Extreme Scale Visual Analytics

Chad A. Steed
Oak Ridge National Laboratory
Oak Ridge, TN
csteed@acm.org

Thomas E. Potok
Oak Ridge National Laboratory
Oak Ridge, TN
potokte@ornl.gov

Laura L. Pullum
Oak Ridge National Laboratory
Oak Ridge, TN
pullumll@ornl.gov

Arvind Ramanathan
Oak Ridge National Laboratory
Oak Ridge, TN
ramanathana@ornl.gov

Galen Shipman
Oak Ridge National Laboratory
Oak Ridge, TN
gshipman@ornl.gov

Peter E. Thornton
Oak Ridge National Laboratory
Oak Ridge, TN
thorntonpe@ornl.gov

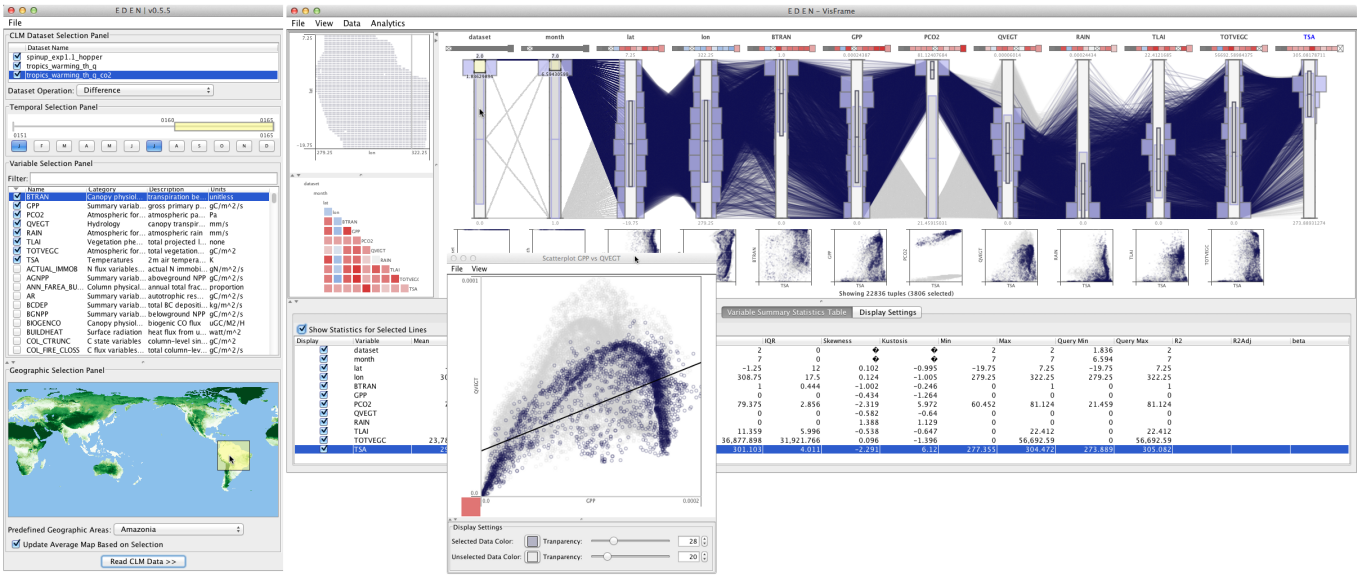

Figure 1. This figure provides an overview of EDEN during analysis of a global CLM4 data set. The CLM4 filter panel (left) facilitates interactive queries into large CLM4 data sets. The VisFrame (right) offers a highly interactive, visual interface to explore multivariate relationships via linked parallel coordinates, scatterplots, correlation matrix, and geographic scatterplot visualizations.

ABSTRACT

Given the scale and complexity of today's data, visual analytics is rapidly becoming a necessity rather than an option for comprehensive exploratory analysis. In this paper, we provide an overview of three applications of visual analytics for addressing the challenges of analyzing climate, text streams, and biosurveilance data. These systems feature varying levels of interaction and high performance computing technology integration to permit exploratory analysis of large and complex data of global significance.

