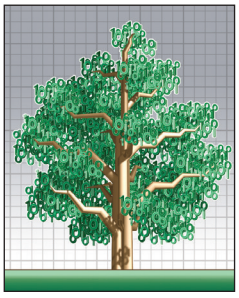
Editor: Munindar P. Singh • m.singh@ieee.org



Rethinking Visual Analytics for Streaming Data Applications

R. Jordan Crouser • *Smith College*

Lyndsey Franklin and Kris Cook • *Pacific Northwest National Laboratory*

Visual analytics is entering a period of renewed growth due to a shift in focus from static to streaming data applications. In this article, the authors illustrate several challenges arising from this pivot and suggest potential avenues for future exploration.

In the age of data science, the use of interactive information-visualization techniques has become increasingly ubiquitous. From online scientific journals to the *New York Times* graphics desk, the utility of interactive visualization for both storytelling and analysis has become ever more apparent. Many visual analytics systems employ an overview first, zoom-and-filter, details-on-demand model, which enables the reader to first get a big picture view and then dig deeper into the data. As these techniques have become more readily accessible, the appeal of combining interactive visualization with computational analysis continues to grow.

Arising from a need for scalable, human-driven analysis, a primary objective of visual analytics systems is to capitalize on the complementary strengths of human and machine analysis, using interactive visualization as a medium for communication between the two. These systems leverage developments from the fields of information visualization, computer graphics, machine learning, and human-computer interaction to support insight generation in areas where purely computational analyses fall short.

Over the past decade, visual analytics systems have generated remarkable advances in many historically challenging analytical contexts. These include areas such as modeling political systems,[1] detecting financial fraud,[2] and cybersecurity.[3] In each of these contexts, domain expertise and human intuition is a necessary component of the analysis. This intuition is essential to building trust in the analytical products, as well as supporting the translation of evidence into actionable insight.

In addition, each of these examples also highlights the need for scalable analysis. In each case, it's infeasible for a human analyst to manually assess the raw information unaided, and the communication overhead to divide the task between a large number of analysts makes simple parallelism intractable. Regardless of the domain, visual analytics tools strive to optimize the allocation of human analytical resources, and to streamline the sensemaking process on data that are massive, complex, incomplete, and uncertain in scenarios requiring human judgment.

Streaming Data: A New Frontier

The analysis of streaming data (data that are generated continuously rather than collected in a single pass) presents an altogether new set of challenges for the designers of visual analytics tools. In a streaming context, the user expends much time and cognitive effort trying to stay abreast changing conditions in a complex data environment that's ripe for misinterpretation. Sampling and filtering mean that data are incomplete in the best of circumstances. Because

 Published by the IEEE Computer Society 2469-7087/19/$33.00

of this, traditional visual analytics systems fall short in many streaming data contexts, as it becomes impossible to maintain an up-to-date overview without derailing the analyst's working model of the situation.

The power grid represents a case study of one of the largest, continuously operated streaming-data machines in the US. It produces enormous volumes of data from a multitude of in-the-field sensors, which requires real-time situation awareness to maintain safe and efficient operations. Intensive observational studies and task analyses have enabled insight into the complex behavior of trained grid operators (see Figure 1).[4] Successful analysis requires shared awareness of tasks and the current state of the grid, both those directly under the operator's control as well as those adjacent to controlled regions.[5] Specialized decision support systems such as M-DART[6] have shown promise in supporting the offloading of certain low-level tasks, enabling grid operators to interact with information at levels that support better anomaly detection and facilitate causal inference. Approaches such as these are needed to help operators cope with the volume of information and alarms that currently occupy the majority of their time.[5]

The power grid is in constant operation at all hours of the day and night, generating new data to be responded to, logged, and archived for both accountability and future-planning purposes. This volume makes proactive responses to changing conditions difficult: small signals of trouble such as a poorly performing piece of equipment in a remote substation are lost until critical alarms draw the attention of operators. By the time operators are made aware, it might be too late to avert an incident. The ability to find small but important signals in the noise of daily operations could be the difference between small-scale disturbances and fully cascading failures.

Some of the most challenging problems for operators could also benefit from visual analytic systems. For instance, updating grid reliability models to handle real-time data and making them accessible to control rooms would support the mitigation of cascading substation losses. In this context, the effect of even modest improvements in reaction time can't be understated: the time elapsed in responding to power grid events can mean the difference between a brief flicker of the lights and the devastating, potentially life-threatening effects of a full scale blackout.

