7

Human-Guided Visual Analytics for Big Data

Morteza Karimzadeh, Jieqiong Zhao, Guizhen Wang, Luke S. Snyder, and David S. Ebert

In this chapter, we provide an overview of the research and practice in visual analytics with a specific focus on decision-support systems that facilitate generating useful information from big, unstructured, and complex data. We first define what is usually referred to as *big data* and its unique characteristics. We then define visual analytics and human–computer collaborative decision-making (HCCD) environments, compare and contrast human-in-the-loop analysis methods with automated algorithms such as machine learning models, and explain how these approaches complement each other for real-world problem solving. To ground our discussions, we provide an overview of four exemplar visual analytics systems with applications in various domains, including humanitarian relief, social media analytics, critical infrastructure vulnerability modeling, resource allocation, and performance evaluation using multidimensional data.

MOTIVATION AND OPPORTUNITY

Advanced analytics and computational algorithms enable the transformation of the evolving deluge of digital data into useful and actionable information. However, as data sets continue to increase in size and complexity in the digital

We wish to acknowledge the work of Calvin Yau, Junghoon Chae, Jiawei Zhang, Sungahn Ko, Abish Malik, Ross Maciejewski, Kelly Gaither, William Ribarsky, and Isaac Cho for their contributions to the systems and research overviewed in this chapter.

http://dx.doi.org/10.1037/0000193-008
Big Data in Psychological Research, S. E. Woo, L. Tay, and R. W. Proctor (Editors)

age, analytics become more computationally demanding, time consuming, and less clear to human analysts, and the analytics output produces large amounts of information that can overwhelm the human user. Some complex algorithms, such as machine learning models, are designed to reduce the massive amounts of complex data to manageable sizes and dimensions. However, the complexity and, at times, lack of transparency of the algorithms result in humans being unable to understand and trust the results (Burrell, 2016). This contradicts the original value of computing, as noted by Hamming (1962): The ultimate purpose of computing is to gather insights into the dynamic processes of the world instead of merely generating numbers.

These problems are exacerbated with big data, where the data is large or complex in one or more of three aspects: volume (size), variability (number of variables or types of data), and velocity (rate of incoming data—e.g., real-time, streaming; Zikopoulos & Eaton, 2011; see Chapter 2, this volume). Big data poses additional challenges for analysis techniques and human ability to synthesize, explore, and distill big data into significant and relevant information. Visualization that is combined and interlinked with data analytics can help alleviate these challenges. Moreover, visualization that is integrated within the analytics pipeline can help confirm the expected and discover the unexpected (Thomas & Cook, 2006). As pointed out by Tay et al. (2017), visualization is key to solving many big data analysis problems if the following four issues are carefully considered in the design: (a) identification (isolating and highlighting relevant data and patterns), (b) integration (combining different data sources and different models to reveal new insights), (c) immediacy (streaming, real-time, and time-sensitive data), and (d) interactivity (user manipulation and exploration to inductively uncover and identify new patterns). The field of visual analytics expands on previous work in these areas to assist researchers, analysts, and decision makers in their use of data for effective discovery, monitoring, analysis, and decision making. In this chapter, we explore the background, potential, challenges, exemplar techniques, and applications of human-guided big data visual analytics, specifically in HCCD environments.

WHAT IS BIG DATA?

The term *big data* has emerged in the last decade to describe data that can be characterized by large volume, variety, or velocity (Zikopoulos & Eaton, 2011). *Volume*, intuitively, refers to the large size of the data that have to be stored, queried, analyzed, and visualized. Large volumes make storing and querying on traditional system architectures challenging. Infrastructures such as Apache Hadoop[1] are used to distribute computational operations on a network or cluster of computers to enable processing large amounts of data. Modern

[1]https://hadoop.apache.org/

graphical processing units (GPUs) also enable speeding up compute-intensive algorithms by parallelizing computations simultaneously into hundreds of thousands of computational threads. Approximate query techniques in visual analytics (Fisher, Popov, Drucker, & Schraefel, 2012; A. Kim et al., 2015) enable interactive data exploration by reducing the data volume in the computation process and providing users approximate results with bounded errors.

Variety in big data refers to the heterogeneity of the data being collected, such as text, numeric, geographic location, and temporal data. Different data structures that are collected at different speeds and sampling rates make drawing connections and identifying patterns a challenging task, and traditional automatic data analysis makes the fusion of information hard, whereas visual analytics systems take advantage of human ability to find patterns and identify connections.

Velocity refers to the speed at which data is being collected (e.g., streaming social media data). Data collected at streaming rates require methods with low computational complexity that can process the incoming data at the same speed. Visual interfaces should be able to use open (and usually two-way, between the user interface and the server) communication technologies such as WebSocket[2] to seamlessly update the user interface with incoming data (or the real-time result of analysis and processing of the incoming data) to enable visual analytics of data with high velocity.

Big data includes items that are interdependent, such as social network data with links including follower and followee, repost, quote, spatial proximity, or topical relatedness. Interdependence of data items are usually recorded in different data structures with different analytical needs (e.g., social network data or spatial coordinates data for social media users and posts), making pure computational analysis more challenging. Humans, however, can find patterns and relationships in heterogeneous data while connecting it to the context that might not necessarily be captured in data, especially if human users are presented with appropriate visualizations.

WHAT IS VISUAL ANALYTICS?

Visual analytics is defined as the science of analytical reasoning facilitated by interactive visual interfaces (Thomas & Cook, 2006). Visual analytics enhances the cognitive abilities of humans by maximizing the use of their perceptual and cognitive capabilities in an integrated visual analysis and exploration environment (Eick & Wills, 1993; S. Kim et al., 2013; Stasko, Görg, & Liu, 2008; Zhao, Chevalier, Pietriga, & Balakrishnan, 2011). The primary goal of visual analytics is to provide insight into various phenomena to enable more effective research, analysis, and decision making. As data size and complexity have grown in the era of big data, the role of visual analytics has become

[2]https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API

increasingly important. HCCD environments effectively and efficiently combine the experience, contextual information, and expertise of the human user with the power of human-guided computational analysis, which, in turn, enhances the human-centered decision-making process.

Visual analytics is the intelligent evolution of visualization, bringing the fundamental understanding of perception and human cognition used in visualization to the realm of analysis of data using and through interactive visualization, instead of merely using visual techniques to communicate the analytical results to users. To achieve this, visual analytics incorporates the principles of design and cognitive science to identify appropriate visual metaphors for data or analytical results, with a strong emphasis on creating perceptually effective representations at the appropriate cognitive level for each analytical task (see the examples in the Opportunities and Examples subsection).

As discussed earlier, new analytical techniques and technologies are being adopted to gain insights and steer decision making in various fields, leveraging the vast amounts of complex data, which are growing at exponential speeds since the emergence of the Internet. In particular, machine learning and artificial intelligence algorithms are being applied to generate information from data and predict future states. Generally, these methods involve sophisticated calculations and numerous input parameters. Visual analytics helps incorporate human domain knowledge through the users' iterative refinement of inputs using visual interfaces to improve the calculated results. More important, without visual techniques, it is difficult and at times impossible (depending on the models used) for users to understand the causality relationship between inputs and derived results. Oftentimes, users may suspect the reliability of the generated results due to the overly complex design of the algorithms. Visual analytics can bridge the gap between the results derived by these automatic algorithms and reasonable interpretation through model-integrated visualization techniques. In other words, visual analytics not only improves data analytics (through the incorporation of human domain knowledge, expertise, and analytical abilities) but also increases trust in, and therefore, the adoption of, the analytical results.

The usefulness of a visual analytics system can be characterized by its utility and usability (Ellis & Dix, 2006). *Utility* refers to the ability of the system to support users in completing the required tasks, and *usability* describes the ease of use of the system in completing the same required tasks. Therefore, utility is more or less an objective measure, whereas usability is related to the subjective satisfaction and user experience, describing the success of a system in terms of intuitive design, ease of learning, efficiency of use, and memorability (Usability.gov, n.d.-a). To ensure a system's utility and usability, visual analytics researchers usually adopt the user-centered design paradigm (Usability.gov, n.d.-b), and work closely with stakeholders at various stages of design and development. *User-centered design* usually entails identifying the context of use, specifying requirements, and creating design solutions. This last stage itself is typically an iterative process in which multiple design ideas are presented to users (via sketches, mockups, or actual implementations of the system), feedback is sought and intermediate "formative" evaluations are conducted,

leading to refining the design and presenting the system again for more feedback (Roth, Ross, & MacEachren, 2015). After a system implementation is finished, researchers conduct final "summative" evaluations through various evaluation protocols (Ellis & Dix, 2006) to scientifically report on the usability and/or utility of the system (for the particular target purpose and particular target users).

Visual analytics has been successfully applied and reported on to support (a) advance research and scientific activities or (b) domain users for practical needs outside academia or industry. For instance, MacEachren, Stryker, Turton, and Pezanowski (2010) reported on HEALTH GeoJunction, a visual analytics application for exploring health-related scientific publications using place–time–theme queries (e.g., studies about Ebola in Africa in 2010). Diakopoulos, Naaman, and Kivran-Swaine (2010) created and evaluated a system for journalists to sift quickly through large amounts of social media traffic about events of interest to identify public sentiment. Wade and Nicholson (2010) reported on the successful use of visual analytics in the aviation safety engineering industry, leading to changes in flight training manuals. Jaiswal et al. (2011) used GeoCAM in computational linguistics research to interpret human-generated route directions. Karimzadeh, Pezanowski, MacEachren, and Wallgrün (2019) described the successful application of a semiautomatic visual analytics platform to create annotated textual data sets (Wallgrün, Karimzadeh, MacEachren, & Pezanowski, 2018) to support the development of automated algorithms for geolocating (i.e., mapping) textual documents. Wagner et al. (2019) reported on the successful evaluation and application of KAVAGait, a system for clinicians to support clinical analysis of patients' gait using complex data sets while incorporating clinicians' domain knowledge. Throughout the rest of this chapter, we also address the design and capabilities of a few other visual analytics systems, elaborating on their use by end-users.

VISUAL ANALYTICS TO TACKLE BIG DATA

Traditional big data analytics may leave out some context in modeling the complex world. Data is rarely complete, and it does not incorporate all the relevant information necessary in decision making (Brooks, 2013). Decision making (by humans) always happens within context; policy makers or executives rarely rely merely on numbers to make decisions. They contextualize analytical results within the broader context of society, risks, and long-term outcomes and, at times, may even go against the analytical (quantitative) results to have more favorable broader impacts when considering every aspect that may not be reflected in analysis. Visual analytics enhances computational algorithms by incorporating humans' extensive information, experience, and domain knowledge that may not be collected in the data used for analysis.