Author Keywords

visualization; analysis; exploratory; knowledge discovery; climate; text

INTRODUCTION

Most existing analysis and visualization tools permit the investigation of gigabyte to terabyte data only through drastic reductions informed by the domain expert's experience and intuition. However, in the process of subsetting the data, information is severely truncated and discarded, which makes analysis more difficult, uncertainty quantification less tractable, and serendipitous discoveries nearly impossible. Meanwhile, technological advances are enabling higher resolution, fidelity, and complexity in domains such as climate simulations, unstructured information stream analysis, and bio-surveillance, fueling an escalation to extreme scale data sets. Therefore, we continue to disproportionately outpace our ability to analyze data in a systematic, exploratory manner, thereby postponing the next round of great scientific breakthroughs.

It is clear that a new approach is necessary, but what is the

key to enabling change? As stated in the 2013 Department of Energy's Advanced Scientific Computing Advisory Committee (ASCAC) Challenges Report, perhaps the most significant ingredient lies within ourselves: "Analysis and visualization of increasingly larger-scale data sets will require integration of the best computation algorithms with the best interactive techniques and interfaces. We must pay greater attention to human computer interface design and human in the loop workflows." [2] This statement is just one example of the growing recognition that we must evolve from the current machine-centered paradigm to one that is human-centered, thereby exploiting the scientists' extremely high bandwidth visual processing channel and cognitive capabilities. Furthermore, the vast majority of scientific analysis and visualization research has been devoted to scientific visualization, where spatial representations are normally provided. However, abstract attributes and hyper-dimensionality dominate many of the extreme scale science domains and scientific visualization, in its purest form, doesn't accommodate such data. In order to usher in the next round of great scientific advances, a new class of interactive visualization and analysis techniques is needed that effectively and seamlessly couples the strengths of humans with the computational advances of machines for information-assisted, human-centered analysis in extreme scale domains.

Information visualization refers to "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition" [1]. Information visualization formed out of the scientific visualization community to address the rapidly increasing volume and complexity of abstract data and has demonstrated success in enabling more efficient analysis in such fields as biomedical, cyber-security, and intelligence. There is tremendous potential in applying information visualization techniques to extreme scale science, particularly applications that harness the massive parallelism of emerging HPC technologies.

A related field is visual analytics, which refers to the "science of analytical reasoning facilitated by interactive visual interfaces" [12]. The fundamental goal of visual analytics is to turn information overload into opportunity by visually representing the information and allowing humans to directly interact with it to gain insight, draw conclusions, and make better decisions. The advantage of visual analytics is that users can focus their full cognitive and perceptual capabilities on the analytical process, while simultaneously applying advanced computational capabilities to augment the discovery process [5].

Both information visualization and visual analytics have demonstrated great promise in amplifying knowledge discovery and cognition in a variety of fields such as intelligence and bioinformatics, but the application of these techniques in extreme scale science is rare. A significant proportion of today's scientific data exhibit abstract attributes, hyper-dimensionality, and unstructured attributes that are all ideally suited to an information visualization approach. Furthermore, calls for processing methods (especially those that involve in-situ analysis) that integrate the "the best computational algorithms" and "interactive techniques and interfaces" can only be addressed through a superior visual analytics approach [2]. Algorithmic advances offer much promise to scientific analysis, but the disruptive impact of said techniques in extreme scale science is necessary to realize the full potential of the abundant data and high-end computing resources in DOE science mission areas.

