Data Visualization and Social Data Analysis

Jeffrey Heer
Stanford University
jheer@cs.stanford.edu

Joseph M. Hellerstein
UC Berkeley
hellerstein@cs.berkeley.edu

Background

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.

In order to produce real value from data, we must make sense of it. Such sensemaking – turning data sets into knowledge – is the basic motivation for query processing and data mining research. But beyond 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 subjective and contextualized human judgments about the meaning and significance of patterns in the data.

Visualization technologies have proven essential for helping people understand data, leveraging the human visual system to analyze large amounts of information. Visualization provides one means of combating information overload, as a well-designed visual encoding can supplant cognitive calculations with simpler perceptual inferences and improve comprehension, memory, and decision making. Visual representations may also help engage more diverse audiences in the process of analytic thinking.

In recent years, researchers and entrepreneurs have introduced online services for data collection and analysis, developing interactive visualizations that enable mass interaction with data. These sites represent the first step in what looks to be a growing online phenomenon: *social data analysis*, that is, collective analysis of data supported by social interaction. By engaging crowds of both experts and non-experts in the process of data exploration, new data analysis applications are being developed in areas ranging from political transparency to business intelligence to citizen science. Achieving this vision, however, will require further innovation in the design of systems and user interfaces for collaborative data management.

VLDB '09, August 24-28, 2009, Lyon, France

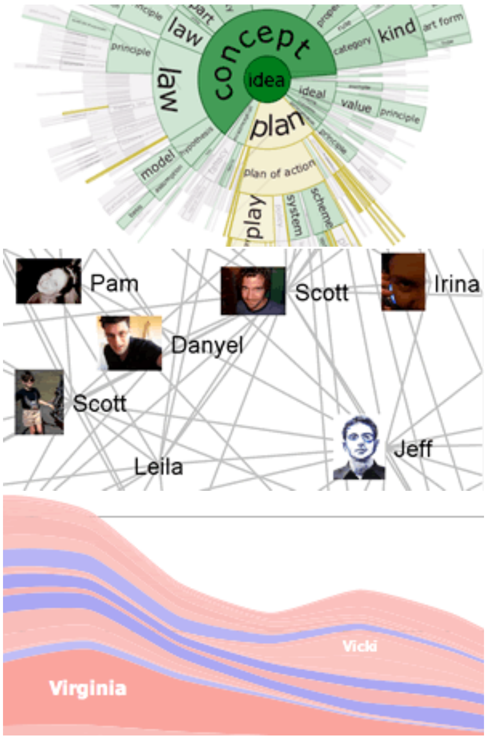


Figure 1: Data visualizations for (a) word relations in WordNet, (b) online social networks and (c) time-series analysis of name popularity. Attendees will learn how to evaluate the use of visual encodings of data.

Tutorial

In this tutorial we begin by surveying techniques and algorithms for creating effective visualizations based on principles from graphic design, perceptual psychology, and cognitive science. We examine how visualization technology can be applied to support data analysis and sensemaking. We discuss techniques for integrating visualization with query languages and large-scale data processing. Finally, we discuss emerging developments in social data analysis.

The intended audience is database researchers and practitioners who are interested in understanding visualization theory and techniques. No prior knowledge of visualization is assumed, only familiarity with data modeling and management techniques. The learning goals of the tutorial are for attendees to:

