in uncertainty. In this case, the early estimates shown in the figure reveals that that colleges `D` and `R` are almost certainly outliers and merit further study. Underlying the interface is an *integrated combination of data sampling, sample-aware query processing algorithms, and statistical estimators* for the aggregates and confidences.

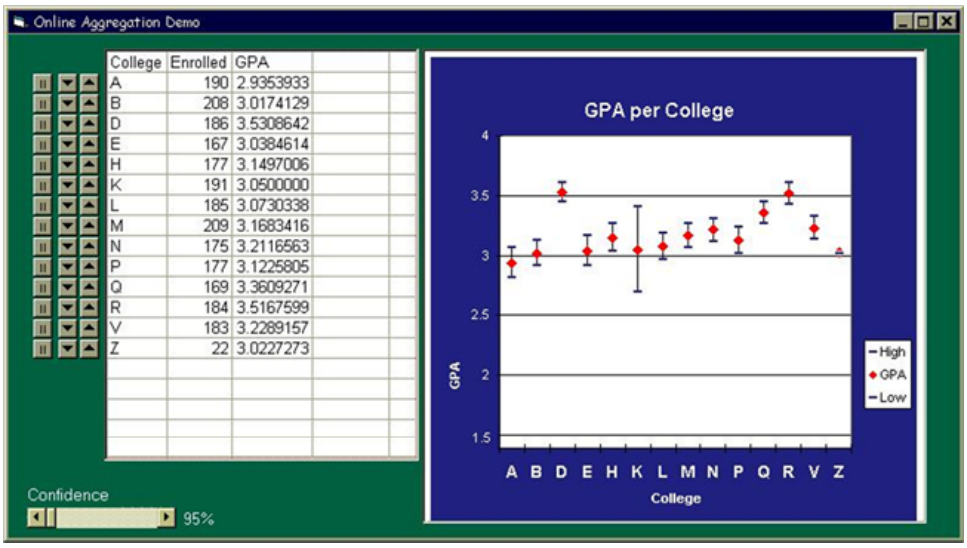

Figure 2: Online aggregation interface. The query requests a breakdown of average GPAs by college. Each dot in the plot represents the approximate average GPA and the confidence bars represent uncertainty in the estimates of the averages. For example, the the average GPA in College A is within 0.2 points of 2.935 with 95% probability. The query progressively refines the estimate and continuously updates the visualization to show the tightening estimates.

Although Online Aggregation has received renewed attention in recent years [24, 37, 42], SQL-style aggregates are only one example of the kind of long-running interactions that need to handle big data – especially in collaborative settings. We plan to extend the interactive approaches to more complex user interactions and statistical methods, as envisioned by algorithmic work on Anytime Algorithms in AI [79]. Beyond the scale of individual algorithms, the full lifecycle of data in an organization involves iteration and communication across tasks and groups. Approximation techniques used to make individual analyses interactive raise higher-level questions: how will users use a partially-completed analysis to discuss data and data products?, how will they reason and communicate about uncertainty?, and what tools do they need to refine, alter and continue discussing analyses along multiple steps of an analytic chain?

Research Goal 2 *In this grant we will develop tools and techniques that enable a scalable interplay of interaction, discussion, and analysis of Big Data, both within individual algorithmic steps, and in the larger context of full data lifecycles.*

2.3 Surfacing Social Context and Activity

Analysis tools must also scale in terms of people. Such scaling is especially important for large databases, which are rarely analyzed by only one person with only one goal. As statistician John Tukey noted [64], "There is often more analysis than there was data." We posit that *it is equally important for analysts to not only make sense of the data, but also make sense of the analysis process itself.* Moreover the *analysts must be aware of the social network and context brought to the process by their collaborators*.

The FOX warehouse (Section 2.2) is shared by many stakeholders with different skills and goals: statisticians designing sophisticated analysis algorithms, sales and marketing professionals looking for intuition and justification for decision-making, and executive staff producing reports for corporate leadership. The statisticians design algorithms to clean and summarize the raw data, and thereby generate *data products* for the analysts in the other groups. These analysts in turn provide feedback and requests based on their own data exploration and understanding [20]. We argue that social data analysis tools should help coordinate

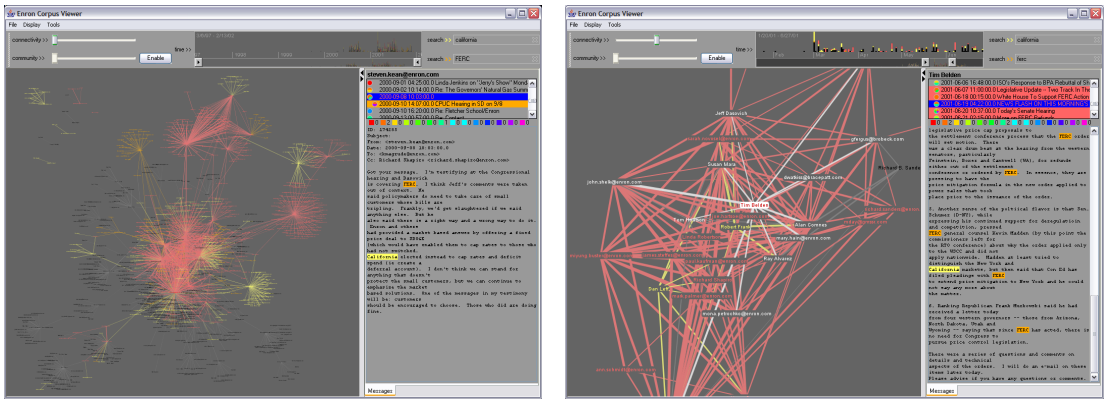

Figure 3: Enron e-mail corpus viewer, consisting of a social network view, timeline, and inbox inspector. (a) An overview of e-mail exchanges showing search hits for the terms "California" (yellow), "FERC" (orange), and their intersection (red). (b) A zoomed-in view showing highly-connected actors.

such analysis efforts, by *assisting in the allocation of tasks*, *facilitating awareness of others' actions*, and *managing visibility* to ensure that people see the data and context that is appropriate to their tasks.

Successful collaborations often depend on understanding the social context in which the work is conducted [14, 18, 26]. Analysts review past activity and roles of other analysts to assess the earlier findings [52]. How regularly do other participants post on similar topics, and how often do they engage in argument? Within an organizational context, is another actor a boss, a co-worker, a (possibly internal) customer of the data, or a consultant? Where are they geographically located? There are many such questions one might ask, and the appropriate questions depend on the context of the particular analysis task.

In previous work, we have built visual tools for exploring social communication patterns in online social networks [34] and in e-mail archives [32]. For example, we developed a tool to analyze the network structure and textual content of the 1.5 million internal emails of the Enron corporation (Figure 3). Using this tool we quickly discovered an anomaly in the communication patterns to a high-level Enron employee named Tim Belden. Belden received all the reports on Congressional meetings concerning Enron, but unlike other executives, never replied to these reports. Running a Google search on Belden we learned that the government had investigated him and determined that Belden was the "mastermind" behind the manipulation of California's markets. He was found guilty on charges of conspiracy. By tracing the social communication structure we were able to observe discrepancies and quickly hone in on one actor of interest among thousands.

Our aim is to extract and utilize such communication structures that naturally arise within a collaborative analysis of any dataset to further the analysis of that data. As a team of analysts generates observations, hypotheses and evidentiary structures, meta-analysis of the analysts' communication behavior and past actions may be used (a) to establish social context to better interpret others' contributions and diagnose potential issues, (b) to determine what entities or datasets have received the most attention—and conversely, which have been relatively neglected—and (c) to connect the dots between analysts doing related work but are unaware of each other. We posit that *incorporating social context and activity as full-fledged data sets amenable to analysis in their own right will improve the quality and coverage of collaborative analysis*.

Research Goal 3 *In this grant we will develop general tools for representing and analyzing the social communication, structure and context that both supports and is created by sensemaking activity.*

3 Data Model for Scalable Collaborative Analysis

In any discussion, the participants must share a common ground [18] of objects and concepts being discussed. In our context of collaboration, the mental models underlying the data analysis and sensemaking tasks need to be captured explicitly in software and reflected back unambiguously to other participants.

Humans have evolved a rich ability to verbally identify items or concepts in conversation by *naming* them and by referring to or *deictically* pointing to these entities via physical gestures [8, 17]. Both naming and deictic reference (pointing) offer important design challenges for visualization [38]. Participants must be able to unambiguously point to visual elements ("that bump in the curve"), as well as name or identify the data that the elements represent ("the stock price at Friday's closing bell").

A key contribution of our work will be the development of a *social analysis model* (SAM) that is designed to handle such naming and deictic reference issues. We are developing the SAM as a concrete software model aside from that of the data, which captures the entities and relationships involved in the full data lifecycle in an organization. The goal of our SAM is not to innovate in data models but rather to use traditional models such as Relational or RDF to catalog various aspects of the data lifecycle including the data and data products, the people and processes involved in manipulating that data, and the computational and social interactions between the data, processes and people. We consider two examples of the types of interactions and analyses our SAM will enable:

1. Capturing social interactions in the SAM: Earlier collaborative visualization systems such as ManyEyes.com and our own work on Sense.us treated comments, annotations and users' interactions as completely separate entities from the data being visualized. The techniques used to render the data visualizations could not be used to render this user-generated data. This approach hinders analysts from using the system to analyze their own analysis behavior. For example, to perform social network analysis of comments analysts would have to first extract the comments from the system via screen-scraping or access to system internals, and reload them into the site as first-class data. The lack of flexibility in analyzing the analysis process undercuts a core goal of social data analysis – to take advantage of the digitization of social interaction. *User communication and analytic activity is all mediated by software, and an analytic system should use its trace of that activity to enrich the data analysis process.* Our SAM will *reify* this social data as first-class data objects with explicit representation of the analysts and their roles, their interactions with the data, and their interactions with each other.