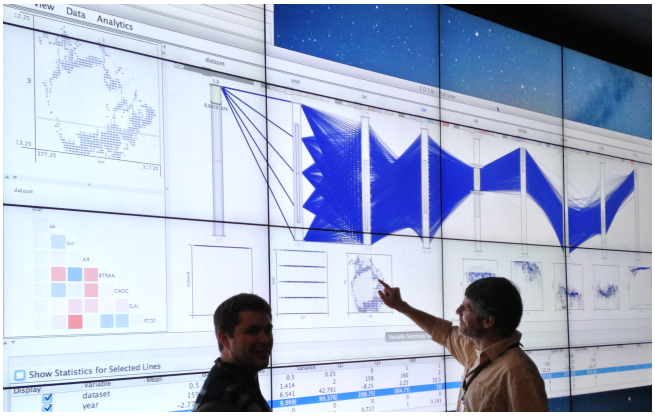

Figure 2. ORNL climate researchers Dan Ricciuto (left) and Peter Thornton (right) are shown using the EDEN visual analytics tool to explore multivariate relationships in Community Land Model Version 4 (CLM4) simulations. The analysis is being performed using the 35 mega-pixel display wall in the ORNL EVEREST visualization laboratory, a unique resource of the Oak Ridge Leadership Computing Facility (OLCF).

In this paper, we will provide an overview of three visual analytics approaches that rely heavily on interactive information visualization techniques for human-centered analysis of large scale data. First, we will discuss a novel climate visual analytics system, called EDEN, that works with HPC architectures and fosters more creative analysis of earth system simulations and ensemble analysis. Next, we will give an overview of a text visual analytics system that permits interactive analysis of high throughput unstructured information streams for situation awareness of global events. Then, we will conclude with an overview of a bio-surveillance project for analyzing epidemiological model data sets to understand and control disease spread and outbreak. These systems feature various levels of interaction with high performance computing (HPC) platforms and deal with so-called "big data" problems of global significance.

CLIMATE VISUAL ANALYTICS

The Exploratory Data analysis ENvironment (EDEN) is a visual analytics system that is designed to support hypothesis formulation and testing with complex, multivariate climate simulation data sets, particularly data produced by the Com-

4th SC Workshop on Petascale (Big) Data Analytics, Nov. 17, 2013, Denver, Colorado, USA.

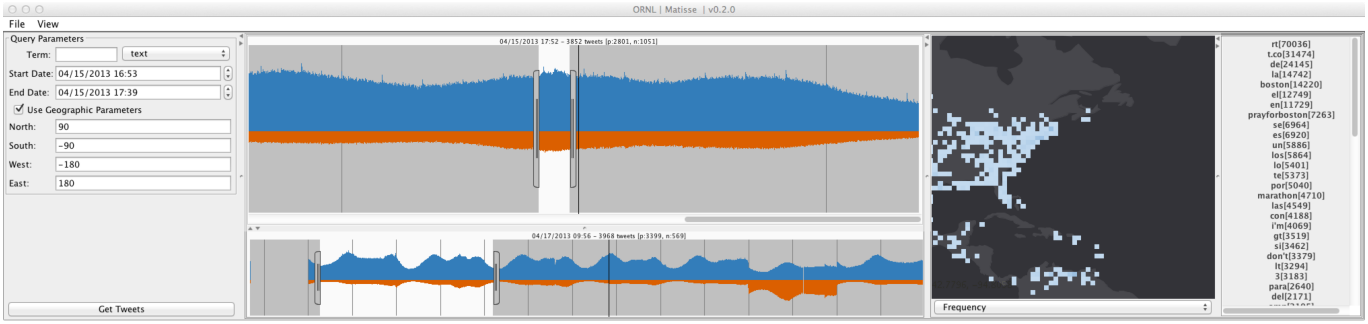

Figure 3. This figure shows an overview of the Matisse visual analytics framework for interactively exploring streaming unstructured information using visual representations. The framework supports the ability to drill down to higher resolution details in the data using mouse gestures and is fed by a near real-time analytics engine that consumes and summaries the streaming content. Matisse has been utilized for analyzing social media, news feeds, and similar text streams.

munity Land Model Version 4 (CLM4) [6]. Unlike conventional climate analysis tools, EDEN employs a highly interactive, information visualization canvas to connect the scientist to the data behind the representations, as shown in Figure 1. EDEN has been developed in close collaboration with leading climate scientists (one of which is a co-author of the current work) and together we have published comprehensive case studies that corroborate the notion that a visual analytics approach leads to more efficient data analysis in real-world climate studies [10, 11]. In Figure 2, two ORNL climate researchers are shown using EDEN for collaborative analysis on the Oak Ridge Leadership Computing Facility's (OLCF) EVEREST display wall.

