

CHAPTER TWELVE

The Design of Sense.us

Jeffrey Heer

I MUST CONFESS THAT I DON'T BELIEVE IN BEAUTIFUL DATA. AT LEAST NOT WITHOUT CONTEXT.

Prior to World War II, the government of the Netherlands collected detailed civil records cataloging the demographics of Dutch citizenry. A product of good intentions, the population register was collected to inform the administration of government services. After the German invasion, however, the same data was used to effectively target minority populations (Croes 2006). Of the approximately 140,000 Jews that lived in the Netherlands prior to 1940, only about 35,000 survived.

Though perhaps extreme, for me this sobering tale underscores a fundamental insight: the “beauty” of data is determined by how it is used. Data holds the potential to improve understanding and inform decision-making for the better, thereby becoming “beautiful” in action. Achieving value from data requires that the right data be collected, protected, and made accessible and interpretable to the appropriate audience. The fiasco in which AOL released insufficiently anonymized search query data is a recent failure of protection.

Fortunately, most examples are not nearly as tragic as these tales. A more common occurrence is data wasting away: collected and stored in data warehouses—sometimes at great infrastructural cost—but left underutilized. For companies and governments alike, languishing data represents a lost opportunity and poor return on investment. The value of data is proportional to people’s ability to extract meaning and inform action.



Somewhat paradoxically then, some data collections possess more (potential) beauty than others. Clearly the choice of data to collect and the design of storage infrastructures, schemas, and access mechanisms shape the potential of data to inform and enlighten while avoiding harm. However, the “last mile” in this climb toward beauty is the problem of human-information interaction: the means by which data is presented to and explored by people to support analysis and communication.

This chapter presents a case study on the use of interactive visualization to help foster beautiful data-in-action: the design of sense.us, a web application for collaborative exploration and sense-making of 150 years of United States census data. I will cover the steps we took in taking a large, government-collected data set—the U.S. census—and making it accessible to a general audience through a suite of interactive visualizations. I will also describe sharing and discussion mechanisms we devised to engage a community of data voyagers in social interpretation and deliberation. Our goal was to realize the potential beauty of data by fostering collective data analysis.

Visualization and Social Data Analysis

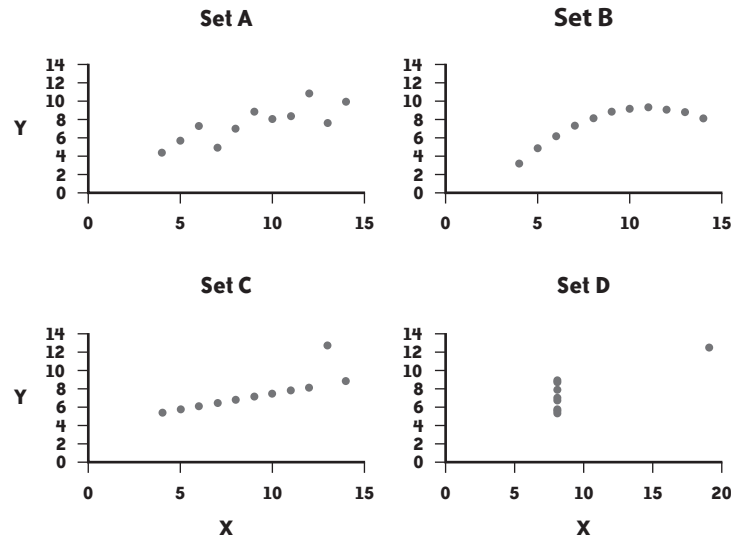
Visualizations are regularly used to construct meaning from data by facilitating comprehension, enabling exploration, and communicating findings. A large part of the human nervous system has evolved to process visual information; in the human brain, over 70% of the receptors and 40% of the cortex are implicated in vision processing (Ware 2004). Visualization design leverages the capabilities of this visual processing system to enable perception of the trends, patterns, and outliers residing within data.

Note that this is not an issue of crafting “fancy graphics.” Often a simple table or bar chart (sans 3-D frills and specular highlights) can provide an effective presentation. The trick is choosing the right visual representation(s) for the data and tasks at hand.

An instructive example is Anscombe’s Quartet, a collection of four data sets created by the statistician Francis Anscombe to illustrate the importance of visualizing data (Anscombe 1973). Each data set appears identical according to common descriptive statistics (Figure 12-1). However, plotting the data immediately reveals salient differences between the sets.

Other writers detail the ways in which effective visual design aids interpretation, communication, and decision-making. Tufte (1997) famously argues that the disaster of the space shuttle *Challenger* might have been avoided had engineers created better visual depictions of rocket damage data (though this is not without some controversy; see Robison et al. 2002).

Visualization researchers have catalogued the space of “visual variables”—such as position, length, area, shape, and color hue—that can be used to encode data in a visual display (Bertin 1967, Card et al. 1999). They have also studied how accurately humans decode these visual variables when applied to different data types, such as categorical (names), ordinal (rank-ordered), and quantitative (numerical) data (Cleveland & McGill 1984, Ware 2004). For instance, spatial position, as used in a bar chart or scatterplot, facilitates decoding




Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary statistics		Linear regression	
$\mu_X = 9.0$	$\sigma_X = 3.317$	$Y = 3 + 0.5X$	
$\mu_Y = 7.5$	$\sigma_Y = 2.03$	$R^2 = 0.67$	

FIGURE 12-1. Anscombe's Quartet, a collection of statistically similar data sets illustrating the use of visualization to aid understanding.

for each of these data types, while color hue ranks highly when used for category labels but poorly when used to convey quantitative values.


In this spirit, most visualization research focuses on the perceptual and cognitive aspects of visualization use, typically in the context of single-user interactive systems. In practice, however, visual analysis is often a social process. People may disagree on how to interpret data and may contribute contextual knowledge that deepens understanding. As participants build consensus, they learn from their peers. Moreover, some data sets are so large that thorough exploration by a single person is unlikely. This suggests that to fully support sense-making, visualizations should also support social interaction.



These observations led myself and my colleagues Martin Wattenberg and Fernanda Viégas of IBM Research to investigate how user interfaces for visualizing data might better enable the “social life of visualization.” We embarked on a research project in which we designed and implemented a website, *sense.us*, aimed at group exploration of demographic data. The site provides a suite of visualizations of United States census data over the last 150 years, coupled with collaboration mechanisms to enable group-oriented data analysis.

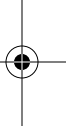
In the rest of this chapter, I will share the design process for *sense.us*: how we selected and processed the data, developed a suite of visualizations, and designed collaboration features to enable social data analysis. I conclude by looking at the ways people worked together through the system to construct insight from the data.