2. Capturing visual elements and mappings in the SAM: Underlying every data visualization system is an implicit analysis model that captures the data values being visualized, the visual marks used for display, and the mappings between the two [13]. There is a body of work that captures these visual elements and mappings in a formal language [5, 19, 48, 50, 62, 74], but that work has not been well integrated either with social data analysis or with data transformation and query languages. *A key piece of this proposal is to capture data, visual elements and the behaviors of code and people that generate and manipulate those elements.* Each visual element will have a unique identity, traceable via the SAM back through the visual and data-centric transformations to the raw input data that generated it. Our goal is to create interactive, graphical data annotation techniques that extend prior visualization work on brushing [3, 33, 47, 51] and interactive querying [1, 23, 28, 39, 54] to support annotation of specific data elements, logical data regions, and data transformations. When users leave comments on a graph in this framework, they will form an unambiguous relationship between a piece of text (their comment) and a specific data product. This relationship can then be rendered and analyzed in a host of ways that go beyond the original context in which it was first generated.

To realize these various goals, **we will design a unified social analysis model (SAM) for data, transfor-**

mation, visualization and collaboration. This SAM will enable expressive, unambiguous reference and naming of semantic entities including raw data, derived data products, social actors and the relationships between the above entities. We will realize this SAM in an open source implementation based in a reusable data-centric software framework such as Ruby on Rails.

4 Surfacing Social Context

A critical dimension of the common ground [18] underlying successful collaboration is an understanding of social context, including the skills, inclinations, past actions, and relationships among collaborators. For instance, *awareness* of the activities of others helps collaborators gauge what work has been done and where to allocate effort next [14,26]. While we and others have demonstrated benefits for collaborative exploratory data analysis [35, 69, 75], research has also shown that communication overhead and biases such as groupthink and cognitive tunneling can potentially degrade the performance of collaborative teams [2, 7]. Thus careful design and study is required when incorporating social interaction into analysis tools.

As discussed in the previous section, our *social analysis model* (SAM) will capture the artifacts, actors, and actions involved in a collaborative analysis effort. It will provide an infrastructure for experimenting with various means of surfacing social context in order to scale collaboration. We will explore the inclusion of social context within analysis tools through both explicit meta-analysis of comments, annotations and other social activity data as well as selective presentation of social activity cues within the data analysis interface.

1. Visualizing Social Structures: Because our proposed model treats social activity as first-class data, it will *enable analysts to visualize and explore social activity just like another other data set* within the system. To facilitate such exploration, **we will include data transforms** (e.g., social network extraction, text entity extraction), **analysis routines** (e.g., betweenness centrality, linkage-based clustering, activity and diversity metrics), **and visualization methods** (e.g., argumentation trees, timelines, node-link diagrams, matrix displays) **applicable to social communication networks**. Our prior experience with social network and e-mail visualization [32, 34] provides a starting point for this space of analysis tools. Our network analysis tools will also **support the coexistence of multiple social networks (edge sets) among an overlapping set of people (nodes)**, enabling analysts to overlay networks such as organization charts, communication graphs, and friend relationships. Real-world applications that could benefit from improved social context include inter-agency collaboration (e.g., DHS and CIA; FBI and police) and ad-hoc analysis teams (e.g., emergency response or search-and-rescue efforts, such as the 2007 e-mail-coordinated search for missing computer scientist Jim Gray [31]).

2. Group Management and Access Control: The division of analysis work often involves the formation and management of teams. Intelligence analysis provides examples of both *cooperative and competitive models* of work [63]. In cooperative scenarios, information is pooled such that collaborators can immediately make use of others' work. Examples include finding relevant information sources, identifying connections between sources, and positing hypotheses. In competitive scenarios, work is not integrated until a later stage of sensemaking, when detailed, evidence-backed hypotheses or recommended actions are made. While lacking the benefits of resource pooling, this approach encourages individual assessment and can reduce groupthink bias. Accordingly, it may benefit collaborative analysis systems to support both fine-grained and coarse-grained work parallelization. In conjunction with our data model, **we will construct group management and access control mechanisms to support the coordination of a work group around specific tasks**. Specific mechanisms we plan to explore include *notification updates* of group activity (c.f., [10]), *selective visibility of activity* to combat groupthink bias, and *lightweight meta-data such as tagging and linking* to indicate areas of interest or flag action items.

3. Social Activity Cues: User activity can be aggregated and abstracted to provide additional forms of social context. Social navigation [27] involves the use of activity traces to provide additional navigation options, allowing users to purposefully navigate to past states of high interest or explore less-visited regions [71]. For example, the system might highlight views with low visitation rates or action items such as unanswered questions and unassessed hypotheses. Our own work [75] provides evidence that social navigation cues can simultaneously promote revisitation of popular or controversial views while also leading to a higher rate of unique discoveries. We will use our data model and analysis applications as a testbed for developing and evaluating awareness cues for collaborative analysis. **We will study how the inclusion of social activity data impacts the quality and social scalability of analysis**.

4. Markers of Identity and Reputation: Within a sensemaking context, interpersonal assessments affect how people value, and respond to the contributions of others. Other things being equal, a hypothesis suggested by a more trusted or reputable person will have a higher probability of being accepted as part of the group consensus [52]. The challenge is to understand and design effective markers of identity, trust and reputation. Even a cue as simple as one's e-mail address can lead to a number of inferences about identity and status [25]. Accordingly, **we will leverage the SAM and appropriate algorithms to explore different presentations of profile and reputation information, and assess their impact on collaborative analysis**.

5. Group Diversity: Another issue in understanding social context is the diversity of group members, such as the distribution of domain-specific knowledge among potential participants and differences in attributes such as geographical location, culture, and gender. Organizational studies [21, 59] find that increased diversity can lead to greater coverage of information and improved decision making. However, diversity can also lead to increased discord and longer decision times. We will develop diversity metrics by analyzing differences between user profiles and structural features of the social networks of the participants [11]. Such networks may be explicitly articulated or inferred from communication patterns, such as the co-occurrence of commenters across discussion threads. Wu et al.'s [77] study of organizational information flow found that information spreads efficiently among homophilous (similar) group members but not across community boundaries, further suggesting the value of identifying structural holes and directing bridging individuals in the social network towards particular findings. By constructing user profiles based on demographic data, social connectivity, and prior usage, **our analysis tools will help users make sense of social context and suggest relevant tasks to appropriate community members**.

5 Collaborating with Big Data

Human abilities including visual perception, attention, cognition and communication are relatively fixed resources. As a result, analysis often focuses on relatively modest amounts of data that people can name, point to, and reason about; graphical elements in a chart, parameters of a statistical model, top-10 lists of matches, and so on. Current collaborative data analysis systems like Sense.us and Many-Eyes.com have centered on the visualization of "spreadsheet-sized" datasets, typically containing fewer than a thousand data points. In many cases these spreadsheets were developed as summaries of much more extensive datasets.

The core challenge in collaboratively analyzing Big Data is managing the human bottleneck of *transforming* the massive datasets into more useful *data products*, usually in an iterative cycle of making a transformation, visualizing and discussing the result, and making another transformation. For analysts to collaborate efficiently on the sensemaking process, we believe it is essential to scale this iterative process to work interactively and to provide immediate visual feedback. Our work will tackle two main issues: *interaction with individual analytic methods*, and *collaboration in the process of developing data products*.

5.1 Interacting with Scalable Statistics

Interactivity and feedback at human timescales is key to sensemaking. Online Aggregation (Section 2.2) in SQL is a step in this direction, integrating system design, approximation and interaction techniques into a query language. Recently, it has been shown that this kind of aggregation can be used in SQL and MapReduce to implement fairly complex statistical methods over massive datasets [16,20]. But these more complex batch jobs are not directly amenable to the techniques used in prior work on Online Aggregation. For example, our work on implementing SVM classifiers via the Conjugate Gradient method in SQL is an iterative approach with a massive SQL query in each iteration [20]. **We will explore methods for leveraging Online Aggregation techniques within scalable statistical algorithms such as those described in [16] and [20].** Making these complex multistep algorithms more interactive requires new work at every level: data access, query processing, statistical estimation, visualization and steering.

The use of approximation to scale up to Big Data complicates the task of scaling collaboration in a big organization. To be useful in a large organization, users of varying sophistication must be able to perceive and reason about the uncertainty in the data products. We are especially interested in capturing the way analysts *discuss* this uncertainty, and the way they collaboratively *refine* their queries and analysis in the context of such uncertainty. As a simple example, a social version of our Online Aggregation system (Figure 2) could allow analysts to name three separate concepts related to uncertainty: the *estimator for the data point* (2.935... for College A), the *range of the confidence interval* for the estimator (plus or minus 0.2), and the *probability that the estimate falls into the confidence interval* (95%). To some users these concepts and their interplay are elementary, but to others they can be confusing. Such analysts might prefer not to discuss the refinement of the statistical metrics but to instead explain the more intuitive behaviors shown for example in the animations where "the red dots initially bounce up and down a lot and the bars shrink quickly, but after a while things settle down." **We will develop visual vocabularies for approximation interfaces and continuously-refining animated data visualizations**. These vocabularies will be targeted at enabling efficient asynchronous discussion of data and analysis process between analysts from a variety of backgrounds. We will realize these vocabularies as software components or language elements in a data-centric framework like Ruby on Rails, and explore their use in a variety of applications (Section 6.)