Further, data analysis relies heavily on quantifiable data. For instance, in population dynamics modeling, projections of the population in the future are generated on the basis of the spatiotemporal measurements and dynamics

of the current population (Uhl et al., 2018; Zoraghein, Leyk, Ruther, & Buttenfield, 2016). Qualitative information, fuzzy data, and social aspects (decisions, emotions, connections, or opinions), although intuitive to humans, are difficult and, at times, impossible to quantify and analyze. As another example, novel sentiment analysis methods may underperform in determining the affective states and subjective information of statements that include exaggerations, sarcastic remarks, jokes, and negations (e.g., "The whole house is flooded. How great!" may confuse a sentiment detection algorithm because the author is using *great* sarcastically). Such naive cases are easier for human readers to identify; therefore, humans in the loop can improve the computational results and alleviate the potential of biased results.

Big data introduces another challenge in data analysis: As the number of data items increases, so does the number of statistically significant relationships. Many of such significant relationships may be misleading or irrelevant. This overwhelms an analyst's ability to find meaningful relationships due to a high ratio of noise to signal. Visual analytics provides interactive querying, sorting, detail on demand, and contextual information to help users focus on actionable data and patterns that matter the most.

Finally, real-world big problems are complex and multifaceted with multiple parameters and interdependencies. Whereas in classic statistics, a controlled or observational study is conducted, many real-world problems cannot be solved by trial and error or analysis of retroactive observations, and results from one experiment are not generalizable to another case. For instance, no two natural events are the same. No "earthquake drill" can simulate the impacts of an earthquake; therefore, traditional data analysis cannot be used to simulate the impacts of one event. Real-time, context-enabled, multifaceted sensing, modeling, and decision-making environments are necessary for human users to evaluate and respond to any specific earthquake and natural disaster.

CHALLENGES IN USING MACHINE LEARNING

Machine learning-based approaches have certain limitations for use in some real-world problem-solving scenarios, where visual analytics is well-positioned to make significant contributions. In this section and the next one, we review the potential and limitations of machine learning and point out situations in which visual analytics can help remedy some of the limitations for real-world use cases.

Machine learning approaches can be generally categorized into supervised or unsupervised learning methods (Alpaydin, 2009). In *supervised learning*, a model is trained with labeled data for different prediction tasks such as classification (e.g., pictures classified into "dog" or "cat" categories) or regression. The goal of supervised learning algorithms is to find the relationships or structures in the input data that allow a model to generate correct output labels. These correct outputs are determined according to training data. *Training data* includes

examples for which input and output are known, usually through the process of human manual annotation of input data (e.g., labeling a cat picture with *cat*). The way the training data is sampled and the method using it is annotated influence the generated automated models and may introduce the sampler's assumptions or bias, especially if such a sample represents a snapshot in time, space, or event type. Most important, training data that do not reflect the real world may lead to erroneous results that may go unnoticed (unless the assumptions and results are visually displayed to users with domain knowledge), and any sampling, by nature, introduces the biases and perceptions of the samplers (Wallgrün et al., 2018). Furthermore, dynamic phenomena, such as various characteristics of human behavior, do not lend themselves to a one-off training of a machine learning model because such characteristics change due to human agency and interdependence of actions. Models generated for one particular event, time, or place may not work as well in other places. Overfitting to training data is always a challenge, too, meaning that the model can predict excellent results for the test data set but not for unseen input data.

Machine learning, like traditional data analysis, may struggle to model human and social contexts that cannot be easily collected in data and, therefore, lacks the ability to generate the narratives that a human analyst can produce using sequences of events, external forces, their relationships, and context. Once such context changes, machine learning algorithms still perform according to the initial model training, whereas humans can base their understanding on the existing model results but also draw the necessary connections with the new context, identify the potential differences and significance in the outcome, and make appropriate inferences or decisions.

Unsupervised learning does not require labeled data or pretrained models. Instead, the training algorithm directly learns from current data. For example, the K-means clustering algorithm finds the natural categories of data by maximizing "within-cluster" similarity and minimizing "inter-cluster" similarity. It can be used, for instance, to identify clusters of grades earned in a class (i.e., the natural grouping of grades that are similar to each other). Still, K-means requires the upfront knowledge of the number of clusters (information that humans with domain knowledge may have a better understanding of, even in the case of some unsupervised methods that can identify a purely computationally optimum number of clusters). Also, the generated clusters are shifted significantly if outliers exist in the data. Again, humans, if equipped with the right tools, visuals, and information, are more reliable at identifying erroneous outliers or natural extreme values depending on the context.

Regardless of whether supervised or unsupervised methods are used, human involvement can ensure relevant results for changing context or dynamic phenomena. Visual analytics provides the infrastructure for human experts to adjust input and hyperparameters (e.g., model configurations and structure, as in Das, Cashman, Chang, & Endert, 2019), monitor a model performance (for precision or speed, as in Zhao et al., 2019), compare results against context,

and correct the erroneously generated output labels to provide real-time examples for online (real-time) learning (e.g., retraining or incremental training of models, where new labels are used to improve the existing models, as in Snyder, Lin, Karimzadeh, Goldwasser, & Ebert, 2019). Most important, visual analytics enables the use of machine learning within what-if scenarios, where users can see the outputs based on different input parameters that reflect different human decisions, assumptions, or policies. We describe examples of such cases throughout the rest of this chapter.

THE DEEP LEARNING PROMISES AND OVERPROMISES

Deep learning is a specific type of artificial neural network with many layers (thus called *deep*) that has partly been revitalized due to the recent advancements in hardware (LeCun, Bengio, & Hinton, 2015). Specifically, input values (e.g., pixel values in images) are multiplied by weights (which are ultimately optimized) and added many times with constants to generate the desired output values (e.g., digit labels for images containing handwritten digits). With the advent of strong GPUs and even commercial deep-learning accelerators (e.g., Nvidia DGX-1), it is possible more than ever to apply deep learning to various domains for classification purposes. Deep learning has provided much better performance in some fields such as computer vision and has shown great promise in other domains, such as natural language processing, though not to the same level of maturity yet.

Deep learning relies heavily on large amounts of training data. The gold standard (ground truth) examples are used in optimizing the weights and constants in the neural network and generating a model that can predict labels for the "testing" data with acceptable accuracy (and, therefore, unseen data). Testing data also usually are manually annotated by humans to ensure that the generated models can produce labels for examples that were not used during the training phase of optimization.

Deep learning models require a high number of training examples for acceptable outputs (much higher compared with statistical machine learning), given that many weights and constants in all the layers have to be optimized. In other words, deep learning models' performance depends heavily on training and testing gold standard data, which is neither cheap nor easy to generate. However, as introduced in the previous section, training and testing data may only represent a snapshot of a time, space, phenomenon, or event; for example, a traffic congestion detection model that works for particular modes of transportation may suffer inaccuracies if new modes of transportation are introduced or new policies are put in place. Building up a new representative training data set is costly and laborious and, again, would only capture the variation of the real world made on hard assumptions of sampling at the time.

In certain scenarios, deep learning alone may suffer from the issues discussed in the previous section: the inability of users to adjust input parameters

dynamically (to account for what-if scenarios of dynamic phenomena that need flexible inputs), detaching results from the context, and being specific to the training data instead of accommodating spatial or temporal variability in the phenomenon. Visual analytics systems integrated with online learning models provide a great opportunity for alleviating these problems. We discuss systems adopting this approach in the following sections.

EXPLAINABLE ARTIFICIAL INTELLIGENCE

Deep learning models are essentially classification methods, where input values are mapped to output values or labels. Unlike traditional statistics, deep learning models are not geared for "explaining" the relative importance or significance of input parameters, and therefore, deep learning models are not "explanatory." For instance, a simple linear regression model can explain the contribution of the "number of cars" or "price of gas" (as independent variables) to the "number of traffic jam incidents" (as the dependent variable). After the regression model is solved, the analyst can examine the generated coefficients and significance values of independent variables and infer how much a unit increase—for instance, in the price of gas—would translate into a decrease (or increase) in traffic jams and if that value is in fact significant. Deep learning models, however, primarily focus on predicting the number of traffic jams without directly explaining the relative contribution of independent variables. This poses a problem for decision makers and policymakers who do not just have to use the classification system but have to understand the underlying phenomenon for planning and policy making.

Moreover, a user's ability to trust the conclusions of machine learning models may be affected negatively by the lack of transparency in the models. Although deep learning and other statistical machine learning models have made significant progress on many challenges, many are opaque black boxes with limited explanatory capabilities.

Explainable artificial intelligence (XAI) is an emerging field of research that seeks to enhance traditional machine learning techniques with explanatory metrics. For instance, current classification models and neural networks can be difficult to understand and unclear with regard to how classification and clustering decisions are made. In other words, it is unclear which specific characteristics of the input data (or independent variables) cause an item to be classified into a certain class. As a result, users may struggle to trust AI outputs. XAI seeks to address this problem by providing explainable models that directly indicate what decisions were made and why, allowing users to more effectively understand and act on the models' outputs (Gunning, 2017). Such explanatory models can be presented and generated in many forms. For instance, some approaches use auxiliary integrated machine learning models that seek to identify the discriminatory features of input data (that distinguish a certain class) and assign a natural language or visual cue to such discriminatory features, effectively explaining (in human language or visual cues) what

specific parameters in the input data led to the machine learning model decision. Such approaches either use ground truth training data that humans have annotated with both labels and (natural language or visual) explanations of the discriminatory features (Rajani & Mooney, 2018) or automatically harvested corpora of such explanations (e.g., image captions sourced from the web; Venugopalan, Hendricks, Mooney, & Saenko, 2016).

Other XAI approaches leverage interactive visualizations to focus on the computational components of a machine learning model with the goal of (a) increasing performance through hyperparameter[3] optimization or (b) reducing the computational time necessary for computation-heavy models (Zhao et al., 2019). Such models do not necessarily need extra annotated explanations for training. Instead, they focus on model parameters, model structure, computation time for each stage, bottlenecks, or functions that lead to higher model performance (Kahng, Andrews, Kalro, & Chau, 2018; Wongsuphasawat et al., 2018).

HUMAN–COMPUTER COLLABORATIVE MACHINE LEARNING FOR BIG DATA

In the first half of this chapter, we discussed the importance of human involvement in analytical tasks to incorporate domain knowledge, social and changing context for dynamic phenomena, adjusting input variables, and monitoring model performance. However, human involvement in big data analytics problems can be beneficial from a computational standpoint, as well. Processing the entire big data to get accurate results could be a lengthy process, even with advanced computational architectures that use more computational resources, because the growth of data has significantly surpassed that of hardware resources (Mozafari, 2017).

