# Modeling in Big Data Environments

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

Samantha Szymczak
Resilient Cognitive Solutions
Pittsburgh, PA, USA
sszymczak@resilientcognitivesolutions.com

Dave Gunning
PARC
Palo Alto, CA, USA
dave.gunning@parc.com

John Gersh
Johns Hopkins University Applied Physics Laboratory
Laurel, MD, USA
john.gersh@jhuapl.edu

## INTRODUCTION

Human-Centered Big Data Research (HCBDR) is an area of work that focuses on the methodologies and research areas focused on understanding how humans interact with "big data". In the context of this paper, we refer to "big data" in a holistic sense, including most (if not all) the dimensions defining the term, such as complexity, variety, velocity, veracity, etc. Simply put, big data requires us as researchers of to question and reconsider existing approaches, with the opportunity to illuminate new kinds of insights that were traditionally out of reach to humans.

The purpose of this article is to summarize the discussions and ideas about the role of models in HCBDR at a recent workshop. Models, within the context of this paper, include both computational and conceptual mental models. As such, the discussions summarized in this article seek to understand the connection between these two categories of models. In this article, we enumerate challenges and opportunities categorized under:

- Types of models involved in the holistic process of using a system to use big data to gain insight about the world.
- Bi-directionality of visual representations leveraged for model steering and visual data exploration
- Impact on analytic culture and policy

## INCORPORATING DIFFERENT TYPES OF MODELS

In our discussion of models at the workshop, we discussed various *kinds* of models,, different instances of those models, and developed some insights about the relationships among these models to achieve better analyst::technology coupling (or a Joint Cognitive System).


*HCBDR '14,* April 01 - 03 2014, Raleigh, NC, USA
ACM 978-1-4503-2938-5/14/04.
http://dx.doi.org/10.1145/2609876.2609890

In this context, "computational model" refers to a computer model of human cognition. Such models are used both to help understand actual human cognitive processes and to provide artificial cognition within systems that aid or replace human perception, analysis, or sensemaking. The second use is of primary interest here.

"Mental model" refers to the conceptual structure that a person uses to describe and reason about aspects of the world. Knowledge about such models can help explain and predict human behavior and aid in the design of the interaction between people and technology [1].

Ultimately, as we build systems to aid in people's work, the intent is for users' mental models of the world to accurately reflect the reality of the world. However, analysts perceive the world through the technology we provide them as system designers and engineers. The design of human interaction with these systems is a significant factor in shaping users' models of the world and of the systems themselves. There are steps we can take as system designers to make mental models better reflect how the world works and the nature of their system's interaction with it.

One hypothesis is that interaction between analysts and computational models is most effective when such dialogue is based on a conceptual structure that is based on analysts' mental models of the domain of analysis. The theory here is that if the computation model "talks like the analyst," then the analyst and the system will work better together as a team. Some challenges with this approach are creating a computational model that can evolve as the user's understanding of the domain evolves.

In addition to analysts having mental models of the world, they also have a mental model of the data and systems that provide them with information about the world and help them make sense of it. Of particular interest in HCBDR is analysts' mental models of the automation necessary to acquire, process, and analyze big data. The analyst has certain beliefs, expectations, and assumptions about the automation's capabilities, functions, and actions; these are reflected in analysts' mental models of the systems they use. (labeled Mental Models of the Computation in Figure 1). It is important that the analyst's mental model of the

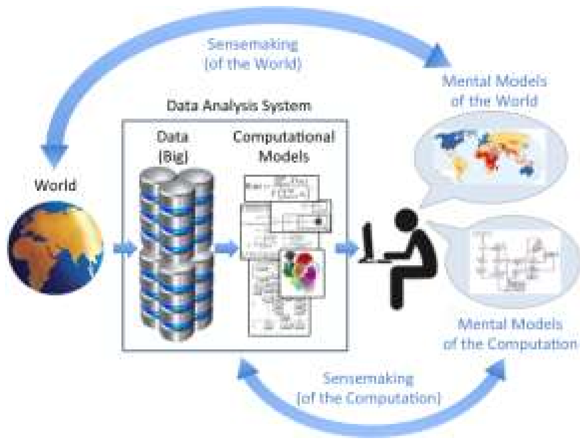


**Figure 1. Users have to maintain an understanding of many types of models when working with systems that help them gain insight into big data. The balancing and synchronization of these models is a challenge, as well as the communication and translation between these forms of models.**

automation lines up with the actual functions & limitations of the automation for the analyst and the automation to work together. Breakdowns in that teamwork can come from users' assuming the automation is doing one thing, while it is doing another (many aviation accidents that are attributed to "human error" are actually human-automation misunderstandings). System designers can close the analyst:automation gap by building system features to aid analysts making sense of the automation, help them trust it, and enhance human-sysem teamwork. (Cognitive systems engineering principles base this on designing systems that are observable and directable).)

