Ganging Up on Big Data
Computer-Intermediated Collaborative Analysis

Mark Stefik and Hoda Eldardiry
Intelligent Systems Laboratory
Palo Alto Research Center (PARC)
Palo Alto, California, USA
{mark.stefik,hoda.eldardiry}@parc.com

***Abstract*—Understanding complex situations is difficult. Intelligence analysis has long been the work of teams including subject matter specialists. Today collaborative analysis takes place in the context of "big data", where information comes from a variety of human, communications, and sensor sources. Understanding the big picture is both about how analysts interact and combine their insights together and with how they engage with data at scale. In this paper we consider opportunities for next generation analysis systems for teams, focusing on the computer-intermediated functions that support and coordinate analytic activities around big data.**

***Keywords*—Planning; generalization; lessons learned knowledge management; collaborative analytics; anomaly detection**

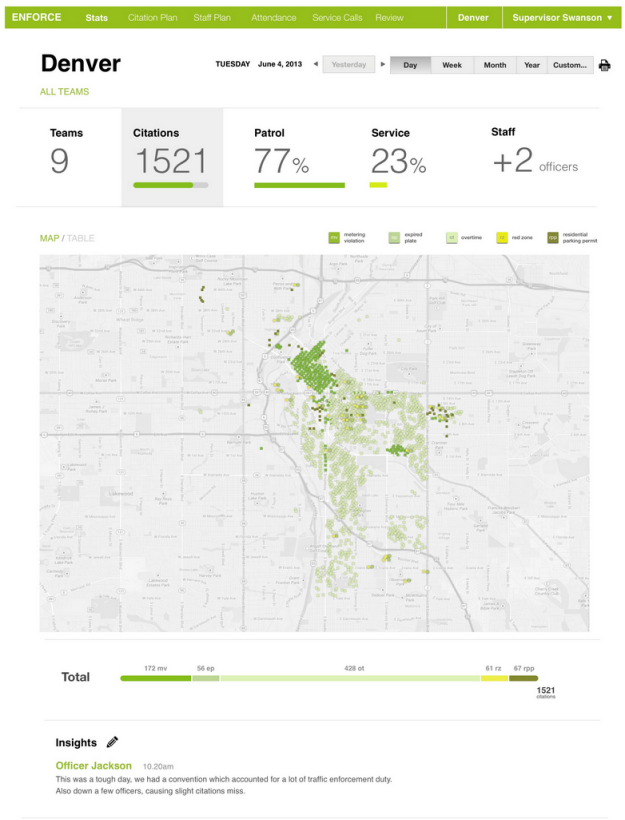

Figure 1. Activity Analytic Map Citations Teams

I. INTRODUCTION

Figure 1 shows a descriptive visual analytic from one of our projects. This dashboard highlights activities of a traffic and parking enforcement organization; and provides insights to the organization on its own activities and their interactions with an urban smart city environment. Beyond such descriptive analytics, the opportunity and challenge for impact requires analyzing the work inside individual and team activities.

We describe three central functions for sustainable impact in many next generation collaborative analytics settings:

- Connecting activities. People engage in different activities, bringing different knowledge and expertise. How can a system leverage big data and this diversity to coordinate and amplify their performance?
- Automating Tasks. People working with analytics and big data always know salient things not yet represented in the system. How do we engage them together with automatic processing to improve performance by guiding foraging, monitoring and interpretation of big data?
- Generalizing learning. Situations evolve, leaving traces in collected big data. How can lessons from the past be updated to keep up with the emerging future?

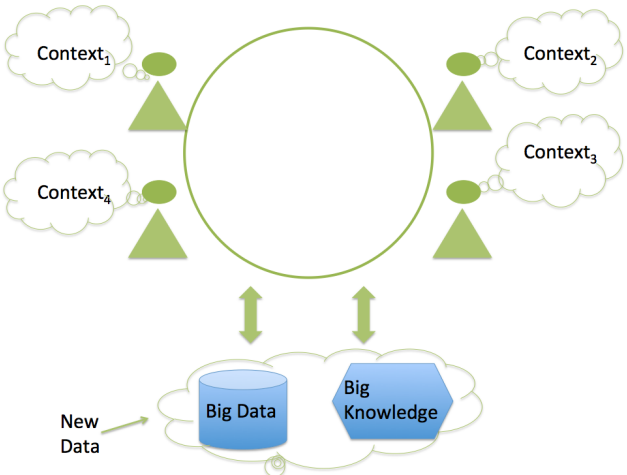

Figure 2. Framework Template

These three functions are exemplars of elements of activities on collaborative analytic tasks. Other examples include guiding information foraging and identifying trends. Our approach is to deconstruct tasks into elements that can be accelerated, improved or coordinated with computers. Figure 2 shows our template for illustrating such functions. The figure emphasizes members of a team together with big data and big knowledge. The circle is for showing the computation support for each function.