5.2 Modeling and Discussing Data Transformations and Provenance

By definition, when data is big, people end up discussing *data products*: outputs of analysis and transformation tasks. These products may be data that has been *cleaned*, *integrated* from multiple sources, and/or *analyzed* via some queries or algorithms. To enable cogent dialog about data products, a system needs to support naming and deixis both for data products, and for all the elements that were brought together to produce them. For example, a common topic of discussion in analysis is the choice not only of algorithms and data, but of tuning parameters used when running the algorithms. This agenda relates both to software management (of transformation scripts and analysis algorithms) [12, 45, 72], and to data provenance (lineage) from input data to output products [15].

We intend to extend the SAM (Section 3) to help analysts track, evolve and discuss the software scripts they use, the data they apply these scripts to, and the data products they generate. The SAM will not prescribe tools, languages or methodologies as we recognize that data analysts commonly use many tools including SQL and MapReduce, R and Matlab, Python and Perl, etc. in varying combinations [20]. Rather, the SAM will capture behaviors and processes in the environment in which these tools are used.

One very common task in large-scale data analysis is to determine the source of a data entity that is produced as a result of a long pipeline or transformations – e.g. a set of outlier data points seen after many stages of transformation involving multiple analysts and groups within an organization. To discuss this process

intelligently requires a representation of that pipeline; the input data, transformations to the data along the way, and the people with relevant knowledge or expertise. It is essential to maintain mappings between data inputs, analytic scripts and systems that run them, and data products at their output. For example, in a fully opaque "black box" transformation, it may only be feasible to capture a three-way association between an input data set, an output data set, and the code that was run to convert one to the other. In a declarative language like SQL, this three-way association may lead to richer derived associations via *data provenance* analyses [15] that can map specific input and output records together and explain their associations. Various intermediate levels of approximate forward and inverse mappings may also be available [76]. In addition to this software-oriented data provenance, a social trail is also important for answering questions like who chose the relevant input data, who wrote, chose and applied the various transformation scripts, and what are the social relationships between these parties? **We will study the ways that various methods of data provenance and transformations can be captured in the SAM, and build software interfaces to surface this information for analysts examining data products and developing transformations.** We will take care to ensure compatibility with a variety of programming models for Big Data, including SQL, MapReduce, and various sub-languages for scripting routines. We will explore the integration of our software into open-source revision control systems like Subversion or Git, as well as Integrated Development Environments like Eclipse and its data presentation layer, BIRT.

These lightweight tools we build will inform and improve current practice, by remaining open and flexible with respect to languages and development patterns. But we are also interested in the potential benefit of a tightly unified stack of software and practices. Thus, we will also explore a deep integration of a query language like SQL or SPARQL with a rich logical visualization language in the spirit of Wilkinson's Grammar of Graphics [74] and Stanford's Polaris [62]. This approach would enable powerful data provenance reasoning to be applied from the pixels of a visualization back to source data, resulting in many potential improvements in visualization and sensemaking: rich naming of data inputs by pointing at data outputs, fluid re-rendering of associations between comments and visual elements as visualizations are modified, and potentially cleaner reuse and evolution of both data transformations and visual mappings.

We posit that *social interaction around analytic processes can lead to a powerful feedback loop*: visualizations of data transformation processes enrich the discussion of data products, and the digitized discussion leads to further curation of the processes in the data lifecycle.

6 Applications

In the preceding sections of this proposal we have outlined a number of goals for our work on scalable, social data analysis. Based on these goals we are developing three software applications designed from the ground-up to facilitate and inform scalable, social data analysis. CommentSpace is a system for asynchronous web-based collaborative visual analysis of large data sets. We are also developing a social e-mail analysis tool to bootstrap our research on social analysis models and surfacing social context. Finally, Bellhop is a relational database profiling system that will be integrated with Comment Space, to support interactive summarization of massive databases as an entry point for deeper analysis.

6.1 CommentSpace

One of the lessons we learned in building and deploying Sense.us is that the ability to place freeform comments on views of a visualization enables people to collaboratively identify patterns in the data, form questions and hypotheses, marshal evidence supporting claims and leave to-do items for others. Yet, the Sense.us commenting model was fundamentally limited in two ways; 1) the comments could only be attached to views of the visualization and 2) the comments themselves consisted of freeform text and did not include any cat-

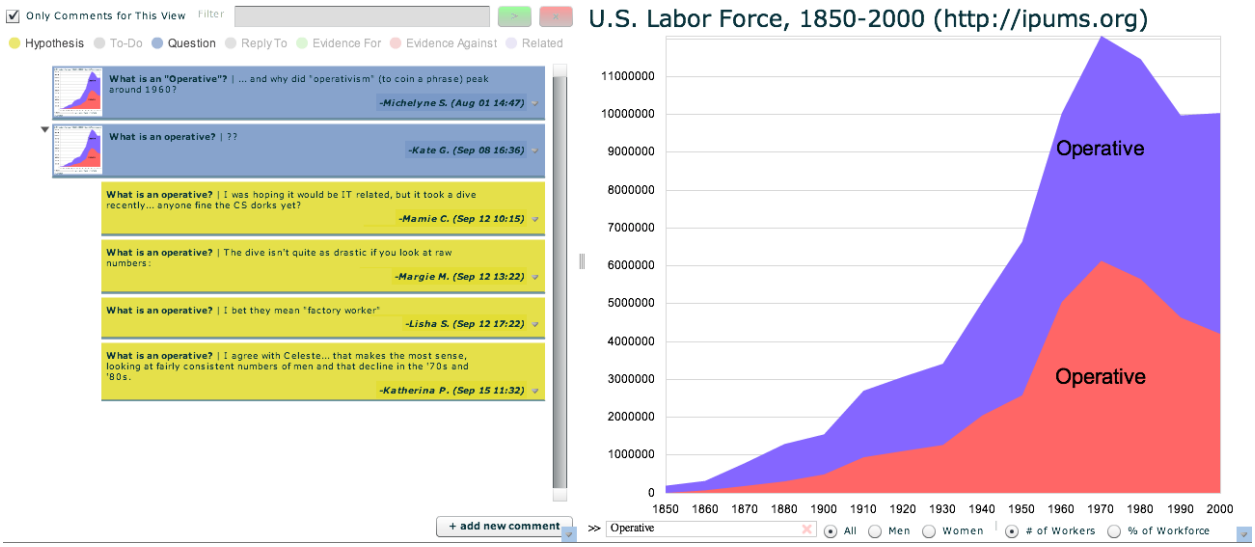

Figure 4: A mockup of the interface we envision for CommentSpace. Users can attach comments any nameable entity, including subsets of data, views of the data, transforms of the data and other comments. Comments and the links between them are color-coded based on tag-types (blue is a question, yellow is a hypothesis) and comments that related to one another (e.g. reply-to) are laid out in an indented format. If a comment is linked to a view a thumbnail of the view appear with the comment and users can enlarge the view to interact with it in the window on the right.

egorical metadata to provide structure. Sense.us did not allow users to attach comments to subsets of the data or to functional transformations of the data, and therefore made it difficult for users to discuss many aspects of the sensemaking process. Moreover, the unstructured nature of the comments left it up to users to develop their own schemas and vocabularies for organizing the comments if they wished to do so.

Our goal with CommentSpace is to develop a collaborative data analysis tool that supports rich conversations throughout every aspect of the sensemaking process. **We will build CommentSpace using our SAM as the underlying model and treat comments as just another form of data.** In contrast to Sense.us users will be able to link together any pair of nameable entities including subsets of the data, views of the data or transformations of the data. Users will be able to create additional structure by tagging the nameable entities, or the links between them. For example, a user browsing a scatterplot might notice a subset of outliers and author a comment asking "Why do these these elements fall outside the normal range?" The user could link the comment to the subset of outliers and tag the comment itself as a *question*. Another user might then filter the comments for such questions and try to find an answer to it. We will experiment with limited vocabularies and schemas for the tags so that for example analysts consistently mark comments as representing *questions, hypotheses, evidence-for, evidence-against* and *to-do items*.

We believe that enabling the creation of such structured links and tags will significantly facilitate collaborative analysis. Analysts will be able to more easily leave signposts marking interesting views, data or transformation. CommentSpace will visually depict the links and tags in a variety of ways so that collaborators can quickly find and navigate to entities of interest, answer questions, discuss data transformation strategies, verify provenance, marshal evidence supporting or refuting hypotheses, browse unexplored areas of the data, etc.

Figure 4 shows a mock-up example of the interface we envision. Comments and the links between them are color-coded based on tag-types and comments that related to one another are laid out in an indented format to indicate this relationship. Users can filter the comments by folding and unfolding the tree-structured

hierarchy. Users can also search through the comments by a specific tag-type, search for an arbitrary text string within the comments or show only the comments attached to the view or the data shown in the visualization window on the right.

Because CommentSpace comments are first class data in the SAM, we will develop a variety of visualizations for depicting the comments themselves. The tree-structured layout shown in Figure 4 is just one form of comment visualization and many other visual representations are possible. In earlier work we found that visual cues indicating where users have recently left comments can facilitate navigation and social analysis [75]. Yet our previous work did not have access to the rich linking and tagging structure of CommentSpace. Thus in CommentSpace we might produce a bar chart of the number of comments left by each user to help everyone better understand who is contributing the most to the analysis. Or, a timeline of the number of comments left on each view could indicate where the most discussion is taking place. Similarly a chart of the number of questions or hypothesis could help users navigate to interesting views and a chart of the number of evidence-for versus evidence against comments could indicate contentious issues.

