> *Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the flexibility of the informed human mind... Accordingly, both [analysis] approaches and techniques need to be structured so as to facilitate human involvement and intervention.* – John W. Tukey & Martin B. Wilk, *Data Analysis & Statistics*, 1966

Though voiced nearly 50 years ago, the sentiments of Tukey & Wilk ring true today: *to facilitate effective human involvement at all stages of data analysis* remains a grand challenge. Advances in computing and statistics provide new opportunities for data-driven discovery, but advances in science ultimately lie with the ability of empowered investigators to pursue questions, uncover domain-specific patterns of interest, identify errors, and assess model outputs. Accordingly, it is essential that we develop tools that put accessible, effective, and interrogable data analysis methods in the hands of more domain scientists.

In keeping with this vision, our five-year goal is to fundamentally improve interactive data preparation and visual exploration tools across multiple natural sciences. Our aim is not a singular scientific breakthrough. Rather, we seek to develop methods and tools that increase the efficiency and scale of human-centered analysis practices, fueling breakthroughs by a wide audience of investigators. Toward this goal, we will partner with scientists to design and evaluate novel interactive analysis tools for specific domain problems. Concurrently, we will identify recurring data types, workflows and analysis patterns; design corresponding interaction and visualization techniques; and build enabling system architectures. Our experiences collaborating with scientists and releasing open-source tools (e.g., D3.js) suggests that this is an achievable goal with the potential to make substantial contributions to scientific practice.

ENGAGING WITH SCIENTISTS

I believe the path to broadly useful analysis tools starts with iterative, bottom-up development driven by real-world projects. In response to being a DDDI semi-finalist, and to augment existing collaborations, I have engaged in a series of lab visits and interviews with scientists across the University of Washington. I found that many groups are already using software developed by our group (often unbeknownst to us). All groups express a strong desire for improved tools for interactive visualization, data quality assessment and transformation. To illustrate these needs, here is a subset of past successes and identified opportunities:

Astronomy: The standard model of galaxy formation involves hierarchical accretion via the merging of smaller galaxies. Astronomers would like to know: how much do these merger histories vary depending on a galaxy's final mass, proximity to other galaxies, and other factors? To help address these questions, we connected UW astronomers with computer science students to develop interactive visualizations of galaxy merger trees. The visualizations allow researchers to view merger histories in detail, analyze similarities among trees, and assess correlations with mass and particle count distributions. The resulting tool and supporting database system is now in use by astronomers and will be presented at SIGMOD 2014, a leading database conference. Meanwhile, Prof. Andrew Connolly described to us his need for visual analysis and modeling tools for astronomical survey data. His lab is already using our group's software to build web-based visualization tools; however, improved systems are needed to better scale – both computationally and perceptually – to their high-volume, high-dimensional data.

Earth Sciences: Prof. John Vidale of UW Earth & Space Sciences investigates earthquakes and improved early warning systems. Early warning is supported by a network of seismometers, with automated analysis methods triggering warnings that must be manually verified by John or a few select colleagues. On average only 1 in 5 triggered events actually warrant follow up. Like on-call doctors, at a moment's notice the scientists must go online to view seismic activity, determine if an earthquake has occurred, and potentially notify authorities. Manual assessment involves viewing time-series charts backed by ancillary geographic maps and spectrographs. Perhaps surprisingly, these central visualizations have largely been unevaluated. Improved designs – in terms of perceptual effectiveness, improved composition, and interactivity – could facilitate seismic data analysis and accelerate critical triage and emergency response efforts.

Biology: Working with Prof. Deborah Gordon at Stanford, we recently investigated the life and death of harvester ant colonies in the American Southwest. We designed novel interactive maps for 20 years of observations, combining spatial position, time-varying status (colony birth and death), and ancestral networks determined via genetic analysis of colony samples. In addition to visual analysis, our maps improved data collection: Prof. Gordon uses the system to author annotated maps that serve as both instructions and data entry forms for her research assistants. She estimates that this has led to an order of magnitude time savings and increased data accuracy. These tools contributed to novel findings that we published in the Journal of Animal Ecology. Meanwhile, the development of these maps provided use cases and requirements for the iterative design of our open-source Protovis and D3.js systems.

We have identified analogous opportunities here at UW. Prof. Julia Parrish manages a network of ~1,000 citizen scientists who report data on bird deaths and debris on Pacific beaches. The collected data raises the challenge of visualizing data for volunteers in a way that encourages improved accuracy and sustained engagement. In environmental biology, Professors Lauren Buckley and Jannecke Hille Ris Lambers study the ecological effects of changing climate. For example, how will a 3° temperature increase impact animal and plant populations? Answering these questions involves integrating data sets from internal (e.g., lab experiments, models) and external (e.g., weather stations, phylogeny databases) sources. Acquiring and wrangling these data sets is manually intensive and potentially error-prone, presenting opportunities for interactive data preparation tools to accelerate and systematize these procedures.