At the heart of EDEN is a highly interactive variant of parallel coordinates—a popular multivariate visualization technique that is well-suited to the analysis of large multivariate data sets. The parallel coordinates technique was initially popularized by Inselberg [3] as an approach for representing hyper-dimensional geometries, and later demonstrated in multivariate analysis by Wegman [13]. In general, the technique yields a compact 2-dimensional representation of even large multi-dimensional data sets by representing the N-dimensional data tuple C with coordinates $(c_1, c_2, \ldots, c_N)$ by points on N parallel axes which are joined with a polyline [4].

EDEN augments the classical parallel coordinates plot by providing cues to guide the analyst's exploration of the information space. This approach is akin to the concept of the scented widget described by Willett et al. [14]. Scented widgets are graphical user interface components that are augmented with an embedded visualization to enable efficient navigation in the information space of the data cases. In EDEN, the parallel coordinates plot is extended with a number of capabilities that facilitate exploratory data analysis and guide the scientists to the most significant relationships in the data. Correlation mining, coordinated scatterplots, and embedded visualization are also used to graphically encode key statistical quantities. For a more detailed discussion of the techniques introduced in EDEN the reader is directed to our previous work [10].

EDEN is currently used by climate researchers to analyze global simulations and point simulations ensembles. In several practical case studies [10, 11] we have demonstrated how EDEN improves the efficiency of climate analysis. EDEN permits exploration of high dimensional spaces without the loss of information. In some cases, EDEN has enabled levels of analysis that were previously impossible, such as the simultaneous display of 88 dimensions from a CLM4 ensemble. EDEN has also helped climate researchers find new relationships that they never anticipated or considered exploring before such as particular input parameter sensitivities [11]. EDEN has been adopted by a wide range of climate scientists throughout the world and it is available for free download[1]. Although it has been evaluated most extensively in climate analysis, EDEN is a general multivariate analysis tool that supports any domain through a general CSV file ingest feature.

TEXT STREAM VISUAL ANALYTICS

With more than 140 million active users, each day Twitter produces more than 340 million posts (called *tweets*), which are limited to 140 characters. Twitter is just one example of many social media systems that are transforming the way our society functions. Techniques for analyzing broad trends over social media data or detailed analysis within small subsets have been demonstrated in recent years, but state-of-the-art tools are mostly inadequate at supporting near real-time analysis of these high-throughput streams of unstructured information.

We have developed a framework, accessible by a highly interactive visual analysis tool called Matisse (see Figure 3), to help detect and analyze global events and trends in large scale text streams like those produced by social media systems. The framework integrates a unique collection of modules that support sentiment analysis, change detection, identification of key associations, and automatic discovery of spatio-temporal patterns.

Efficient management of streaming textual information is paramount to our objective of enabling interactive visualization and analysis. Our system continuously consumes and indexes streams from multiple platforms using a variety of modern database formats that support interactive processing. The system can consume Twitter sample streams and

[1] http://cda.ornl.gov/projects/eden/

RSS news feeds directly, but is sufficiently general to support any unstructured information stream (e.g. system status logs). These streams are characterized by high velocity, high throughput information flow. Our system processes new items in near real-time using a fault-tolerant, stream processing engine that identifies similarities, trends, and estimates sentiment in the stream. This backend processing engine can segment the stream into sub-topics that are defined by a unique vector of terms or phrases for more focused awareness of key concepts.

To analyze the information, Matisse utilizes a highly interactive canvas (see Figure 3) for graphically depicting the current (and past) state of activity in the stream. Matisse utilizes a coordinated multiple view approach whereby changes in one display are propagated to the other displays, appropriately. Furthermore, the display supports focused analysis while also providing higher level contextual views for more particular investigations. In the remainder of this section, we will provide detailed descriptions of the various visualization techniques.