Challenges in Streaming Data Analysis

This example demonstrates the need for clear communication regarding changes in data over time, as well as how these changes might alter a user's understanding of the past and expected future. When analysts are working with streaming data, a substantial portion of their time and cognitive effort is spent trying to stay on top of changing situations within a complex data environment that's ripe for misinterpretation. Because of the incoming data's overwhelming scale, sampling and filtering are a necessity: this means that even in the best of circumstances, the data that ultimately reach the analyst are incomplete. Initial results generated using fast heuristics might be contradicted by more computationally intensive, more accurate analysis that comes in later. As a result of these simultaneous, asynchronous processes, data might arrive out of order, with information about temporally later phenomena becoming available before data about precursor events that provide important context. These complexities in change and uncertainty add to a user's cognitive load, presenting several new challenges within the streaming visual analytics lifecycle (see Figure 2).

Figure 1. A transmission control dispatcher monitors multiple data streams simultaneously, bringing in external context while trying to keep pace with a rapidly evolving data landscape. The North American Electric Reliability Corporation oversees the reliability of a grid that provides electricity to 334 million people through eight regional entities. The Western Electricity Coordinating Council alone oversees 121,200 circuit-miles of transmission lines connecting 350 US and 34 international entities serving a population of more than 80 million. (Photograph by Eric Andersen at Pacific Northwest National Laboratory.)

Challenge 1: Orientation

Imagine an operator working in the control room of a transmission utility in the power grid. Every day, this operator receives a schedule of planned maintenance, anticipated impacts to the grid, and the scheduled use of transmission resources. The operator combines this information with daily weather forecasts and historical grid performance data to complete a mental model of their day and possible scenarios for grid operations. These activities, as well as any other preparatory tasks, are what we refer to as *orientation*.

In a general streaming context, the analyst often faces the daunting

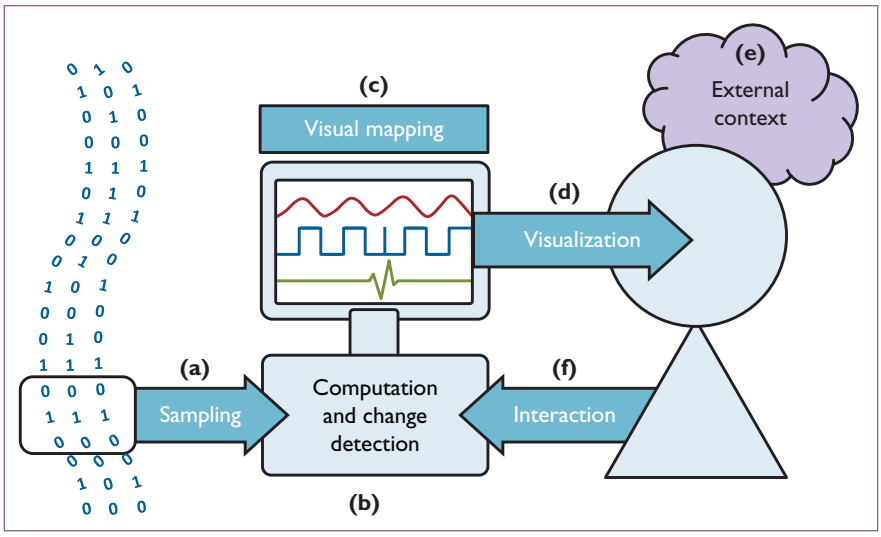


Figure 2. Streaming visual analytics lifecycle. (a) Data are sampled from various streams; (b) sampled data are processed, relevant changes detected; (c) the processed sample is mapped onto various visual dimensions; and (d) the visualization is interpreted by the analyst (e) in the context of domain knowledge and other external information, which drives (f) interaction with the visualization and underlying model.

task of testing competing lines of reasoning on multiple data streams in tandem. These streams might be sampled, filtered, error-prone, and uncertain. They might not be sampled at the same resolution, or they might come in at irregular intervals. Moreover, the pace at which the data are changing could preclude the use of traditional exploratory data analysis strategies for orientation. By the time the computation of even a modest statistical model terminates, the data landscape could be completely different. Analysts of streaming data are thus presented with a conundrum: they must address problems requiring human intellect, but they must also adapt to machine speeds. They must build a robust understanding of the current state of the data, as well as any relevant changes, and they must do so without the luxury of unlimited hindsight. Meeting this challenge will require a reimagining of the overview first model, perhaps with an emphasis on the communication of critical change rather than the communication of full history.