Involving humans in big data analysis can reduce computational latencies efficiently and effectively because, in many situations, approximate results of analysis on fewer representative data points can satisfy the analytical requirements of end users (Fisher et al., 2012). One preliminary experiment (Wu & Nandi, 2015) indicated that a query to estimate the average of a data set can eliminate the need for sampling 10^4 more data through a reduction in the perceptually perceived error by 10−5. Researchers across both the database and visualization fields have devised a series of approximate query methods and integrated visual analytics approaches to facilitate the decision making of end users using approximate query processing. Researchers in the database field have explored novel approximate data query techniques to generate samples with the specific consideration of human perception of the generated samples. For instance, Park, Cafarella, and Mozafari (2016) proposed a spatial

[3]Hyperparameters is a model configuration parameters (e.g., number of hidden layers in a deep learning approach) whose value is set before the learning process.

sampling method that improves the perceptual accuracy of generated visualizations. In a similar vein, Ding, Huang, Chaudhuri, Chakrabarti, and Wang (2016) proposed an approach to assist end users in specifying the approximate query accuracies in "Group By" aggregation queries (in which an approximate aggregate of an attribute for several data items is retrieved). A. Kim et al. (2015) advocated a ranking-aware sampling method to generate data samples through which approximate results of "Group By" queries keep the same ranking order as the exact results from the analysis of the entire data.

In the visual analytics field, a series of proposed approaches allow end users to analyze big data through providing quick approximate results using fewer data items and progressively improving the accuracy of the analytical results until the accuracy satisfies the analytical requirements (Fisher et al., 2012; Mozafari, 2017). Considering the uncertainty of approximate results for end users to make decisions, these visual analytics approaches assist end users to understand better the accuracy of approximate results via customized visual designs that encode statistical measurements of approximate accuracies. In one of our preliminary works in this area, for instance, we proposed a user-driven spatiotemporal big data sampling approach for data residing in remote servers (Wang et al., 2017). Through the well-designed spatial and temporal data indexing, our method focuses on data within the spatial and temporal query ranges expected by users to avoid sampling data outside the query range as much as possible. As a result, a visual analytics system built on this approach loads data from the servers in real time and reduces data transfer and sampling latencies.

Aside from improving the computation time, human-in-the-loop machine learning can help circumvent the need for training data. Attributed to the mixed-initiative user interfaces, these approaches seek to aid the interactive visual exploration process through a combination of machine learning and user domain knowledge. The premise is to enable users to provide interactive feedback to the system for retraining the underlying machine learning model parameters (Badam, Zhao, Sen, Elmqvist, & Ebert, 2016; Wall et al., 2018). Users play a central role in guiding the workflow, and auxiliary machine learning algorithms provide shortcuts by generating candidate results from which users can choose. Over time, the mixed-initiative visual analytics systems leverage newly added (e.g., sensed) data and user feedback into the simulation models (active learning). The interoperation of data and the various machine learning models, along with user feedback, enhances the underlying models and reduces the overall reliance on the need for additional data.

Mixed-initiative visual analytical systems use direct human-user manipulation of the desired results to update the parameters of automatic computational models through interfaces while trying to achieve fluid interactions to represent cognitive processes with externalized cognitive artifacts (Elmqvist et al., 2011; Horvitz, 1999). Users can investigate the model outputs dynamically and iteratively. The analysis process is a reasoning procedure that involves information foraging, formulating hypotheses, and validating results. The system provides instant feedback based on user input. Thus, users have

opportunities to correct their outcome if the results do not match their expectations. However, it is challenging to understand users' actions through learning of trivial and repetitive user interactions. Human-in-the-loop and mixed-initiative visual analytics create new opportunities to collect user behavioral information (e.g., eye gazing, mouse interactions) and analyze that information with machine learning algorithms to understand the semantics of the interactions (intention of performing a specific interaction; Endert, Chang, North, & Zhou, 2015) or predict future interactions (Heer, Hellerstein, & Kandel, 2015).

It is important to recognize that the human-in-the-loop may introduce their own biases in the analysis process. The core idea central to visual analytics is to use the complementary powers of humans and machines. Thus, it is imperative that bias is quantified, visualized (i.e., externalized), and present to the human user as much as possible. For instance, Zhao et al. (2019) demonstrated how an expert human user could modify the features and (hyper) parameters in machine learning models and view the resulting accuracies on the fly with the ultimate goal of selecting optimal models, which are results of data-driven models and the human's contextual knowledge. In such a scenario, if the human user relies too much on their assumptions (i.e., a misconception on strong discriminatory power of a certain hyperspectral index), the resulting performance drop—as quantified and visualized by the system—indicates to the users that their knowledge is biased or not applicable to the current context. Related to the idea of bias and systematic errors is that visualizing uncertainty is an active area of research that strives to address the positivist aspects of visualization by explicitly conveying the risks ensuing in the use of imprecise data or human-introduced errors during analysis (Spiegelhalter, Pearson, & Short, 2011), using which users can steer the analysis and decision-making path.

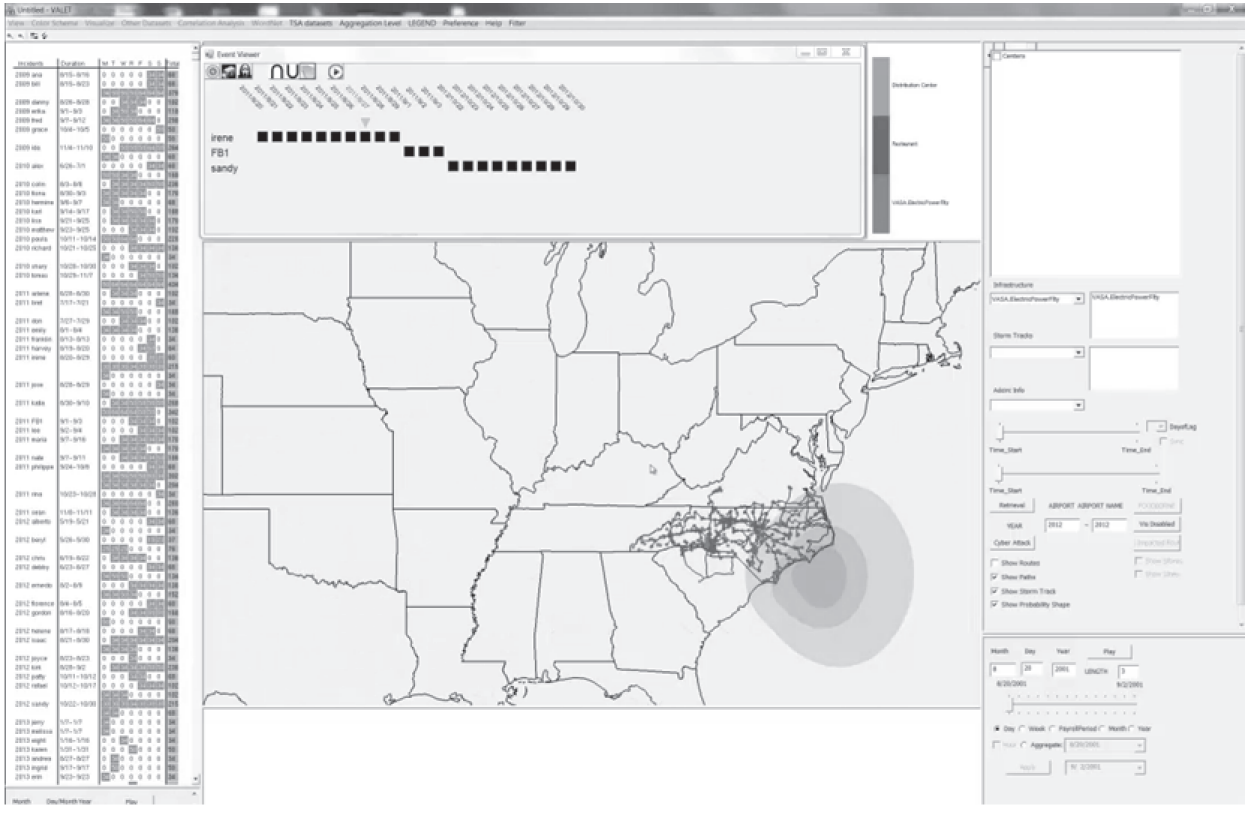
HUMAN–COMPUTER COLLABORATIVE DECISION-MAKING ENVIRONMENTS

An HCCD typically consists of an interactive front-end application (through which the user interacts with the system), as well as data-driven and model-integrated back ends. HCCDs usually include simulation capabilities for what-if scenarios that enable end users to see the outcomes, patterns, and trends in the information based on the decisions and assumptions they introduce into the system. In this section, we describe exemplar systems that exhibit HCCD characteristics with simple integrated models.

The Visual Analytics for Simulation-based Action (VASA) system (Ko et al., 2014) shown in Figure 7.1,[4] is a visual analytics platform for modeling the

[4]The images in this chapter were generated in systems that use color, though they are printed in black and white here. For color images with more information, see http://pubs.apa.org/books/supp/woo

FIGURE 7.1. VASA System Overview, With Calendar View on the Left, Event View on the Top, Map (Geographical) View in the Center, and Advanced Filtering and Querying Panels on the Right (Ko et al., 2014)



This figure shows the simulation of the landfall of Hurricane Irene in North Carolina. In this simulation, power generation units are hit by up to 34-knot winds. The hurricane proxy estimates the impacted restaurants and distribution centers. The system also allows identifying the power outage areas and out-of-service roads, which can be used in computing new food delivery paths. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

impact of various kinds of threats (e.g., natural threats such as hurricanes or human-caused threats such as cyberattacks) on critical infrastructure, such as supply chain systems, road networks, cyber networks, and power grids. VASA includes a set of components encapsulating high-fidelity (i.e., realistic) simulation models for each type of threat and infrastructure that together form a system of systems of individual simulations.

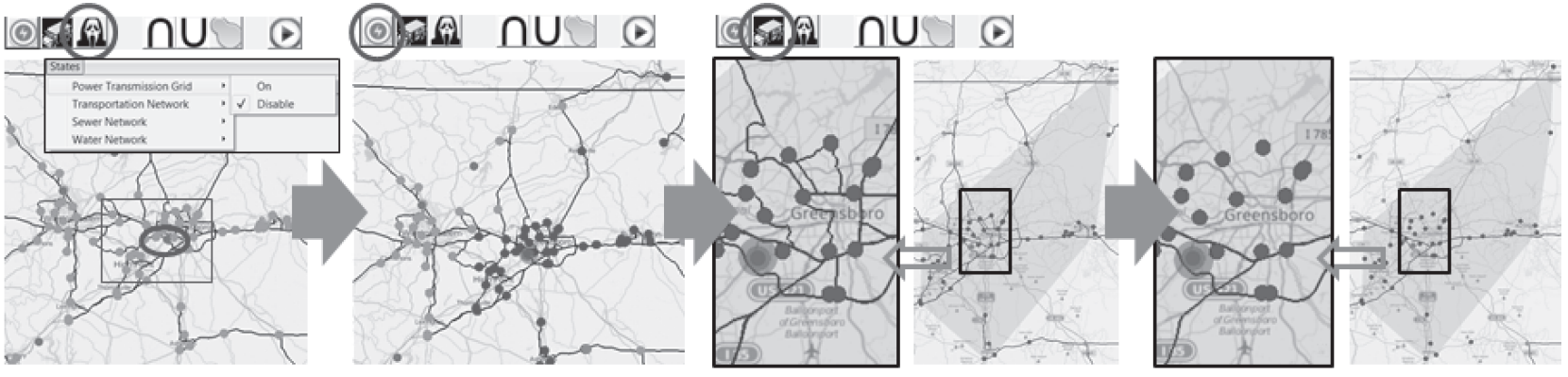
VASA (Figure 7.1) provides a critical infrastructure module that allows analysts to identify vulnerabilities in critical infrastructures (e.g., maritime supply chain networks, power plants, management command centers) in case they are compromised by adverse elements (e.g., cyberattacks). This component, shown in Figure 7.2, models a hierarchical network (e.g., power grid, supply chain), where the nodes (e.g., power generators) are connected with edges (e.g., transmission lines). This model simulates the impacts of the closure of a node and provides information on the other impacted regions in the network (e.g., power outage areas). It helps analysts answer questions such as, "When a main network node is compromised, how do the effects propagate through the network? What other nodes connected to the affected node are impacted, and thus, which critical areas are vulnerable to threats?" The VASA framework provides analysts with the ability to input different models and can be used to study the effects of different cyber threat vectors on critical maritime infrastructures to detect vulnerabilities. Further, the incorporation of multiple displays in the visual analytics environment enhances monitoring capabilities. For instance, cyberattacks combined and cascaded with other natural events (e.g., severe weather) and cyber threats (e.g., attacking systems to disable ports) could drastically exacerbate damages, and multiple displays help users integrate the information from different simulations.