Data



Martin, Fernanda, and I came to this project in the mindset of researchers: we wanted to understand how best to support social interaction in the visual analysis process. Our choice of data set was not predetermined, though it was clear that a good data domain would satisfy some specific properties: we wanted a large, real-world data set, relevant to a general audience, and rich enough to warrant many different analyses. According to these criteria, census data seemed ideal. I had also long been interested in making census data more publicly accessible: I believe it is an important lens through which we might better understand ourselves and our history.

I started by rummaging through the U.S. census bureau’s website (<http://www.census.gov>). This proved only mildly productive. The census bureau provides a number of data sets at various levels of aggregation (e.g., by zip code, metro area, region), but this rich data is only available for recent census decades. I also realized that I was in a bit over my head. I had much to learn about the ins and outs of how census data has been collected and modeled over the decades. For example, the questions and categories used by the census bureau have evolved over the decades, meaning that even if one has data for every year, it does not guarantee that the data can be easily compared.



In general, one should not dive into visualization design before gaining at least a basic familiarity with the data domain. So my next step was to meet with domain experts: my colleagues in the Sociology and Demography departments at UC Berkeley, where I was a graduate student at the time. Through these discussions I gained a deeper appreciation of how the census works and which data sources demographers use to study the population. In the process, I was introduced to a valuable resource: the Integrated Public Use Microdata Series (IPUMS) databases maintained by the University of Minnesota Population Center (<http://www.ipums.org>).

The IPUMS-USA database consists of United States census data from 1850 to 2000. Data from each decade in this period is included, with the exception of 1890, the records for which were destroyed by a fire in the Commerce Building in 1921. For each decade, the IPUMS data consists of representative sample data, either a 1% or 5% sample of that decade’s census records.

Each record represents a characteristic person sampled from the population. In some cases, persons and households with certain characteristics are over-represented, and so different weights are associated with the records.

What makes this database particularly attractive is that the IPUMS project has developed uniform codes and documentation for all demographic variables, facilitating analysis of change over time. This “harmonization” is a monumental service, enabling comparative analyses and, by extension, new insights. However, the process of fitting disparate data into a shared schema inevitably introduces artifacts, an issue that will surface again later.

All told, the IPUMS-USA database contains 413 demographic variables, ranging over common categories such as gender, age, race, marital status, and occupation, down to estimates of how many households have washers, dryers, flush toilets, and televisions (Figure 12-2). In many cases, variables are recorded only in a subset of decades; in other years, the variables simply were not measured.

Detailed Version		General Version	Variable	Label	Case Selection	Attach Variables	1850	1860	1870	1880	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
<input type="checkbox"/>	<input type="checkbox"/>		RELATE	Relationship to household head	<input type="checkbox"/>	<input type="checkbox"/>	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
<input type="checkbox"/>	<input type="checkbox"/>		IMPEL	Imputed relationship to household head	<input type="checkbox"/>	<input type="checkbox"/>	X	X	X	X	X	X	X	X							
<input type="checkbox"/>	<input type="checkbox"/>		AGE	Age	<input type="checkbox"/>	<input type="checkbox"/>	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
<input type="checkbox"/>	<input type="checkbox"/>		SEX	Sex	<input type="checkbox"/>	<input type="checkbox"/>	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
<input type="checkbox"/>	<input type="checkbox"/>		MARST	Marital status	<input type="checkbox"/>	<input type="checkbox"/>				X	X	X	X	X	X	X	X	X	X	X	X
<input type="checkbox"/>	<input type="checkbox"/>		AGEMONTH	Age in months	<input type="checkbox"/>	<input type="checkbox"/>	X	X	X	X	X	X	X	X							
<input type="checkbox"/>	<input type="checkbox"/>		BIRTHMO	Month of birth	<input type="checkbox"/>	<input type="checkbox"/>			X	X	X	X			X						
<input type="checkbox"/>	<input type="checkbox"/>		BIRTHQTR	Quarter of birth	<input type="checkbox"/>	<input type="checkbox"/>			X	X					X	X	X	X			
<input type="checkbox"/>	<input type="checkbox"/>		BIRTHYR	Year of birth	<input type="checkbox"/>	<input type="checkbox"/>					X	X									
<input type="checkbox"/>	<input type="checkbox"/>		AGEMARR	Age at first marriage	<input type="checkbox"/>	<input type="checkbox"/>							X	SL			X	F1	X		
<input type="checkbox"/>	<input type="checkbox"/>		DURMARR	Duration of current marital status	<input type="checkbox"/>	<input type="checkbox"/>				X	X					SL					
<input type="checkbox"/>	<input type="checkbox"/>		MARRNO	Times married	<input type="checkbox"/>	<input type="checkbox"/>					X				SL	SL	X	F1	X		
<input type="checkbox"/>	<input type="checkbox"/>		MARRIN1Y	Married within the past year	<input type="checkbox"/>	<input type="checkbox"/>	X	X	X	X					SL	SL	X	F1	X		
<input type="checkbox"/>	<input type="checkbox"/>		MARRMONTH	Month married	<input type="checkbox"/>	<input type="checkbox"/>			X												
<input type="checkbox"/>	<input type="checkbox"/>		MARRQTR	Quarter of first marriage	<input type="checkbox"/>	<input type="checkbox"/>											X	F1	X		
<input type="checkbox"/>	<input type="checkbox"/>		WIDOW	Marriage ended by death	<input type="checkbox"/>	<input type="checkbox"/>														F1	X
<input type="checkbox"/>	<input type="checkbox"/>		CHBORN	Children ever born	<input type="checkbox"/>	<input type="checkbox"/>				X	X				SL	SL	X	X	X	X	
<input type="checkbox"/>	<input type="checkbox"/>		CHSURV	Children surviving	<input type="checkbox"/>	<input type="checkbox"/>				X	X										
<input type="checkbox"/>	<input type="checkbox"/>		All Demographic Variables																		

FIGURE 12-2. An excerpt of the available demographic variables in the IPUMS-USA database.

The motto of the IPUMS project is “use it for good, never for evil.” Fortunately, the enforcement of this maxim extends beyond the obligatory checkbox one must click when downloading a data extract. To protect individual privacy, the availability of some data has been restricted. For example, religious affiliation is not included, and the availability of detailed geographic data is highly limited, particularly for low-population areas.

We decided to use IPUMS-USA as the primary data source for sense.us. Using the IPUMS web interface, we first selected samples for the years 1850–2000 (excluding 1890) and then selected a set of variables to extract. The vast majority of variables are only available for a subset of census decades. To enable visual exploration of long-term change, we selected the variables that were available for at least a century. This set consisted of 22 variables, including age, sex, marital status, birthplace (either a U.S. state/territory or a foreign country), occupation, race, school attendance, and geographic region. Due to privacy constraints, geographic data was limited to coarse-grained regions such as New England and the west coast. The resulting data extract was a 520-megabyte GZIP file that decompressed into a 3.3-gigabyte text file.