II. CONNECTING ACTIVITIES

Different people focus on different tasks and bring different knowledge and expertise. How can a system leverage big data and this diversity to coordinate and amplify their performance?

Well-known crowdsourcing examples that answer this question include crowdsourced interpretations of craters in satellite images [1], crowdsourced language translation [2], and crowdsourced building of ancestry trees [3]. These examples have in common the idea that a big data system links together a large solution from small contributions by different people.

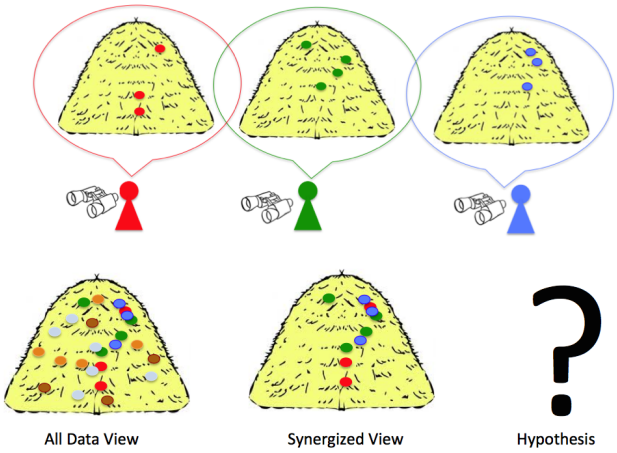

Figure 3. Contextual lenses enabling three analysts looking at shipments of farm materials, the organizational structure of a terrorist organization, and local movements of particular actors near places where large crowds will gather for a political event. The viewed dots can be seen as consistent with a hypothesis pattern (?) where explosives will be deployed at an important political gathering.

We offer an example of a team of intelligence analysts working on different tasks in the same part of the world. Figure 3 shows three intelligence analysts (red, blue, and green) working on their tasks. They each use a visual analytic tool to give them spatial, temporal, and relational presentations of data they have selected in their area of interest. In the example, the analysts look separately at shipments of farm materials (red), the organization of a local terrorist group (green), and local movements of particular actors near places where large crowds will gather for a political event (blue). The system uncovers a combined big-picture hypothesis (shown as "?") that reveals a terrorist plot involving procurement of dual-use materials by separate members of a terrorist organization in order to make explosives that will disrupt a political event.

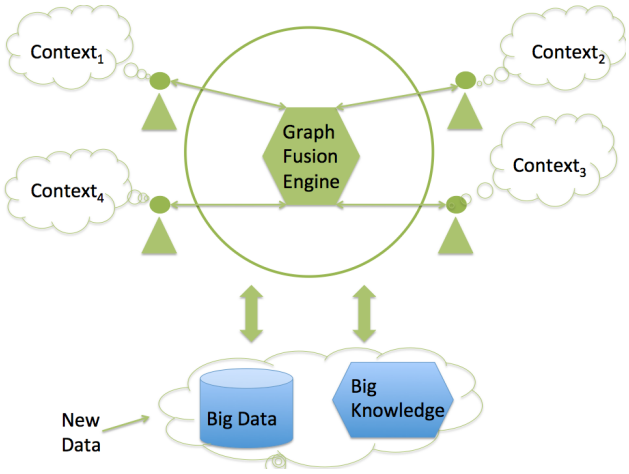

Figure 4. High performance graph fusion and reasoning system.

In an early computer-mediated collaboration system [4] we explored how a multi-user interface could foster coordination and efficient team brainstorming. The computer system provided participants with contextual awareness of each others ideas in a multi-user workspace. Similarly in our haystack analyst example, as each analyst adds interpretations or conclusions to the data, the system computes and distributes relevant consequences to other analysts.

This computation is carried out largely by HG (HiperGraph), a high-performance graph fusion and reasoning system at PARC (see figure 4). HG stores large amounts of heterogeneous entity and relationship types in a graph representation. Its native graph representation enables answering complex graph queries very efficiently. HG draws on a library of patterns to check and propagate inferences enabled by the special knowledge of one user to others.

III. AUTOMATING TASKS

People working with analytics and big data typically know salient facts not yet represented in the system. How can we engage this knowledge to improve performance by guiding foraging, monitoring and interpretation of big data?

Consider this question for a team of analysts and experts that are monitoring an evolving world situation. The analysts look for events that need attention. Figure 5 illustrates the learning needed in these activities. The light shaded upper half of the figure is about discovering new patterns in the situation. The darker lower half is about detecting known patterns. The matcher of known patterns is the workhorse of the system, automatically and tirelessly triggering alerts when patterns appear in the data. Initially a human domain expert provides the known patterns.

