Visualization Viewpoints

Editor:
Theresa-Marie Rhyne

The Top 10 Challenges in Extreme-Scale Visual Analytics

Pak Chung Wong
Pacific Northwest National Laboratory

Han-Wei Shen
Ohio State University

Christopher R. Johnson
University of Utah

Chaomei Chen
Drexel University

Robert B. Ross
Argonne National Laboratory

In this issue of *CG&A*, researchers share their R&D findings and results on applying visual analytics (VA) to extreme-scale data. Having surveyed these articles and other R&D in this field, we've identified what we consider the top challenges of extreme-scale VA. To cater to the magazine's diverse readership, our discussion evaluates challenges in all areas of the field, including algorithms, hardware, software, engineering, and social issues.

Background

The September/October 2004 issue of *CG&A* introduced the term visual analytics to the computer science literature.[1] In 2005, an international team of multidisciplinary panelists consensually and collectively defined the then newly established area as "the science of analytical reasoning facilitated by interactive visual interfaces."[2] The means and targets of VA have since evolved and expanded significantly, covering both scientific and nonscientific data of different types, shapes, sizes, domains, and applications. As extreme-scale datasets began revolutionizing our daily working life,[3] researchers looked to VA for solutions to their big-data problems.

Today's extreme-scale VA applications often combine high-performance computers for computation, high-performance database appliances and/or cloud servers for data storage and management, and desktop computers for human-computer interaction. Sources of extreme-scale data often come from models or observations, arising from different scientific, engineering, social, and cyber applications. Although many petabyte (10^{15}) or even terabyte (10^{12}) data analytics problems remain unsolved, scientists have begun analyzing exabyte (10^{18}) data.

The Top 10 Challenges

Addressing the top challenges has profound, far-reaching implications, for not only fulfilling the critical science and technical needs but also facilitating the transfer of solutions to a wider community. We thus evaluate the problems from both technical and social perspectives. The order of the following challenges doesn't reflect their relative importance but rather the content correlation among individual challenges.

1. In Situ Analysis

The traditional postmortem approach of storing data on disk and then analyzing the data later might not be feasible beyond petascale in the near future. Instead, in situ VA tries to perform as much analysis as possible while the data are still in memory. This approach can greatly reduce I/O

Top Interaction and UI Challenges in Extreme-Scale Visual Analytics

Human-computer interaction has long been a substantial barrier in many computer science development areas. Visual analytics (VA) is no exception. Significantly increasing data size in traditional VA tasks inevitably compounds the existing problems. We identified 10 major challenges regarding interaction and user interfaces in extreme-scale VA. The following is a brief revisit of our study of the topic.[1]

1. In Situ Interactive Analysis
In situ VA tries to perform as much analysis as possible while the data are still in memory. Major challenges are to effectively share cores in the hardware execution units and alleviate the overall workflow disruption due to human-computer interaction.

2. User-Driven Data Reduction
Traditional data reduction approaches via compression can become ineffective when the data are overwhelmingly large. A challenge is to develop a flexible mechanism that users can easily control according to their data collection practices and analytical needs.

3. Scalability and Multilevel Hierarchy
Multilevel hierarchy is a prevailing approach to many VA scalability issues. But as data size grows, so do the hierarchy's depth and complexity. Navigating an exceedingly deep multilevel hierarchy and searching for optimal resolution are major challenges for scalable analysis.

4. Representing Evidence and Uncertainty
Evidence synthesis and uncertainty quantification are usually united by visualization and interpreted by human beings in a VA environment. The challenge is how to clearly represent the evidence and uncertainty of extreme-scale data without causing significant bias through visualization.

5. Heterogeneous-Data Fusion
Many extreme-scale data problems are highly heterogeneous. We must pay proper attention to analyzing the interrelationships among heterogeneous data objects or entities. The challenge is to extract the right amount of semantics from extreme-scale data and interactively fuse it for VA.

6. Data Summarization and Triage for Interactive Query
Analyzing an entire dataset might not be practical or even necessary if the data size exceeds petabytes. Data summarization and triage let users request data with particular characteristics. The challenge is to make the underlying I/O components work well with the data summarization and triage results, which enable interactive query of the extreme-scale data.