I will largely spare you the details of what happened next. A straightforward yet tedious process of data processing, cleaning, and import ensued, ultimately resulting in a MySQL database containing the census data extract in queryable form. To facilitate analysis, we organized the data using a star schema (http://en.wikipedia.org/wiki/Star_schema): we stored the census measures in a large fact table containing a column for each demographic variable, with compact keys used to indicate categorical variable values. A collection of dimension tables then stored the text labels and descriptions for the values taken by each demographic variable.

This setup provided a base for conducting exploratory analysis. We generated data summaries by issuing queries that “rolled up” the data along chosen dimensions. For example, we could isolate the relationships between age, gender, and marital status by summing the number of people across all the other variables. In short, we had a foundation from which we could explore the data and prototype visualizations.

Visualization

Given the size and scope of the census data, we realized early on that trying to fit all the data within a single visualization design would be a recipe for disaster. Where we could, we wanted to boil down the data into the simplest forms that could support a range of analyses. As we were designing for a general audience, we settled on the approach of creating a collection of visualizations that present selected slices of the data. In essence, we wanted to make our visualizations as simple as possible while remaining useful, but no simpler.

Our design philosophy thus required that we figure out which data dimensions would be of greatest interest and which visualization designs and interaction techniques would best support active exploration of those dimensions. To do this, we began simultaneously exploring the data itself and the space of visualization designs.

Before crafting an interface to help others explore data, I wanted to ensure that the data was interesting enough for others to even bother. I used a number of methods to conduct my exploration, including SQL queries, Excel, and visualization systems. The most useful tool was Tableau, a database visualization system. Using Tableau, one can map database fields to visual encodings in a drag-and-drop fashion; the application then queries the database and visualizes the result. (Full disclosure: I have worked as a consultant for Tableau Software.) We were thus able to prototype a number of different approaches (for example, Figure 12-3). We generated a large collection of visualizations and shared them with our colleagues to collect feedback. The ability to quickly evaluate visualization approaches using the actual data saved us countless hours of experimentation and allowed us to conserve our development efforts for our final system design.

As I explored the data, I kept track of the interesting trends, patterns, and outliers I discovered. In some cases, interesting stories were found nested within a combination of dimensions. For example, stratifying marital status by both age and gender over time revealed

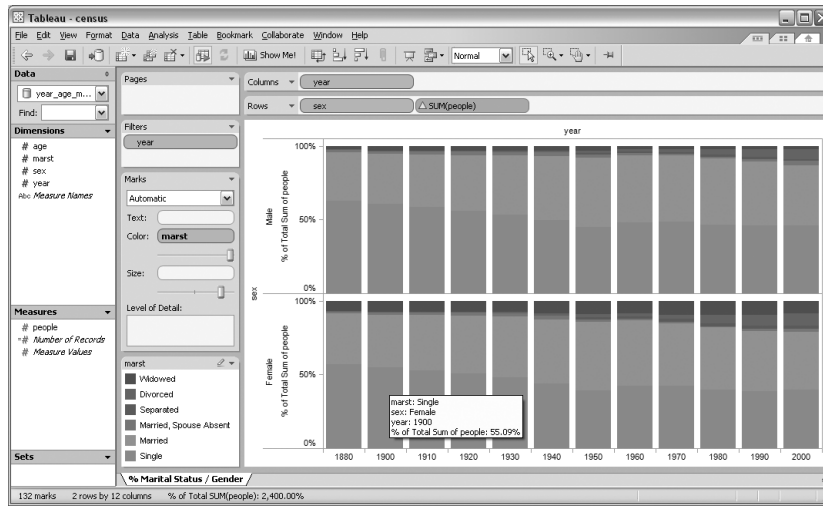


FIGURE 12-3. A prototype visualization built using Tableau showing the distribution of marital status over multiple decades. (See Color Plate 36.)

that the average age at which females (but not males) first get married has increased by about five years over the last century. In other cases, a single type of data plotted over time revealed a number of interesting stories, such as the wax and wane of farmers in the labor force and different waves of immigration from around the globe. I often found it useful to transform the data, alternately viewing the data as absolute population numbers or as percentages within a census decade.

Design Considerations




Prototypes in hand, Martin, Fernanda, and I then collaboratively designed the interactive visualizations for sense.us. To do so, we first outlined a set of design considerations.

Foster personal relevance

If no one cares about the data, no one will explore it. We hypothesized that familiar dimensions such as geography and time enable users to quickly look for themselves (or people like them) in the data and form narratives. Given that the geographic data available to us was limited, we focused on the presentation of data over time. We also tried to use interaction techniques that let users quickly search for data records of interest to them, such as particular occupations or countries of origin.

Provide effective visual encodings


Naturally, we wanted to use visual encodings that would facilitate comprehension of the data. In some cases, this task is straightforward: if I want to examine the correlation between two numerical values, I would be hard pressed to find a better representation than a scatterplot. In this case, we had to balance a number of trade-offs.




A common way to visualize change over time is to use a line graph. However, displaying over 200 occupations in a line graph results in a cluttered mess of occluding lines. We instead chose stacked graphs, which visually sum multiple time series by stacking them on top of one another (see Figures 12-4 and 12-5, later in this chapter). Our choice was influenced by Martin's Baby Name Voyager visualization, a stacked graph of baby name popularity that became surprisingly popular online (Wattenberg and Kriss 2006). Stacked graphs show aggregate patterns clearly and comfortably support interactive filtering, but do so at the cost of obscuring individual trends—perception of a trend is biased by the contour of the series stacked beneath it. In response, we ensured that clicking a series would filter the display so that the trend could easily be viewed in isolation.

Furthermore, we followed established cultural conventions to encode data in a fashion familiar to many viewers (e.g., blue for boys, pink for girls). When considering how to visualize the interaction between a number of demographic variables, rather than try to invent something completely novel, we instead augmented a chart type already in common use by demographers: the population pyramid.

Make each display distinct



In some instances we used the same visualization type to show different data types. For example, we used stacked graphs for both occupation and birthplace data. However, we wanted to make each display visually distinct, so that users could recognize them at a glance. Consequently, we constructed a unique color palette for each visualization.



Support intuitive exploration

To foster interactive exploration, we wanted to make manipulating the interface as simple as possible. For stacked graphs, we let users type keyword searches to query for items of interest. In other cases, we provided a collection of drop-down menus to select or filter dimensions. We also included controls for selecting between absolute people counts and normalized percentages. Although more advanced manipulations are possible, we found that providing this level of control enabled a range of exploration with an uncluttered, easy-to-learn interface.

Be engaging and playful

In addition to fostering personal relevance, we wanted interaction with our system to be engaging and enjoyable. We thus strove for an aesthetic as well as effective presentation of data. We designed interaction techniques and animated transitions to promote a feeling of responsiveness and dynamism, but we did not want to take such stylistic features too far. We wanted to enhance, not disrupt, data exploration. We varied animation styles and timing until our designs “felt right.” An animation duration of ~1 second provided transitions that viewers could follow without slowing down the analysis process.