Genomics: In ongoing research, we have been investigating visual analysis techniques for population-scale genetic variation. Aided by Prof. Arend Sidow and Prof. Serafim Batzoglou at Stanford (full disclosure: I am on the advisory board of their company DNAnexus), we initially focused on *in vitro RNA selection*. This method simulates evolutionary processes by chemically mutating RNA strands and then selecting for resulting sequences based on testable binding or catalysis properties, allowing scientists to synthetically create large pools of sequences in which they then search for functional molecules. To aid analysis, we developed *invis*, a visual tool that interactively links population overviews with sequence-level maps. Invis enables biologists to identify related sequences, compare sequence populations over varying conditions and visualize likely pathways of genetic evolution. Invis is now in use by collaborating biologists and we are continuing to refine and scale our approach. Exciting new challenges come from our colleague Prof. Virginia Armbrust in UW Oceanography: can a future variant of invis aid exploration of metagenomic data, involving the results of simultaneous sequencing of a host of ocean microorganisms?

ADVANCING DATA SCIENCE METHODS

These scenarios illustrate scientific pursuits that would benefit from new interactive analysis tools. Our plan is to advance visual, exploratory analysis of scientific data through a combination of empirical research and novel systems design. We will begin by conducting an expanded survey of analysis practices in the natural sciences. Through lab visits, interviews, and large-scale online questionnaires, we will survey the scientific goals, data sets, and analysis methods of our colleagues, then delineate bottlenecks and opportunities for interactive analysis tools. Based on the results (which we will publish as a shared community resource), we will identify tiers of candidate projects: for example, problems addressable by existing tools, problems well-scoped for projects by students in a visualization course, and problems requiring novel computational or interactive methods. While project selection will ultimately be data-driven – and informed by consultation with the Moore Foundation – possible candidates include:

Evaluating Scientific Visualization Practices. Visual representations are regularly used throughout the sciences, but (ironically?) given little scientific treatment. Encodings known to be problematic, such as quantitative rainbow color scales and the extraneous use of 3D, are commonplace. Novel methods are adopted without systematic evaluation, e.g., Circos plots and sequence maps in biology. We will identify common visualization practices in need of further assessment, explore alternative designs, and perform task-driven perceptual experiments to assess effectiveness. We will use the results to distill design guidelines and inform ongoing tool design, including automated design methods.

Scalable Visualization Methods. Colleagues using visualization tools such as D3.js report that they still struggle with high-volume and high-dimensional data. Building on our *imMens* system, we will develop improved methods for providing responsive interaction with large data sets, including appropriate data summarization, pre-computation and predictive pre-fetching methods. We will also explore interaction techniques for high-dimensional data. Assessing genetic variation, evaluating regression models and understanding the spectra of stars all share the problem of finding salient features in high-dimensional data. Current visualization methods typically show at most a dozen or so dimensions (e.g., using parallel coordinates) or rely on dimensionality reduction techniques that can prove difficult to interpret. We will investigate methods for linking overviews produced with varied reduction methods to interpretable feature-specific displays. For example, selecting points in an overview plot might trigger retrieval and visualization of the dimensions that best discriminate selected points from the remainder of the data set.

Interactive Visualization Design Tools. Though visualization tools continue to improve, domain-specific design work remains tedious, often requiring programming. Extending our work on declarative visualization languages, we will develop graphical design environments that minimize the need for textual programming. Analogous to spelling and grammar checking in word processors, we will investigate the integration of automated design methods based on perceptual models to foster more effective visualization design choices. Also of central importance is specifying interaction techniques, including selection, cross-filtering, zooming, and variable transformation. We will research improved abstractions for interaction design using concepts from functional reactive programming, which models input events (e.g., mouse, touch and keyboard events) as composable streams of data. Our goal will be to develop declarative models for interaction that can simplify development, aid reuse, and enable optimized execution – while also being amenable to visual specification. In addition, scientific data analysis often involves complex data ranging from common semantic types (date-times, geographic locations) to specialized structures (galaxy merger trees, phylogenetic trees, genomic sequences). We will explore the inclusion of extensible semantic data types to support domain-specific operations and guide visual representation choices.

RESEARCH MANAGEMENT AND EVALUATION

If awarded, Moore Foundation support will be used to scale our efforts and focus them on challenges of scientific data analysis. Funds will support post-doctoral, graduate and undergraduate student researchers who also will participate in the labs of our scientific collaborators. To aid cross-cutting design and software development tasks, I intend to hire a research staff member. In addition, we will conduct educational and outreach activities, including a new studio-based class pairing computer science and natural scientists in project teams to collaboratively prototype and evaluate new interactive analysis tools.

We will evaluate progress on multiple fronts. First, we will solicit ongoing feedback from collaborators on the applicability and effectiveness of our tools. Are our tools being productively used to advance scientific discovery? What types of insights are gained, and at what rate? How are these insights reflected in ensuing models and published results? To aid assessment, we will instrument developed applications to capture usage data and model interaction histories. This usage data can itself be a valuable subject of visualization and analysis. Second, we will conduct controlled experiments to evaluate the visualization and interaction techniques produced by our research. Likely experimental subjects include scientists, university students, and participants from online crowdsourcing platforms. Third, all software will be released under an open-source license and made publicly available on platforms such as GitHub. As our software matures, we will produce the necessary documentation, examples, and tutorials to enable dissemination and uptake. We will also seek out ways to integrate our software into shared scientific computing tools, such as R and the iPython notebook. For example: *Vega*, our declarative visualization language, is already being used as a foundation for *ggvis*, the successor of the popular *ggplot2* library for creating visualizations in R. As with our prior work, we will provide the requisite support and maintenance for other labs to adopt our tools and contribute back to them. Monitoring community development and software uptake will provide an important avenue for assessing the long-term impact of our research.