7. Analytics of Temporally Evolved Features
An extreme-scale time-varying dataset is often temporally long and spectrally (or spatially, depending on the data type) narrow. The key challenge is to develop effective VA techniques that are computationally practical (for time streams) and that exploit humans' cognitive ability to track data dynamics.

8. The Human Bottleneck
Experts predict that all major high-performance computing (HPC) components—power, memory, storage, bandwidth, concurrence, and so on—will improve performance by a factor of 3 to 4,444 by 2018.[2] Human cognitive capability will certainly remain constant. One challenge is to find alternative ways to compensate for human cognitive weaknesses.

costs and maximize the ratio of data use to disk access. However, it introduces an array of design and implementation challenges, including interactive analytics, algorithms, memory, I/O, workflow, and threading.

Some of these technical challenges could be solved theoretically even today. However, the potential solution would require a radical change in the high-performance computing (HPC) community's operation, regulation, and policy and in commercial hardware vendors' systems and engineering support. We'll revisit this issue later.

2. Interaction and User Interfaces

The roles of interaction and UIs in VA are increasingly prominent in the front line of the extreme-scale-data campaign. Whereas data sizes are growing continuously and rapidly, human cognitive abilities remain unchanged. We performed an in-depth study focusing on the interaction and UI challenges in extreme-scale VA.[4] (The "Top Interaction and UI Challenges in Extreme-Scale Visual Analytics" sidebar provides an extended summary.)

The human-centric challenges of interaction and UIs are deep, multifaceted, and overlapping on several fronts. Machine-based automated systems might not address some challenges involving natural human bias in the analytical process. Other challenges that are rooted in human cognition and that push human performance to the limit might not be entirely solvable.

3. Large-Data Visualization

This challenge focuses primarily on data presentation in VA, which includes visualization tech-

9. Design and Engineering Development
Community-wide API and framework support on an HPC platform still isn't an option for system developers. The HPC community must establish standardized design and engineering resources for interaction and UI development on HPC systems.

10. The Renaissance of Conventional Wisdom
Perhaps the most important challenge is to spark a renaissance of conventional wisdom in VA as applied to extreme-scale data. (For example, Ben Shneiderman's information visualization mantra—overview, zoom and filter, details on demand—falls short when applied to many of the aforementioned challenges.) Successfully returning to the solid principles found in conventional wisdom and discovering how to apply them to extreme-scale data will most likely foster solutions to many of the problems we described.

References

1. P.C. Wong, H.-W. Shen, and C. Chen, "Top Ten Interaction Challenges in Extreme-Scale Visual Analytics," *Expanding the Frontiers of Visual Analytics and Visualization*, J. Dill et al., eds., Springer, 2012, pp. 197–207.
2. S. Ashby et al., *The Opportunities and Challenges of Exascale Computing—Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee*, US Dept. of Energy Office of Science, 2010; http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf.

niques and the visual display of information. Recent R&D in abstract visualization, highly scalable data projection, dimension reduction, high-resolution displays, and power wall displays has helped overcome aspects of this challenge.

However, more data projection and dimension reduction in visualization also mean more abstract representations. Such representations require additional insight and interpretation for those performing visual reasoning and information foraging. In addition, although we can build ever-larger and higher-resolution visual displays, the limitations of human visual acuity hinder the effectiveness of the large-screen approach in extreme-scale VA. The challenge of large-data visualization, which involves both human and machine limitations, will remain relevant in the foreseeable future.

4. Databases and Storage

The emergence of cloud services and applications has profoundly affected the extreme-scale database and storage communities. The prevailing Apache Hadoop framework supports applications working with exabytes of data stored in a public cloud. Many online vendors, including Facebook, Google, eBay, and Yahoo, have developed Hadoop-based extreme-scale-data applications.

The exabyte data challenge is real and demands attention. A cloud-based solution might not meet the needs of data analytics challenges such as those set forth by the US Department of Energy (DOE) scientific-discovery community.[5] One ongoing concern is that the cost of cloud storage per gigabyte is still significantly higher than hard drive storage in a private cluster. Another concern is that a cloud database's latency and throughput are still limited by the cloud's network bandwidth.

