Provenance for Sensemaking

Sense making refers to “how we structure the unknown so as to be able to act in it”. In the context of data analysis, it involves understanding the data, generating hypotheses, selecting analysis methods, creating novel solutions, and critical thinking and learning wherever needed. Due to its explorative and creative nature, sensemaking is arguably the most challenging part of any data analysis. While the field has a theoretical understanding of the sensemaking progress, such as the Pirolli and Card's model of intelligence analysis and Klein’s et al’s Data-Frame model, the research and development for applicable approaches and techniques to support sensemaking lags far behind the fast-growing user needs. As a result, sensemaking is often performed manually and the limitations of human cognition can become a bottleneck.

One of the recent advances in sensemaking research is the capture, visualization, and analysis of provenance information. Provenance is the history and context of sense making, including the data processing and analysis workflow and the users’ critical thinking process. Provenance has been studied in many fields, often under different names. For example, the Human-Computer Interaction community relies on the analysis of logs to understand user behaviours and intentions; the WWW and database community has been working on data lineage to understand uncertainty and trustworthiness; and finally, reproducible science heavily relies on provenance to improve the reliability and efficiency of scientific research. These examples illustrate the diverse types of provenance data involved (user, data, analysis workflow) and the wide range of sensemaking tasks provenance can support (e.g., understanding users, evaluating uncertainty, and authenticating results).

Provenance can support a variety of sensemaking tasks. For example, provenance can be used for recalling the analysis process by for example visualising the provenance information that includes the sequence of the investigations performed and their context information (such as parameters and motivation). Similar information can be used to assist the presentation of the analysis outcome, as the provenance consists of not only the results of each analysis stage (including the final results), but also the process that leads from data to conclusion. Other examples include providing an overview of what has been examined, revealing gaps such as unexplored data or solution possibilities, and supporting collaborative sensemaking and communication by sharing the rich context of the analysis process.

There are still many open research questions in provenance research, particularly in the context of data analysis. The three papers in this special issue address three different aspects: the practitioner perspective, a theoretical framework for provenance, and the design process of a visual provenance tool.

In “Analytic Provenance in Practice: The Role of Provenance in Real-World Visualization and Data Analysis Environments”, Madanagopal et al. present the results of interviews with provenance practitioners from domains such as intelligence analysis, cyber-security, and geospatial intelligence. The interviews examine the role of provenance for a variety of analysis scenarios central to national security and considered practical work-place implications related to organizational roles and the level of analyst experience. The results demonstrate how needs for provenance differ depending on analysis roles such as data analyst, task manager, data analyst trainer, and quality control analyst.

In “A Provenance Task Abstraction Framework”, Bors et al. introduced a conceptual framework that leverages a hierarchical provenance structure to generate effective task abstraction across multiple levels of provenance. Such structure can be significant in addressing one of the central provenance research challenges: how to derive from lower-level automatically captured information (such as data provenance and user interaction) higher-level semantically rich provenance such as user intention and the nature of

analysis task. The creation of this provenance structure consists of three stages: initialization of the provenance hierarchy, the parsing of it into a task abstraction hierarchy, and the leveraging of this task abstraction hierarchy to aid users of visual analytics systems. Furthermore, the framework is designed to accommodate iterative refinement, context, variability, and uncertainty during all three stages.

In “Capturing and Visualizing Provenance from Data Wrangling and Cleansing”, Bors et al. present DQProv Explorer, a visual analytics approach for capturing provenance data from data wrangling with annotations in the form of data quality metrics and descriptive measures. DQProv Explorer enables users to explore the provenance graph of wrangling operations and assess the impact of those operations on the overall quality of a dataset. Their in-depth discussion of the design process should provide useful insights and lessons for similar efforts.

We hope the papers in this special issue demonstrate the wide range of challenges provenance research faces and the great potential it can deliver to support sensemaking. There are still many other promising research questions not covered here, such as the application of machine learning techniques to help decipher provenance data. Provenance is not limited to data visualization or visual analytics; it will benefit greatly from research progress in the closely related fields such as HCI, databases, and reproducible science, and provide valuable contributions in turn.

Editor bio

Jean-Daniel Fekete is Senior Research Scientist at INRIA, Scientific Leader of the INRIA Project Team AVIZ that he founded in 2007. He received his PhD in HCI from Université Paris-Sud in France. His main research areas are Visual Analytics, Information Visualization, and HCI. He is a member of the Eurographics publication board, and Associate Editor in Chief of IEEE Transactions on Visualization and Computer Graphics. He was the chair of the IEEE Information Visualization Conference Steering Committee, member of the IEEE VIS Executive Committee, member of the Eurographics EuroVis Steering Committee, an ACM Distinguished Speaker, the General Chair of the IEEE VIS Conference in 2014, the first time it was held outside of the USA in Paris. Contact him at Jean-Daniel.Fekete@inria.fr.

T.J. Jankun-Kelly is an Associate Professor of Computer Science & Engineering in the Bagley College of Engineering, Mississippi State University, USA. His research lies at the intersection of visualization methods, visualization science, and visualization usage. Jankun-Kelly’s PhD is from the University of California, Davis. Contact him at tjk@acm.org.

Melanie Tory is a user research manager at Tableau Software. Her research interests include visual analysis workflows and human-data interaction. Tory has a Ph.D. in visualization from Simon Fraser University. She is an associate editor of IEEE Computer Graphics and Applications and has served as papers co-chair for the IEEE InfoVis and Interactive Surfaces and Spaces conferences. Contact her at mtory@tableau.com.

Kai Xu is an Associate Professor in Data Analytics at Middlesex University, London, UK. He received his PhD in Computer Science from the University of Queensland, Australia. His research interests include data visualization, provenance, sensemaking, and machine learning, with a focus on integrating human and machine intelligence. Contact him at k.xu@mdx.ac.uk.