Visualization Designs

By first engaging in data exploration of our own, we were able to determine the data dimensions we found most interesting. We applied these observations in conjunction with our design considerations to design a suite of visualizations of the census data.

Job Voyager

The Job Voyager is a stacked graph showing the composition of the U.S. labor force over the last 150 years (Figure 12-4). Each series represents an occupation, subdivided by sex: blue indicates male, pink indicates female. Users can explore the data by clicking on a series to show only the corresponding occupation, or by typing keyword queries to filter out jobs that do not match the query. We also included drop-down menus to filter by sex and to switch between views of absolute people count and percentage of the labor force. These operations support exploration of both aggregate trends (e.g., the influx of women into the labor force after World War II) and individual patterns (e.g., the rise and fall of locomotive engineers).

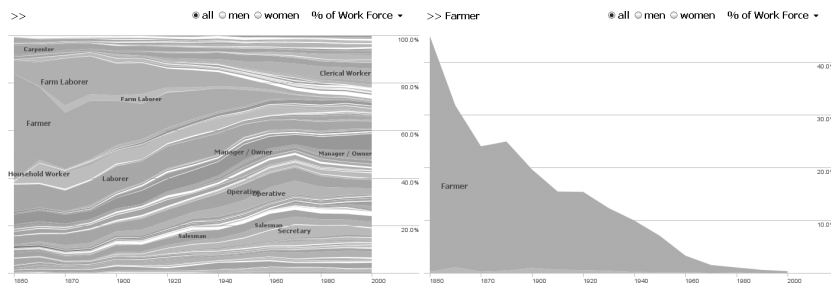


FIGURE 12-4. Job Voyager visualization: (left) an overview showing the constitution of the labor force over 150 years, and (right) a filtered view showing the percentage of farmers. (See Color Plate 37.)

We quickly realized that coloring the series solely on gender was not enough. When we filtered the view to show only males or only females, it became difficult to differentiate individual series. One solution is to enable perceptual discrimination by varying color saturation in an arbitrary fashion. Martin then suggested a clever variation on this approach: rather than vary colors arbitrarily, do so in a meaningful, data-driven way. We subsequently varied color saturation according to socio-economic index scores for each occupation. Thus a series with a higher median income was drawn darker. In practice, this encoding worked well to improve identification of different occupations without adding misleading or meaningless visual features to the display.

Birthplace Voyager

The Birthplace Voyager is similar in design to the Job Voyager, but instead shows the birthplace of U.S. residents in each census year (Figure 12-5). The recorded birthplaces are either U.S. states and territories or foreign countries. The interactive controls enable filtering by keyword query or by continent and the display of both absolute counts and percentages.

The visualization supports investigation of immigration trends across the world (e.g., past waves of immigration from Europe and current waves of immigration from Latin America) and within the U.S. (e.g., the changing proportion of residents in each state).

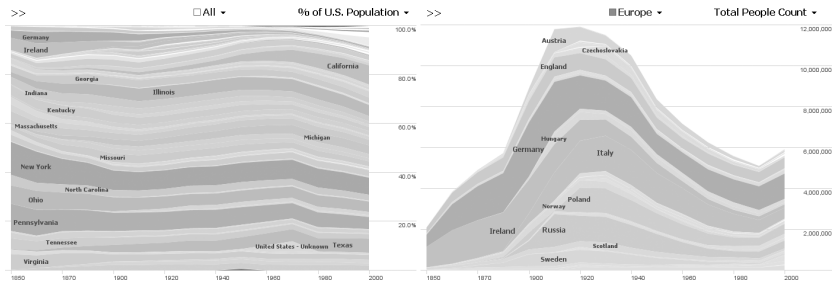


FIGURE 12-5. Birthplace Voyager visualization: (left) an overview showing the distribution of birthplaces over 150 years and (right) a filtered view showing the total number of European immigrants. (See Color Plate 38.)

With the Birthplace Voyager, we encountered similar coloring issues as before. In this case, we assigned color hues according to continent, plus an extra dedicated hue for U.S. states. We then experimented with different means of varying color saturation until we settled on using the total number of people born in the state or country across all time slices as the backing data.

U.S. census state map and scatterplot

While the timelines of the Job and Birthplace Voyagers were designed to engage viewers in historical narrative, we wanted to include more conventional views as well. We provided a colored state map for viewing the distribution of demographic variables for each state. In an annotated map of population change for states between 2000 and 2005 (Figure 12-6, left), one can see substantial growth in the southwest while the population of North Dakota has decreased. We also provided a scatterplot display to examine potential correlations between variables. Users can map demographic variables to the x position, y position, and size of circles representing the U.S. states. For example, one may note a correlation between household income and retail sales across states (Figure 12-6, right). The backing data for these views includes additional statistics we downloaded from the U.S. census bureau website.

Population pyramid

Our most sophisticated visualization was an interactive population pyramid, designed to facilitate exploration of multiple demographic variables at once (Figure 12-7). Population pyramids (sometimes called “age-sex” pyramids) are a chart type introduced to us by our colleagues in demography. The pyramid is divided by a vertical axis into two halves: one for males, one for females. The y-axis represents age, often grouped into 5- or 10-year bins, and the x-axis represents the total number of people in that age group, with values

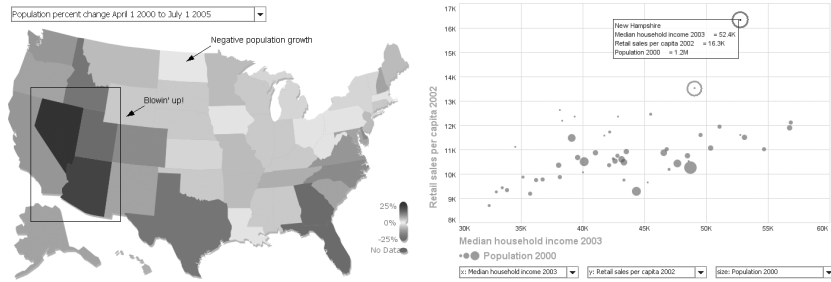


FIGURE 12-6. (Left) Interactive state map showing changes in each state's population from 2000 to 2005, and (right) scatterplot of U.S. states showing median household income (x-axis) versus retail sales (y-axis); New Hampshire and Delaware have the highest retail sales. (See Color Plate 39.)

for males increasing in one direction and values for females increasing in the opposite direction (Figure 12-7, left). The pyramid's shape communicates population dynamics: a steeply tapering pyramid indicates higher mortality rates than a more cylindrical shape.