Finally, not all cloud systems support the requirements for ACID (atomicity, consistency, isolation, and durability) in distributed databases. Regarding Hadoop, the requirements must be addressed in the application software layer. The extreme-scale-databases challenge is thus both a hardware and a software problem.

5. Algorithms

Traditional VA algorithms often weren't designed with scalability in mind. So, many algorithms either are too computationally expensive or can't produce output with sufficient clarity that humans can easily consume it. In addition, most algorithms assume a postprocessing model in which all data are readily available in memory or on a local disk.

We must develop algorithms to address both data-size and visual-efficiency issues. We need to introduce novel visual representations and user interaction. Furthermore, user preferences must be integrated with automatic learning so that the visualization output is highly adaptable.

When visualization algorithms have an immense search space for control parameters, automatic algorithms that can organize and narrow the search space will be critical to minimize the effort of data analytics and exploration.

6. Data Movement, Data Transport, and Network Infrastructure

As computing power's cost continues to decrease, data movement will quickly become the most expensive component in the VA pipeline. To make the matter even more challenging, as data sources disperse geographically and the data become

overwhelmingly large, applications' demand to move data will increase.

Computational-science simulations have led the way in using HPC systems to tackle large-scale problems. One challenge in using HPC systems for such parallel computation is efficient use of the system network.

The computational-science community has devoted much effort to this challenge, with the message-passing-interface standard and high-quality implementations of this standard forming the basis for most large-scale simulation codes. Similarly, as VA computations execute on ever-larger systems, we must develop algorithms and software that efficiently use networking resources and provide convenient abstractions that enable VA experts to productively exploit their data.

Our list concentrates particularly on human cognition and user interaction issues raised by the VA community.

7. Uncertainty Quantification

Uncertainty quantification has been important in many science and engineering disciplines, dating back to when experimental measurements generated most data. Understanding the source of uncertainty in the data is important in decision-making and risk analysis. As data continue to grow, our ability to process an entire dataset will be severely limited. Many analytics tasks will rely on data subsampling to overcome the real-time constraint, introducing even greater uncertainty.

Uncertainty quantification and visualization will be particularly important in future data analytics tools. We must develop analytics techniques that can cope with incomplete data. Many algorithms must be redesigned to consider data as distributions.

Novel visualization techniques will provide an intuitive view of uncertainty to help users understand risks and hence select proper parameters to minimize the chance of producing misleading results. Uncertainty quantification and visualization will likely become the core of almost every VA task.

8. Parallelism

To cope with the sheer size of data, parallel processing can effectively reduce the turnaround time for visual computing and hence enable interactive data analytics. Future computer architectures will likely have significantly more cores per processor. In the meantime, the amount of memory per core will shrink, and the cost of moving data within the system will increase. Large-scale parallelism will likely be available even on desktop and laptop computers.

To fully exploit the upcoming pervasive parallelism, many VA algorithms must be completely redesigned. Not only is a larger degree of parallelism likely, but also new data models will be needed, given the per-core memory constraint. The distinction between task and data parallelism will be blurred; a hybrid model will likely prevail. Finally, many parallel algorithms might need to perform out-of-core if their data footprint overwhelms the total memory available to all computing cores and they can't be divided into a series of smaller computations.

9. Domain and Development Libraries, Frameworks, and Tools

The lack of affordable resource libraries, frameworks, and tools hinders the rapid R&D of HPC-based VA applications. These problems are common in many application areas, including UIs, databases, and visualization, which all are critical to VA system development. Even software development basics such as post-C languages or debugging tools are lacking on most, if not all, HPC platforms. Unsurprisingly, many HPC developers are still using `printf()` as a debugging tool.

The lack of resources is especially frustrating for scientific-domain users. Many popular visualization and analytics software tools for desktop computers are too costly or unavailable on HPC platforms. Developing customized software is always an option but remains costly and time-consuming. This is a community-wide challenge that can be addressed only by the community itself, which brings us to the final challenge.

10. Social, Community, and Government Engagements

Two major communities in the civilian world are investing in R&D for extreme-scale VA. The first is government, which supports the solution of scientific-discovery problems through HPC. The second is online-commerce vendors, who are trying to use HPC to tackle their increasingly difficult online data management problems.

