Knowledge Exploration, Analysis and Discovery (KNEAD) Challenge Workshop

Mark Maybury and Penny Chase
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
{maybury, pc}@mitre.org

Keywords: massive data, knowledge exploration, analysis, multiperspective collaboration, tailorable environments, contextual processing, cultural processing

Abstract

This paper summarizes results from the government sponsored Knowledge Exploration, Analysis and Discovery (KNEAD) Challenge Workshop. The problem focus was to create tools and methods to enable groups of interdisciplinary computer forensic analysts to organize, access, and "mine" maximally relevant information from large volumes of continuously evolving multimedia, multilingual, and multicultural data. This document summarizes the activities, findings, and recommendations from the workshop.

1. The Challenge Workshop

The objective of the Knowledge Exploration, Analysis and Discovery (KNEAD) Challenge Workshop was to identify tools and methods to enable groups of interdisciplinary analysts to organize, access, and "mine" maximally relevant information from large volumes of continuously changing multimedia, multilingual, and multicultural data. This challenge is found in many application domains such as computer and network forensics, financial fraud detection, and competitive intelligence. Following much planning, the workshop took place over two days at The MITRE Corporation in McLean, VA and involved participants from industry academia, and government.

On the first day, a combination of hands-on exercises, presentations, and discussions enabled the participants to gain a more complete picture of the challenge problem. The first exercise, called the *MicroExperiment*, was a hands-on exercise in which participants were divided into several multidisciplinary teams, given a single collection, and asked to review it and report as much as possible about it in a short period of time. After regrouping and reporting out, the teams were then asked to engage in a larger group exercise, the *MacroExperiment*, to consider how their experience with a single collection would scale up 100 or 1000 times. Reust et al (2005) reference multi terabyte collections as common, with one collection 450TBs large. Participants also gave short presentations on the contributions their respective disciplines bring to the problem, what could be done today, what would be possible in the future, and what gaps must be bridged.

On the second day, informed by the experiences and knowledge gained during day one, the workshop participants focused on key issues that arose during the experiments and developed recommendations to achieve the workshop's goals. For example, for analysis they identified the need to:

- Find relevant information quickly from hundreds to thousands of collections
- Filter irrelevant noise
- Turn noise into information as knowledge evolves
- Correlate entities (e.g., people, organizations, locations, events) across many sources
- Counter denial and deception
- Deal with evolving information and tradecraft
- Manage multiple uses of results (e.g., investigation vs. intelligence, tactical vs. strategic)

Workshop results were refined via virtual collaboration for several months following the workshop.

One of the basic assumptions of the workshop was that discovering valuable intelligence buried within large amounts of multimedia, multilingual, multicultural data requires a multifaceted approach. A comprehensive understanding would need to incorporate expertise from many fields, including, but not limited to, computer science, data mining, information retrieval, information extraction, knowledge discovery, digital libraries, human-computer interaction, collaboration and computer-supported cooperative work, psychology, ethnography, organizational behavior, forensics, information security, competitive intelligence, and cognitive science, among others. Several key aspects of the challenge include the need to address data, architecture and tools, analytic methods, collaboration, and evaluation methodology.

2. Problem

Today's information analysts need to organize, access, and "mine" maximally relevant information from large volumes of extremely diverse data. This need cuts across many problem domains, including computer forensics, financial fraud detection, and competitive intelligence. Key features of the data that analysts must deal with include:

- *Massive*: The data sets are large, complex, and heterogeneous (e.g., unstructured, semi-structured, and structured data).
- *Multimedia*: The data consist of a variety of objects – including text, images, and audio – in a variety of file formats.
- *Multilingual*: The data are in many natural languages, not just English.
- *Multicultural*: The data come from and are analyzed by users with distinct cultural backgrounds, cognitive styles, and purposes; they may use information and tools very differently, which in turn suggests distinct methods of storage, organization, and retrieval.
- *Multiscale*: Significant information may be derived at many levels of scale – subdocuments, documents, collections or repositories
- *Streaming*: Some data are not disk-resident, but are streaming, and therefore require real-time analysis.
- *Heterogeneous purposes*: The purposes of analysis differ among different users; they may be investigative (e.g., law enforcement), tactical (e.g., military), and strategic (e.g., national intelligence).
- *Continuously changing tradecraft/evolving analysis techniques*: Analysis techniques evolve continuously and analysis results today will by definition not be the same as results from yesterday or tomorrow. Both information seekers and hiders modify their behaviors to achieve and/or avoid discovery.
- *Denial & Deception*: The adversary may have intentionally hidden information among the data or distort data (e.g., using a malicious root kit).