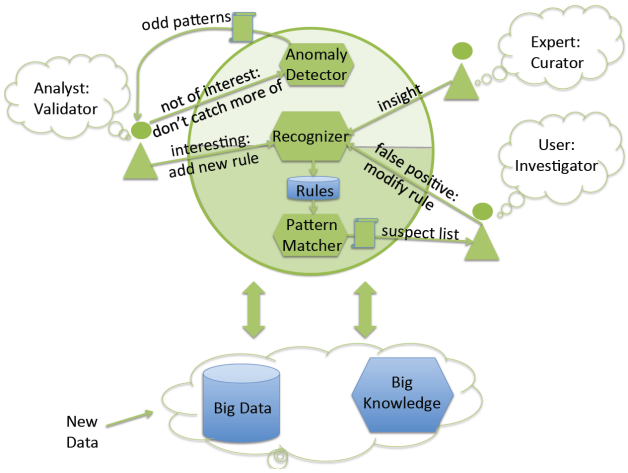

Figure 5. Automating Tasks for interactive foraging, monitoring and learning.

Learning takes place through feedback and training from an investigator, who tunes the rules by flagging any false positives that occur in the alerts. Along these lines we have previously built a system for delivering personalized news [5], which grew its strength by combining the light work of readers, the harder work of curators, and the "tireless" work of a set of computers. From a perspective of collaborative analysts, this approach shows how that knowledge of subject matter experts can be captured and repeatedly applied by a system that recognizes and supports this function.

Going beyond known patterns, it is important to identify potential unknown patterns that may appear. In the upper half of figure 5, an anomaly detector checks new data and generates a set of odd patterns that may be of interest. An analyst/validator checks the odd patterns set and provides two types of feedback. If the odd pattern is not of interest, the analyst alerts the anomaly detector to modify its algorithm so that it does not catch similar instances in the future. If the odd pattern is of interest, and the analyst would like the system to catch similar instances in the future, the analyst alerts the recognizer to learn and construct a new rule that encodes a generalization of the discovered instance.

Overall this system is an example of collaborative learning for an organization. It combines human judgment in training with automation. Perhaps the most novel aspect of this system is the anomaly detector for identifying potential new patterns of interest. In our projects at PARC, we have developed an anomaly detector that uses unsupervised discovery techniques to find statistically rare patterns and supervised learning techniques to detect unpredictable events.

Our previous work on anomaly detection includes insider threat detection [6] and fraud detection [7]. In this work, we presented new definitions for anomalies that go beyond the straightforward 'outlier', 'rare', 'temporal' and 'structural' anomalies. These include 'blend-in' and 'rare-change' anomalies. We developed novel suspicion indicators and fusion methods for processing multiple sources of information. We also presented ranking and visualization schemes that provide explanations to aid and direct analysis efforts.

Figure 6. Yom Kippur Sensitivity Analysis.

IV. GENERALIZING LEARNING

The world situation is always evolving, leaving traces in collected big data. How can lessons from the past be updated to keep up with the emerging future?

We consider this question in the context of an intelligence team wanting to avoid strategic surprise. In 1973 Israeli intelligence failed to provide warning before the surprise attack starting the Yom-Kippur War [8], [9].

Figure 6 summarizes the main elements of the original Israeli analysis. Red (or dark) coloring in the figure shows wrong hypotheses and yellow shows misleading ones. In the Yom Kippur case, observations of the military "faux practice" exercises might have revealed technology changes that gave the Egyptians surprising advantages in the early stages of the war, specifically the use of RPG-7 rockets, RPG-43 grenades, and a novel use of high-pressure water canons to breach sand walls to undermine Israeli defenses using water from the Suez canal.

Another issue was that the Israeli analysis of the political situation assumed that Egypt's goal would be complete conquest, and that they would not attack if the prospects for victory were very slim. Israel did not credit the possibility that Anwar Sadat would be satisfied with a very narrow victory or even a political boost from standing up to Israel. What lessons should be drawn from the case? In a too specific lesson characterization, the new situation must involve Israel, Egypt, and Syria where there is a build up of Egyptian forces to invade Israel. In a too general characterization, a lesson should be considered any time an enemy threatens an attack. The right generalization falls somewhere between these extremes.

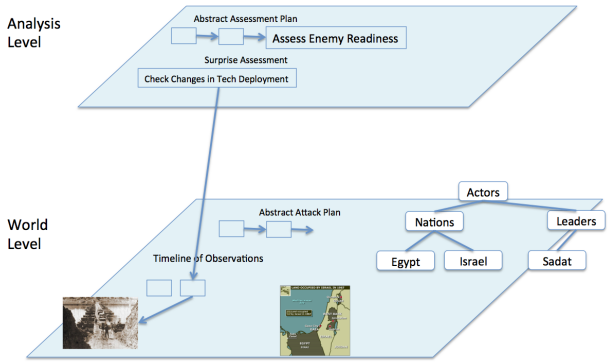

Figure 7. Planning

Figure 7 shows elements of a lesson generalizer. It is based on abstract plan representations and the logic of plan monitoring [10], [11]. These representations capture the narrative structure and causality of stories and formal representations of abstraction. At the world level, the model includes concept hierarchies, plans and abstract plans ("attack plans") with links to temporal and spatial data about unfolding events. The analysis level holds plans for analysis ("assessing enemy readiness"). The figure shows a link from an analytic step for assessing readiness to a world step to check an observation.