Because users will be able to link comments to transformations of the data CommentSpace will support sharing of analysis procedures and provide important context to enrich traditional computational trails of data provenance. One analyst might transform the data to produce a derived product such as a mean and another could annotate the transform to suggest "Might be better off using a robust statistic such as the median." Such comments could serve as a mechanism for analysts to better review one another's transformations and thereby provide greater visibility on the analysis process itself.

6.2 Using E-mail to Bootstrap Analysis of Social Context

Assessing the social scalability of our approach in an ecologically valid fashion requires a large-scale deployment of suitably mature analysis tools. Such a deployment will be infeasible in the early stages of our research. We intend to bootstrap this process by using e-mail as a testbed for designing social communication analysis tools. This application will serve as a proxy (and adjunct) for analysis and visualization of social context. Prior research has investigated usage practices surrounding e-mail [4, 73] and developed new interfaces intended to improve e-mail management [41, 53, 57, 78]. These systems aim to enhance personal productivity, for example by surfacing one's regular correspondents, depicting temporal patterns, or surfacing aging mail in need of a response. Other systems attempt to visualize communication patterns [22, 32, 55, 66–68] in order to analyze those structures or reflect on personal relationships. Our goals are related, but with a shifted focus: **we will develop an e-mail analysis tool as a milestone project to seed our work on surfacing communication and activity patterns within analysis**.

E-mail is an attractive data type for bootstrapping our efforts. There is a great deal of available data in our own inboxes, providing non-trivial data sets to work with. Our inboxes contain messages regarding a number of analysis tasks; for Co-PI Hellerstein this includes his logs of the time-constrained, multi-party amateur search and rescue effort to find missing computer scientist Jim Gray [31]—a process involving rapid organizational development, role assignment, and matching of volunteers to tasks. Working with email has several benefits including (a) forcing us to prototype our interactive approach to data cleaning, entity resolution, and social network extraction, (b) as our tools mature, we can easily integrate e-mail archives within the SAM as an additional source of social context, and (c) by releasing an e-mail client plug-in, we can gain user feedback on our designs as we concurrently build an analysis suite.

We will import data into the SAM using ingesters for common e-mail transfer protocols (POP, IMAP) and a plug-in for the open source Thunderbird mail client. Collected e-mail will serve as a natural testbed for investigating approaches to surfacing and analyzing social context (Section 4). As the ingestion process will require us to transform data and write scripts, e-mail analysis will also bootstrap our code management tools

(Section 5). For example, we have already developed scalable, database-managed scripts for computing network metrics such as PageRank and clustering coefficients, and have begun scripts for text analysis methods such as q-gram similarity.

6.3 Bellhop: Interactive, Collaborative Data Profiling

As with our discussion of social context, we need to bootstrap our assessment of our proposal to enable collaboration with interactive statistics via online aggregation. To this end, **we will build a collaborative, interactive data profiling tool we are calling Bellhop** – named after a batch-oriented data profiler called Bellman developed in the 1990's [43].

It is common practice for analysts to begin exploratory data analysis in relational databases by computing *data profiles* that provide summaries of coarse statistics across the databases. Common profile elements include column-summary statistics like means, medians, standard deviations, and frequent values. They also include identification of potential keys per table (row identifiers) and likely foreign keys (references) across tables, as well as likely "join paths" that can be used to connect tables. A number of these profile statistics can be directly estimated via known online aggregation techniques, others have been less studied but look tractable; development of online profile algorithms fits into the goals of Section 5.

The advantage of building Bellhop is that data profiles are a common "entry path" into different data analyses. So on one hand they are likely to receive shared attention over time from many analysts, but on the other hand are only of casual interest in many cases, since analysts are likely to proceed deeper into the data after a glance at the profile. This is an ideal setting for beginning our evaluation of ways that analysts manage issues around approximation, impatience with processing, and collaboration over time. For example, the first analyst to play with a data set may run profiling queries for a short time, stop and save their outputs, tag a feature of interest and move on to a deeper analysis. The next analyst may distrust the degree of uncertainty in certain stored profile results and continue their execution, tagging the refined results before moving on to their own analyses. A third analyst may let certain profiling tasks run to absolute completion. Over time, the refinement of profile results and their discussion should provide a historical record of shared analysis and discussion.

7 Evaluation: Metrics and Methodology

The applications we develop will serve as testbeds for experimenting with a variety of approaches and building blocks to support scalable social data analysis. A key feature of our approach is that *it ties together issues of scalability and performance* commonly addressed by database researchers, *with issues of utility, usability and quality of analysis results* that are usually studied by HCI researchers. As a result we believe it is crucial to evaluate the effectiveness of our application designs using methodologies from both communities.

Our work aims to emphasize real-time interaction with massive data sets. Thus, our applications and underlying software components will be designed to use statistical approximations that progressively refine the quality of transformations and visualizations. Regardless of data set size, the applications will immediately display initial results and continuously refine the results based on statistical samples of the data, when possible providing confidence intervals or other statistical assessments of uncertainty. We will measure responsiveness of the applications in terms of latency (how quickly first results are displayed), update rates (how quickly results are updated as they are refined), quality (confidence of estimate) and speed with which a given quality is achieved (how quickly a confidence interval threshold is reached with 95% confidence). Scalability will be evaluated on Cloud Computing infrastructure, using data sets whose sizes vary by orders of magnitude, and include a variety of underlying statistical properties generated both synthetically and from real data.

In addition to the core system measurements of our component technologies, we plan to measure the utility of our approaches and tools in human sensemaking. As a first step, we wish to gain more insight into the needs and practices of real-world analysts. To do so, we plan to conduct fieldwork observing and interviewing analysts at a number of collaborating institutions, including FOX Audience Network, Yahoo!, and the NYSE (see letters of support). Our *contextual inquiry* [6] will investigate data collection and curation processes, and the analysis approaches used by various groups within these institutions. We suspect that a key challenge of scaling tools across an organization is not just the sheer number of people, but also the diversity of roles. We will attempt to observe the social dynamics and group incentives within these organizations, including the collaborations (or "hand-offs") between groups with potentially diverse organizational roles. We believe that this investigation will be highly informative of our subsequent design and development efforts.

In order to develop useful and effective applications, we will rapidly iterate using the three stage *design–prototype–evaluate* cycle. Our aim is to produce prototype applications quickly and then observe and interview analysts as they work with these prototypes to solve real-world analysis problems. Initial evaluations will be performed using both informal and controlled deployments with university students engaged in analysis tasks; we plan to evaluate more mature prototypes with analysts at collaborating institutions, ideally leveraging "informants" established in our contextual inquiry. Based on these observations and feedback we will iteratively refine the design of our algorithms, interfaces and visualizations to better serve analysts.

To assess the quality of analysis we will develop a set of analysis tasks and scenarios that are based on the datasets used in each application. For example, with CommentSpace we can use U.S. census data (e.g., 15 decades of data from ipums.org, also used in our Sense.us system) and ask analysts to "Find evidence of gender bias in US jobs." Other data sets of interest include those used in the annual IEEE VAST analysis contests [60]. These sets are derived from real-world intelligence data, and are released in conjunction with well-specified analysis tasks whose outcome can be compared against those of past contest submissions.

We will observe analysts as they use our applications to perform such analyses and then apply *content analysis* techniques [46] to the resulting social interactions (e.g., comments, annotations) in which trained experimenters classify comments (e.g., is it a hypothesis, a question, etc.) and rate the quality of each comment and annotation left by the analysts. We will measure the degree to which our tools, such as the linking and tagging structures supported by CommentSpace, lead to more hypotheses, better evidence gathering and more efficient distribution of tasks. We will also compare our applications to earlier state-of-the-art applications (e.g. Sense.us, Outlook, Thunderbird, business intelligence tools, Excel, etc.) to check whether our applications increase the quality, efficiency and depth of analysis.

A key goal of our work is to understand how well our social data analysis tools scale as the number of analysts grows and the social structure of the analysts changes. A small group of three to five analysts in a single company is likely to work in a very different manner from a group of hundreds of casual analysts collaborating via the Internet to analyze a dataset. To study the effects of group size and structure we will work with a variety of different sized groups. Initially we will study how small groups of graduate students work with the system and then scale up the system to work with larger groups of students (at the graduate, undergraduate and eventually high-school levels) within a classroom. With students we will focus on the *educational benefits* of our tools and again use content analysis methods [46] to *examine how they increase the level of discourse and analysis about the data in the classroom*. We will also work with teams of analysts in companies such as the Fox Audience Network (see letter of support) to compare their analysis approaches to those of the students. In parallel we will provide public access to many of our applications via the Web and open source and study how they are used *in the wild* as varied networked users work together to make sense of the data.

Throughout each of these studies we will use our observations, interviews with analysts, recorded usage data, and content analysis to assess the effectiveness and malleability of our social analysis model. While we believe that surfacing provenance and activity data can aid analysis, this process inevitably introduces additional complexity. To assess such trade-offs, our evaluations will also include (a) analyses of what forms of social activity data analysts access and utilize as part of their work and (b) controlled experiments that vary the types and richness of contextual data provided in order to assess their effect on analysis outcomes.