Analysts' skills in dealing with these data vary, as do their strategies for retrieval and analysis. They may work individually or collaboratively. They will have individual biases based on their training, abilities, experiences, beliefs, and history.

Given the nature of the problem, it is likely that the solution will need to draw on multiple, heterogeneous disciplines. Point solutions exist, but how can they work together? Also, what kind of architecture/framework do analysts need to bypass "noise," navigate data, and locate interesting/relevant information? What gaps need to be filled to achieve an integrated solution? The following high-level issues would need to be addressed by a solution encompassing both humans and machines:

- *Architecture*: The system should have an open, plug-and-play architecture supporting heterogeneous, possibly asynchronous operations.
- *Privacy, Security, Information Sharing*: The system should both preserve, if not enhance, privacy and security and enable sharing of information across organizational boundaries, particularly the promotion and amplification of trust.
- *Evaluation*: The system must be evaluated on multiple measures (e.g., timeliness, correctness, quality, usability).

A solution should have the following features:

- *Noise Reduction*: The ability automatically to filter out the irrelevant "noise" in large volumes of diverse data and isolate only the information of value for the immediate task.
- *Speed*: The ability quickly to isolate the relevant information from this sea of data.
- *Accuracy*: The ability to strike the proper balance between recall and precision[1] and employ as many automated techniques as possible, thus minimizing the chance that the analyst will miss anything of importance.
- *Retention*: The ability to allow noise to be saved and become information in the future by enabling analysts to look at old data in new ways as new thinking and knowledge about particular topics evolve.
- *Longevity*: The ability to allow data to be saved for the long term to support retrospective analysis.

KNEAD Problem Statement: To create tools and methods to enable groups of interdisciplinary analysts to organize, access, and "mine" maximally relevant information from large volumes of continuously changing multimedia, multilingual, and multicultural data.

[1] Recall is the ability to retrieve *all* of the relevant results whereas precision is the ability to retrieve *only* relevant results.

- *Cross correlation*: The ability to compare different discoveries from one datatype and/or dataset with/to another.

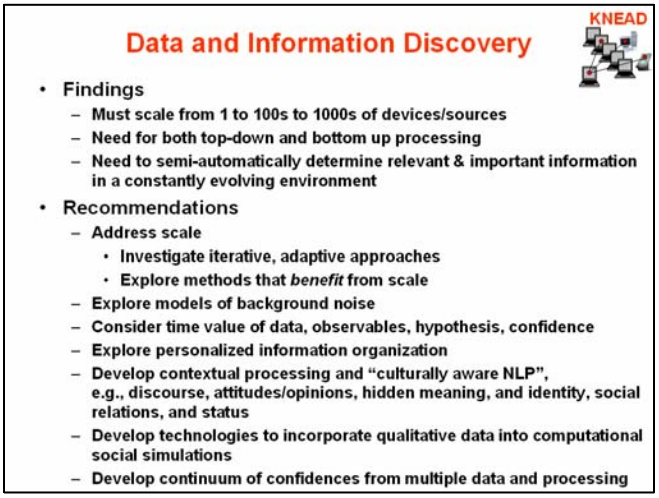
- *Process*: An environment that supports a workflow without impeding the analyst's flexibility. How does this workflow support individual, team, and enterprise activities? Is there support for structured, unstructured, and/or *ad hoc* workflow? The analyst should easily be able to view and work with the same data in different analysis tools.

- *Automation*: In inherently errorful processes likely complicated by cascading errors, what is the appropriate tradeoff between manual analyst and automated computer support?

The remainder of the report captures the findings and recommendations in each key area including data and information discovery, architectures and tools, analytic methodologies, analyst collaboration, and evaluation.

3. Data and Information Discovery

Figure 1 summarizes the findings and recommendations with respect to data and information discovery. The workshop considered how analysts are challenged to move from the processing of single, homogeneous collections to dealing with hundreds and thousands of heterogeneous collections. Accordingly, methods should seek to *benefit* from scale. Data processing requires addressing issues of confidence, trust, novelty, redundancy, context, and culture within an integrated information space that embraces live and evolving data. Heterogeneity will come from many sources including multiple complex data types, platforms (e.g., computers, routers, PDAs, cell phones), operating systems, and applications.