The final challenge is for these two communities to jointly provide leadership to disseminate their extreme-scale-data technologies to society at large. For example, they could influence hardware vendors to develop system designs that meet the community's technical needs—an issue we

mentioned earlier. They must engage the academic community to foster future development talent in parallel computation, visualization, and analytics technologies. They must not just develop problem-solving technologies but also provide opportunities for society to access these technologies through, for example, the Web. These goals require the active engagement of the technical community, society, and the government.

Discussion

The previous top-10 list echoes challenges previously described in a 2007 DOE workshop report.[6] Compared to that report, our list concentrates particularly on human cognition and user interaction issues raised by the VA community. It also pays increased attention to database issues found in both public clouds and private clusters. In addition, it focuses on both scientific and nonscientific applications with data sizes reaching exabytes and beyond.

Although we face daunting challenges, we're not without hope. The arrival of multithreaded desktop computing is imminent. Significantly more powerful HPC systems that require less cooling and consume less electricity are on the horizon. Although we accept the reality that there is no Moore's law for human cognitive abilities, opportunities exist for significant progress and improvement in areas such as visualization, algorithms, and databases. Both industry and government have recognized the urgency and invested significant R&D in various extreme-scale-data areas. Universities are expanding their teaching curricula in parallel computation to educate a new generation of college graduates to embrace the world of exabyte data and beyond.

The technology world is evolving, and new challenges are emerging even as we write. But we proceed with confidence that many of these challenges can be either solved or alleviated significantly in the near future.

Acknowledgments

The article benefited from a discussion with Pat Hanrahan. We thank John Feo, Theresa-Marie Rhyne, and the anonymous reviewers for their comments. This research has been supported partly by the US Department of Energy (DOE) Office of Science Advanced Scientific Computing Research under award 59172, program manager Lucy Nowell; DOE award DOE-SC0005036, Battelle Contract 137365; DOE SciDAC grant DE-FC02-06ER25770; the DOE SciDAC Visualization and Analytics Center for Enabling Technologies; DOE SciDAC grant DE-AC02-06CH11357; US National Science Foundation grant IIS-1017635; and the Pfizer Corporation. Battelle Memorial Institute manages the Pacific Northwest National Laboratory for the DOE under contract DE-AC06-76R1-1830.

References

1. P.C. Wong and J. Thomas, "Visual Analytics," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, 2004, pp. 20–21.
2. J.J. Thomas and K.A. Cook, eds., *Illuminating the Path—the Research and Development Agenda for Visual Analytics*, IEEE CS, 2005.
3. B. Swanson, "The Coming Exaflood," *The Wall Street J.*, 20 Jan. 2007; www.discovery.org/a/3869.
4. P.C. Wong, H.-W. Shen, and C. Chen, "Top Ten Interaction Challenges in Extreme-Scale Visual Analytics," *Expanding the Frontiers of Visual Analytics and Visualization*, J. Dill et al., eds., Springer, 2012, pp. 197–207.
5. "ASCR Research: Scientific Discovery through Advanced Computing (SciDAC)," US Dept. of Energy, 15 Feb. 2012; http://science.energy.gov/ascr/research/scidac.
6. C. Johnson and R. Ross, *Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale*, US Dept. of Energy, Oct. 2007; http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Doe_visualization_report_2007.pdf.

Pak Chung Wong *is a project manager and chief scientist in the Pacific Northwest National Laboratory's Computational and Statistical Analytics Division. Contact him at pak.wong@pnnl.gov.*

Han-Wei Shen *is an associate professor in Ohio State University's Computer Science and Engineering Department. Contact him at hwshen@cse.ohio-state.edu.*

Christopher R. Johnson *is a Distinguished Professor of Computer Science and the director of the Scientific Computing and Imaging Institute at the University of Utah. Contact him at crj@sci.utah.edu.*

Chaomei Chen *is an associate professor in Drexel University's College of Information Science and Technology. Contact him at chaomei.chen@drexel.edu.*

Robert B. Ross *is a senior fellow of the Computation Institute at the University of Chicago and Argonne National Laboratory. Contact him at rross@cs.anl.gov.*

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.