8 Results from Prior NSF Funding

PI Maneesh Agrawala is an Associate Professor of Electrical Engineering and Computer Science at the University of California, Berkeley. His research investigates how understanding of human perception and cognition can be used to improve the effectiveness of visualizations. He has received an Okawa Foundation Research Grant in 2006, a Sloan Fellowship in 2007 and a SIGGRAPH Significant New Researcher award in 2008. He also received an NSF CAREER award *#CCF-0643552 CAREER: Design Principles, Algorithms, and Interfaces for Visual Communication, $400,000* in 2007 and an NSF HCC award *#IIS-0812562 HCC-Small: Collaborative Research: Design and Evaluation of the Next Generation of E-book Readers, $75,999* in 2008. While these awards have enabled work that has appeared in a variety of venues including SIGGRAPH, IEEE Information Visualization, and UIST, as well as recent best paper awards at SIGCHI and Graphics Interface, neither of these previous NSF awards overlaps with the work described in this document.

Co-PI Jeffrey Heer has not yet received any funding from NSF.

Co-PI Joseph M. Hellerstein has been involved in a number of NSF awards in the last five years. Three are currently active, but entering their final year, and focused in rather different areas of computing than this proposal: *III-COR; Dynamic Meta-Compilation in Networked Information Systems* (0713661, 09/01/2007 - 08/31/2010, $450,000), *NeTS-NBD: SCAN: Statistical Collaborative Analysis of Networks* (0722077, 01/01/2008 - 12/31/2010, $249,000), *NGNI-Medium: MUNDO: Managing Uncertainty in Networks with Declarative Overlays* (09/01/2008 - 08/31/2010, $303,872). The remainder are completed: *ITR: Data on the Deep Web: Queries, Trawls, Policies and Countermeasures* (0205647, 10/2002 - 9/2007, $1,675,000), *Adaptive Dataflow: Eddies, SteMs and FLuX* (0208588, 08/2002 - 07/2005, $299,998), *Query Processing in Structured Peer-to-Peer Networks* (0209108, 08/2002 – 07/2005, $179,827), and *ITR: Robust Large Scale Distributed System* (5710001486, 10/2002 – 9/2007, $2,840,869). The most relevant of these awards is the ITR involving Data on the Deep Web, which funded Hellerstein's influential Telegraph project on adaptive query processing of Web and Stream data (http://telegraph.cs.berkeley.edu). The work produced many prestigious papers including Best Paper awards at VLDB 2004 and IEEE SPOTS 2006, and a paper at SIGMOD 2006 chosen for the CACM Research Highlights to appear this fall. It led to the education of numerous graduate students (three now faculty members at leading research universities), provided research opportunities for seven undergraduates, and led to the founding of a startup company called Truviso by one of the funded graduate students, which is still in business. Two other influential open source projects came out of that proposal: the P2 Declarative Networking system and the PIER Internet-scale query processor.

E Collaboration Plan

Although Visualization, Databases, HCI and Social Computing are strongly interrelated areas of Computer Science, they are usually studied independently. A key feature of this proposal is that it brings together ideas from all of these areas, with PIs that have expertise in one or more of each of them: Visualization (Agrawala, Heer), Databases (Hellerstein), HCI (Agrawala, Heer) and Social Computing (Heer).

The PIs already have a history of working together. Heer recently graduated from Agrawala's group and is now an Assistant Professor at Stanford. They have co-authored many papers together on many different topics in Visualization and Collaborative Data Analysis. Hellerstein has developed a wide variety of core Database technologies and all three PIs have had joint weekly research meetings for the past two years to discuss our shared interests in Scalable Social Data Analysis. Hellerstein taught a graduate seminar course on Data Management for Collaborative Environments in Spring 2009 based on these discussions.

The primary mechanism for collaborating on research projects and discussing research ideas is the joint advising of students. Mentoring graduate students is the key to vibrant and successful academic collaboration. The PIs envision working closely with graduate students at both Berkeley and Stanford in co-advising relationships. Already the PIs are playing co-advisory roles for Berkeley Ph.D. students Wesley Willett, Kuang Chen and incoming Stanford Ph.D. student Diana MacLean. Our experience is that the close proximity of Berkeley and Stanford significantly facilitates such mentoring and collaboration between the PIs and their students. Hellerstein has a long track record of co-advising students, including students who are now faculty at MIT (Sam Madden) and University of Pennsylvania (Boon Thau Loo). Loo was co-advised across the areas of databases and networking, with Prof. Ion Stoica. Alexandra Meliou is another Ph.D. alumna beginning a post-doc at U. Washington, co-advised by Hellerstein and CMU Prof. Carlos Guestrin, whose area is Machine Learning.

All three PIs are deeply committed to mentoring and producing students with skills that span Visualization, Databases, HCI and Social Computing. As we have noted, research in each of these areas has largely been conducted independently and instances of sharing insights and approaches between the communities have been rare. We aim to fill this gap as we believe that the best way to develop new, scalable tools for analyzing very large datasets is to train students in this combination of areas.

E.1 PI Roles and Responsibilities

This project will be directed by PI Agrawala through the Department of Electrical Engineering and Computer Science at the University of California, Berkeley. Agrawala will be responsible for overall management of the project including maintaining the schedule of weekly meetings of the entire team and submitting NSF progress reports.

All three PIs will work together closely with graduate students at Berkeley and Stanford to develop the *social analysis model* (SAM) (Section 3). However, each PI will also take primary responsibility for developing one of the applications described in Section 6. More specifically, Agrawala will lead the effort to develop CommentSpace (Section 6.1). Co-PI Heer will lead the effort to develop the tools required to use email for bootstrapping the analysis of social context (Section 6.2). Co-PI Hellerstein will lead the effort to develop the Bellhop data profiler. Each PI will also direct the effort to evaluate their respective applications. Nevertheless, we expect to collaborate closely on developing all of the applications as they will be built using the same underlying SAM and share analytic libraries and visual components.

E.2 Project Management Across Investigators, Institutions and Disciplines

While PI Agrawala will be responsible for the overall management of the project, we believe that the close proximity of Berkeley and Stanford as well as the strong ties and working relationships of the PIs will enable extensive cross-fertilization and deep collaboration. We expect that many of the graduate students will be co-advised by various combinations of or even all three of the PIs. The tightly integrated research plan is also designed to further foster the collaboration as all of the applications we build will be based on the SAM and many of the tools and techniques we develop will be designed to work with all of the applications.

All three PIs have a long and successful history of interdisciplinary projects across multiple institutions, often involving faculty from a variety of other disciplines.

E.3 Specific Coordination Mechanisms

Weekly Team and Project Meetings: The entire team of PIs and graduate students will meet weekly to discuss progress and identify next steps. In addition to meeting as a complete team, each PI will also meet with their graduate students to discuss progress on the specific applications. The team meetings will be held using conference calls and shared desktop/whiteboard software. At least once a month the complete team will meet at either Berkeley or Stanford rather than teleconference, to provide better face-to-face discussion.

Yearly Workshop on Scalable Social Data Analysis: In order to provide better connection between the Visualization, Database and HCI communities, we will hold a day-long workshop once a year on the topic of scalable social data analysis. We will invite team members as well as researchers and data analysts from a variety of organizations (including our industrial partners – see letters of support) to give talks or present posters. The workshop will be open to the public and based on past experience we believe there is great interest in having such a workshop. Agrawala and Heer previously organized such a workshop at Berkeley in Fall Spring 2006 on the general topic of Visualization and many participants have encouraged us to continue to hold such meetings.

E.4 Budget Line Items Supporting Coordination Mechanisms

The weekly team and project meetings do not require additional budget because of the proximity of the institutions. We have allocated $1625 of the yearly budget as *Workshop Expense* (see Berkeley budget) to hold the workshop.

References Cited

[1] Christopher Ahlberg and Ben Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 313–317, New York, NY, USA, 1994. ACM.

[2] Aruna Balakrishnan, Sara Kiesler, and Susan R. Fussell. Do visualizations improve synchronous remote collaboration? In *ACM Human Factors in Computing Systems (CHI)*, pages 1227–1236, 2008.

[3] Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[4] Victoria Bellotti, Nicolas Ducheneaut, Mark Howard, and Ian Smith. Taking email to task: the design and evaluation of a task management centered email tool. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 345–352, New York, NY, USA, 2003. ACM.

[5] J. Bertin. *Semiology of graphics*. University of Wisconsin Press, 1983.

[6] Hugh Beyer and Karen Holtzblatt. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, 1998.

[7] D Billman, G Convertino, J Shrager, P Pirolli, and J P Massar. Collaborative intelligence analysis with cache and its effects on information gathering and cognitive bias. In *Human Computer Interaction Consortium*, 2006.

[8] Susan E Brennan. How conversation is shaped by visual and spoken evidence. In Trueswell and Tanenhaus, editors, *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 95–129, Cambridge, MA, 2005. MIT Press.

[9] Susan E Brennan, K Mueller, G Zelinsky, I V Ramakrishnan, D S Warren, and A Kaufman. Toward a multi-analyst, collaborative framework for visual analytics. In *IEEE Symposium on Visual Analytics Science and Technology*, 2006.

[10] A J Brush, D Bargeron, J Grudin, and A Gupta. Notification for shared annotation of digital documents. In *Proc. ACM Conference on Human Factors in Computing Systems (CHI'02)*, 2002.

[11] R. S. Burt. Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399, 2004.

[12] S.P. Callahan, J. Freire, E. Santos, C.E. Scheidegger, C.T. Silva, and H.T. Vo. VisTrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 745–747. ACM New York, NY, USA, 2006.

[13] Stuart Card, Jock D. Mackinlay, and Ben Shneiderman, editors. *Readings In Information Visualization: Using Vision To Think*. Morgan Kauffman Publishers, Inc., San Francisco, USA, 1999.

[14] J Caroll, Mary Beth Rosson, G Convertino, and C H Ganoe. Awareness and teamwork in computer-supported collaborations. *Interacting with Computers*, 18(1):21–46, 2005.

[15] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Found. Trends databases*, 1(4):379–474, 2009.