Data and Information Discovery

KNEAD

- Findings
  - Must scale from 1 to 100s to 1000s of devices/sources
  - Need for both top-down and bottom up processing
  - Need to semi-automatically determine relevant & important information in a constantly evolving environment
- Recommendations
  - Address scale
    - Investigate iterative, adaptive approaches
    - Explore methods that *benefit* from scale
  - Explore models of background noise
  - Consider time value of data, observables, hypothesis, confidence
  - Explore personalized information organization
  - Develop contextual processing and "culturally aware NLP", e.g., discourse, attitudes/opinions, hidden meaning, and identity, social relations, and status
  - Develop technologies to incorporate qualitative data into computational social simulations
  - Develop continuum of confidences from multiple data and processing

Figure 1. Data and Information Discovery

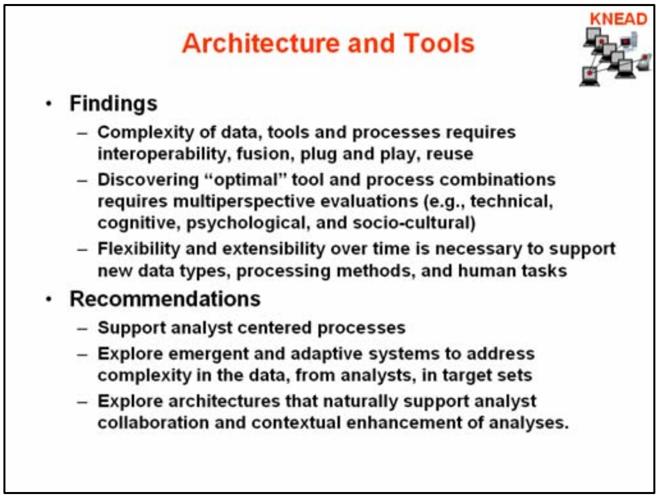
There is also a need to semi-automatically determine relevant and important information in this constantly evolving environment. This might imply exploring dynamic models of background noise since what is considered noise now could later become information. Accordingly, one method that was discussed was the ability to explicitly annotate and evaluate the time value of data, observables, hypotheses, and confidences. In the latter case, we will need to develop a continuum of confidences arising from multiple data sources and processing methods (e.g., to assess and combine confidences from multiple automated processes and human ones). Some efforts have explored using outlier analysis and existing evidence to automat digital evidence target definition (Carrier and Spafford 2005).

There is also a need for both top-down and bottom up processing as well as personalized information organization. One method of accessing the richness underlying data is to develop contextual (e.g., spatial, temporal, topical) processing as well as "culturally aware NLP". Examples of the latter include the automated extraction of discourse, attitudes/opinions, hidden meaning, identity, social relations, and status. Preliminary efforts in this direction have focused on the modeling of concepts, attributes, and relationships (Bogan and Dampier, 2005). Finally, we will need methods and technologies to develop capabilities to incorporate qualitative data into computational social simulations.

4. Architectures and Tools

Figure 2 summarizes the findings and recommendations related to tools and architectures. The size of collections, their complexity and dynamicity demand a flexible approach to data processing and architecture. A challenge is to develop tools that can fit interchangeably into a multianalyst, asynchronous process that supports the contextual and cultural enrichment of data and the multiplexing of automated, non-automated, and semi-automated processes/methodologies. Discovering the "optimal" tool and process combinations will require multiperspective evaluations that will consider multiple dimensions such as technical, cognitive, psychological, and socio-cultural ones. Flexibility and extensibility over time is necessary to support new data types, processing methods, and human tasks. In short, the complexity of the data, tools and processes requires interoperability, fusion, plug and play, reuse, and extensibility.

In light of these findings, the workshop recommended a focus on analyst centered processes. It suggested the exploration of emergent and adaptive systems to address complexity arising from the data, analysts, and target sets. Finally, it encouraged the exploration of architectures that naturally support analyst collaboration and contextual enhancement of analyses.



