From Streaming Data to Streaming Insights: The Impact of Data Velocities on Mental Models

Alex Endert
Pacific Northwest National Laboratory
Richland, WA USA
alex.endert@pnnl.gov

William A. Pike
Pacific Northwest National Laboratory
Richland, WA USA
william.pike@pnnl.gov

Kristin Cook
Pacific Northwest National Laboratory
Richland, WA USA
kris.cook@pnnl.gov

INTRODUCTION

The rise of Big Data has influenced the design and technical implementation of visual analytic tools required to handle the increased volumes, velocities, and varieties of data. This has required a set of data management and computational advancements to allow us to store and compute on such datasets. However, as the ultimate goal of visual analytic technology is to enable the discovery and creation of insights from the users, an under-explored area is understanding how these datasets impact their mental models. That is, how have the analytic processes and strategies of users changed? How have users changed their perception of how to leverage, and ask questions of, these datasets?

An emerging challenge in analytics is enabling sensemaking from streaming data. In many cases, just as the phenomena we wish to understand do not always have a defined start and end, data about these phenomena arrive in stream over time. The challenge is to characterize the phenomena from this stream, allowing models or hypotheses that explain the phenomena to evolve over time as data arrive. In addition, the streaming nature of the analysis problem motivates the use of anytime algorithms, which can provide an approximate solution at any point in their execution – they do not require a complete dataset. This approach is akin to human sensemaking, where we are able to construct explanations of the world around us as we experience it, and update these as we collect additional observations.

In dynamic analysis environments, both computational models and human mental models can be in states of constant evolution. Synchronization between computational models and mental models is therefore critical. That is, just as there is the need to capture and analyze streaming data, there is also the need to consider the streaming hypotheses that users generate during analysis, which if interpreted and addressed correctly by the system, can result in *streaming insights*. Moreover, circumstances when computational models and mental models fall out of alignment provide opportunities for key breakthroughs, so developing the capacity to monitor for this alignment becomes important. When computational models trail human understanding, means of rapidly capturing new, tacit human knowledge and updating models (e.g., classifiers) are needed. When human understanding trails computational models evolving through active learning, methods of helping human operators or decision makers understand the new state (e.g., through visualization or other computer-human communication modalities) are needed.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

RELATED WORK

There is prior work on understanding the analysis processes of users, although much of this posits a world in which human understanding grows while data remains static. In particular, work on sensemaking has illuminated some understanding for how users reason about their data in an intelligence analysis setting. For example, the sensemaking loop introduced by Pirolli and Card, is one model describing the cognitive process intelligence analysts perform as part of their daily tasks [1-2]. Each node within this diagram illustrates a mental stage through which users traverse in order to turn the raw external data source into insight. This process involves internalizing the information, combining and juxtaposing it with one's own understanding of the world, and externalizing it again to solidify the insight.

From Pirolli and Card's perspective, sensemaking can be categorized into two primary phases: foraging and synthesis. Foraging refers to the stages of the process where models filter and users gather collections of interesting or relevant information. This phase emphasizes the computational ability of models, as the datasets are typically much larger than what a user can handle. Then, using that foraged information, users advance through the synthesis stages of the process, where they construct and test hypotheses about how the foraged information may relate to the larger plot. In contrast to foraging, synthesis is more "cognitively intensive", as much of the insights stem from the user's intuition and domain expertise. We contend that streaming data impacts both phases of sensemaking, and likely involves more frequent transitions between

phases as shared understanding between machine and human models emerges.

Leveraging machine learning and active learning techniques to passively adapt to the changing hypotheses and assertions of users has been previously studied [3]–[7]. In general, the approach is to leverage the changing insights and reasoning artifacts (e.g., assertions, hypotheses, etc.) to adapt the underlying mathematical models over time. However, the impact of streaming data for such semantic interaction techniques has yet been explored. It is likely the case that in such streaming applications, the ability to steer the models implicitly can lead to better performance [8].

TOWARD STREAMING INSIGHTS

We contend that to achieve the concept of streaming insights, we must understand the impact that streaming data has on the mental models of users, so that we can gain clarity on how to support their analytic processes. How does insight evolve over time as new data become available, and are there key differences in the sensemaking process between circumstances when all of the data are available for analysis at once, versus those when data continually arrive over time? We are developing an approach in which users can pose hypotheses to streaming models (including the data they draw from) through visual analytic interfaces, and for these hypotheses to serve in some cases as forecasts for future states. We will also use interaction with visual interfaces as a means for human and machine models to co-evolve, enabling faster and more sound machine reasoning and an improved ability for humans to understand the evolving state of complex systems.

Below, we highlight some challenges and open questions for discussion.

Temporal nature of streaming data

Reasoning about streaming information likely has different challenges than static data. For example, in addition to gaining an overview of the dataset, including key topics and relationships between information, users can also pose questions regarding the evolution of topics over time. While some approaches may consider time simply another dimension in typically high-dimensional data, it is more likely that the temporal nature of streaming data represents challenges beyond traditional dimension reduction techniques. Challenges include:

1. How can users reason about information that is streaming?
2. What visual metaphors and encodings allow an overview of the content, as well as a sense for what is changing?