For example, let us examine a scenario where a cyberterrorist has disabled a power generation plant. This scenario is shown in Figure 7.2 (leftmost), where the analyst first disables the plant by selecting a power plant shown as a red rectangle. The model instantly estimates the affected operational facilities in the network (second left, Figure 7.2). The simulation results are rendered by a polygon that represents the area where all facilities are shut down (second right, Figure 7.2). The right-most image provides magnification of the result.

We have obtained initial feedback about the system from various groups. For instance, food chain experts stated that the VASA system helped them identify alternative routes in extreme weather through simulations of hurricanes and the resulting impact on societal infrastructure, as well as the impact on local stores. The regional Federal Emergency Management Agency personnel appreciated the simulation pipeline provided by VASA, which enables proactive planning for severe weather conditions (Ko et al., 2014).

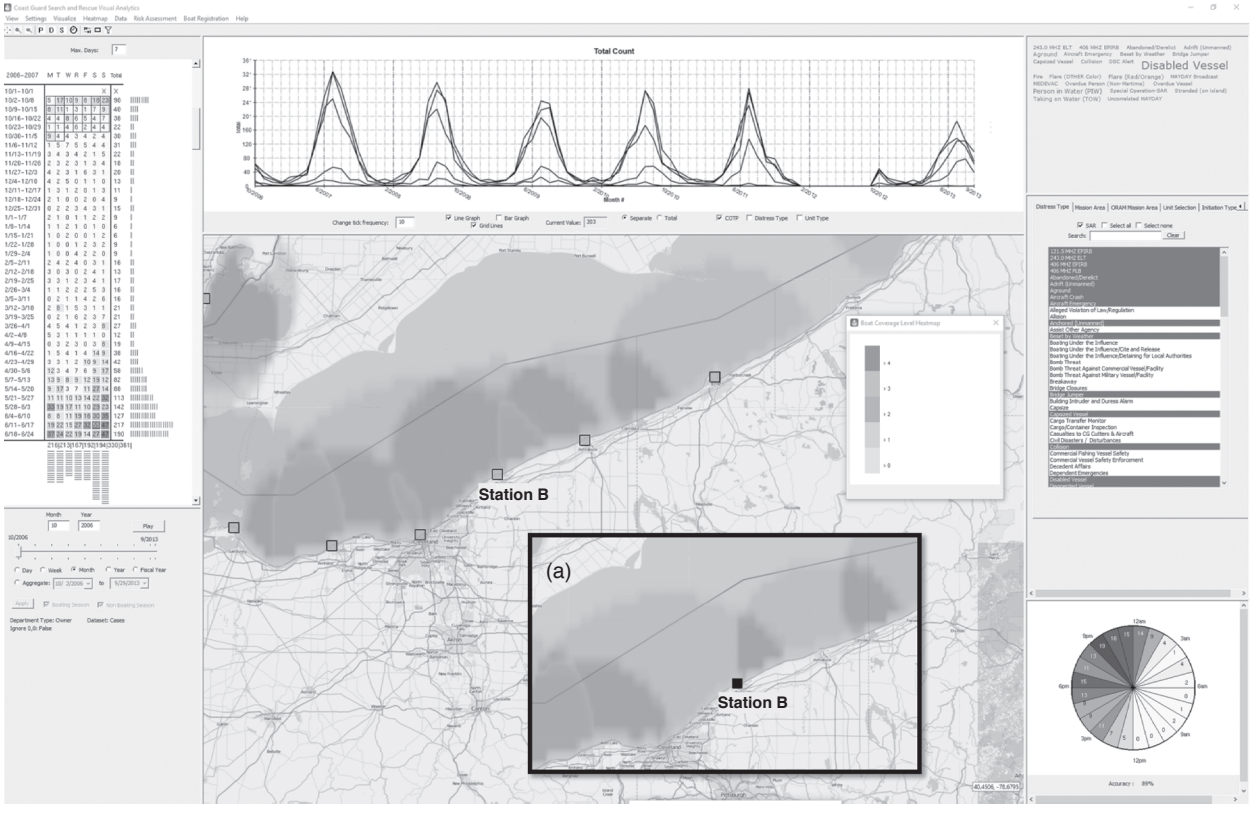
cgSARVA (Malik, Maciejewski, Maule, & Ebert, 2011) is another exemplar visual analytics system that helps users use customized simulation results in a decision-making software environment (seen in Figure 7.3). The system, which was developed for the U.S. Coast Guard Ninth District and Atlantic Area Commands, gives expert analysts a method of interactively analyzing the historical performance of their search and rescue (SAR) operations and

FIGURE 7.2. An Example of a Cyberattack Simulation Using the VASA System



The analyst selects the option to disable a power plant to simulate a cyberattack on that plant and selects a region shown using the red rectangle to indicate the area of interest (left). One main plant (purple dot) falls within this rectangle (second left). The integrated models estimate the affected subsidiary plants and workstations (red dots, second right). Finally, network-disabled regions are represented by a polygon and along with updated routes (right). This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

FIGURE 7.3. cgSARVA User Interface



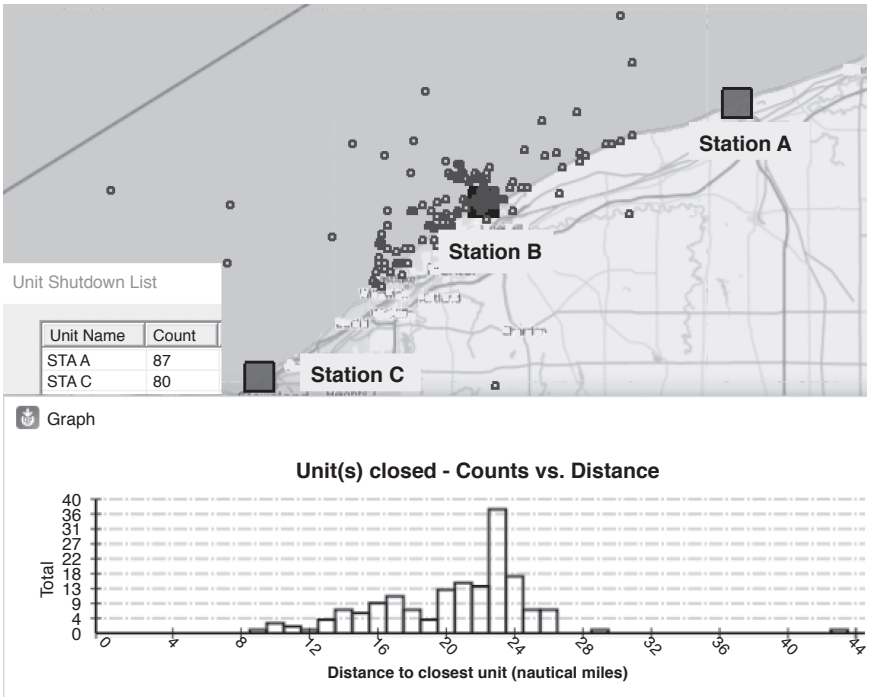
The calendar view at the top left counts incidents in the day-of-week layout to assist temporal analysis of incidents, the time slider at the bottom left allows users to interactively filter data in a variety of temporal granularities, the time series view at the top center counts incidents by the user-specified granularity (e.g., by year, month, or week), the clock view at the bottom right visualizes incident counts per hour. The view at the top right shows the keywords of incident reports. The view in the middle right allows users to filter data by multiple attributes (e.g., incident types or rescue stations). The map view at the bottom center shows the spatial analysis of incidents and rescue resources. Here, the map view visualizes the water area safety by the number of available boats. The subfigure (a) visualizes the water area safety with the assumption that Station B is closed and its boats are out of operation. The colors show boat coverage level, with green showing low and red showing higher boat coverage levels. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

assessing potential risks in the maritime environment (e.g., the spatiotemporal trends of cases or water area safety coverage of stations). For instance, the system provides risk assessment functions such as station closure analysis to identify the potential risks if one or multiple auxiliary stations are closed due to budgetary shortages or natural disasters that render a station dysfunctional. The goal of this analysis is to optimize resource allocation, whose cost depends significantly on where stations are located.

cgSARVA provides several types of risk assessments. Here, we describe two for station closure analysis, namely the distance for rescue teams to arrive at incident scenes and boat coverage in the water area. The distance between the geographical location of a case and its closest station (which provides the rescue assets) contributes to the "level of risk" for a case (because it determines the time and distance required to attend to that case). On the basis of their domain knowledge and potential policies and commands (e.g., to prepare for natural hazards or maintenance or budgetary shutdowns), the analyst interactively selects a target station for a specific temporal range in the visual interface to indicate a hypothetical closure for that station. The system automatically determines the nearest station for each case that would have been handled by the hypothetically closed station and the shortest distance between a case and the newly assigned station. After that, cgSARVA visualizes the distance assessment results in three aspects, including the distance distribution, the number of cases each station would take over, and intuitive representation of cases and related stations on the map (see Figure 7.4). The map view at the bottom center of Figure 7.3 shows the safe areas through the boat coverage in the water. The visualization shows the longest distance that boats can reach with their fuel limits from base stations. Figure 7.3 (a) shows the updated view if Station B is closed and its boats are not operational. The interactive analysis process enables an analyst to assess effectively and efficiently the potential risks caused by a particular station closure.