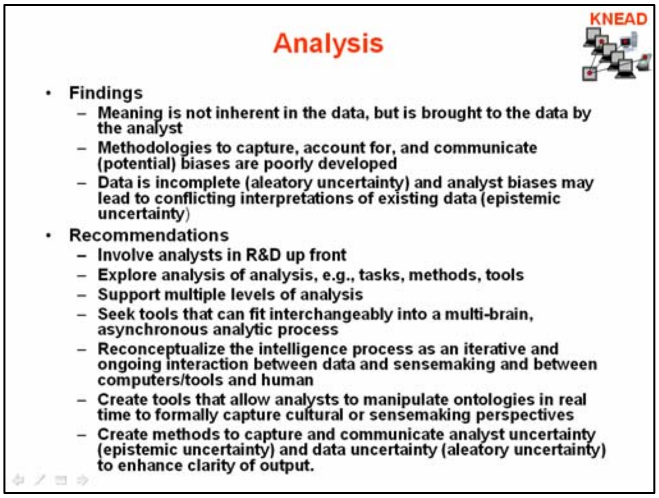
Architecture and Tools

- Findings
  - Complexity of data, tools and processes requires interoperability, fusion, plug and play, reuse
  - Discovering “optimal” tool and process combinations requires multiperspective evaluations (e.g., technical, cognitive, psychological, and socio-cultural)
  - Flexibility and extensibility over time is necessary to support new data types, processing methods, and human tasks
- Recommendations
  - Support analyst centered processes
  - Explore emergent and adaptive systems to address complexity in the data, from analysts, in target sets
  - Explore architectures that naturally support analyst collaboration and contextual enhancement of analyses.

Figure 2. Architecture and Tools

5. Analytic Methodologies

Effective architectures can enable more effective analysis. Figure 3 summarizes the findings and recommendations of the workshop participants with respect to analysis. One key observation is that meaning is not inherent in the data, but is brought to the data by the analyst. Because the analyst plays such a key role, it is important to understand their limitations. Unfortunately, methodologies to capture, account for, and communicate (potential) human biases are poorly developed. A related key observation is the pervasiveness of uncertainty. Data is incomplete (aleatory uncertainty) and analyst biases may lead to conflicting interpretations of existing data (epistemic uncertainty).



Analysis

- Findings
  - Meaning is not inherent in the data, but is brought to the data by the analyst
  - Methodologies to capture, account for, and communicate (potential) biases are poorly developed
  - Data is incomplete (aleatory uncertainty) and analyst biases may lead to conflicting interpretations of existing data (epistemic uncertainty)
- Recommendations
  - Involve analysts in R&D up front
  - Explore analysis of analysis, e.g., tasks, methods, tools
  - Support multiple levels of analysis
  - Seek tools that can fit interchangeably into a multi-brain, asynchronous analytic process
  - Reconceptualize the intelligence process as an iterative and ongoing interaction between data and sensemaking and between computers/tools and human
  - Create tools that allow analysts to manipulate ontologies in real time to formally capture cultural or sensemaking perspectives
  - Create methods to capture and communicate analyst uncertainty (epistemic uncertainty) and data uncertainty (aleatory uncertainty) to enhance clarity of output.

Figure 3. Analysis

Analysts should both have deep expertise and be skeptical, that is be anxious about their own conclusions. Since it is not possible to know in advance what is important, we will always need the human mind, where the imagination together with a “nose” connected to that can “sniff” out the truth to help decide what things are important, particularly where deception is present. In fact, one key limitation of machines is the current inability to detect deceit. And because direct contact with the subject is important, analysts need to get outside of their universe to directly understand the problem. This of course implies original foreign language source is important as it also includes cultural interpretation.

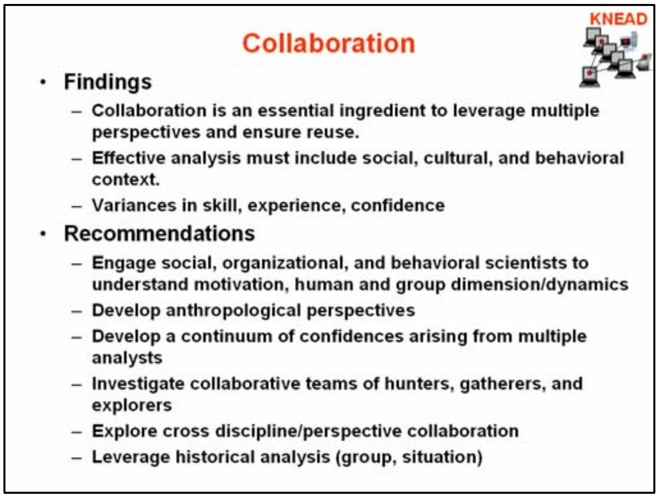
The analysis challenge requires new concepts for analyst-centered processes/methods, including how tools fit interchangeably into a reconceptualized analyst workflow that is symbiotic (not simply faster, better, more accurate) and ensures added value/meaning. For example, the process should be an iterative and ongoing interaction between data and sensemaking and between computers/tools and human. Another recommendation was that analysts should be involved in research and development up front to guide it. Furthermore, research should focus on the analysis of analysis, including tasks, methods, and tools and should support multiple levels of analysis. There also is a need for tools that allow analysts to manipulate ontologies in real time to formally capture cultural or sensemaking discoveries. Finally, methods should be created that can capture and communicate analyst uncertainty (epistemic uncertainty) and data uncertainty (aleatory uncertainty) to enhance the clarity of output.