We created an interactive pyramid incorporating demographic variables in addition to age and sex: geographic region, race, marital status, school attendance, and income level. By default, the two sides of the pyramids split the data by sex. We relaxed this restriction and added drop-down menus with which users could select a demographic variable and map two values to the sides of the pyramid. For example, users looked at geographic region, placing the west coast on one side of the pyramid and New England on the other.

We also introduced a color-encoding menu: selecting a demographic variable turns the pyramid sides into stacked graphs depicting the prevalence of values. For example, users can stratify the pyramid by school attendance to see what segments of the population were in school (Figure 12-7, right). We chose a distinct color palette for each variable, relying on existing cultural conventions where possible (e.g., blue=male, pink=female) and using ColorBrewer (<http://colorbrewer.org>) to determine color choices for the other cases. We used a gray color for bands representing missing or unknown values.

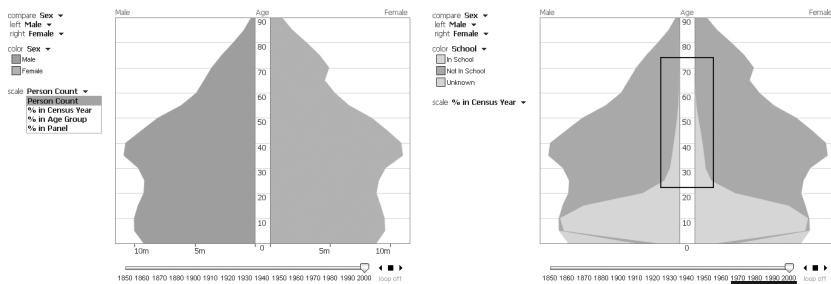


FIGURE 12-7. Population pyramid visualization: (left) a comparison of the total number of males and females in each age group in 2000, and (right) the distribution of school attendees in 2000 (an annotation highlights the prevalence of adult education). (See Color Plate 40.)

A timeline beneath the pyramid enables temporal exploration across census decades, and a playback feature animates changes to the pyramid over time. For example, animating population change over time dramatically shows the baby boom rippling through society in the post-war period. The mixture of layered colors and bubbling animation led users to endearingly rename our population pyramid “Georgia O’Keefe’s lava lamp.”

Finally, we included support for four data measures: total people count, percentage within decade, percentage within panel (useful when the two sides of the pyramid are disproportionate), and percentage within age group (to explore proportional differences across ages). Our own explorations found that each measure helps reveal specific stories. For example, viewing percentage within age group shows that elderly men are more likely to be married than elderly women, presumably because on average women live longer and become widows.

Implementation details

We implemented each visualization as a Java applet so we could embed it on a web page. We chose Java over Flash partially for performance reasons, but mostly due to the availability of visualization frameworks in Java at the time. The stacked graphs and population pyramid were built using the open source *prefuse* toolkit (<http://prefuse.org>). Backing each visualization is a flat text file extracted from our census database. In the case of the population pyramid, we created one file for every possible combination of demographic variables, precomputing all the relevant projections of the data. This approach eliminated the need for data processing on the server, and resulted in a very manageable storage footprint: though we started with a database of over 3 gigabytes, the final deployed data was reduced to little more than 3 megabytes! Of course, this approach does have limitations: it impedes users from exploring novel combinations of demographic variables and complicates the introduction of future visualizations requiring server-side data processing.

Collaboration

We then created the *sense.us* website, which couples the visualizations with collaborative analysis mechanisms (see Figure 12-8). In the left panel is the visualization applet (Figure 12-8a) and annotation tools (Figure 12-8b). The right panel provides a graphical bookmark trail (Figure 12-8c), providing access to views saved by the user, and a discussion area (Figure 12-8d and e), displaying commentary associated with the current view. We augmented the visualizations with a set of collaboration features, described in detail later: view sharing, doubly linked discussions, graphical annotations, bookmark trails, and social navigation via comment listings and user activity profiles.

View Sharing

When collaborating around visualizations, we reasoned that participants must be able to see the same visual environment in order to ground one another’s actions and comments. To this aim, *sense.us* provides a mechanism for bookmarking views. We tried to make

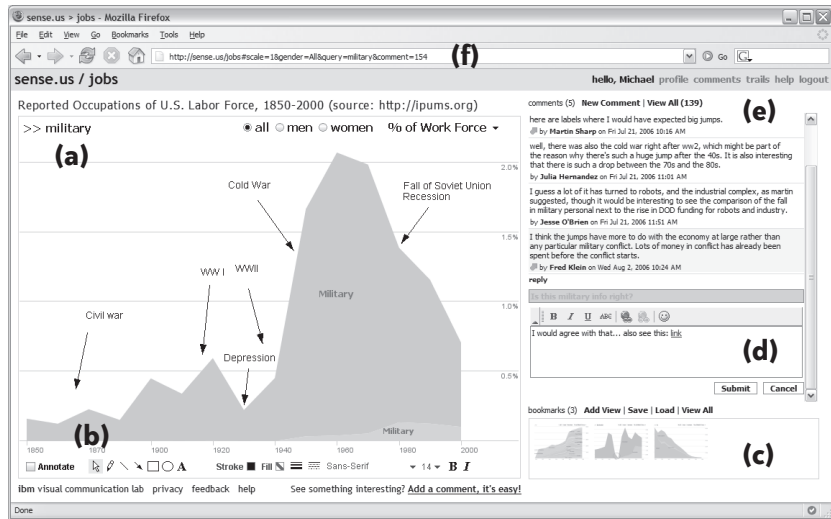


FIGURE 12-8. The sense.us collaborative visualization system: (a) An interactive visualization applet, with a graphical annotation for the currently selected comment. The visualization is a stacked time-series visualization of the U.S. labor force, broken down by gender. Here, the percentage of the workforce in military jobs is shown. (b) A set of graphical annotation tools. (c) A bookmark trail of saved views. (d) Text-entry field for adding comments. Bookmarks can be dragged onto the text field to add a link to that view in the comment. (e) Threaded comments attached to the current view. (f) URL for the current state of the application. The URL is updated automatically as the visualization state changes. (See Color Plate 41.)

application bookmarking transparent by tying it to conventional web bookmarking. The browser's location bar always displays a URL that links to the current state of the visualization, defined by the settings of filtering, navigation, and visual encoding parameters. As the visualization view changes, the URL updates to reflect the current state (Figure 12-8f), simplifying the process of sharing a view through email, blogs, or instant messaging by enabling users to cut-and-paste a link to the current view at any time. To conform to user expectations, the browser's back and forward buttons are tied to the visualization state, allowing easy navigation to previously seen views.

Doubly Linked Discussion

To situate conversation around the visualization, we created a technique we call *doubly linked discussion*. The method begins with an independent discussion interface in which users can attach comments to particular states (or views) of a visualization. Comments are shown on the right side of the web page and grouped into linear discussion threads (Figure 12-8e). Each comment shows the thread topic, comment text, the author's full name, and the time at which the comment was authored. Clicking on a comment takes the visualization to a bookmarked state representing the view seen by the comment's author.