Matisse aggregates summary statistics for a pre-defined unit of time (e.g. seconds, minutes, hours) to form a time series. This time series is used to build a temporal visualization which represents the summary metric as a chart. The chart can encode a single value over time or it can be partitioned to show two or more metrics. For example, in Figure 3 the display is split to show positive tweets as the top graph of blue bars and negative tweets as the bottom graph of orange bars with a common baseline. The height of each bar indicates of the frequency of tweets having positive or negative sentiment at the point in time. Users may select time ranges of interest directly in the temporal graph and the other views will be updated, appropriately.

Matisse also provides an interactive geospatial heatmap for the selected time range. The color scale used in the map represents grid cells with higher tweet counts as darker and more saturated shades of blue and lower tweet counts as lighter and less saturated shades of blue. Areas with higher activity are therefore, presented in a more visually salient manner to highlight relative activity. The map can be used to visualize overall frequency, positive frequency, or negative frequency in the current version. Furthermore, additional derived statistical metrics can be created and visualized in the map view. Users can also select a geospatial region in the map view to set a spatial query for tweets in the area.

At right of the geospatial view, the term view shows the most frequently occurring terms for the selected time and spatial location. These top terms are calculated and stored in a summary object for the time unit of interest (e.g. minutes, hours). To quickly render the top terms for the selected time range, the top term summary information is then used to populate the term view.

We have alluded to the linked views in Matisse whereby selections in the various displays are propagated to the other views. This coordinated multiple view model is combined with the temporal focus+context display which shows the overview of the complete time series with a detailed view of the time unit of interest. In the geospatial map, the view provides additional interactions for zooming in/out of the display and panning the viewpoint. Furthermore, the user can select words of interest in the term view to query specific text that contains the selected word(s). Selections in each of these panels are used to set the filter/search criteria in the left hand filter panel. With this tool, individual tweets can be queried to see general text for tweets and aggregated statistics.

BIO-SURVEILLANCE VISUAL ANALYTICS

We have also applied a visual analytics approach to the domain of bio-surveillance. Using the parallel coordinates based visualization capabilities of EDEN (see Figure 4), we examined compartmental epidemiological models which are used to model how diseases spread as well as strategies for control during epidemics [9]. Compartmental models proceed by segregating a population into distinct groups:

Susceptible (S) – part of the population previously unexposed to the pathogen;

Infected (I) – part of the population affected by the pathogen;

Exposed (E) – part of the population that is infected by the pathogen but not infectious;

Recovered (R) – part of the population that has successfully been cured of the infection.

Compartmental models use ordinary differential equations to model the various aspects of disease spread and control. For a more detailed discussion of the models and our metamorphic testing framework we encourage the reader to review our previous work [9].

The models produce a complex and massive record of data that is difficult to explore efficiently. However, exploration is required to determine relationships between variables and identify the most significant sensitivities and associations. We have applied the parallel coordinates based visual analytics methodology in EDEN to examine the parameter sweeps and metamorphic testing results on the SIR/SEIR models. For example, Figure 4 shows a dynamic visual query that was formed interactively to select tuples with $\beta = 1$ and varying γ from a minimum of 0.2 to a maximum of 0.5. From this selection, it is evident that when the value of γ is higher than β, there is no epidemic, as evidenced by the top lines in the susceptible column of the plot. However, as γ values tend to rise, there is a subsequent increase in the number of infected and recovered populations. Thus, a visual scan of the parameter space allows us to visualize how γ and β are dependent on each other and further allows us to examine the behavioral properties of the SIR model in terms of when an epidemic may prevail in the population. In the case of epidemiological models, we found that parameter scanning along with data exploration and visualization provides novel insights into the behavioral properties of these models.

