C Project Summary

CAREER: Interactive Tools for Data Transformation and Visualization

Jeffrey Heer, Stanford University

The increasing scale and accessibility of digital data—including government records, corporate databases, and logs of online activity—provides an under-exploited resource with which we can better understand and improve governance, business, academic research, and our personal lives. However, for the data to prove broadly useful, people from a variety of backgrounds must be able make sense of it. Facilitating the analysis of large and diverse data sets is a fundamental challenge in both computer systems and human-computer interaction research, and requires the design of new tools for exploring, analyzing and communicating data.

The goal of this research is to enable a broad class of data analysts to more effectively work with data, using novel interactive tools for data transformation and visualization. We propose to target two critical challenges in the data life-cycle: *visualization design* – creating effective visual representations of data to gain insight – and *data wrangling* – assessing data credibility and transforming data into usable forms.

By leveraging human visual processing, visualizations can dramatically aid our understanding of patterns within data; however, creating effective, customized visualizations remains a difficult task. The Protovis project aims to facilitate the design and deployment of interactive data visualizations. We will first develop an expressive declarative language for visualization design. The language will provide a base from which we will create interactive visualization design tools that combine direct manipulation editing with automated design routines and evaluation aids leveraging computational models of human perception.

Another critical impediment to working with data is transforming acquired data into a usable format suitably free of errors. The Wrangler project aims to facilitate "data wrangling" using interactive visual interfaces that integrate data transformation and analysis methods. We will develop new visualization techniques for assessing data quality issues and new interactive approaches for authoring data transformations. Our tool will produce transformation scripts specified in a high-level language and enable transformation reuse, modification, annotation with analysts' rationales, and targeting to multiple runtime environments.

Intellectual Merit: Our research will produce novel visual and interactive techniques for specifying visualizations and authoring reusable data transformations. We will contribute new systems that marry automated analyses (e.g., visual perception models, data mining algorithms) with direct-manipulation graphical interfaces. Through studies and real-world deployments, we will evaluate our approaches and distill best practices for facilitating data analysis. In prior work, the PI has designed novel visualization and interaction techniques, developed software architectures for visualization, and conducted controlled human-subjects experiments and longitudinal deployments. These experiences provide the necessary background skills to successfully conduct this effort.

Broader Impacts: The research will advance our collective ability to discover and communicate useful insights from a variety of data sources. We seek to improve the scale of data with which analysts can work and enable more people to engage with data by lowering the threshold for non-experts, ranging from professionals analysts, to journalists, to students. Tools developed by this research will be released as open-source software and deployed as web applications. Additionally, our systems and insights into effective practices will be incorporated into university courses on visualization and technologies for the data life-cycle.

Education Plan: My overarching educational goal is to develop the next generation of data scientists and interaction designers. I will broaden the current Stanford curriculum by developing a new course on technologies for supporting the data life-cycle. I will integrate real-world problems and stakeholders with course material, connecting students with domain experts with analysis challenges. I will also leverage our software deployments as an educational opportunity: through the careful design of examples, online demonstration videos, and social sharing of visualizations and transformation scripts, I hope to use our research to help educate and empower interested members of the public.