cgSARVA was accredited for use by the United States Coast Guard (USCG). Vice Admiral Robert C. Parker (Ret.), Commander, U.S. Coast Guard in Atlantic Area, described the system as "especially helpful in guiding operations and resource decisions by carefully analyzing data in a way that ensures the best return on investment" (Venere, 2013, para. 6). According to the Government Accountability Office, the permanent closure of stations that duplicate the services of nearby stations (without tangibly improving SAR efficiency) could result in up to $290 million in cost savings over 20 years. cgSARVA was successfully used by the USCG to right size the USCG SAR resources in the Great Lakes region, and it was used to avoid resource relocation costs following Super Storm Sandy along the eastern seaboard. The output from cgSARVA demonstrated that the number of anticipated SAR missions would be low because of colder fall and winter temperatures and the number of private boats damaged during the storm. Although the USCG's ability to respond was diminished due to the storm damage, the requirement for SAR response was also lower. cgSARVA demonstrated that a lower cost solution than shifting

FIGURE 7.4. Distance Risk Assessment in cgSARVA When a Station Is Hypothetically Closed (Malik et al., 2011)



In this case, a simulation is run assuming that Station B is shut down. As a result, all cases in Station B between 2006–2013 would be handled by the other nearest stations: 87 cases in Station A and 80 in Station C (shown in Unit Shutdown List). In the map view, blue dots are Station B's cases. The histogram shows the frequency of case distances for the Coast Guard to have rescue resources on the scene for response from the newly assigned stations. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

USCG assets from other regions was possible. cgSARVA was also used to analyze swimmer deaths and provided information for the USCG swimmer and boating safety public information campaign in 2011. Also, the cgSARVA analysis provided input to determine the number of patrols used in 2011, leading to a significant decrease in deaths in 2011.

cgSARVA was used to determine the allocation of resources during Hurricane Irene, which occurred along the east coast in the summer of 2011. The USCG initially discussed diverting resources from the Great Lakes area to the east coast, but the data from cgSARVA indicated that there was a demonstrable need to keep the Great Lakes region fully resourced at that time and to draw the resources from another region. Similarly, cgSARVA was used to analyze the effects of closing Port Arthur, Texas, in 2011, including the economic impact and the effectiveness of alternative mitigation strategies.

HUMAN–COMPUTER COLLABORATIVE DECISION-MAKING ENVIRONMENTS FOR BIG DATA

HCCDs are interactive and integrated discovery environments that balance human cognition with automated data analytics methods. Computerized analysis is designed and integrated within HCCDs to amplify human cognition (e.g., helping human users identify patterns and relationships among salient data items). The ultimate goal of HCCDs is to enable discovery or facilitate informed decision making by providing transparent, reliable, and reproducible evidence. In an HCCD, the different data flows and parameters for the simulation modules and analytical methods are configurable through the visual interface. A real-time calculation or approximation of each simulation module enables an interactive visual discourse, which allows users to use the HCCD tool even while simulations are computing. In what follows, we expound on the most important characteristics of HCCDs, namely, interactivity and integrated models.

VISUALIZATION TECHNIQUES AND INTERACTIVITY FOR BIG DATA

The analysis of big data poses unique challenges in volume, variety, and velocity that must be accommodated and considered for data and visual analytics. Here, we briefly review how the interactivity of visual analytics interfaces helps address the unique characteristics of big data.

Most notably, big data includes large volumes of data, making identification of particular data points, groups of data points, or patterns difficult. Identification involves isolating or highlighting data that is relevant to the analysis question or phenomena of interest. Visual analytics systems use interactive dynamics such as selection, view coordination, sorting, or real-time querying for identification while enabling the users to correct the system-generated identified data items (unlike traditional sampling). Identification has another purpose too—determining the appropriate scale of analysis for any phenomena. Usually, data is aggregated in various units (spatial units, such as census tracts, counties, or states, or temporal units, such as days, weeks, or months). Using interactive dynamics, automated aggregating mechanisms, and iterative refinement of views, users can identify the relevant scale of analysis according to the data and context at hand (Klein & Kozlowski, 2000).

Interactivity helps with addressing the integration of various types of data (which is another characteristic of big data) though giving the users the ability to select, switch, swap, and combine different data types on visual interfaces for gaining insight into a complex phenomenon of interest. Furthermore, interactivity helps with the analysis of data that have high velocity (e.g., streaming data), by giving the users the ability to view and sort through the incoming data (which is added incrementally to the interactive and dynamic interface) for identifying key insights. Real-time visual analytics enables users to identify important dynamic changes over time using both incoming and historical data.

These interactive dynamics are especially important for big data analytics because the goal of big data analytics is usually not to just validate known hypotheses but to unveil new patterns. In traditional data analysis (e.g., statistics), hypotheses are formed beforehand and tested for validity during analysis, and visualization is used for communicating the analysis results. In big data visual analytics, however, data visualization is used as a means of exploration and pattern identification (Kirk, 2012).

As far as visualization techniques are concerned, individual big data analytics techniques are not inherently different from small data visualization techniques. Given that familiarity and memorability are key to successful visualizations, it is common to renovate and reenvision familiar visualizations and integrate them within complex visual analytics systems with added interactivity, linked views, and bushing techniques (Robinson, 2011). For instance, the same visualizations for summarization (e.g., bar graphs, pie charts, line graphs) are used for big data but with more interactive features that can render additional elements and details on demand. Innovative data visualizations are more common with unstructured novel data sources such as text. Even though the majority of all data in digital form is in unstructured, free-form text, leveraging text in research and analysis has become common only in the social media age (Karimzadeh et al., 2013; Savelyev et al., 2014; Wallgrün et al., 2018). Various novel techniques such as Themerivers (Havre, Hetzler, Whitney, & Nowell, 2002) or overlaid tag-clouds (Bateman, Gutwin, & Nacenta, 2008; Zhang et al., 2018) are used in interactive settings to visualize textual content in time or space, respectively.

Last, interactivity is essential in incorporating user input in what-if scenarios, such as the discussed examples of cgSARVA and VASA. The interactivity of HCCD and visual analytics systems also allows stakeholder feedback for the coupled, data-driven methods for correction or adjustment. In other words, interactive visualizations are not used just for adjusting parameters and introducing what-if scenarios but also to correct the integrated (machine learning or biophysical) models.

OPPORTUNITIES AND EXAMPLES

In this section, we review the visual design, features, and application of two proven visual analytics systems for big and high-dimensional data, leveraging computational algorithms and user input for two different domains: social media analysis for situational awareness and organizational performance evaluation. These systems highlight the capabilities of visual analytics and integrated HCCD environments described throughout this chapter.

Social Media Analytics

Social media data have become increasingly popular due to the ability to provide useful information on people's attitudes, opinions, and behavior. Social media has enabled researchers and practitioners to have access to real-time

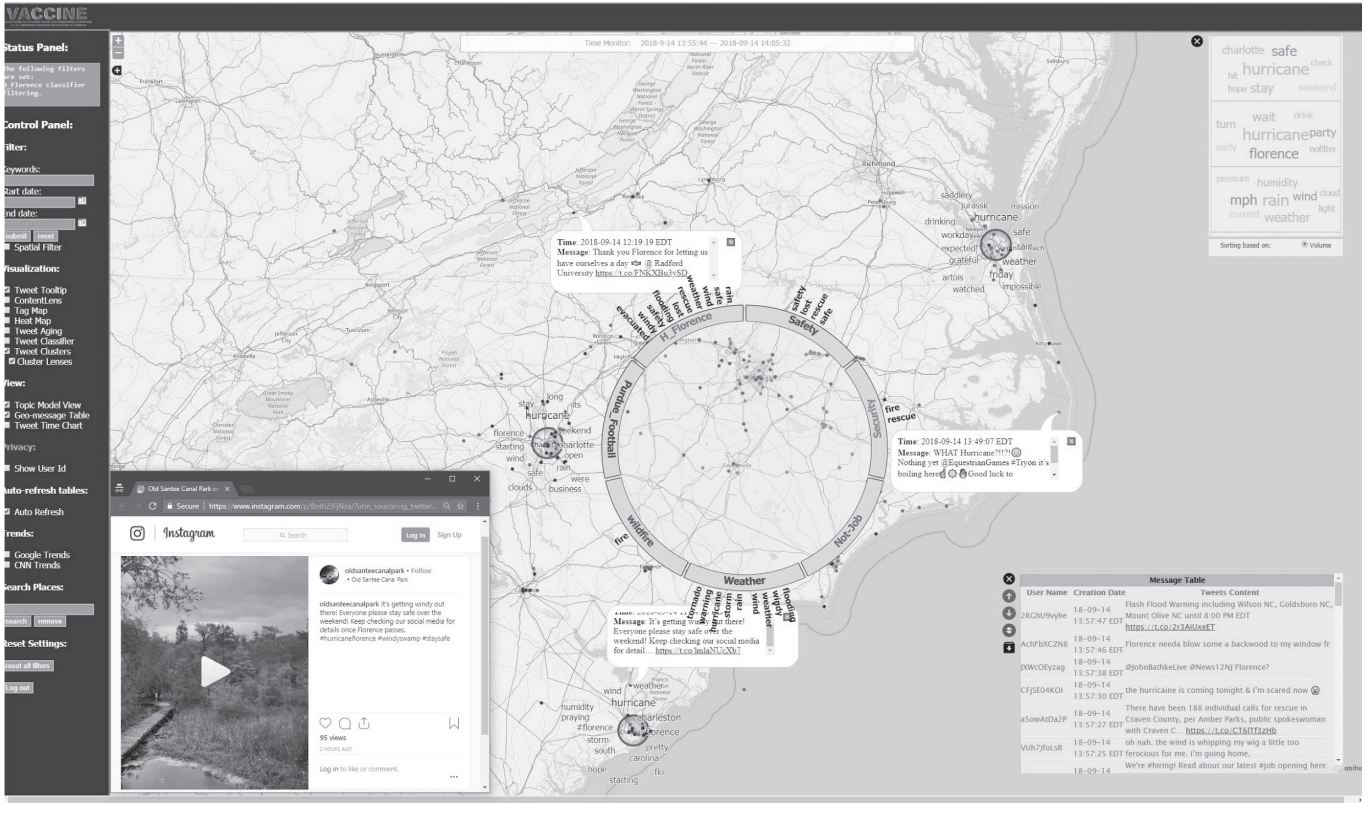
data at unprecedented rates and resolutions, essentially using humans as sensors on the ground, enhancing "situational awareness." Of specific relevance to our discussions on visual analytics of big data is the widely used platform, SMART (Social Media Analytics and Reporting Toolkit—Figure 7.5), which exemplifies our efforts in machine learning-based visual analytics to support humanitarian assistance and disaster response (Zhang, Chae, Surakitbanharn, & Ebert, 2017; Snyder, Karimzadeh, Stober, & Ebert, 2019). SMART provides users with scalable, real-time, and interactive social media data (e.g., Twitter and Instagram) visual analytics. SMART allows analysts to customize classifiers to monitor trending topics as well as unusual anomalies in the online discourse. SMART combines advanced statistical modeling, text analytics, and novel anomaly detection techniques augmented by human expertise. It provides users with the ability to search, examine, and further investigate relevant social media messages from the streaming big social media data by using natural language processing, topic modeling, advanced filtering techniques, and visual summarization techniques. The system uses several semiautomated text analysis and probabilistic event detection tools together with traditional zooming, interaction, and exploration to enable the detection and exploration of tending and abnormal topics. Web and news media sources are also incorporated into the system so that users can search for relevant news articles of interest to further corroborate intelligence acquired from social media data.