[16] Cheng-Tao Chu, Sang Kyun Kim, Yi-An Lin, YuanYuan Yu, Gary R. Bradski, Andrew Y. Ng, and Kunle Olukotun. Map-reduce for machine learning on multicore. In *NIPS*, pages 281–288, 2006.

[17] Herb H Clark. Pointing and placing. In S Kita, editor, *Pointing. Where language, culture, and cognition meet*, pages 243–268. Erlbaum, 2003.

[18] Herb H. Clark and Susan E. Brennan. Grounding in communication. In L. B. Resnick, R. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association, 1991.

[19] W.S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[20] J. Cohen, B. Dolan, M. Dunlap, J.M. Hellerstein, and C. Welton. MAD Skills: New analysis practices for big data. In *VLDB*, 2009.

[21] J. Cummings. Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science*, 50(3):352–364, 2004.

[22] Kushal Dave, Martin Wattenberg, and Michael Muller. Flash forums and forumreader: navigating a new kind of large-scale online discussion. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 232–241, New York, NY, USA, 2004. ACM.

[23] Mark Derthick, John Kolojejchick, and Steven F. Roth. An interactive visual query environment for exploring data. In *UIST '97: Proceedings of the 10th annual ACM symposium on User interface software and technology*, pages 189–198, New York, NY, USA, 1997. ACM.

[24] A. Dobra, C. Jermaine, F. Rusu, and F. Xu. Turbo-charging estimate convergence in dbo. In *VLDB*, 2009.

[25] Judith S. Donath. Identity and deception in the virtual community. In M Smith and P Kollock, editors, *Communities in Cyberspace*, 1998.

[26] Paul Dourish and Victoria Belotti. Awareness and coordination in shared workspaces. In *Proc. ACM Conference on Computer-Supported Cooperative Work*, pages 107–114, Toronto, Ontario, 1992.

[27] Paul Dourish and Matthew Chalmers. Running out of space: Models of information navigation. In *Proc. Human Computer Interaction (HCI'94)*, 1994.

[28] Ken Fishkin and Maureen C. Stone. Enhanced dynamic queries via movable filters. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–420, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[29] J.F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, and A. Toncheva. The diverse and exploding digital universe: An updated forecast of world-wide information growth through 2011. Technical report, IDC White Paper, March 2008. http://www.emc.com/about/destination/digital_universe/.

[30] J.F. Gantz, D. Reinsel, C. Chute, W. Schlichting, J. McArthur, S. Minton, I. Xheneti, A. Toncheva, and A. Manfrediz. The expanding digital universe: A forecast of world-wide information growth through 2010. Technical report, IDC White Paper, March 2007. http://www.emc.com/about/destination/digital_universe/.

[31] Katie Hafner. Silicon valley's high-tech hunt for colleague. New York Times, February 3, 2007.

[32] Jeffrey Heer. Exploring enron. Accessed: August, 2009.

[33] Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. Generalized selection via interactive query relaxation. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 959–968, New York, NY, USA, 2008. ACM.

[34] Jeffrey Heer and danah boyd. Vizster: Visualizing online social networks. In *IEEE Information Visualization*, pages 32–39, 2005.

[35] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1029–1038, New York, USA, 2007. ACM Press.

[36] J. M. Hellerstein, R. Avnur, A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth, and PJ Haas. Interactive data analysis: The control project. *Computer*, 32(8):51–59, 1999.

[37] Joseph M. Hellerstein, Peter J. Haas, and Helen J. Wang. 2007 test-of-time award "online aggregation". In *SIGMOD*, 2007.

[38] Will C Hill and James D Hollan. Deixis and the future of visualization excellence. In *Proc. of IEEE Visualization*, pages 314–319, 1991.

[39] Harry Hochheiser and Ben Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[40] Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. Direct manipulation interfaces. *Human-Computer Interaction*, 1(4):311–338, 1985.

[41] David Huynh. Seek: Faceted browsing for thunderbird. Accessed: August, 2009.

[42] C. Jermaine, S. Arumugam, A. Pol, and A. Dobra. Scalable approximate query processing with the dbo engine. In *SIGMOD*, 2007.

[43] Theodore Johnson, Amit Marathe, and Tamraparni Dasu. Database exploration and bellman. *IEEE Data Eng. Bull.*, 26(3):34–39, 2003.

[44] Paul E. Keel. Collaborative visual analytics: Inferring from the spatial organization and collaborative use of information. In *IEEE Visual Analytics Science and Technology*, pages 137–144, 2006.

[45] Nodira Khoussainova, Magdalena Balazinska, Wolfgang Gatterbauer, YongChul Kwon, and Dan Suciu. A case for a collaborative query management system. In *CIDR*, 2009.

[46] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, 2004.

[47] M. Livny, R. Ramakrishnan, K. Beyer, G. Chen, D. Donjerkovic, S. Lawande, J. Myllymaki, and K. Wenger. Devise: integrated querying and visual exploration of large datasets. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 301–312, New York, NY, USA, 1997. ACM.

[48] Miron Livny, Raghu Ramakrishnan, Kevin S. Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi Myllymaki, and R. Kent Wenger. DEVise: Integrated querying and visual exploration of large datasets (demo abstract). In *SIGMOD*, pages 517–520, 1997.

[49] P. Lyman and H. Varian. How much information? 2003, 2003. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

[50] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

[51] Allen R. Martin and Matthew O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 271, Washington, DC, USA, 1995. IEEE Computer Society.

[52] Susan Mohammed. Toward an understanding of cognitive consensus in a group decision-making context. *The Journal of Applied Behavioral Science*, 37(4):408–425, December 2001.

[53] Carmen Neustaedter, A.J. Brush, Marc Smith, and Danyel Fisher. The social network and relationship finder: Social sorting for email triage. In *Proc. Conference on E-mail and Anti-Spam*, July 2005.

[54] C. Olston, M. Stonebraker, A. Aiken, and J. M. Hellerstein. Viqing: Visual interactive querying. In *VL '98: Proceedings of the IEEE Symposium on Visual Languages*, page 162, Washington, DC, USA, 1998. IEEE Computer Society.

[55] Adam Perer and Marc A. Smith. Contrasting portraits of email practices: visual approaches to reflection and analysis. In *International Conference on Advanced Visual Interfaces (AVI 2006)*, pages 389–395, 2006.

[56] Peter Pirolli and Stuart K. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proc. of International Conference on Intelligence Analysis*, 2005.

[57] Steven L. Rohall, Dan Gruen, Paul Moody, Martin Wattenberg, Mia Stern, Bernard Kerr, Bob Stachel, Kushal Dave, Robert Armes, and Eric Wilcox. Remail: a reinvented email prototype. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 791–792, New York, 2004. ACM.

[58] Dan M. Russell, Mark J. Stefik, Peter Pirolli, and Stuart K. Card. The cost structure of sensemaking. In *ACM Human Factors in Computing Systems (CHI)*, 1993.

[59] S. Schultz-Hart, D. Frey, C. Lüthgens, and S. Moscovici. Biased information search in group decision making. *Journal of Personality and Social Psychology*, 78(4):655–669, 2000.

[60] Semvast: Scientific evaluation methods for visual analytics science and technology. Accessed: August, 2009.

[61] Yedendra B. Shrinivasan and Jarke J. van Wijk. Supporting the analytical reasoning process in information visualization. In *ACM Human Factors in Computing Systems (CHI)*, pages 1237–1246, 2008.

[62] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional databases. *Commun. ACM*, 51(11):75–84, 2008.

[63] James J. Thomas and Kristin A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, August 2005.

[64] John W. Tukey. The future of data analysis. *Annals of Mathematical Statistics*, 33(1):1–67, 1962.

[65] Hal Varian. Hal varian on how the web challenges managers. *McKinsey Quarterly*, January 2009.

[66] Gina Danielle Venolia and Carman Neustaedter. Understanding sequence and reply relationships within email conversations: a mixed-model visualization. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 361–368, New York, NY, USA, 2003. ACM.

[67] Fernanda B. Viégas, Danah Boyd, David H. Nguyen, Jeffrey Potter, and Judith Donath. Digital artifacts for remembering and storytelling: Posthistory and social network fragments. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4*, page 40109.1, Washington, DC, USA, 2004. IEEE Computer Society.

[68] Fernanda B. Viégas, Scott Golder, and Judith Donath. Visualizing email content: portraying relationships from conversational histories. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988, New York, 2006. ACM.

[69] Fernanda B. Viégas, Martin Wattenberg, Matt McKeon, Frank van Ham, and Jesse Kriss. Harry potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools. In *Hawaii International Conference on Systems Science (HICSS)*, 2008.

[70] Fernanda B. Viégas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2007)*, 12(5):1121–1128, November/December 2007.

[71] Martin Wattenberg and Jesse Kriss. Designing for Social Data Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, July–August 2006.

[72] David Whitgift. *Methods and tools for software configuration management*. John Wiley & Sons, Inc., New York, NY, USA, 1991.

[73] Steve Whittaker and Candace Sidner. Email overload: exploring personal information management of email. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 276–283, New York, NY, USA, 1996. ACM.

[74] Leland Wilkinson. *The Grammar of Graphics*. Springer, 2005.

[75] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.

[76] Allison Woodruff and Michael Stonebraker. Supporting fine-grained data lineage in a database visualization environment. In *ICDE '97: Proceedings of the Thirteenth International Conference on Data Engineering*, pages 91–102, Washington, DC, USA, 1997. IEEE Computer Society.

[77] F. Wu, B. A. Huberman, L. A. Adamic, and J. R. Tyler. Information flow in social groups. *Physica A: Statistical and Theoretical Physics*, 337(1-2):327, 2000.