Users can add comments either by starting a new thread or posting a reply to an existing thread. When a "New Comment" or "Reply" link is clicked, a text editor appears at the site where the comment will be inserted (Figure 12-8d) and the graphical annotation tools

(discussed next) become active. Upon submission, the comment text and any annotations are sent to the server and the comment listing is updated.

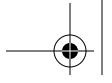
The interface just described is based on links from the commentary into the visualization. Our system also provides links in the other direction: from the visualization into the discussion. As users change parameters and views in the visualization, they may serendipitously happen upon a view that another person has already commented on. When this occurs, the relevant comments will automatically appear in the righthand pane. Our intuition was that this “doubly linked” discussion interface, which combines aspects of independent and embedded discussion, would facilitate grounding and enable the visualization itself to become a social place.

We quickly realized that our bookmarking mechanism was not sufficient to support doubly linked discussions. To see the challenge in linking from a view state back to all comments on that view, consider the visualization in Figure 12-8. When a user types “military” into the top search box (Figure 12-8f), he sees all jobs whose titles begin with the string “military.” On the other hand, if he types only “mili,” he sees all titles beginning with “mili”—but this turns out to be the identical set of jobs. These different parameter settings result in different URLs, and yet provide exactly the same visualization view. More generally, parameter settings may not have a one-to-one mapping to visualization states. To attach discussions to views, we therefore need an indexing mechanism that identifies visualization states that are equivalent despite having different parametric representations.

We solve this indexing problem by distinguishing between two types of parameters: filter parameters and view parameters. Filter parameters determine which data elements are visible in the display. Rather than index filter parameters directly, we instead index the filtered state of the application by noting which items are currently visible, thereby capturing the case when different filter parameters give rise to the same filtered state. View parameters, on the other hand, adjust visual mappings, such as selecting a normalized or absolute axis scale. Our current system indexes the view parameters directly. The bookmarking mechanism implements this two-part index by computing a probabilistically unique SHA-1 hash value based on both the filtered state and view parameters. These hash values are used as keys for retrieving the comments for the current visualization state.

Pointing via Graphical Annotation

In physical collaborations, people commonly use both speech and gesture, particularly pointing, to refer to objects and direct conversation. In the distributed, asynchronous context of the Web, graphical annotations can play a similar communicative role. We hypothesized that graphical annotations would be important for both pointing behavior and playful commentary. To add a pictorial element to a comment or point to a feature of interest, authors can use drawing tools (Figure 12-8b) to annotate the commented view.



These tools allow free-form ink, lines, arrows, shapes, and text to be drawn over the visualization view. The tools are similar to presentation tools such as Microsoft PowerPoint and are intended to leverage users' familiarity with such systems.

Comments with annotations are indicated by the presence of a small icon to the left of the author's name in the comment listing (see Figure 12-8e). When the mouse hovers over an annotated comment, the comment region highlights in yellow and a hand cursor appears. Subsequently clicking the region causes the annotation to be shown and the highlighting to darken and become permanent. Clicking the comment again (or clicking a different comment) will remove the current annotation and highlighting.

The graphical annotations take the form of vector graphics drawn above the visualization. When a new comment is submitted, the browser requests the current annotation (if any) from the visualization applet. The annotation is saved to an XML format, which is then compressed using gzip and encoded in a base-64 string representation before being passed to the browser. When comments are later retrieved from the server, the encoded annotations are stored in the browser as JavaScript variables. When the user requests that an annotation be displayed, the encoded annotations are passed to the applet, decoded, and drawn.

We refer to this approach as *geometric annotation*, which operates like an "acetate layer" over the visualization, in contrast to *data-aware* annotations directly associated with the underlying data. We chose to implement a free-form annotation mechanism so that we could first study pointing behaviors in an unconstrained medium. Aside from the freedom of expression it affords, geometric annotation also has a technical advantage: it allows reuse of the identical annotation system across visualizations, easing implementation and preserving a consistent user experience.

Collecting and Linking Views

In data analysis it is common to make comparisons between different ways of looking at data. Furthermore, storytelling has been suggested to play an important role in social usage of visualizations. Drawing comparisons and telling stories both require the ability to embed multiple view bookmarks into a single comment.

To support such multiview comments and narratives, we created a "bookmark trail" widget. The bookmark trail functions something like a shopping cart: as a user navigates through the site, she can click a special "Add View" link to add the current view to a graphical list of bookmarks (Figure 12-8c). Bookmarks from any number of visualizations can be added to a trail. A trail may be named and saved, making it accessible to others.

The bookmark trail widget also functions as a short-term storage mechanism when making a comment that includes links to multiple views. Dragging a thumbnail from the bookmark trail and dropping it onto the text area creates a hyperlink to the bookmarked view; users can then directly edit or delete the link text within the text editor. When the mouse hovers over the link text, a tool-tip thumbnail of the linked view is shown.

Awareness and Social Navigation

The term *social navigation* refers to our tendency to navigate in the world based on the actions or advice of others. On the Web, such navigation can be achieved by surfacing others' usage history to provide additional navigation options. We designed sense.us to support social navigation through comment listings and user profile pages that display recent activity. Comment listings provide a searchable collection of all comments made within the system, and can be filtered to focus on a single visualization (Figure 12-9). Comment listing pages include the text and a thumbnail image of the visualization state for each comment. Hovering over the thumbnail yields a tool tip with a larger image. Clicking a comment link takes the user to the state of the visualization where the comment was made, displaying any annotations included with the comment. The author's name links to the author's profile page, which includes his five most recent comment threads and five most recently saved bookmark trails. The view also notes the number of comments made on a thread since the user's last comment, allowing users to monitor the activity of discussions to which they contribute.

