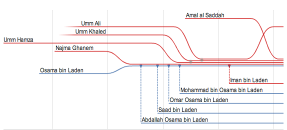
RESEARCH STATEMENT

Jeffrey Michael Heer

Mapping global open source software collaboration using our visualization tools.

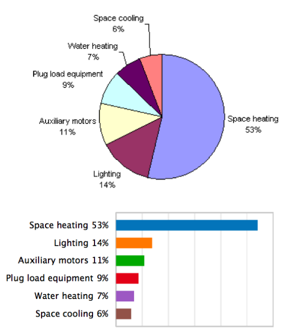
The increasing scale and accessibility of digital data provides an unprecedented resource for informing governance, public policy, business strategy, and our personal lives. Yet, acquiring and storing this data is, by itself, of little value. We must make sense of data in order to produce value from it. Turning data into knowledge is a fundamental challenge for both computer systems and user interface research: it requires integrating analysis algorithms with human judgments of the meaning and significance of observed patterns.

My research group seeks to enhance our collective ability to analyze and communicate data through the design of interactive visual analysis tools. We study the perceptual, cognitive, and social factors affecting data analysis in order to enable a broad class of "analysts" to more effectively work with data: to improve the efficiency and scale at which expert analysts work, and simultaneously lower the threshold for non-experts.



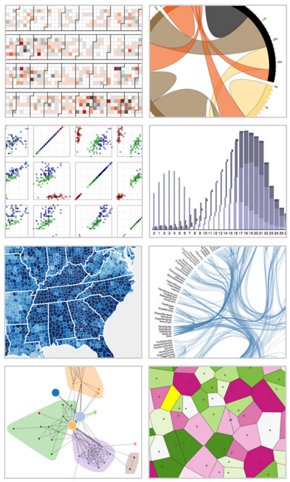
TimeNets show genealogical relations over time, revealing patterns of divorce, remarriage, polygamy and age differences.

My research in *visualization techniques* uses insights from studies of human perception and cognition to design visual representations and interaction techniques for data. Part of this work is empirical: we conduct studies to assess the effectiveness of visual encoding choices. We then apply the findings to create new visualizations and automated design methods. My group has designed and evaluated novel representations for visual analysis of genealogies (AVI 2010), large networks (InfoVis 2011, AVI 2012), and statistical topic models of large text collections (CHI 2012, AVI 2012). In related projects, we have investigated perceptual principles of animation design (InfoVis 2007), chart sizing (CHI 2009, CHI 2010), and color palette design (CHI 2012, EuroVis 2013).



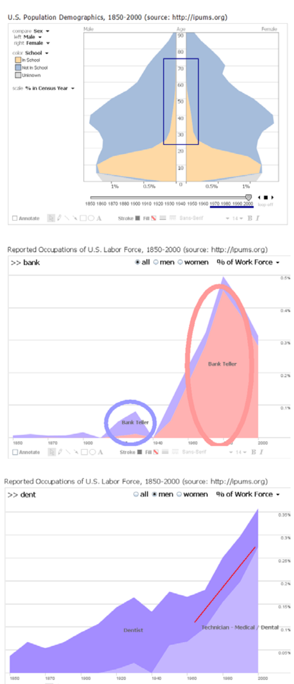
ReVision takes bitmap chart images as input (top), extracts the data table, and creates redesigned graphs (bottom).

Our work in *graphical perception* uses human-subject experiments to evaluate visual encoding choices. We have demonstrated how crowdsourcing systems such as Amazon's Mechanical Turk can be used to generate reliable and novel perceptual findings (CHI 2010). For example, our experimental results (CHI 2010, InfoVis 2010) challenge a decade-held assumption regarding the effects of aspect ratio on rectangular shape comparisons. These results suggest new guidelines for perceptually effective layout algorithms. Building on this research, we have begun to investigate computational models of chart perception. Our *ReVision* system (UIST 2011) uses computer vision methods to classify bitmap images of charts according to the visualization type. Custom extraction procedures then recover the underlying data to enable automated redesign and indexing. Moving beyond low-level perception, recent empirical work (InfoVis 2010) examines the narrative devices employed to effectively "tell stories" with data, and has received wide dissemination among journalists and designers.



Visualizations created with *D3* (Data-Driven Documents).
http://d3js.org/

Of course, improved visualization techniques are of little use if they never make it into the hands of designers and developers. Accordingly, my group also researches software architectures that support the design of novel, customized visualizations. In particular, we are investigating the use of *declarative, domain specific languages for visualization design*. *Protovis* (InfoVis 2009) provides a grammar for authoring expressive visualizations by mapping data to the visual properties of graphical primitives. Protovis statements serve as functional style sheets for data, requiring only limited programming ability. We can also leverage the declarative nature of the language to optimize processing, leading to 20x scalability improvements over previous data visualization frameworks (InfoVis 2010). Building on the Protovis approach, we recently developed *Data-Driven Documents (D3)* (InfoVis 2011). Rather than map data to a specialized lexicon of graphic marks, D3 binds data directly to elements of a web page. Using D3, designers can generate data-driven text and graphics to create interactive browser-based displays. These tools have been widely adopted in both academia and industry, and are actively used by technology firms, design consultancies, and journalists (*e.g.*, the Washington Post and New York Times).