In addition to our work with epidemiological model analysis, we are exploring ways to combine text mining, graph analysis, and visual text analytics to conduct bio-surveillance using

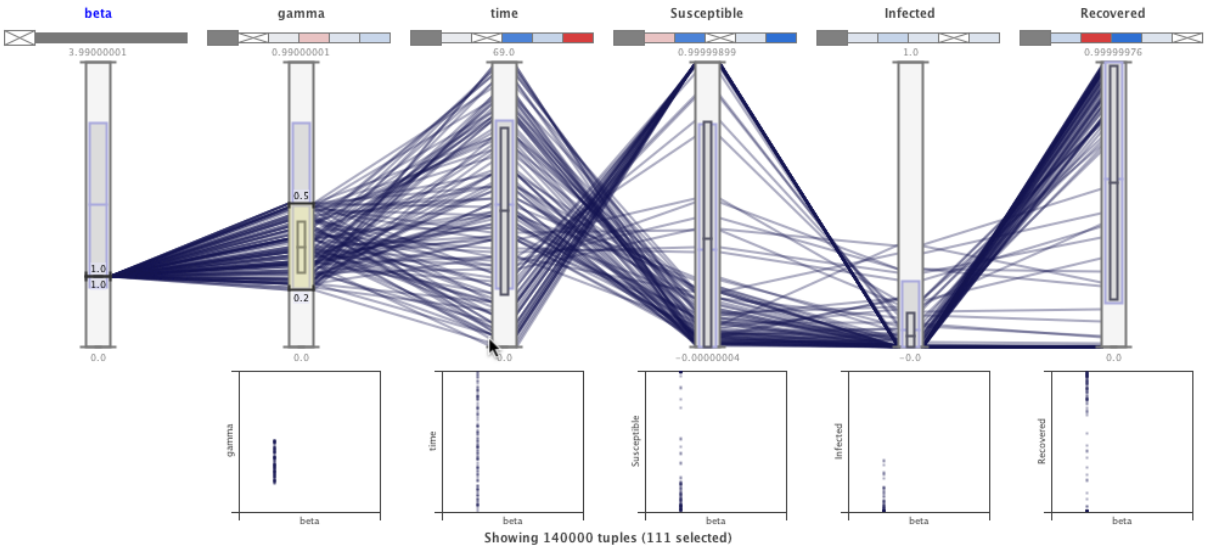

Figure 4. The figure shows a novel visual analytics tool for exploring sensitivities in compartmental epidemiological models. By facilitating interactive visual queries of the complex parameter space, scientists can explore the model outputs and develop insight about the conditions under which an epidemic can occur.

social media [8]. One particular challenge in this work is devising new algorithms to efficiently explore information from diverse data streams that include both text and multi-media. Such tasks are particularly relevant for tasks involving situational awareness. For example, a cursory search on the internet for flu yielded thousands of images (with annotated information), with additional textual information that could be gathered from Twitter (with the same or similar annotations). The Oak Ridge Bio-surveillance toolkit (ORBiT) [7] incorporates new new visual exploration techniques that enables users to link information from multiple data sources. ORBiT will also utilize the advanced text visual analytics algorithms in Matisse.

CONCLUSION

Human-centered analysis is attractive for extreme scale science applications because it takes advantage of the inherent strengths of humans (e.g. high bandwidth visual processing channel, creativity, and background knowledge). If we adopt the visual analytics paradigm, we further increase the effectiveness by combining the human strengths with the tremendous computational capabilities of machines. The resulting system is ideal for extreme scale analysis because the mathematical and statistical tasks are efficiently handled by machines whereas the visual pattern recognition tasks are ideal for humans. Given the scale and complexity of today's data in areas such as climate, bio-surveillance, and social media networks, a visual analytics approach is perhaps the only viable solution for exploratory analysis.

In the current work, we have provided an overview of three key areas where we have applied the visual analytics approach to extreme scale analysis. For climate, the data is large and occupies a complex multivariate space. Working with leading domain scientists, we developed EDEN which has proven to be very effective and is being used today for practical climate studies. We have also applied the visual analytics paradigm to the near real-time analysis of social media text streams. The mining of social media data has demonstrated promise in natural disaster management and observing community resilience, as well as for detecting disease outbreaks. Finally, we have explored the use of EDEN for more general analysis of epidemiological models. Each of these applications share a common challenge of large and complex data. In the case of social media, the data streams also exhibit high velocity and high throughput characteristics with a need to deal with items in near real-time.