User interaction paradigms

User interactions are used to control model and data parameters to afford the visual exploration of datasets. Users are given the opportunity to change parameters of the models (or select data objects) within the system. However, in the context of streaming data, this may likely pose challenges, including:

3. What user interactions fit within the streaming data paradigm?
4. How can one interact with information that is changing more rapidly than the formation and articulation of hypotheses (i.e., the interaction = the hypothesis)?

The third challenge above is particularly interesting, as it raises questions regarding the speed of understanding the phenomena, reasoning about the phenomena, and interacting (or acting upon) the phenomena to perform sensemaking. Therefore, the three-faceted challenge has direct implications for the design of a streaming analytic system.

Streaming Models to Support Sensemaking

While computational models for streaming data exist, such as incremental machine learning algorithms, an underexplored aspect is how to steer these models. From a data-driven perspective, models whose structure evolves over time can detect anomalies and emerging trends in the data. However, it is less clear how users perceive emerging trends and insights over this temporal dimension. Thus, challenges in this area include:

1. How to allow the user to steer the computational models?
    a. What aspects of the data do users find interesting, or worthy of a trend? How does this differ from the methods used to computationally determine trends, themes, etc.? Active learning has been used to support model steering, but we must embed active learning techniques in visual interfaces and ensure that input provided by users (based on their mental models) aligns with expectations of the computational model, and vice versa.
2. What data do users maintain in their working memory while reasoning? How do we emphasize such information in the model?
    a. What are approaches for utilizing the limited data cache for streaming models that reflect the methods utilized by users?
    b. How can users communicate to the system what data to maintain in the cache?
    c. What data can the model ignore as it “ages off”?

3. What types of models lend themselves well to human suggestions/steering? E.g., symbolic reasoning, dimension reduction, information retrieval, deterministic, probabilistic, etc.

Who is ahead?

For streaming data, users reason about information at multiple time scales. For example, they may have hypotheses about phenomena in the data that span long and short timeframes, occur at long of short intervals, as well as singular occurrences. The myriad of combinations of phenomena and how they map to a temporal analysis pose challenges regarding the communication of what kind of (or rather, *when*) the phenomena of interest is hypothesized to occur. The challenge in this particular area comes from the common understanding between the system and the user regarding which part is trying to catch up with the other. For example, if the user has a hypothesis in mind that may pan out in the future, how would she communicate this to the system, and vice versa? Similarly, if the user is investigating an interesting set of data that happened in the past, how does the system learn from this emphasis, and understand the temporal nature to the action? Additional challenges include:

1. How to engage in the process of common ground between the system and the user to understand the temporal context of the hypotheses?
2. How to determine who is "ahead" – i.e., is the user trying to tell the system something, or have the system tell the user something regarding an emerging trend?

CONCLUSION

This position paper is intended to convey a set of challenges and topics for discussion regarding the ways in which to support the streaming reasoning of users with streaming computational models. The questions focus around open questions aimed at understanding the sensemaking process of users analyzing streaming datasets. The approach is centered around how the visual, computational, and sensemaking challenges. The discussion points above contend as users' mental models are impacted by streaming data, the design of visual analytic systems must also be reconsidered.

ACKNOWLEDGEMENTS

PNNL is managed for the US Department of Energy by Battelle under Contract DE-AC05-76RL01830.

REFERENCES

[1] P. Pirolli and S. Card, "Information foraging in information access environments," presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, 223911, 1995, pp. 51–58.

[2] P. Pirolli and S. Card, "Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Though Cognitive Task Analysis," *Proc. 2005 Int. Conf. Intell. Anal. McLean Va.*, p. 6, 2005.

[3] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, "Dis-function: Learning Distance Functions Interactively," presented at the IEEE VAST, 2012.

[4] J. Liu, E. T. Brown, and R. Chang, "Find distance function, hide model inference," presented at the Poster at IEEE Conference on Visual Analytics Science and Technology, 2011.

[5] A. Endert, P. Fiaux, and C. North, "Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering," *Vis. Comput. Graph. IEEE Trans. On*, vol. 18, pp. 2879–2888, 2012.

[6] A. Endert, L. Bradel, and C. North, "Beyond Control Panels: Direct Manipulation for Visual Analytics," *IEEE Comput. Graph. Appl.*, vol. 33, no. 4, pp. 6–13, 2013.

[7] A. Endert, C. Han, D. Maiti, L. House, S. C. Leman, and C. North, "Observation-level Interaction with Statistical Models for Visual Analytics," presented at the IEEE VAST, 2011, pp. 121–130.

[8] A. Endert and C. North, "Interaction Junk: User Interaction-Based Evaluation of Visual Analytic Systems," presented at the BELIV: Beyond Time and Errors - Novel Evaluation Methods for Visualization, 2012.