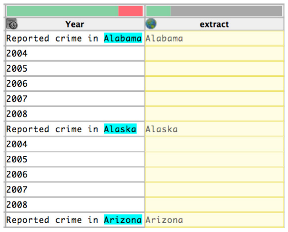
Annotated views from social data analysis in *sense.us*.

1. The rise of adult education from the 1970s onward
2. Reversal of the dominant gender of bank tellers
3. Stratification of dentistry into dentists and hygienists

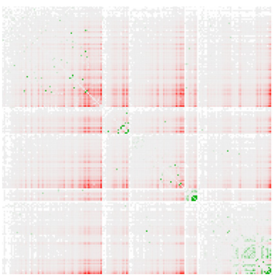
By deploying visual analysis tools, we can also study how users apply them to make sense of data. For example, we observe that analysis is often a social process: the magnitude of data and the diversity of expertise needed to interpret it requires information interfaces that enable us to work together to effectively forage, analyze, point, argue and disseminate. Our research on *collaborative visual analysis* explores how interfaces can catalyze social interpretation and deliberation. One such system is *sense.us*, a site for social exploration of 150 years of U.S. census data (CHI 2007, CACM 2009). Our studies of system usage found that social features can improve hypothesis generation and that visualizing others' activity spurs new explorations by collaborators. This work has informed a number of subsequent data sharing and visualization services, including IBM's Many-Eyes.com and Google's Fusion Tables. Our continuing research examines how to further structure collaboration to improve analysis outcomes. For example, we have found that dividing analysis tasks into explicit stages and providing structured discussion tools can improve both the quantity and quality of findings (CHI 2011, CHI 2012).

We also observe that analysts spend much of their time cleaning and reformatting data to make it suitable for analysis. Domain experts often spend more time manipulating data than they do exercising their specialty, while less technical users (*e.g.*, non-programmers) may be excluded. In response, we are investigating new methods for *interactive data transformation*. With our *Wrangler* system (CHI 2011, UIST 2011), users construct data transformation scripts in a direct manipulation interface. Wrangler uses programming-by-demonstration methods to automatically suggest applicable transforms. The result is not simply transformed data, but a reusable program that can be run on other platforms (*e.g.*, MapReduce) to transform data at scale and be audited to review data provenance.

ONGOING RESEARCH



A user generates text extraction rules by example in *Wrangler*.



Analysis of medical condition co-morbidity in *Orion*, using data from online health forums.

End-user programming for data manipulation. Both prior research and our interviews with analysts (VAST 2012) indicate that data preparation consumes a disproportionate part of most analysis efforts. We are addressing this problem through new interfaces for authoring transformation scripts, including *Wrangler* for tabular data and *Orion* for social network modeling (VAST 2011). Our approach starts with the design of targeted, domain specific languages for data manipulation. We then develop interfaces in which user actions translate into statements in this language. By performing inference over possible statements in the language, we can automatically suggest applicable operations and use visualization techniques to preview their effects. We are currently investigating improved inference methods that search the space of possible programs using both data quality measures and historical usage data. The goal is to create improved interfaces that accelerate transformation and integration of large, heterogeneous data sets.

Intelligent, scalable data profiling. Once source data has been transformed, visualizations can summarize the data, reveal quality issues, and spur hypothesis formation. We are investigating new visual analysis tools that help users identify and correct data quality errors by combining statistical analysis and automatic view suggestion (AVI 2012). To support scalable profiling of massive data sets, we combine data summarization methods with parallel query processing to enable real-time (50 frames-per-second) visual querying of data set sizes ranging from thousands to billions of data elements (EuroVis 2013).

Interactive topic modeling. While modeling is typically the domain of statistics and machine learning, in practice it involves an iterative, time-consuming process of model selection and verification. To facilitate assessment, we are developing a framework that provides diagnostic feedback on how the outputs of statistical topic models differ from experts' organizations of domain concepts (ICML 2013). Our framework enables analysts to systematically assess thousands of models, and compare the effects of various parameter choices, inference algorithms, and automated quality measures.

Visualization design tools. Though data visualization is enjoying increasing popularity and use, custom visualization design still requires programming. We are investigating interactive, graphical techniques that enable non-programmers to create novel designs. Our prior work on *Protovis* provides a formalism for mapping data to the visual properties of graphical marks (*e.g.*, bars, lines, images). We aim to achieve similar expressiveness in a graphical Visualization Design Environment (VDE). We are exploring interaction methods for mapping data attributes to visual properties, specifying interactive and animated behaviors, and supporting semi-automated design using "auto-complete" suggestions informed by existing visualization theory and a corpus of prior designs.

For a complete list of projects, papers, and demos see **http://vis.stanford.edu**