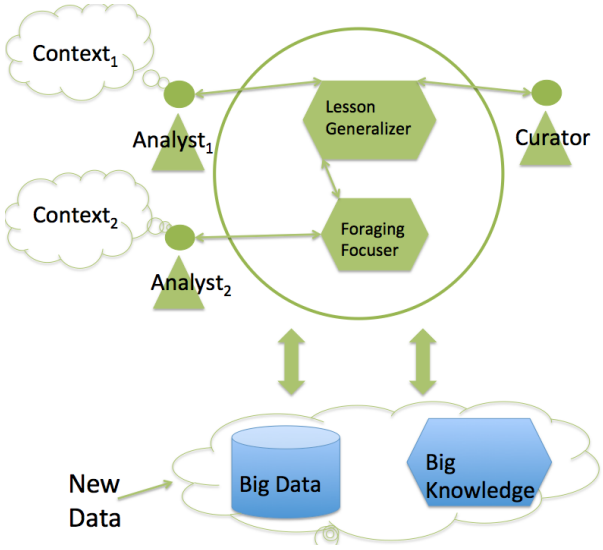

Figure 8. Lessons Manager

Figure 8 shows how a lesson generalizer could work into a system for managing and re-using lessons. A foraging focuser guides foraging and interpretation by a later user when a generalized old lesson seems to fit a current situation. A curator expert checks the validity of automatically proposed generalizations. In this way, abstract lessons are created, curated and applied, while involving different users. Overall, the "learning" is a combination of expert judgment and machine learning technology.

V. Conclusion

Looking ahead to next generation systems for intelligence analysts, we propose a perspective for focusing on computer-intermediated collaborative functions around the use of big data. We gave three examples of functions typically needed in sustainable big data settings and show how to combine complementary human and computational capabilities. In a broad sense, we are interested in augmenting the world knowledge and expertise of people with machine learning and coordination functions of computers on big data.

In our examples, the people not only do their work (e.g., analysis), but they exercise their particular strengths as observers, trainers, experts, and curators. At the same time, the particular strengths of the system are in the tireless carrying out of routine tasks over big data (such as the alert and anomaly detection functions). Overall, the goal is to enable an enhanced level of understanding that transcends in exciting ways what people or computers could do on their own.

Acknowledgment

The authors thank Danny Bobrow, Johan deKleer, Dave Gunning, Peter Pirolli and Bob Price for their insightful comments.

References

[1] J.S.S. van t Woud, Is crowdsourcing in the form of a serious game applicable for annotation in a semantically-rich research domain?, Thesis. Human-Centered Multimedia. Universiteit van Amsterdam. 2010.
[2] O.F. Zaidan, C. Callison-Burch. Crowdsourcing Translation: Professional Quality from Non-Professionals. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Pages 1220-1229, 2011.
[3] J.S. Jacobs. Are you my cousin? The New York Times Sunday Review, page SR1, February 2, 2014.
[4] M. Stefik, G. Foster, D. Bobrow, K. Kahn, S. Lanning and L. Suchman. Beyond the chalkboard: computer support for collaboration and problem solving in meetings. Communications of the ACM. Vol. 30, Issue 1, 1987.
[5] M. Stefik and L. Good. Design and deployment of a personalized news service. AI Magazine. Vol. 33, No. 2. 2012.
[6] H. Eldardiry, E. Bart, J. Liu, J. Hanley, B. Price and O. Brdiczka. Multi-domain information fusion for insider threat detection. IEEE Workshop on Research for Insider Threat. 2013.
[7] H. Eldardiry, J. Liu, Y. Zhang and M. Fromherz. Fraud detection for healthcare. Knowledge Discovery and Data Mining Workshop on Data Mining for Healthcare. 2013.
[8] M. Chorev. Surprise attack: the case of the Yom-Kippur war. Washington D.C. National Defense University. 1996.
[9] A. Shlaim. Failures in national intelligence estimates: the case of the Yom Kippur war. World Politics, Vol. 28. 1976.
[10] C. Fritz and S. A. McIlwraith. Generating optimal plans in highly dynamic domains. In Proceedings of The 25th Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, Canada, June 18-21, 2009.
[11] M. Stefik. Planning and meta-planning. (MOLGEN: Part 2). Artificial Intelligence. Vol. 14, No. 2, 141-169, 1980.