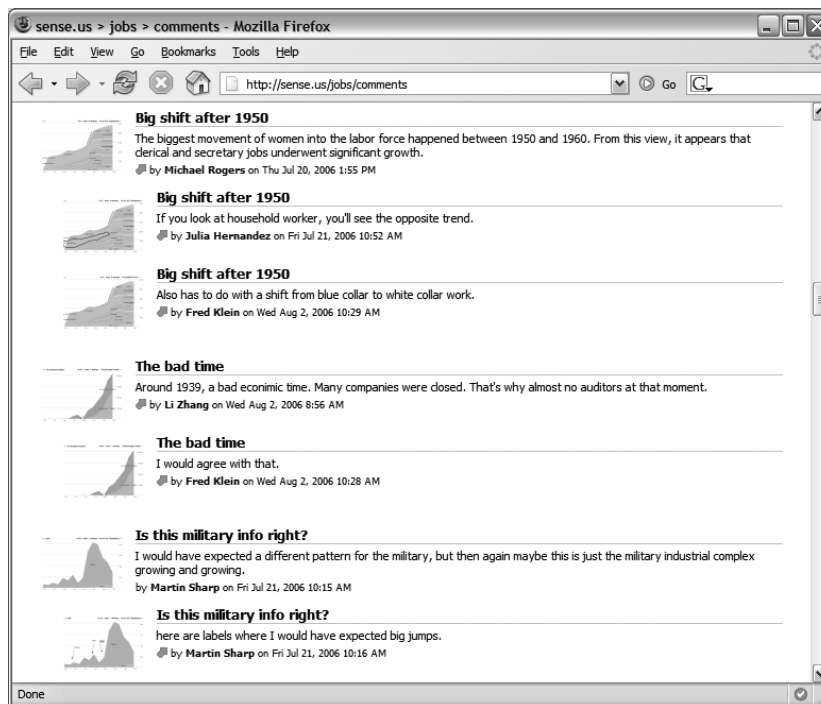


FIGURE 12-9. The sense.us comment listing page; comment listings display all commentary on visualizations and provide links to the commented visualization views. (See Color Plate 42.)

Although more elaborate social navigation mechanisms are possible, we wanted to observe system usage with just these basic options. We were particularly interested in observing the potential interplay between data-driven exploration and social navigation.

By allowing discussions to be retrieved unobtrusively while a user explores the data, potentially relevant conversation can be introduced into the exploration process. Meanwhile, comment listings and indications of recent posts may help users find views of interest, making social activity a catalyst for data exploration.

Unobtrusive Collaboration

We also followed a common design guideline from the field of computer-supported cooperative work: collaborative features should not impede individual usage. As a result, we do not litter views with annotations by default. Rather, comments for a visualization are displayed unobtrusively on the right side of the screen, and graphical annotations are displayed “on demand” by the user.

Voyagers and Voyeurs

After these steps of data acquisition, design, and system implementation, we now had a running website and were ready to do “field tests” with users. We deployed the system in a set of user studies to observe how people would react to our system, what insights they might produce, and how we might improve the site.

We invited 30 people into our lab to observe how they explored data with sense.us. Each person could view what the previous participants had contributed to the site. We also ran a live deployment on the IBM corporate intranet that all employees in the company could access. From these studies, we investigated how people engaged with the visualizations and how the collaboration features impacted their explorations. Next, I summarize some of the more interesting usage patterns we observed.

Hunting for Patterns

Most users’ first instinct was to engage in “scavenger hunts” for interesting and amusing observations, often driven by personal context. For example, users would search for jobs they or their friends or family members have held, or look at birthplace data for the countries of their ancestors. Along the way, people often left comments documenting the trends they found most interesting.

For example, participants noticed that the number of bartenders dropped to zero around the 1930s and posted comments attributing the drop to alcohol prohibition. One person found a peak and then steady decline in Canadian immigration as a percentage of the population in 1800, and posted a question pondering what may have contributed to the trend. Yet another user noticed a drop in stockbrokers in the Great Depression, leading to the visual commentary in Figure 12-10.

Users also found interesting trends via the population pyramid. For example, users explored changes in marital status over time (see Figure 12-11). The green and purple bands indicate the prevalence of separation and divorce, which increases dramatically after 1960. One user investigated school attendance and commented that adult schooling noticeably rises from 1960 onward (the right panel of Figure 12-7, shown previously).

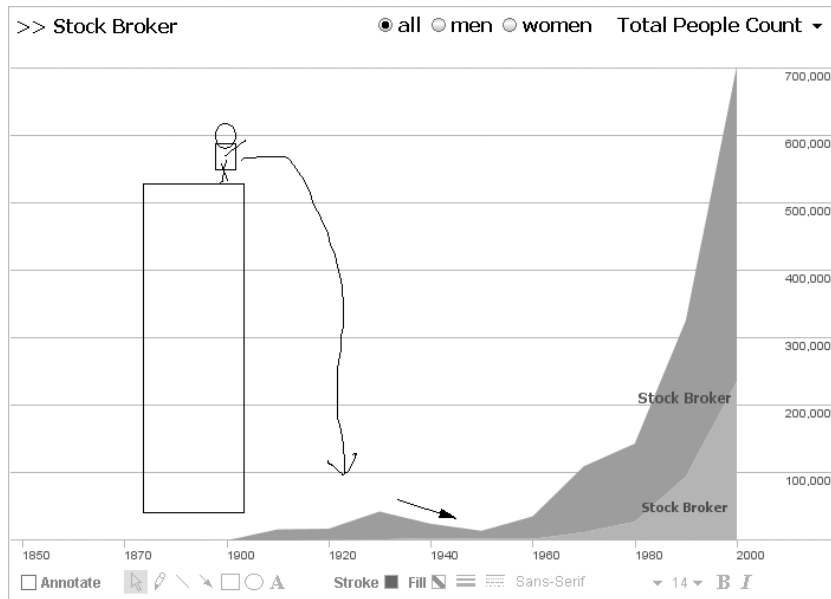


FIGURE 12-10. Annotated view of stockbrokers; the attached comment reads “Great depression ‘killed’ a lot of brokers.” (See Color Plate 43.)

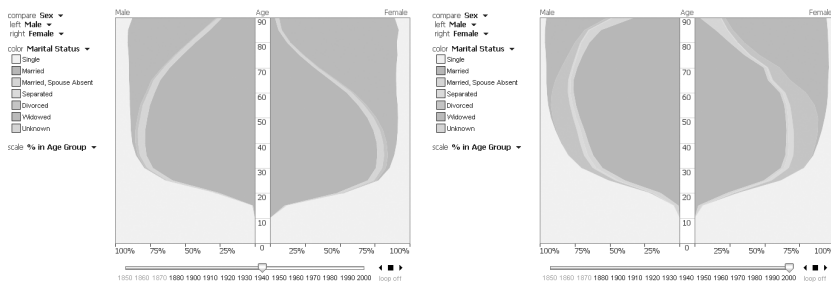


FIGURE 12-11. Population pyramid showing the distribution of marital status for each age group in (left) 1940, and (right) 2000. (See Color Plate 44.)

In another instance, a user mapped the two sides of the pyramid to the populations of the mid-Atlantic (i.e., New York, Pennsylvania, and New Jersey) and the west coast (see Figure 12-12). In 1850, the population of the Gold Rush–era west coast is decidedly different from the east, being dominated by young and middle-age males. Seen 90 years later, the demographics are more closely aligned, though a user noted that the west coast skewed about 10 years older.

Some users were less interested in specific views than in recurring patterns. One user was interested in exploring careers that were historically male-dominated, but have seen increasing numbers of females in the last half-century. The user systematically explored the data, saving views in a bookmark trail later shared in a comment named “Women’s Rise.”