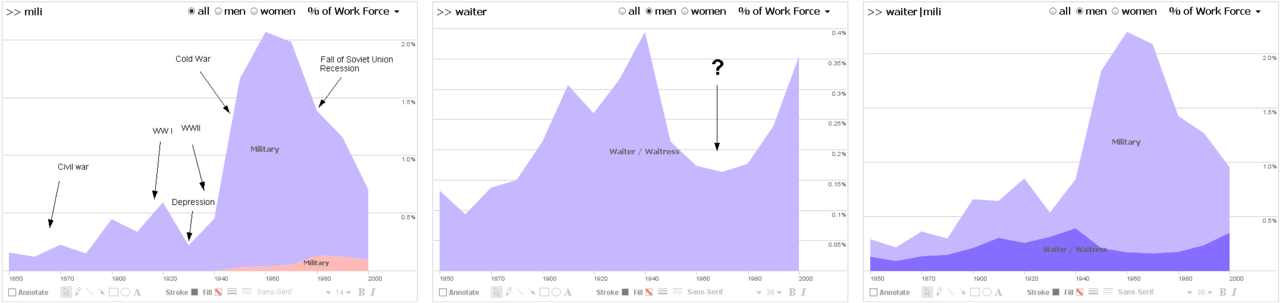


Figure 2: Collaborative sensemaking in the sense.us system. (a) Multiple users debated the causes of military build-up while (b) another noted a drop and subsequent rise in male waiters in the late 1900's. (c) A different participant sought to connect the two, noting an inverse correlation between the occupations.

1. Gain an overview of the landscape of visualization research, including recent developments as well as milestone achievements in visualization theory and techniques.

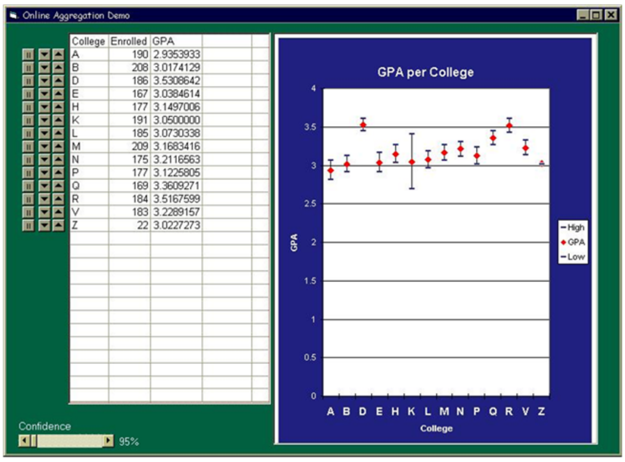
2. Be able to evaluate visualization techniques for application in data analysis and management tasks.

3. Gain perspective on current synergies in database and visualization/HCI research, with an eye towards further cross-disciplinary fertilization and collaboration.

Speakers

Jeffrey Heer is an Assistant Professor of Computer Science at Stanford University, where his research focuses on human-computer interaction, interactive visualization, and social computing. His work has produced novel visualization techniques for exploring data, software tools that simplify visualization creation and customization, and collaborative analysis systems that leverage the insights of multiple analysts. He has guided the design of the prefuse, flare, and protovis open-source visualization toolkits, currently in use by the visualization research community and numerous corporations. Over the years, he has also worked at Xerox PARC, IBM Research, Microsoft Research, and Tableau Software. He holds B.S., M.S., and Ph.D. degrees in Computer Science from the University of California, Berkeley.

Joseph M. Hellerstein is a Professor of Computer Science at the University of California, Berkeley, whose research focuses on data management and distributed systems. His work has been recognized via awards including an Alfred P. Sloan Research Fellowship, MIT Technology Review's inaugural TR100 list, and two ACM-SIGMOD "Test of Time" awards – including one for his work on Online Aggregation to support interactive visualization of large-scale data processing. He has served as Director of Intel Research Berkeley and Chief Scientist at Cohera Corporation, and now advises a number of companies, including Swivel, a social data visualization website, and Greenplum, a parallel database system vendor.



College	Enrolled	GPA
A	190	2.9353933
B	208	3.0174129
D	186	3.5308642
E	167	3.0384614
H	177	3.1497006
K	191	3.0500000
L	185	3.0730338
M	209	3.1683416
N	175	3.2116563
P	177	3.1225805
Q	169	3.3609271
R	184	3.5167599
V	183	3.2289157
Z	22	3.0227273

Figure 3: An online aggregation interface for interactive visualization of a long-running approximate query, computing average GPAs for students enrolled in various colleges of a university. As sample sizes increase over time, rows in the answer table are updated, the corresponding dots in the graph oscillate up and down more slowly, and interval boundaries shrink nearly monotonically to invisible width. Widgets on the left give users intuitive, fine-grained control over statistical tradeoffs between processing time and estimation accuracy. By manipulating the slider on the lower left, users can understand the connection between the probability of the intervals and their width: moving the slider to the right will cause the intervals on the graph to grow accordingly.