[78] Xobni: Outlook plugin to search people, email, and attachments. Accessed: August, 2009.

[79] S. Zilberstein. Using anytime algorithms in intelligent systems. *AI magazine*, 17(3):73, 1996.

Biographical Sketch: Maneesh Agrawala

Associate Professor, EECS Computer Science Division
University of California, Berkeley
URL: http://vis.berkeley.edu/∼maneesh

PROFESSIONAL PREPARATION:

June 1994	**Stanford University** Bachelor of Science in Mathematics
January 2002	**Stanford University** Doctor of Philosophy in Computer Science Dissertation Topic: *Visualizing Route Maps*

APPOINTMENTS:

2009–Present	**University of California, Berkeley** Associate Professor, EECS Department
2005–2009	**University of California, Berkeley** Assistant Professor, EECS Department
2006–Present	**Swivel.com** Technical Advisor
2002–Present	**University of Washington** Affiliate Faculty, Department of Computer Sciences and Engineering

FIVE MOST RELEVANT PUBLICATIONS:

1. "Perceptual Interpretation of Ink Annotations on Line Charts" by Nicholas Kong and Maneesh Agrawala. *Proc. ACM User Interface Software Technology [UIST 2009], To Appear.*

2. "Design Considerations for Collaborative Visual Analytics" by Jeffrey Heer and Maneesh Agrawala. *Information Visualization Journal, 7(1), pp. 49-62, 2008.*

3. "Generalized Selection via Interactive Query Relaxation" by Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. *Proc. ACM Human Factors in Computing Systems (CHI), pp. 959-968, Apr 2008.*

4. "Scented Widgets: Improving Navigation Cues with Embedded Visualizations" by Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis'07), 13(6), pp. 1129-1136, Nov/Dec 2007.*

5. "Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations" by Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. *Proc. ACM Human Factors in Computing Systems (CHI), pp. 1303-1312, Apr 2009.*

FIVE OTHER PUBLICATIONS:

1. "Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation" by Jeffrey Heer, Jock D. Mackinlay, Chris Stolte, and Maneesh Agrawala. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis'08), 14(6), pp. 1189-1196, Nov/Dec 2008.*

2. "Generating Photo Manipulation Tutorials by Demonstration" by Floraine Grabler, Maneesh Agrawala, Wilmot Li, Mira Dontcheva and Takeo Igarashi. *ACM Transactions on Graphics 28(3) [SIGGRAPH 2009], 66:1–66:9.*

3. "Multi-Scale Banking to 45 Degrees" by Jeffrey Heer and Maneesh Agrawala. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis'06), 12(5), pp. 701-708, Nov/Dec 2006.*

4. "Software Design Patterns for Information Visualization" by Jeffrey Heer and Maneesh Agrawala. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis'06), 12(5), pp. 853-860, Nov/Dec*

2006.

5. "Designing Effective Step-By-Step Assembly Instructions" by Maneesh Agrawala, Doantam Phan, Julie Heiser, John Haymaker, Jeff Klingner, Pat Hanrahan and Barbara Tversky. *ACM Transactions on Graphics 22(3) [SIGGRAPH 2003], pp. 828–837.*

EDUCATIONAL ACTIVITIES:

1. Re-developed and taught class on User Interfaces at UC Berkeley (Fall 06, Spr 08). Co-developed and taught graduate class on Visualization at Stanford (Spr 02), Univ. of Washington (Spr 05), and UC Berkeley (Spr 06, Fall 07).

2. Co-organized half-day course on Computation and Journalism at SIGGRAPH 2008.

SYNERGISTIC ACTIVITIES:

1. Wrote LineDrive route mapping system that has been available at *www.mappoint.com* since October 2000. Ideas developed for Interactive Photomontage project have been incorporated into Microsoft's Digital Picture Pro and GroupShot software, as well as Adobe Photoshop CS3.

2. Panelist at first NIH/NSF workshop on Visualization Research Challenges. Worked to identify short-term and long-term research objectives in visualization – October 2004.

3. Invited by National Academy of Engineering to speak about new research on visual communication at the German-American Frontiers of Engineering Symposium – May 2005.

COLLABORATORS:

Aseem Agarwala (Adobe), Georg Apitz (Univ. of Maryland), Ravin Balakrishnan (Univ. of Toronto), Martin S. Banks (Berkeley), Connelly Barnes (Princeton), Patrick Baudisch (Microsoft), Pravin Bhat (Univ. of Washington), Nicholas Chen (Univ. of Maryland), Michael Cohen (Microsoft), Alex Colburn (Microsoft), Maxime Collomb (Montpellier Univ.) Brian Curless (Univ. of Washington), Edward Cutrell (Microsoft), Mary Czerwinski (Microsoft), Doug DeCarlo (Rutgers), Morgan Dixon (Univ. of Washington), Mira Dontcheva (Adobe), Steve Drucker (Microsoft), Raanan Fattal (Hebrew Univ.), Adam Finkelstein (Princeton), Dan B Goldman (Adobe), Tovi Grossman (Autodesk), Francois Guimbretiere (Univ. of Maryland), Jeff Heer (Stanford), Aaron Hertzmann (Univ. of Toronto), Ken Hinckley (Microsoft), Hugues Hoppe (Microsoft), David E. Jacobs (Stanford), Sing Bing Kang (Microsoft), Cassandra Lewis (Univ. of Maryland), Wilmot Li (Adobe), Jock Mackinlay (Tableau), Mark Pauly (ETH), Georg Petschnigg (Microsoft), Marc Pollefeys (ETH), Ravi Ramamoorthi (Columbia), Gonzalo Ramos (Microsoft), Lincoln Ritter (Univ. of Washington), Dan Robbins (Microsoft), George Robertson (Microsoft), Szymon Rusinkiewicz (Princeton), David Salesin (Adobe and Univ. of Washington), Jason Sanders (Nvidia), Anthony Santella (Rutgers), Raman Sarin (Microsoft), Steve Seitz (Univ. of Washington), Sudipta Sinha (Univ. of North Carolina), Noah Snavely (Cornell), Drew Steedly (Microsoft), Chris Stolte (Tableau), Robert W. Sumner (Disney), Rick Szeliski (Microsoft), Desney Tan (Microsoft), Dan Vogel (Univ. of Toronto), Ian Vollick (Univ. of Toronto), Jue Wang (Adobe), Andrew Wilson (Microsoft), Shengdong Zhao (Nat. Univ. of Singapore), Ke Colin Zheng (Univ. of Washington), C. Lawrence Zitnick (Microsoft)

PH.D. ADVISOR:

Pat Hanrahan (Stanford)

ADVISEES AT UC BERKELEY:

Robert Carroll, Floraine Grabler, Kenrick Kin, Nicholas Kong, James O'Shea, Wesley Willett

Biographical Sketch: Jeffrey Heer

Assistant Professor, Computer Science Department
Stanford University
URL: http://hci.stanford.edu/jheer

PROFESSIONAL PREPARATION:

Jun 2001	**University of California, Berkeley** Bachelor of Science in Electrical Engineering and Computer Science
Dec 2004	**University of California, Berkeley** Master of Science in Computer Science
Dec 2008	**University of California, Berkeley** Doctor of Philosophy in Computer Science Dissertation: *Supporting Asynchronous Collaboration for Interactive Visualization*

APPOINTMENTS:

2009–Present	**Stanford University** Assistant Professor, Computer Science Department
2008–Present	**OurGroup.org** Technical Advisor

FIVE MOST RELEVANT PUBLICATIONS:

1. "Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization" by Jeffrey Heer, Fernanda Viégas, and Martin Wattenberg. *Communications of the ACM, 52(1), pp. 87-97, Jan 2009.*

2. "Design Considerations for Collaborative Visual Analytics" by Jeffrey Heer and Maneesh Agrawala. *Information Visualization Journal, 7(1), pp. 49-62, 2008.*

3. "Generalized Selection via Interactive Query Relaxation" by Jeffrey Heer, Maneesh Agrawala, and Wesley Willett. *Proc. ACM Human Factors in Computing Systems (CHI), pp. 959-968, Apr 2008.*

4. "Scented Widgets: Improving Navigation Cues with Embedded Visualizations" by Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis'07), 13(6), pp. 1129-1136, Nov/Dec 2007.*

5. "Vizster: Visualizing Online Social Networks" by Jeffrey Heer and danah boyd. *Proc. IEEE Symposium on Information Visualization (InfoVis), pp. 32-39, Oct 2005.*

FIVE OTHER PUBLICATIONS:

1. "Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation" by Jeffrey Heer, Jock D. Mackinlay, Chris Stolte, and Maneesh Agrawala. *IEEE Transactions on Visualization and Computer Graphics (Proc. InfoVis'08), 14(6), pp. 1189-1196, Nov/Dec 2008.*

2. "Prefuse: A Toolkit for Interactive Information Visualization" by Jeffrey Heer, Stuart K. Card, and James A. Landay. *Proc. ACM Human Factors in Computing Systems (CHI), pp. 421-430, Apr 2005.*

3. "Protovis: A Graphical Toolkit for Visualization" by Michael Bostock and Jeffrey Heer. *IEEE Transactions on Visualization and Computer Graphics (InfoVis'09), To Appear.*

4. "TimeTree: Exploring Time Changing Hierarchies" by Stuart K. Card, Bongwon Suh, Bryan Pendleton, Jeffrey Heer, and John W. Bodnar. *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST), pp. 3-10, Oct 2006.*

5. "Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations" by Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. *Proc. ACM Human Factors in*

Computing Systems (CHI), pp. 1303-1312, Apr 2009. (CHI 2009 Best Paper Award)

EDUCATIONAL ACTIVITIES:

1. Co-developed and taught class on Visualization at Stanford (Win 09, Fall 09) and UC Berkeley (Fall 05, Spr 06). Re-developed and taught class on Research Topics in Human-Computer Interaction at Stanford (Spr 09).

