FRONTIERS OF VISUALIZATION II: DATA WRANGLING

WORKSHOP SUMMARY

Prepared by the

HUMAN COMPUTER INTERACTION AND
INFORMATION MANAGEMENT TASK FORCE

BIG DATA INTERAGENCY WORKING GROUP

NETWORKING & INFORMATION TECHNOLOGY
RESEARCH & DEVELOPMENT SUBCOMMITTEE

COMMITTEE ON SCIENCE & TECHNOLOGY ENTERPRISE

of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

JULY 2018

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal research and development enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at http://www.whitehouse.gov/ostp/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at https://www.whitehouse.gov/ostp.

About the Big Data Interagency Working Group

The Big Data (BD) Interagency Working Group (IWG) coordinates Federal agency research and development (R&D) to extract knowledge and insight from large, diverse, and disparate data sources, including mechanisms for data capture, curation, management, access, analysis, and visualization. The BD IWG works under the auspices of the Networking and Information Technology Research and Development (NITRD) Subcommittee of the NSTC's Committee on Science and Technology Enterprise to coordinate current big data R&D activities across the Federal Government and to enhance collaboration among agencies, academia, and the private sector. The Human Computer Interaction and Information Management (HCI&IM) Task Force (TF) reports to the BD IWG and organizes workshops on the science of visualization. More information about the BD IWG and the HCI&IM TF is available at https://www.nitrd.gov/groups/bd and https://www.nitrd.gov/groups/hciim, respectively.

Copyright Information

Key Takeaways

The *Data Visualization Workshop II: Data Wrangling* was a web-based event held on October 18, 2017. [1] Workshop discussions highlighted the following R&D gaps:

- There is a need for a cross-disciplinary community of visualization experts to facilitate collaboration and drive innovative methodologies that exist independent of the data or domain.
- Effective visualizations of big data require a team of experts with wide-ranging skills and expertise.
- Automated data wrangling methods are required to cope with the increasing need to combine large, diverse, and heterogeneous data sets.
- To improve visualization effectiveness, visualization literacy must be taught at all levels of education and become a part of life-long learning.

Background

"Humans are visual creatures: our brain processes images 60,000 times faster than text, and 90 percent of information sent to the brain is visual. Visualization is becoming increasingly useful in the era of big data, in which we are generating so much data at such high rates that we cannot keep up with making sense of it all."[2] At its heart, data visualization is used to effectively communicate information through graphic elements that represent relevant attributes of the data under review.

This workshop report summarizes the individual perspectives of a group of visualization experts from the public, private, and academic sectors who met online to discuss how to improve the creation and use of high-quality visualizations. The specific focus of this workshop was on the complexities of "data wrangling". Data wrangling includes finding the appropriate data sources that are both accessible and usable and then shaping and combining that data to facilitate the most accurate and meaningful analysis possible. The workshop was organized as a 3-hour web event and moderated by the members of the Human Computer Interaction and Information Management Task Force of the Networking and Information Technology Research and Development Program's Big Data Interagency Working Group.

The 22 workshop participants from multiple domains were provided with perspectives on the importance of data visualizations for furnishing new understanding and insights in scientific disciplines and for facilitating communication in a healthcare context. Discussions revealed many important issues and suggestions for how to approach key challenges of data wrangling in these two areas.

Visualizations for New Understanding and Insights in Scientific Disciplines

The first session of the workshop featured a presentation on the AlloSphere Research Facility and demonstrated by means of a video, "Stunning Data Visualization in the AlloSphere",[3] how individuals use the AlloSphere tools to develop interactive multimodal representations of data. The video included medical applications in the form of 3D interactive travel through a human brain, a simulation of the

[1] For more information, see https://www.nitrd.gov/nitrdgroups/index.php?title=Frontiers_of_Visualization.

[2] Visualizing Scientific Big Data in Informative and Interactive Ways, https://www.bnl.gov/newsroom/news.php?a=212074.

[3] https://www.nitrd.gov/presentations/presentationdetail.aspx?pid=PID-10-18-2017-0001-02.

biological systems of the St. Lawrence estuary, and nanoscale displays of metals to help scientists develop new alloys that are not degraded by exposure to radiation in nuclear reactors.

The resulting discussion identified major attributes that are shared by high-quality scientific visualizations: these projects all include a team of experts with a broad range of skills in areas such as design, statistics, data analytics, and machine learning, as well as domain knowledge. Such teams share a visualization vocabulary to facilitate collaboration, and a common set of instructions to mitigate the time and effort involved in data cleaning and preparation.

Visualizations for Communication in a Healthcare Context

The second session featured a presentation, "Science in the Trenches: Why Data Wrangling and Visualization are Crucial,"[4] that described how in the mid-to-late-1990s, scientists struggled with "memory wrangling" (i.e., to have sufficient memory to do the computation) rather than data wrangling. The mid-2000s saw the rise of larger datasets and distributed computing, which led to an increasing need for collaboration. Today, 95% of the effort to create visualizations is in data wrangling, but improved automation may reduce the time involved in data cleaning and preparation.

Examples for electronic health records (EHRs) and related health data demonstrated the challenges in visualizing big data. Topics included the tensions between individualization of data to treat patients, data fusion to look at healthcare across a spectrum of topics, and the need to educate stakeholders regarding the overall state of healthcare. To illustrate the growing volume of data to be wrangled, the accumulation of health data for a population of approximately 30 million people (roughly the population of Texas) was used. It was estimated that the data would grow by almost 1.6 zettabytes[5] per year.

The discussion that followed resulted in ideas for how to automate the data wrangling process and increase visualization literacy:

- Automate the data wrangling process:
  - Improve mechanisms to share and reuse data across disciplines.
  - Build datasets that are easy to connect (e.g., enable integration of data from disparate EHR systems).
  - Find or develop data cleaning and preparation methods that scale well.
  - Understand the importance of using visualization early in the data wrangling process.
  - Understand the security and privacy risks inherent in the process (e.g., unintended disclosure of anonymized data).
- Increase visualization literacy:
  - Develop a data-driven science curriculum that includes data analytics, visualization, and related skills.
  - Train primary and secondary students to create and understand the results of visualizations.

[4] https://www.nitrd.gov/nitrdgroups/images/b/b0/Data-Wrangling-Gaither.pdf.

[5] 1 zettabyte = 10^{21} or one sextillion bytes.

Conclusion

This workshop report built upon the Frontiers of Visualization Workshop I, which identified several topics critical to the development of a science of visualization.[6] This report outlines insights and issues discussed by this engaged community of experts on one of those topics, data wrangling. The takeaways focused on how to improve data wrangling and enable effective data visualization for decision making. Potential topics for future R&D collaborations and workshops were suggested in such areas as measurement of visualization effectiveness, promotion of visualization literacy, establishment of design and presentation basics, and automation of the data wrangling process.

[6] https://www.nitrd.gov/nitrdgroups/index.php?title=Frontiers_of_Visualization#Frontiers_of_Visualization_Workshop_I.