Challenge 2: Reorientation

Generally, once analysts have successfully oriented themselves, they must then be able to efficiently identify and react to new developments that violate their assumptions and expectations. They must rapidly refine their understanding, and then generate and test new hypotheses. They must swiftly interpret and reinterpret incoming and historical data in light of these changes, and ultimately use these new models to consider potential futures states they didn't or couldn't previously anticipate.

Early research into how humans build and use mental models[7] provides an incisive window into why this presents such a daunting challenge. Faced with an overwhelming barrage of competing signals and relatively expensive biological computation, we build abstracted, highly simplified models of the world to cope. This results in some fairly predictable behavior. For example, we know that mental models form quickly but change slowly, and that we tend to see what we expect to see.[8] We know that new information gets incorporated into the existing mental model. Moreover, we know that when presented with competing information, the brain will go to extraordinary lengths to avoid recalibration.

Returning to our example grid operator, we can imagine a number of scenarios that he or she must deal with that weren't a part of earlier orientation tasks. For instance, the unexpected loss of a transmitter at a substation would impact the operator's ability to stick to the day's scheduled outages. This would trigger reorientation activities that could include re-evaluating upcoming plans for feasibility, cancelling low-priority maintenance, contacting adjacent utilities for help, and triggering repair efforts to resolve the situation and return to normal. To support reorientation at the frequency required in streaming data applications, we must investigate more effective mechanisms for alerting the operator to changing conditions as well as triage support technologies that go beyond rigid, rule-based methods.

Challenge 3: Summary Statistics

During the course of a regular workday, grid operators must carefully match the supply of electricity with consumer demand. Generate too much and resources are wasted, too little and customers experience blackouts. This estimation is re-evaluated every hour in transmission utilities and is based on operators' ability to make inferences from actual electricity usage on similar days.[5] Accurate statistics and summarization of all available history is critical in this process, as the operator might have to decide which of several similar historical usage patterns is the best predictor of current conditions.

Summary statistics are the backbone of many data visualization techniques that highlight anomalies and other interesting events in the data as deviations from a baseline. However, as data volume increases, standard summary statistics and aggregate measures start to lose their descriptive power. This is especially apparent when large, sudden changes occur in the baseline. In such cases, summary statistics that appear stable in the short term might be significantly misleading in the long term. When working with streaming data, we might need to evaluate alternative techniques such as moving averages and local regression.

Challenge 4: Scalability and Approximation

To be most effective, an interactive system needs to update and render information at a rate of at least 12 frames per second. In many streaming data applications, the scale of the data to be processed and visualized makes this benchmark difficult or impossible to achieve. As the refresh rate starts to fall below 10 frames per second, the delay between a user's action and the system's response begins to disrupt the user's ongoing cognitive processes, which greatly diminishes their ability to explore the data and test hypotheses. Thus, to achieve the benefits of interactive visual analysis on streaming data, we might need to gradually, but intelligently, degrade the data sampling, analytics, and visual representations in favor of more efficient approximations.

In many cases, we can use what we know about human perception to help inform these approximation strategies. For example, we know that information density in visual analytics is bounded by the discrimination power of the human visual system, which is determined by well-understood perceptual limits. Therefore, when continued refinement won't lead to a human-detectable change to the visual display, we know that we can terminate any additional computation without degrading the visual display's accuracy. Bounds such as these can inform an approximate/adaptive computing strategy that enables us to intelligently trim extraneous computation, maintaining locally optimal performance under changing computational circumstances.

Discussion

Although we framed these challenges in the context of power grid resilience, these same issues emerge wherever streaming data are in play; for example, cybersecurity, medicine, climate change, and more. The blind reapplication of established strategies for visualizing static data might not succeed when applied to streaming data, even when the systems were designed for similar tasks in similar domains. Such challenges aren't restricted to the domains highlighted previously; in the wake of ever-evolving data landscapes and human intelligence that proves difficult to scale, the visualization community faces pressing issues as a whole.