FIGURE 12-12. Population pyramid comparing the populations of the west coast and mid-Atlantic regions in (left) 1850, and (right) 1940. (See Color Plate 45.)

Similarly, a more mathematically minded participant was interested in patterns of job fluctuations, creating a trail showcasing recurring distributions. Another searched for jobs that had been usurped by technology, such as bank tellers and telephone operators. In each of these cases, the result was a tour or story winding through multiple views.

Making Sense of It All

As users made observations of the data, they commonly sought out explanations by posing questions or hypotheses that might make sense of a trend. Many of these questions and hypotheses attracted responses from other users, initiating a cyclic process of social interpretation.

In our live deployment, one user commented on a scatterplot view, asking why New Hampshire has such a high level of retail sales per capita (Figure 12-6). Another user noted that New Hampshire does not have sales tax, and neither does Delaware, the second highest in retail sales. In this fashion, discussion regularly involved the introduction of contextual information not present in the visualization. For instance, users iteratively constructed a timeline of events to annotate military build-ups (Figure 12-8), while another user annotated a graph of teachers with the introduction of compulsory education.

One instance of social data analysis occurred around a rise, fall, and slight resurgence in the percentage of dentists in the labor force (Figure 12-13). The first comment noted the trends and asked what was happening. One subject responded in a separate thread, "Maybe this has to do with fluoridation? But there's a bump...but kids got spoiled and had a lot of candy??" To this another subject responded, "As preventative dentistry has become more effective, dentists have continued to look for ways to continue working (e.g., most people see the dentist twice a year now v. once a year just a few decades ago)." Perhaps the most telling comment, however, included a link to a different view, showing both dentists and dental technicians. As dentists had declined in percentage, technicians had grown substantially, indicating specialization within the field. To this, another user asked, "I wonder if school has become too expensive for people to think about dentistry, or at least their own practice when they can go to technical school for less?" Visual data analysis, historical knowledge, and personal anecdote all played a role in the sensemaking process, explicating various factors shaping the data.



FIGURE 12-13. Annotated job voyager views highlighting (left) a decline in dentists after 1930, and (right) an overall increase in dentistry due to the rising ranks of dental technicians. (See Color Plate 46.)

Another role of comments was to aid data interpretation, especially in cases of unclear meaning or anomalies in data collection. Despite the hard work of the IPUMS project, missing data and obscure labels still occur. To enable comparison across census decades, a shared classification scheme has to be formed. In the case of the job data, a 1950s schema is used. The schema does not include modern jobs such as computer programmer, and some labels are vague.



One prominent occupation was labeled “Operative,” a general category consisting largely of skilled labor. This term had little meaning to users, one of whom asked, “*What the hell is an operative?*” Others responded to reinforce the question or to suggest an explanation, e.g., “*I bet they mean factory worker.*” Another subject agreed, noting that the large number of workers and the years of the rise and fall of operatives seemed consistent with machine-operators in factories.

In this fashion, users collectively engaged in data validation and disambiguation, often planting “signposts” in the data to help aid interpretation by others. Overall, about 16% of the comments referenced data naming, categorization, or collection issues.

Crowd Surfing

We observed that most users initially explored the data driven by their own interests or by items of interest found in the overview (e.g., “*Wow, look how the poor farmers died out!*”). Eventually, users would run out of ideas or tire of exploration. At this point, every user we observed then left the visualizations to explore the comment listings. Some felt that by doing so they would find interesting views more quickly. Remarks to this effect included, “*I bet others have found even more interesting things*” and “*You get to stand on the shoulders of others.*” Other subjects were interested in specific people they knew or discovering what other people had investigated. One user said, “*I feel like a data voyeur. I really like seeing what other people were searching for.*”


Switching between data-driven exploration and social navigation was common: views discovered via comment listings often sparked new interests and catalyzed more data analysis



in the visualizations. After some exploration, participants routinely returned to the listings for more inspiration. Thus we observed a positive feedback loop between data-driven exploration and social navigation: surfacing social activity helped catalyze exploration of new analysis questions. In other words, users fluidly switched between the roles of voyager and voyeur.

Conclusion

Based on the results of the sense.us project, we observed that the combination of interactive visualization and social interpretation can help an audience more richly explore a data set. However, as a research prototype, the sense.us site was never publicly released. Instead, my colleagues at IBM succeeded sense.us with the launch of Many-Eyes.com: a public website where users can upload their own data sets, visualize data using a variety of interactive visualization components, and engage in discussion on-site or embed visualization views in external blogs and wikis.

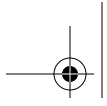


In a similar spirit, web services such as Swivel.com and Data360.org, and commercial products such as Spotfire Decision Site Posters and Tableau Server, now enable users to post visualizations to the Web and engage others in the process of social data analysis. In parallel with the larger movement toward web-scale social computing, there remains much to learn about how to catalyze and support social forms of data exploration. Many exciting research questions regarding how to integrate data analysis and social activity remain to be addressed. Open problems include the design of better social navigation cues, richer annotation techniques, and new methods for combining users' observations, questions, and hypotheses into a reasoned analysis story.

Though the forms of analysis we observed in sense.us were exploratory in nature, the system had a clear educational benefit and users reported that using sense.us was both enjoyable and informative. Furthermore, many of the observations, questions, and hypotheses generated by users invite follow-up by trained analysts. Accessible presentations of data, coupled with social interaction, helped a population turn data into a richer understanding of society. I find that rather beautiful.

References

- Anscombe, Francis J. (1973). "Graphs in Statistical Analysis." *American Statistician*, 27, 17–21.
- Bertin, Jacques. (1967). *Sémiologie Graphique*, Gauthier-Villars. English translation by W.J. Berg as *Semiology of Graphics*, University of Wisconsin Press, 1983.
- Card, Stuart K., Ben Shneiderman, and Jock D. Mackinlay. (1999). *Readings in Information Visualization: Using Vision To Think*, Morgan-Kaufmann.
- Cleveland, William S. and Robert McGill. (1985). "Graphical Perception and Graphical Methods for Analyzing Scientific Data." *Science*, 229(4716), 828–833.



Croes, Marnix. (2006). "The Holocaust in the Netherlands and the Rate of Jewish Survival." *Holocaust and Genocide Studies*, 20(3), 474–499.

Robison, Wade, Roger Boisjoly, David Hoeker, and Stefan Young. (2002). "Representation and Misrepresentation: Tufte and the Morton Thiokol Engineers on the Challenger." *Science and Engineering Ethics*, 8, 59–81.

Tufte, Edward R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*, Graphics Press.

Ware, Colin. (2004). *Information Visualization: Perception for Design*, Second Edition, Morgan-Kaufmann.

Wattenberg, Martin and Jesse Kriss. (2006). "Designing for Social Data Analysis." *IEEE Transactions on Visualization and Computer Graphics*, 12(4), 549–557.