SMART leverages text classifiers to sift through large amounts of social media posts (with a low signal-to-noise ratio) for advanced yet intuitive visualizations to present users with the most relevant information on the disaster or event in question. To ensure that these classifiers do not generate false positives (posts that are not relevant to the analyst's interests) due to content that includes various meanings of certain keywords (e.g., "I feel on fire tonight; everything is going great"), we have supplemented the system with human-in-the-loop deep learning classifiers that leverage context to identify relevant (or irrelevant) content (Snyder, Lin, et al., 2019).

One of SMART's visualizations is the topic model view (see Figure 7.6). The topic model view uses latent Dirichlet allocation (Blei, Ng, & Jordan, 2003) to discover, define, and prioritize the primary topics in the social media data. Within each topic, associated keywords are displayed in a word cloud visualization, where each word's size and color jointly encode its usage frequency. By clicking on any of the words, the user can immediately view the tweets containing them on the map and in the geo-message table, allowing for the rapid discovery of social media posts on particular trending topics.

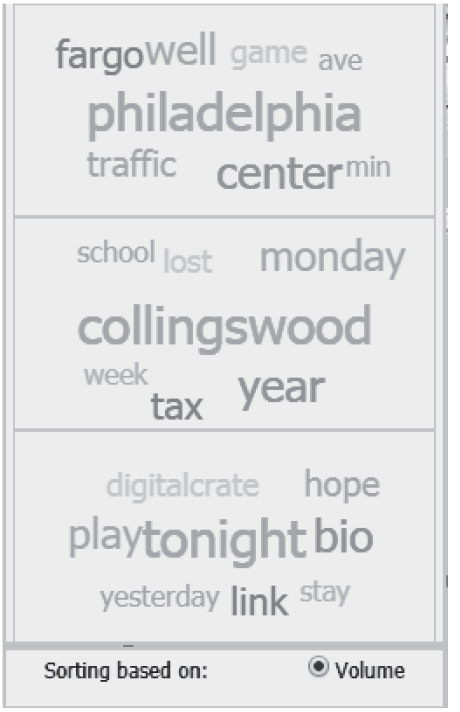
In addition, the content lens feature (see Figure 7.7) complements the topic model view by allowing the user to hover over an area on the map and view the most frequently used words among the tweets within that area. The topic model view and content lens together enable users to quickly detect and learn important content information about specific areas and topics. This shows the strong potential of coupled computational models such as latent Dirichlet allocation, natural language processing, and text and spatial data visualization for sifting through massive amounts of social media to identify important insights during various events of interest.

FIGURE 7.5. SMART's Main User Interface Summarizing Hurricane-Related Twitter and Instagram Posts for Hurricane Florence



The topic lens (center) helps analysts identify different categories of posts (e.g., weather, Hurricane Florence, safety) according to user-defined classifiers and machine learning models, content lenses (three lenses placed on the map) showing different human-generated first-hand reports in the north (e.g., "WHAT Hurricane?!!?!") and the south (e.g., "It's getting windy out here!"), and relevant information extracted through Instagram posts such as real-time videos and reports on the weather. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

FIGURE 7.6. Topic View, Showing Words in Each Detected Topic Using Latent Dirichlet Allocation



The size and color of words show their frequency in each topic. Users can click on any word to narrow down the visualization in other views to topical contents represented by those words. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

To narrow down the sea of incoming text data from social media, SMART also allows users to interactively define and apply semantic text classification filters, such as ones for "Safety," "Security," and "Weather" (see Figure 7.8), which can provide increased detection of contextually relevant data on the fly. Moreover, the cluster lens visualization (see Figure 7.9) aggregates a specific geographic area and populates the topical keywords that the social media posts used for each classifier.

Together, SMART's classification and cluster lens are incredibly powerful tools that provide users with situational awareness for effective decision making based on the situation on the ground, as sensed by human users. Particularly, emergency responders can rapidly identify user posts pertaining to disaster-related, life-threatening, or hazardous events for further investigation or resource deployment. When combined with geographic coordinates and supplementary human knowledge, responders can use information from SMART to efficiently determine the best course of action and act appropriately.

FIGURE 7.7. Content Lens of SMART, Combining Spatial Aggregation With Latent Dirichlet Allocation for Topic Modeling, Showing the Top Topics and Keywords for Any Area of Interest Represented by User-Placed Circles on the Map View

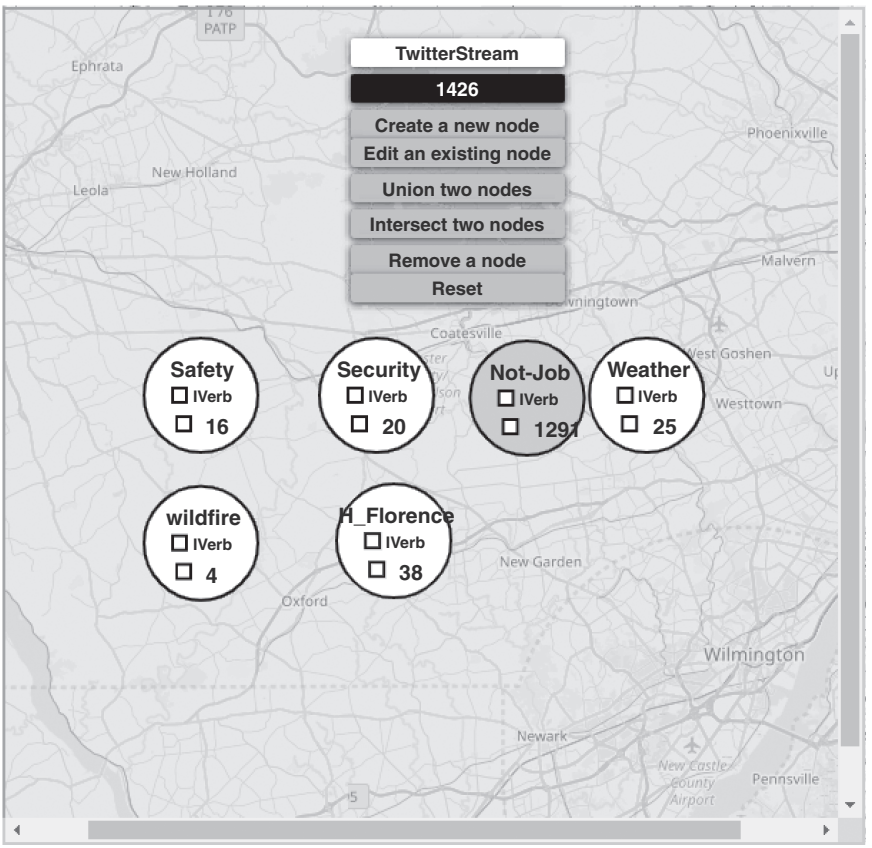
This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

SMART has been used successfully by over 50 agencies across the country, including first responders, nongovernmental organizations, and government agencies for situational awareness and hurricane response and recovery (Snyder, Karimzadeh, et al., 2019). The USCG has used SMART to maintain situational awareness for safety during several significant events (e.g., San Francisco Fleet Week, Cincinnati Riverfest 2017 and 2018, Thunder-Over-Louisville, multiple hurricanes during the 2017 and 2018 hurricane seasons, and the Republican National Convention held in Cleveland in July 2016). Overall, SMART provides a wide range of use cases, such as monitoring planned events or detecting unexpected issues, that might otherwise be difficult due to the vast amount of social media data.

Performance Evaluation for Law Enforcement Agencies

The performance evaluation of individuals, teams, and organizations requires the combination of multidimensional performance metrics that are objectively

FIGURE 7.8. Creating New Classifiers in SMART

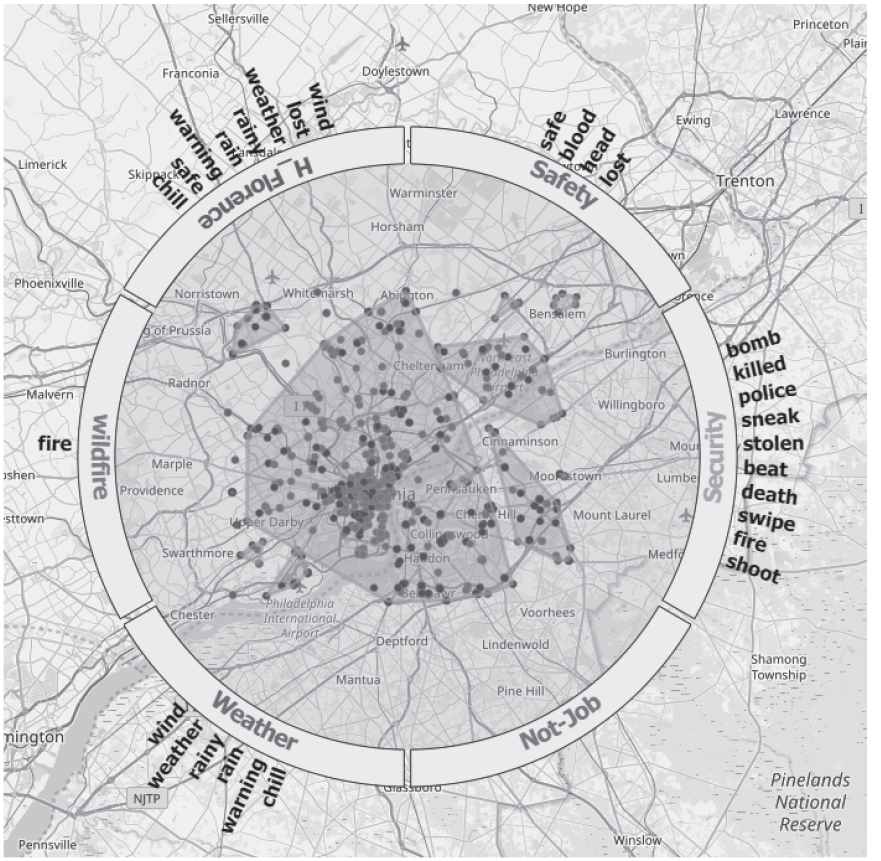


Users can associate certain keywords with different categories for further filtering down of posts related to those classifiers, creating a union or intersection of different classifiers. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

measured, flexible structure for incorporating various kinds of tasks and organizational dynamics, and supervisors' domain knowledge of the importance of various metrics. The visual analytics approach has been applied in the field of organizational performance evaluation, which is a fundamental topic in organizational psychology (Cleveland, Murphy, & Williams, 1989). In this section, we provide an example of such an approach from our previous work on a performance evaluation tool kit designed for medium to large-sized law enforcement agencies. The approach and the visualization techniques presented, however, are generalizable to other kinds of organizations.