In the future, we will investigate tighter linkages to the massively parallel facilities such as the OLCF, effectively connecting emerging HPC architectures to interactive visualizations for more intuitive, human-centered analysis. We will also explore ways to harness newer, high resolution display technologies, such at the 35 mega-pixel display wall in the ORNL EVEREST visualization laboratory, a resource of the OLCF. Visualization represents the interface between humans and machines and anything we can do to improve that interface will directly increase our ability to make sense of the vast volumes of complex information that we face today.

ACKNOWLEDGMENTS

This work is sponsored by the Laboratory Directed Research and Development Program (LDRD #6427) of Oak Ridge National Laboratory and the U.S. Department of Energy's Office of Biological and Environmental Research. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy. The United States Government retains and the publisher, by accepting this article for publication, acknowledges that the United States Government retains non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. This research used resources of the Center for Computational Sciences at Oak

Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DEAC0500OR22725.

REFERENCES

1. Card, S. K., Mackinlay, J. D., and Shneiderman, B. Information visualization. In *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Diego, CA, 1999, 1–34.
2. DOE ASCAC Data Subcommittee. Synergistic challenges in data-intensive science and exascale computing. Tech. rep., U.S. Department of Energy Office of Science, March 2013.
3. Inselberg, A. The plane with parallel coordinates. *The Visual Computer 1*, 4 (1985), 69–91.
4. Inselberg, A. Parallel coordinates: Interactive visualization for high dimensions. In *Trends in Interactive Visualization*, E. Zudilova-Seinstra, T. Adriaansen, and R. Liere, Eds. Springer-Verlag, London, UK, 2009, 49–78.
5. Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. Visual analytics: Scopes and challenges. In *Visual Data Mining*, S. J. Simoff, M. H. Böhlen, and A. Mazeika, Eds. Springer-Verlag, Berlin, Germany, 2008, 76–90.
6. Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zong-Liang Yang, Levis, S., Sakaguchi, K., Bonan, G. B., and Slater, A. G. Parameterization improvements and functional and structural advances in version 4 of the community land model. *Journal of Advances in Modeling Earth Systems 3*, M03001 (2011), 27 pp.
7. Ramanathan, A., Pullum, L. L., Steed, C. A., Quinn, S. S., and Chennubhotla, C. S. Oak Ridge Bio-surveilance Toolkit. In *IEEE VAST Workshop on Public Health's Wicked Problems: Can InfoVis Save Lives?* (Atlanta, GA, October 2013).
8. Ramanathan, A., Pullum, L. L., Steed, C. A., Quinn, S. S., Chennubhotla, C. S., and Parker, T. Integrating heterogeneous healthcare datasets and visual analytics for disease bio-surveillance and dynamics. In *IEEE Workshop on Interactive Visual Text Analytics* (Atlanta, GA, October 2013).
9. Ramanathan, A., Steed, C. A., and Pullum, L. L. Verification of compartmental epidemiological models using metamorphic testing, model checking, and visual analytics. In *Workshop on Verification and Validation of Epidemiological Models* (Washington, D.C., December 2012).
10. Steed, C. A., Ricciuto, D. M., Shipman, G., Smith, B., Thornton, P. E., Wang, D., and Williams, D. N. Big data visual analytics for earth system simulation analysis. *Computers & Geosciences 61* (2013), 71–82.
11. Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D., and Branstetter, M. Practical application of parallel coordinates for climate model analysis. In *Proceedings of the International Conference on Computational Science* (Omaha, NE, June 2012), 877–886.
12. Thomas, J. J., and Cook, K. A., Eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, Los Alamitos, CA, 2005.
13. Wegman, E. J. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association 85*, 411 (1990), 664–675.
14. Willett, W., Heer, J., and Agrawala, M. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov.-Dec. 2007), 1129–1136.