The desire to build effective visual-analytics systems for streaming data will require us to develop novel ways to represent change. These representations must not only be accurate, but must also present identified patterns in the context of the analyst's understanding of the evolving situation. This necessity for context and clarity has been a driving force in the development of visual analytics as a field, and so we might look past solutions for inspiration. For example, consider the utility of video keyframes in facilitating rapid orientation to a lengthy video's content. Could analytic keyframes provide a user with succinct change points in both data and analytic thinking? How would those points be identified and kept up to date, and how would we deal with branching analytical paths? Could they be used to provide both fast orientation, as well as a compact representation of change over a long window of time?

Historically, the analysis of complex data has been an offline process, in which the data's dynamic nature is ignored during analysis. This simplifying assumption is acceptable when the analysis is fast relative to the rate of change of the data. However, in big data environments in which data and conditions continually change, ignoring this change is insufficient. Acknowledging the dynamic nature of real-world problems compels a new line of research to study the effects of streaming data in visual analytics systems. This article is a call to action to address these challenges and enable users to benefit fully from the capability to gain insight from their data in real time. ◻

Acknowledgments

Many of the themes described in this article emerged as a result of a 2016 Workshop on Streaming Visual Analytics, organized by the Laboratory for Analytic Sciences at North Carolina State University, and various academic collaborators. We thank the organizers as well as the more than 40 workshop attendees for their contributions to the workshop and support in developing a guiding vision for streaming visual analytics. Some of the research described here was sponsored by the US Department of Energy through the Analysis in Motion Initiative at the Pacific Northwest National Laboratory. The views and conclusions contained in this document are those of the authors and shouldn't be interpreted as representing the official policies, either expressed or implied, of the US government.

References

1. R.J. Crouser et al., "Two Visualization Tools for Analyzing Agent-Based Simulations in Political Science," *IEEE Computer*

Graphics and Applications, vol. 32, no. 1, 2012, pp. 67–77.
2. R. Chang et al., "Scalable and Interactive Visual Analysis of Financial Wire Transactions for Fraud Detection," *Information Visualization*, vol. 7, no. 1, 2008, pp. 63–76.
3. L. Harrison et al., "NV: Nessus Vulnerability Visualization for the Web," *Proc. 9th Int'l Symp. Visualization for Cyber Security*, 2012, pp. 25–32.
4. J.H. Obradovich, "Understanding Cognitive and Collaborative Work: Observations in an Electric Transmission Operations Control Center," *Proc. Human Factors and Ergonomics Society Annual Meeting*, vol. 55, no. 1, 2011, pp. 247–251.
5. J. Scholtz et al., "Cybersecurity Awareness in the Power Grid," *Advances in Human Factors in Cybersecurity*, Springer, 2016, pp. 183–193.
6. N. Lu et al. "A Multi-Layer, Data-Driven Advanced Reasoning Tool for Intelligent Data Mining and Analysis for Smart Grids," *Proc. IEEE Power and Energy Society General Meeting*, 2012.
7. H.A. Simon, "A Behavioral Model of Rational Choice," *The Quarterly J. Economics*, vol. 69, no. 1, 1955, pp. 99–118.
8. R.J. Heuer, *Psychology of Intelligence Analysis*, Center for the Study of Intelligence, 1999.

R. Jordan Crouser is an assistant professor in the Department of Computer Science at Smith College. His research interests include human-computer interaction and visual analytics, with a focus on human computation and human-computer teams. Crouser has a PhD in computer science from Tufts University. Contact him at jcrouser@smith.edu.

Lyndsey Franklin is a user experience research scientist at the Pacific Northwest National Laboratory. Her research interests include information visualization and visual analytics in domains such as energy and environment, healthcare, and cybersecurity. Franklin has an MS in computer science from the University of Maryland, College Park. Contact her at lyndsey.franklin@pnnl.gov.

Kris Cook is the Analytical Insights Technical Team lead at the Pacific Northwest National Laboratory. Her research interests include the use of mixed initiative techniques to support analysis and the exploration of visual analytics techniques to support sensemaking using streaming data. Contact her at kris.cook@pnnl.gov.

This article originally appeared in IEEE Internet Computing, *vol. 21, no. 4, 2017.*