With the assistance of computer-aided dispatch systems, law enforcement agencies can take advantage of digitized incident logs to analyze officer response. Importantly, digitized records can allow police department chiefs and supervisors to examine more effectively officer productivity for performance

FIGURE 7.9. Cluster Lens in SMART, Using Spatial Aggregation and Topic Modeling to Show Important Keywords for Each Category (Resulting From a User-Defined Text Classifier) on Any Area of Interest



The cluster lens in this example signifies the presence of important social media content for the Security, Weather, and Safety classifiers. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

improvement across departments. To facilitate this process, we designed a visual analytics application called MetricsVis that supports data-driven, multi-criteria performance evaluation of employees (Zhao et al., 2017). Specifically, the system allows supervisors to both interactively customize evaluation metrics by defining what data characteristics constitute exemplar performance and discover influential factors that can improve resource allocation, strategic planning, and operational decision making.

MetricsVis uses multi-attribute vectors, which are obtained from stored relational database records to represent employee performance. Within the context of law enforcement, different types of incidents (e.g., theft, murder, arson) each represent an attribute, and an officer's number of responses to a given incident represents the numeric value for that attribute. Overall

performance can then be computed as the sum of an officer's attributes, with each attribute weighted by the user according to its importance. However, the large number of dimensions and extensive data can cause difficulty in comparing performance between employees or for specific tasks. Thus, MetricsVis adopts an interactive reorderable matrix (see Figure 7.10b) that effectively demonstrates the details of each data item in one view. Users can dynamically adjust weights and filter attributes to understand performance from different perspectives and gain insight into improving and maintaining organizational achievements. As a result, supervisors can better understand employee performance on an individual, team, and organizational level. The police chief and commanders at the Lafayette police department have confirmed the usefulness of the matrix visualization for obtaining a holistic view of all officers' effectiveness for each incident category, as well as the overall performance of the entire department.

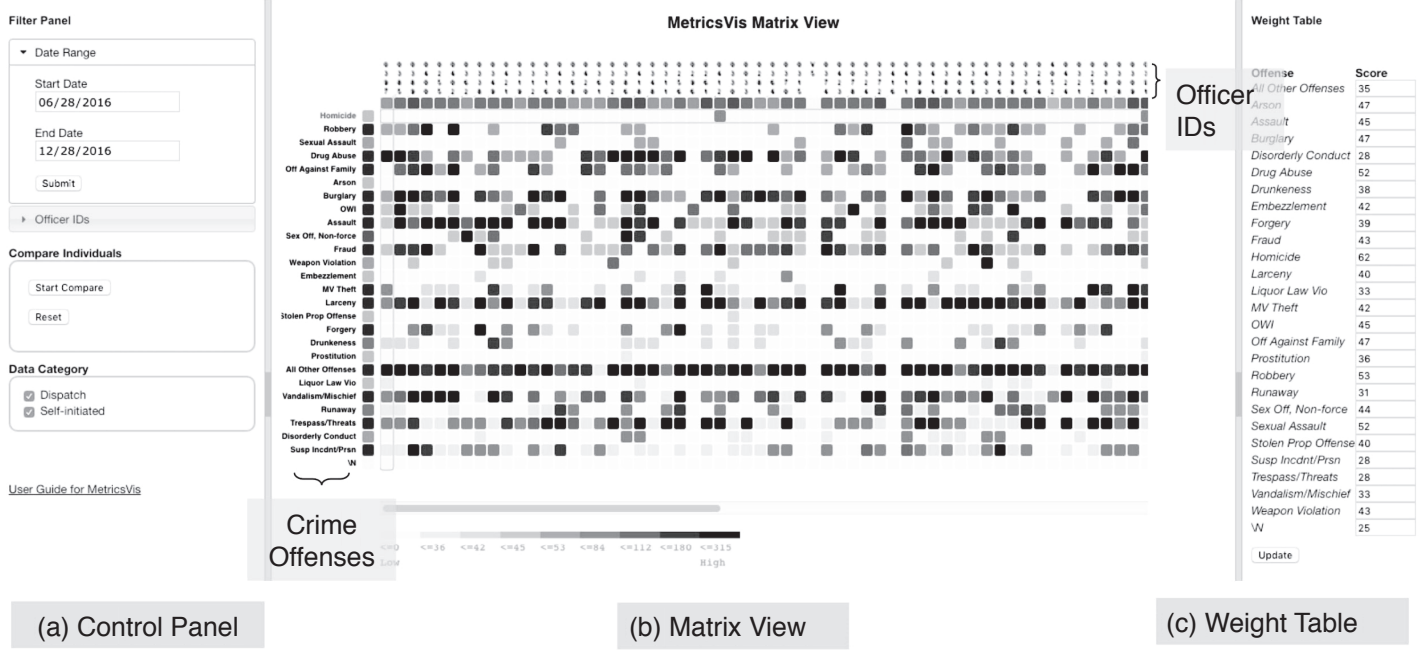
CONCLUSION

Advanced automated methods, such as deep learning models, provide great promise in gleaning insight from big data. Dynamic, changing, and context-dependent phenomena, however, require human knowledge, expertise, and reasoning for decision making. Humans' natural ability to identify patterns in disparate sources of data and to contextualize analytical results in a broader social context complement automated methods in generating useful, actionable information for real-world problems. Extending the traditional visualization paradigm that communicates the analysis results through storytelling, visual analytics enables further exploration and analysis of data and discovering unknown stories and patterns in data. For these systems to be successful, though, they have to ensure that they amplify the cognitive and analytical processes of the human while not increasing the user's cognitive load or reducing their effectiveness.

Machine learning models generate results that depend heavily on the training data or configuration parameters, potentially leading to biased results that reflect the choices made in the sampling of training data (or the real-world conditions captured in the training data as a snapshot) or the choice of configuration parameters. Integrating machine learning models within interactive systems allows researchers to elicit feedback from human users to correct erroneously generated results and provide additional training data, resulting in models that reflect real-world conditions. This interactive, visual, and explainable machine learning offers the greatest promise for the successful adoption and use of machine learning.

In addition, HCCD environments enable the integration of various computational models with interactive user interfaces for generating simulation results that facilitate testing various what-if scenarios for optimal decision making. HCCDs integrate sensed data, models (e.g., environmental, energy, or decision models), and interactive analysis, exploration, and prediction

FIGURE 7.10. The Overview of the Organizational Performance Evaluation Visual Analytics Application



(a) A control panel includes options for selecting a temporal range and filtering by incident types (e.g., dispatched vs. self-initiated). (b) A matrix visual representation shows the crime offense categories in rows and individual officers in columns. The colored blue cells demonstrate the product of [number of incidents that were responded to by an officer in one crime offense category] and the [weight assigned to the particular crime offense category by the system user]. Users can change the weight for each category interactively according to the perceived importance and policies for each organization. The red cells in the top headings show the total score of an officer, and the red cells on the left show the total score of a crime offense category. Darker colors mean higher values. (c) A weight table lists the crime offense categories and the corresponding manually assigned weights. The weights usually reflect the priorities of an organization. This image was generated in a system that uses color, though it is printed in black and white here. For a color image with more information, see http://pubs.apa.org/books/supp/woo

capabilities. Users can visualize and manipulate intermediate and final results of the different data-driven and theoretical models for key decision points (within what-if scenarios) to identify optimal solutions, such as balancing resource allocation and response time in disaster management.

As demonstrated in this chapter, visual analytics systems have been in active use and have great potential for problem solving in various domains such as social media analytics, humanitarian relief, disaster preparedness response and mitigation, and resource allocation. Furthermore, visual analytics provides pathways for researchers in various fields, including psychology, to engage in inductive or abductive approaches using the wealth of available data to generate new hypotheses (see Chapters 1 and 12).

Research in computational methods, visualization, and cognitive science can help advance visual analytics by finding solutions that leverage both machines' computational power and humans' cognitive abilities. Specifically, behavioral and cognitive studies can identify the tendencies and biases in ways humans view and use information, approach analysis, and make decisions. This line of research is essential not only to visual analytics but also to computational methods because automated methods are also marked by the choices (and potentially biases) humans introduce when designing algorithms, sampling data, and interpreting the results.

REFERENCES

Alpaydin, E. (2009). *Introduction to machine learning*. Cambridge, MA: MIT press.

Badam, S. K., Zhao, J., Sen, S., Elmqvist, N., & Ebert, D. (2016). *Timefork: Interactive prediction of time series*. Retrieved from http://users.umiacs.umd.edu/~elm/projects/timefork/timefork.pdf

Bateman, S., Gutwin, C., & Nacenta, M. (2008). *Seeing things in the clouds: The effect of visual features on tag cloud selections*. Retrieved from http://www.hci.usask.ca/publications/2008/fp039-bateman.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 933–1022.

Brooks, D. (2013, February 18). What data can't do. *The New York Times*. Retrieved from https://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. http://dx.doi.org/10.1177/2053951715622512

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130–135. http://dx.doi.org/10.1037/0021-9010.74.1.130

Das, S., Cashman, D., Chang, R., & Endert, A. (2019). BEAMES: Interactive multimodel steering, selection, and inspection for regression tasks. *IEEE Computer Graphics and Applications, 39*(5), 20–32. http://dx.doi.org/10.1109/MCG.2019.2922592

Diakopoulos, N., Naaman, M., & Kivran-Swaine, F. (2010). Diamonds in the rough: Social media visual analytics for journalistic inquiry. In A. MacEachren & S. Miksch (Eds.), *2010 IEEE Symposium on Visual Analytics Science and Technology* (pp. 115–122). Piscataway, NJ: IEEE. http://dx.doi.org/10.1109/VAST.2010.5652922

Ding, B., Huang, S., Chaudhuri, S., Chakrabarti, K., & Wang, C. (2016, June–July). Sample + seek: Approximating aggregates with distribution precision guarantee. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 679–694). New York, NY: ACM. http://dx.doi.org/10.1145/2882903.2915249

Eick, S. G., & Wills, G. J. (1993). Navigating large networks with hierarchies. In D. Bergeron & G. Nielson (Eds.), *Proceedings of the 4th Conference on Visualization '93* (pp. 204–209). Washington, DC: IEEE Computer Society.

Ellis, G., & Dix, A. (2006). An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (pp. 1–7). New York, NY: ACM. http://dx.doi.org/10.1145/1168149.1168152

Elmqvist, N., Vande Moere, A., Jetter, H., Cernea, D., Reiterer, H., & Jankun-Kelly, T. J. (2011). Fluid interaction for information visualization. *Information Visualization*, *10*, 327–340. http://dx.doi.org/10.1177/1473871611413180

Endert, A., Chang, R., North, C., & Zhou, M. (2015). Semantic interaction: Coupling cognition and computation through usable interactive analytics. *IEEE Computer Graphics and Applications*, *35*(4), 94–99. http://dx.doi.org/10.1109/MCG.2015.91

Fisher, D., Popov, I., Drucker, S., & Schraefel, M. C. (2012). Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1673–1682). New York, NY: ACM. http://dx.doi.org/10.1145/2207676.2208294

Gunning, D. (2017). *Explainable artificial intelligence (XAI)*. Retrieved from https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Hamming, R. W. (1962). *Numerical analysis for scientists and engineers*. New York, NY: Dover.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, *8*, 9–20. http://dx.doi.org/10.1109/2945.981848

Heer, J., Hellerstein, J. M., & Kandel, S. (2015). *Predictive interaction for data transformation*. Retrieved from http://cidrdb.org/cidr2015/Papers/CIDR15_Paper27.pdf

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 159–166). New York, NY: ACM.