6. Analyst Collaboration

To be successful analysts need to collaborate. Figure 4 summarizes the findings and recommendations regarding collaboration of analysts. The workshop found that collaboration was necessary in order to leverage multiple perspectives and ensure reuse across analysts and investigations. Collaboration services that securely support both synchronous (e.g., chat, voice/video/data conferencing) and asynchronous (e.g., file sharing, email) interactions, as well as mechanisms for information persistence and shared awareness of individual and group expertise and activities will be essential for efficient group interactions.

Characterizing multiple alternative viewpoints and where possible finding a common ground among competing views is an important activity. An advantage of the latter is that sometimes apparent conflicting views disappear over time. This implies when selecting analytic teams the importance of knowing which individuals work well in groups. However, while diversity is important for discovery, analytic findings should be attributed to an individual or two as opposed to a group to ensure accountability, to increase transparency, and to expose credibility.

Multiple collaborating analysts implies, however, that there will be variances in skill, experience, and confidence. Success of interdisciplinary teams will require not only infrastructure for team support such collaboration services and spaces (e.g., team rooms) but also services such as team training and expert facilitation. For example, Majchzrak et al. (2004) studied 54 successful teams in 26 companies who rarely if ever met as a whole face-to-face. They found that far-flung teams can actually be more productive than their face-to-face counterparts if they keep three practices: exploit diversity, use simple technology (e.g., teleconference calls and shared websites) to simulate reality, and hold the team together via lots of communication. And while information and technical analysis is foundational, the workshop group asserted that incorporation of social, cultural, and behavioral context is important to a more comprehensive understanding. More generally, the group recognized that the KNEAD challenge requires new concepts that explore how cross-disciplinary collaboration can influence analyst perspectives to enable breakthroughs.


KNEAD

Collaboration

- Findings
  - Collaboration is an essential ingredient to leverage multiple perspectives and ensure reuse.
  - Effective analysis must include social, cultural, and behavioral context.
  - Variances in skill, experience, confidence
- Recommendations
  - Engage social, organizational, and behavioral scientists to understand motivation, human and group dimension/dynamics
  - Develop anthropological perspectives
  - Develop a continuum of confidences arising from multiple analysts
  - Investigate collaborative teams of hunters, gatherers, and explorers
  - Explore cross discipline/perspective collaboration
  - Leverage historical analysis (group, situation)

Figure 4. Collaboration

Accordingly, the group recommended incorporating social, organizational, and behavioral scientists during analyses to understand motivation, human and group dimension/dynamics. It also recommended the development of anthropological perspectives to support digital forensics analysis. It recommended the development of tools and methods to manage the continuum of confidences that arise from collaborating analysts. Also necessary are mechanisms to leverage analyses of groups or situations that occur over time or across analytic teams.

Finally the group discussed the challenge of simultaneously supporting multiple analytic methodologies using collaborative teams. Just as successful computation will require heterogeneous methods and integration, so too human collection and analytic activities will require a multidisciplinary and multifaceted approach. One useful metaphor could be that of collaborative teams of hunter, gatherer, and explorer (see Figure 5). Hunters stereotypically chase specific prey or targets using specialized tools that extend their range and effectiveness whereas gatherers typically go to pre-known locations to collect materials. By contrast, explorers typically travel to unknown locations seeking interesting or valuable artifacts, and along the way map out territory, react to local conditions, and act opportunistically.

In summary, the group identified the following necessary actions to advance collaboration:

- Understand and model collaboration within and across disciplines
- Evaluate alternative collaborative work structures
- Develop shared mental models
- Identify and reduce barriers to successful collaboration
- Develop tools and methods to foster collaboration
- Integrate analytic tools across disciplines