There are many research questions that came out of the discussion at the workshop:

- What methods are available for capturing / documenting analyst's mental models (both domain and automation)?
- What are the pros and cons of the various methods for capturing / documenting analyst mental models?
- How can the researcher / system designer detect when the analyst's mental model and the computational model no longer align?
- What methods are effective for aiding sensemaking about the automation: getting the users' mental models and the automation into alignment in a way that does not expect the analyst to be an expert in computation over big data, and provides observability and directability using analytic concepts. Ultimately, in a world of big data, the layer between the analyst and the raw data is thicker than ever. Researchers and system designers must consider the end user analyst when designing both automation and user interfaces, and align the design of both the automation and the UI to better match the analyst's understanding of the domain. Additionally, system designers must consider that the automation may at time mis-align with the analyst's intent and needs, and the system designer must build features in the system for the analyst to first detect the mis-alignment, and also be able to correct the mis-alignment.

## BI-DIRECTIONAL VISUAL REPRESENTATIONS

Through visual data exploration, analysts engage in a dialog with their data via a specific system of visualization. This analytic discourse between the machine and the human is fundamental to mixed-initiative systems, and more generally systems intended to leverage both computation and human capabilities. For example, a system's communication with the user can be via visualization.

Information visualization provides users with powerful and effective visual representations of their data [2]. The representation provides the system a method of visually communicating characteristics, patterns, and other features about data to the user. Further, in visual analytic applications for big data, it is often the case that one or more models is used to calculate or transform data prior to visualizing the results. For example, dimension reduction models can be used to reduce similarity of high-dimensional data to two-dimensional spatializations [3]. As a result, the model's calculated similarity between the information is shown as relative distance between the points. Graphical user interfaces are typically used for the communication between the user and the system. This leads to potential fundamental usability concerns when the complexity of the models utilized by the system (and thus requiring user control and training) are outside the expertise of the user [4].

Instead, we contend that the visual representation of the information can act as a bi-directional medium for the analytic discourse. Prior work in this area has shown how incremental machine learning methods have been applied to spatializations of high-dimensional data, where the user interaction is directly within the visualization (not on graphical controls for the model parameters, such as sliders) [5], [6].

Challenges in this creating this bi-directional communication for steering and training a system include:

- How can users understand how their training has impacted a system's computational methods?
- How can systems listen to the user interactions of users and learn, evolve, and adapt?
- What visual representations afford such bi-directionality?

- How to foster trust through the communication? This includes the system trusting a user's guidance, and a user trusting a system's recommendations and visual representations of the data.

### IMPACT ON ANALYSIS CULTURE AND POLICY

The analytic processes and culture surrounding the analysis of information has been documented by prior work [7], [8]. These works found that a significant amount of formality and structure exists with regards to showing evidence for assertions and conclusions, as well as other structured analytic techniques and methods to ensure proper analysis.

Most relevant to big data (and the technology developed to aid users with gaining insights) is how to show evidence of findings. Traditionally, evidence may come in the form of a set of documents, network packets, or other data exemplars that help frame the knowledge that was uncovered during the investigation [7]. However, in analyzing data of a larger scale, the sheer amount of evidence that may be needed to support a finding may become overwhelming. For example, many statistical techniques can be used to determine trends in data, where most (and at times, all) of the data is used for the calculation. For such methodologies, what can be used as evidence when reporting a finding to another person? Further, it becomes decreasingly reasonable to store and maintain *all* of the data in many of these scenarios. These, as well as other challenges with big data, suggest a change in the analytic culture and policy in order to leverage the advantages big data can provide to understanding a given phenomena.

### ACKNOWLEDGEMENTS

We would like to thank the Laboratory for Analytic Sciences for hosting this workshop and all of the stakeholders representing the intelligence community who brought insight and real-world applications to this discussion.

### REFERENCES

[1] Norman, Donald, "Some Observations on Mental Models," in *Mental Models*, Gentner, Dedre and Stevens, Albert, L., Eds. Lawrence Erlbaum Associates, 1983, pp. 7–14.

[2] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999.

[3] A. Skupin, "A Cartographic Approach to Visualizing Conference Abstracts," *IEEE Comput. Graph. Appl.*, vol. 22, pp. 50–58, 2002.

[4] 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.

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

[6] 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.

[7] D. T. Moore, "Critical Thinking and Intelligence Analysis," Mar. 2007.

[8] 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.