Jaiswal, A., Pezanowski, S., Mitra, P., Zhang, X., Xu, S., Turton, I., . . . MacEachren, A. M. (2011). GeoCAM: A geovisual analytics workspace to contextualize and interpret statements about movement. *Journal of Spatial Information Science*, *2011*(3), 65–101. http://dx.doi.org/10.5311/JOSIS.2011.3.55

Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H. (2018). ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 88–97. http://dx.doi.org/10.1109/TVCG.2017.2744718

Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., . . . MacEachren, A. M. (2013). GeoTxt: A web API to leverage place references in text. In C. Jones & R. Purves (Eds.), *Proceedings of the 7th Workshop on Geographic Information Retrieval* (pp. 72–73). New York, NY: ACM. http://dx.doi.org/10.1145/2533888.2533942

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, *23*, 118–136. http://dx.doi.org/10.1111/tgis.12510

Kim, A., Blais, E., Parameswaran, A., Indyk, P., Madden, S., & Rubinfeld, R. (2015). Rapid sampling for visualizations with ordering guarantees. *Proc. VLDB Endow*, *8*, 521–532. http://dx.doi.org/10.14778/2735479.2735485

Kim, S., Maciejewski, R., Malik, A., Jang, Y., Ebert, D. S., & Isenberg, T. (2013). Bristle Maps: A multivariate abstraction technique for geovisualization. *IEEE Transactions on Visualization and Computer Graphics*, *19*, 1438–1454. http://dx.doi.org/10.1109/TVCG.2013.66

Kirk, A. (2012). *Data visualization: A successful design process*. Birmingham, England: Packt.

Klein, K. J., & Kozlowski, S. W. J. (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. San Francisco, CA: Jossey-Bass.

Ko, S., Zhao, J., Xia, J., Afzal, S., Wang, X., & Abram, G., . . . Ebert, D. S. (2014). VASA: Interactive computational steering of large asynchronous simulation pipelines for societal infrastructure. *IEEE Transactions on Visualization and Computer Graphics*, *20*, 1853–1862. http://dx.doi.org/10.1109/TVCG.2014.2346911

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May 27). Deep learning. *Nature*, *521*, 436–444. http://dx.doi.org/10.1038/nature14539

MacEachren, A. M., Stryker, M. S., Turton, I. J., & Pezanowski, S. (2010). HEALTH GeoJunction: Place-time-concept browsing of health publications. *International Journal of Health Geographics*, *9*(1), 23. http://dx.doi.org/10.1186/1476-072X-9-23

Malik, A., Maciejewski, R., Maule, B., & Ebert, D. (2011). A visual analytics process for maritime resource allocation and risk assessment. In *IEEE Conference on Visual Analytics Science and Technology* (pp. 221–230). Washington, DC: IEEE. http://dx.doi.org/10.1109/VAST.2011.6102460

Mozafari, B. (2017). Approximate query engines: Commercial challenges and research opportunities. In *Proceedings of the 2017 ACM International Conference on Management of Data* (pp. 521–524). New York, NY: ACM. http://dx.doi.org/10.1145/3035918.3056098

Park, Y., Cafarella, M., & Mozafari, B. (2016). Visualization-aware sampling for very large databases. In *2016 IEEE 32nd Conference on Data Engineering* (pp. 755–766). Washington, DC: IEEE. http://dx.doi.org/10.1109/ICDE.2016.7498287

Rajani, N. F., & Mooney, R. J. (2018). Ensembling visual explanations. In H. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. van Gerven (Eds.), *Explainable and interpretable models in computer vision and machine learning* (pp. 155–172). Cham, Switzerland: Springer. http://dx.doi.org/10.1007/978-3-319-98131-4_7

Robinson, A. C. (2011). Highlighting in geovisualization. *Cartography and Geographic Information Science*, *38*, 373–383. http://dx.doi.org/10.1559/15230406384373

Roth, R. E., Ross, K. S., & MacEachren, A. M. (2015). User-centered design for interactive maps: A case study in crime analysis. *ISPRS International Journal of Geo-Information*, *4*, 262–301. http://dx.doi.org/10.3390/ijgi4010262

Savelyev, A., MacEachren, A. M., Pezanowski, S., Karimzadeh, M., Luo, W., Nelson, J., & Robinson, A. C. (2014). *Report on new methods for representing and interacting with qualitative geographic information, Stage 2: Task Group 4: Message-focused use case.* Retrieved from https://apps.dtic.mil/dtic/tr/fulltext/u2/a616154.pdf

Snyder, L. S., Karimzadeh, M., Stober, C., & Ebert, D. S. (2019, November). *Situational awareness enhanced through social media analytics: A survey of first responders*. Paper presented at the IEEE International Symposium on Technologies for Homeland Security, Woburn, MA.

Snyder, L. S., Lin, Y. S., Karimzadeh, M., Goldwasser, D., & Ebert, D. S. (2019). Interactive learning for identifying relevant tweets to support real-time situational awareness. *IEEE Transactions on Visualization and Computer Graphics*, *1*, 1. Advance online publication. http://dx.doi.org/10.1109/TVCG.2019.2934614

Spiegelhalter, D., Pearson, M., & Short, I. (2011, September). Visualizing uncertainty about the future. *Science*, *333*(6048), 1393–1400. http://dx.doi.org/10.1126/science.1191181

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, *7*, 118–132. http://dx.doi.org/10.1057/palgrave.ivs.9500180

Tay, L., Ng, V., Malik, A., Zhang, J., Chae, J., Ebert, D., . . . Kern, M. (2017). Big data visualizations in organizational science. *Organizational Research Methods*, *21*, 660–688. http://dx.doi.org/10.1177/1094428117720014

Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *IEEE Computer Graphics and Applications*, *26*, 10–13. http://dx.doi.org/10.1109/MCG.2006.5

Uhl, J. H., Zoraghein, H., Leyk, S., Balk, D., Corbane, C., Syrris, V., & Florczyk, A. J. (2018). Exposing the urban continuum: Implications and cross-comparison from an

interdisciplinary perspective. *International Journal of Digital Earth*. http://dx.doi.org/10.1080/17538947.2018.1550120

Usability.gov. (n.d.-a). *Usability evaluation basics*. Retrieved from https://www.usability.gov/what-and-why/usability-evaluation.html

Usability.gov. (n.d.-b). *User-centered design basics*. Retrieved from https://www.usability.gov/what-and-why/user-centered-design.html

Venere, E. (2013, April 22). U.S. Coast Guard accredits analytical system developed at Purdue. *Purdue University News*. Retrieved from https://www.purdue.edu/newsroom/releases/2013/Q2/u.s.-coast-guard-accredits-analytical-system-developed-at-purdue.html

Venugopalan, S., Hendricks, L. A., Mooney, R., & Saenko, K. (2016). Improving LSTM-based video description with linguistic knowledge mined from text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1961–1966). Austin, TX: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/d16-1204

Wade, A. T., & Nicholson, R. (2010). *Improving airplane safety: Tableau and bird strikes*. Retrieved from http://de2010.cpsc.ucalgary.ca/uploads/Entries/Wade_2010_InfoVisDE_final.pdf

Wagner, M., Slijepcevic, D., Horsak, B., Rind, A., Zeppelzauer, M., & Aigner, W. (2019). KAVAGait: Knowledge-Assisted Visual Analytics for Clinical Gait analysis. *IEEE Transactions on Visualization and Computer Graphics*, *25*, 1528–1542. http://dx.doi.org/10.1109/TVCG.2017.2785271

Wall, E., Das, S., Chawla, R., Kalidindi, B., Brown, E. T., & Endert, A. (2018). Podium: Ranking data using mixed-initiative visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *24*, 288–297. http://dx.doi.org/10.1109/TVCG.2017.2745078

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, *32*, 1–29. http://dx.doi.org/10.1080/13658816.2017.1368523

Wang, G., Malik, A., Surakitbanharn, C., Florencio de Queiroz Neto, J., Afzal, S., Chen, S., . . . Ebert, D. (2017). *A client-based visual analytics framework for large spatio-temporal data under architectural constraints*. Retrieved from https://www.interactive-analysis.org/papers/2017/Wang-Spatiotemporal-2017.pdf

Wongsuphasawat, K., Smilkov, D., Wexler, J., Wilson, J., Mané, D., Fritz, D., . . . Wattenberg, M. (2018). Visualizing dataflow graphs of deep learning models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics*, *24*, 1–12. http://dx.doi.org/10.1109/TVCG.2017.2744878

Wu, E., & Nandi, A. (2015). *Towards perception-aware interactive data visualization systems*. Retrieved from https://pdfs.semanticscholar.org/0ebf/291393632cc89a786c39a1423892407993ab.pdf

Zhang, J., Chae, J., Surakitbanharn, C., & Ebert, D. S. (2017). *SMART: Social media analytics and reporting toolkit*. Retrieved from https://pdfs.semanticscholar.org/3ec1/49c0cd87e3af3269558b7fc9584e0fb3c334.pdf

Zhang, J., Surakitbanharn, C., Elmqvist, N., Maciejewski, R., Qian, Z., & Ebert, D. (2018). *TopoText: Context-preserving text data exploration across multiple spatial scales*. Retrieved from http://users.umiacs.umd.edu/~elm/projects/topotext/topotext.pdf

Zhao, J., Chevalier, F., Pietriga, E., & Balakrishnan, R. (2011). Exploratory analysis of time-series with ChronoLenses. *IEEE Transactions on Visualization and Computer Graphics*, *17*, 2422–2431. http://dx.doi.org/10.1109/TVCG.2011.195

Zhao, J., Karimzadeh, M., Masjedi, A., Wang, T., Zhang, X., Crawford, M. M., & Ebert, D. S. (2019). *FeatureExplorer: Interactive feature selection and exploration of regression models for hyperspectral images*. Retrieved from https://arxiv.org/pdf/1908.00671.pdf

Zhao, J., Malik, A., Zu, H., Wang, G., Zhang, J., Surakitbanharn, C., & Ebert, D. (2017). MetricsVis: A visual analytics framework for performance evaluation of law enforcement officers. In *2017 IEEE International Symposium on Technologies for Homeland Security* (pp. 221–227). Piscataway, NJ: IEEE.

Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. New York, NY: McGraw-Hill Osborne Media.

Zoraghein, H., Leyk, S., Ruther, M., & Buttenfield, B. P. (2016). Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems*, *58*, 19–28. http://dx.doi.org/10.1016/j.compenvurbsys.2016.03.004