2. Co-organized half-day course on Visualization and Social Data Analysis at VLDB 2009.

3. Co-organized half-day course on Computation and Journalism at SIGGRAPH 2008.

SYNERGISTIC ACTIVITIES:

1. Led the development of the open source *prefuse*, *flare*, and *protovis* visualization toolkits (see prefuse.org, flare.prefuse.org, and protovis.org). The tools have collectively been downloaded over 100,000 times, referenced in over 500 research publications, and are actively used by the visualization research community and numerous corporations.

2. Invited keynote speaker at Conference on Innovative Data Systems Research (CIDR), to speak about recent trends in visualization and collaborative data analysis – January 2009.

3. Co-organized full-day workshop on Social Data Analysis at CHI 2008.

COLLABORATORS:

Maneesh Agrawala (Berkeley), Chris Beckmann (Google), danah boyd (Microsoft), Stuart Card (PARC), Sheelagh Carpendale (Univ. of Calgary), Scott Carter (FXPAL), Daniel Chang (Stanford), Ed Chi (PARC), John Christensen (Stanford), Nicole Coleman (Stanford), Marc Davis (Yahoo!), Anind Dey (CMU), Mira Dontcheva (Adobe), Dan McFarland (Stanford), Sue Fussell (Cornell), Yuankai Ge (Stanford), Nathan Good (PARC), Kenji Hakuta (Stanford), Pat Hanrahan (Stanford), Marti Hearst (Berkeley), Joe Hellerstein (Berkeley), Jason Hong (CMU), Petra Isenberg (Univ. of Calgary), Sara Kiesler (CMU), Aniket Kittur (CMU), Nicholas Kong (Berkeley), James Landay (Univ. of Washington), Scott Lederer (Google), Wilmot Li (Adobe), Jock Mackinlay (Tableau), Chris Manning (Stanford), Jennifer Mankoff (CMU), Tara Matthews (IBM), Alan Newberger (Google), Jeff Pierce (IBM), Bryan Pendleton (CMU), George Robertson (Microsoft), Lynn Robertson (Berkeley), Noam Sagiv (Brunel), Shiwei Song (Stanford), Chris Stolte (Tableau), Bongwon Suh (PARC), Frank van Ham (IBM), Fernanda Viégas (IBM), Martin Wattenberg (IBM), Chris Weaver (Univ. of Oklahoma), Greg Wientjes (Stanford), Wesley Willett (Berkeley)

ADVISORS:

Maneesh Agrawala (Ph.D., Berkeley), James A. Landay (M.S., Berkeley)

ADVISEES AT STANFORD:

Michael Bostock, Jason Chuang, Hyung Suk Kim, Nam Wook Kim, Nathan Sakunkoo

Biographical Sketch: Joseph M. Hellerstein

Professor, EECS Computer Science Division
University of California, Berkeley
URL: http://db.cs.berkeley.edu/jmh

AREAS OF SPECIAL INTERESTS

Database systems, interactive data analysis, distributed systems, sensor networks.

PROFESSIONAL PREPARATION

Harvard University	Computer Science	A.B. (1990)
University of California, Berkeley	Computer Science	M.S. (1992)
University of Wisconsin, Madison	Computer Science	Ph.D. (1995)

APPOINTMENTS

2005 –	Professor, UCB EECS Dept.
2006 –	Technical Advisor, Swivel.com
2008 –	Technical Advisor, Greenplum
2003 – 2005	Director, Intel Research Berkeley
2000 – 2005	Associate Professor, UCB EECS Dept.
2001 – 2003	Consultant and Visiting Faculty, Intel Berkeley
1997 – 2001	Chief Scientist, Cohera Corporation
1995 – 2000	Assistant Professor, UCB EECS Dept.
1995	Postdoctoral Fellow, Hebrew University, Jerusalem
1990 – 1991	Pre-Doctoral Intern, IBM Almaden Research Center

FIVE MOST RELEVANT PUBLICATIONS:

1. J. M. Hellerstein, P. J. Haas, H. J. Wang. "Online Aggregation". In *Proc. ACM-SIGMOD*, 1997. *(Winner of the SIGMOD Test-of-Time award, 2007*).

2. J. M. Hellerstein, *et al*. "Interactive Data Analysis: The Control Project." *IEEE Computer*, August, 1999.

3. V. Raman and J. M. Hellerstein. "Potter's Wheel: An Interactive Framework for Data Cleaning and Transformation". In *Proc. VLDB*, 2001.

4. D. Z. Wang, E. Michelakis, M. N. Garofalakis, and J. M. Hellerstein. "BayesStore: Managing Large, Uncertain Data Repositories with Probabilistic Graphical Models". In *Proc. VLDB*, 2008.

5. J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein and C. Welton. "MAD Skills: New Practices for Big Data". In *Proc. VLDB*, 2009.

FIVE OTHER PUBLICATIONS:

1. J. M. Hellerstein and M. Stonebraker and J. Hamilton. "Architecture of a Database System". *Foundations and Trends in Databases* 1(2), 2007.

2. B. T. Loo, T. Condie, J. M. Hellerstein, P. Maniatis, T. Roscoe, and I. Stoica. "Implementing Declarative Overlays". In *Proc. 20th SOSP*, 2005.

3. S. Madden, M. Franklin, J. M. Hellerstein and W. Hong. "TinyDB: An Acquisitional Query Processing System for Sensor Networks." *ACM TODS*, 2005.

4. J. M. Hellerstein, E. Koutsoupias, D. Miranker, C. Papadimitriou, and V. Samoladas. "On a Model of Indexability and its Bounds for Range Queries." *Journal of the ACM*, 2002.

5. J. M. Hellerstein, J. F. Naughton, and A. Pfeffer. "Generalized Search Trees for Database Systems". *VLDB*, Zurich, September, 1995.

EDUCATIONAL ACTIVITIES

1. Co-designed and taught undergraduate course on Database System Implementation, 1997-2009.

2. Co-designed and taught graduate course on Advanced Topics in Computer Systems to combine operating systems and database systems viewpoints, 1999-2007.

3. Co-editor, *Readings in Database Systems*, Third Edition (Morgan Kaufmann, 2000) and Fourth Edition (MIT Press, 2005).

SYNERGISTIC ACTIVITIES

5 1. Advisory Board member: Swivel, LLC (2006 – present), Greenplum Corp (2006 – present).

1. Director of Intel Research Berkeley, 2003-2005.

2. Led the development of the Generalized Search Tree (GiST) extension to PostgreSQL, a key component in the PostGIS Geographic Information System.

3. Invited Keynote, *International Conference on Data Mining* 2007 on social data analysis.

4. Invited speaker, Computer Community Consortium symposium on Data-Intensive Computing.

5. Invited Guest Blogger on trends in "Big Data": Computing Community Consortium (2008), O'Reilly Radar (2008), GigaOM (2008).

6. Editor-in-Chief, *Foundations and Trends in Databases*, 2006 – present.

Alphabetical list of collaborators (last 48 months), students and advisors. Eric A. Brewer, UCB; Harr Chen, MIT; Kuang Chen, UCB; David Chu, UCB; Jeff Cohen, Greenplum; Tyson Condie, UCB; Amol Deshpande, U. Maryland; Brian Dolan, Fox Audience Network; Mark Dunlap, Evergreen Technologies; Michael J. Franklin, UCB; Minos Garofalakis, Tech University of Crete; David Gay, Intel; Johannes Gehrke, Cornell; Carlos Guestrin, CMU; Wei Hong, Intel; Ling Huang, UCB; Ryan Huebsch, UCB; Garrett Jacobsen, UCB; Ankur Jain, U. Washington; Shawn R. Jeffery, UCB; Michael I. Jordan, UCB; Anthony D. Joseph, UCB; Andreas Krause, CMU; Philip Levis, Stanford; Boon Thau Loo, UCB; Samuel R. Madden, MIT; Petros Maniatis, Intel; Alexandra Meliou, UCB; Eirinaios Michelakis, UCB; Tapan Parikh, UCB; Lucian Popa, UCB; Raghu Ramakrishnan, Yahoo! Research; Frederick Reiss, UCB; Timothy Roscoe, ETH Zurich; Mehul A. Shah, HP Labs; Scott Shenker, UCB; Arie Shoshani, Lawrence Berkeley Lab; Kurt Stockinger, Lawrence Berkeley Lab; Ion Stoica, UCB; Michael Stonebraker, MIT; Nina Taft, Intel Research; Arsalan Tavakoli, UCB; Daisy Zhe Wang, UCB; Wei Wang, UCB; Caleb Welton, Greenplum; Kesheng Wu, Lawrence Berkeley Lab; Aydan Yumerefendi, Duke.

ADVISORS

Jeffrey F. Naughton (Ph.D.), Michael R. Stonebraker (M.S.)

ADVISEES

Ph.D.: Marcel Kornacker, Vijayshankar Raman, Megan C. Thomas, Sam Madden, Mehul A. Shah, Amol Deshpande, Fred Reiss, Boon Thau Loo, Ryan Huebsch, Alexandra Meliou, Tyson Condie, David Chu, Kuang Chen, Peter Alvaro, Neil Conway

M.S.: Ashima Atul, Varun Kacholia, Ron Avnur, Mohanakrishna Lakhamraju, Owen Cooper