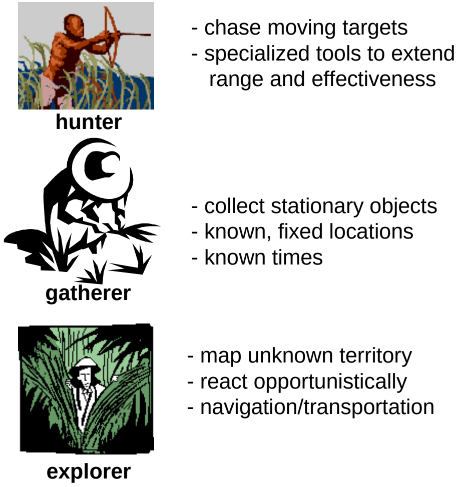

- chase moving targets
- specialized tools to extend range and effectiveness
hunter
- collect stationary objects
- known, fixed locations
- known times
gatherer
- map unknown territory
- react opportunistically
- navigation/transportation
explorer

Figure 5. Collaborative teams of analysts who are hunters, gatherers, and explorers

7. Evaluation

To motivate and measure rapid progress given data complexity and the continuously evolving analytic challenge will require task-based and task-situated evaluation methods. Evaluation of novel algorithms, tools, and analytic methods in KNEAD will require careful evaluation methods, metrics, and measures. Some requirements for evaluations include:

1. The results must be valid, reliable, and objective. Metrics must be simple to specify and straightforward to measure through a standard method. They should be objective and replicable and, ideally, automatable to support evaluation of large data sets. Preferable also

are those that are independent of (natural) language, theory, and development paradigm.

2. The process must be cost effective to administer in as many resource dimensions (time, cost, data, human) as possible.
3. The results must be useful to the consumer of the evaluation weather they be for users, program managers, developers or systems integrators).

The group recognized the EAGLES (Expert Advisory Group for Language Engineering Standards) panel's design of a task-based approach to evaluation which could serve as a model for evaluation of analytic tools. Also identified was the AQUAINT program which has explored two novel evaluation approaches. One involves task-based cross-evaluation of the production of draft reports, implemented using factor analysis. This has proven capable of extracting statistically significant differences among different ways of supporting the analytic process, while rigorously correcting for the (usually very large) effects of task complexity and analyst skill. The process itself was developed in connection with the DARPA-funded AntWorld project evaluation (Sun and Kantor, to appear) and was refined at the ARDA AQUAINT 2004 Challenge Workshop.

8. Summary

In summary, KNEAD challenge is characterized by large-scale, complex, multilingual, multimedia, multiparty, multicultural, operationally relevant dirty data. Among important findings with respect to solution specifications:

- The solution must address the need to scale, eliminate noise, process heterogeneous sources, support multidisciplinary analysis, and manage uncertainty.
- The research must be (realistic) data driven and analyst/operator driven.
- The research as well as analysis discovery process must be iterative and rapid.

Other key findings related to the process were:

- Small experiments are necessary to converge on progress.
- Investments must be differentiated and leveraged, and many existing investments in other government agencies can contribute to the solution.
- Both unclassified and sensitive/classified data sets are needed to effectively evaluate performance of tools and methods. However, both share many common features implying a public dataset is possible which could accelerate scientific discovery.
- A jump start demonstration and experiment would help accelerate progress.

To advance research, the workshop recommended to:

- Employ multidisciplinary research teams (including ethnographers, computer scientists, psychologists, and other domain experts as needed) given the interdisciplinary nature of the challenge.
- Augment existing programs with investments to advance KNEAD-specific gaps as opposed to launching an entirely new program.

Finally, the group identified important areas for further research including:

- Scaling from single collection to hundreds or thousands of collections, to include cross-device and cross-collection analysis.
- Enhanced processing, to include reduction of noise from massive data, detecting deception, and context and cultural enrichment of data to discover attitudes and opinions, hidden meaning, and social analysis of relations and status.
- Collaborative and multiperspective analysis.
- Hypothesis, evidence, and uncertainty management to enable multiple analysts to deal with a continuum of data quality and confidence levels.
- Tailorable architectures and environments that support configuration of data, processing, and presentation/interaction and to support the changing nature of the threat, data, and analytic process.

Acknowledgments

Special appreciation goes to challenge workshop participants including Brian Carrier (Basis/Purdue), Jeremy Christianson (IRS), LeeEllen Friedland (MITRE), Susan Fussell (CMU), Paul Kantor (Rutgers University), Sue Lee (Johns Hopkins University), Flo Reeder (MITRE), Sarah Taylor (Lockheed Martin), Jessica Glicken Turnley (Galisteo Consulting Group), Jay Wallin (MITRE). We also thank Brant Cheikes, Chris Elsaesser, Frank Stech, Guido Zarelli, and Linda Smith for their post workshop contributions. We also received feedback from many including but not limited to Fred Roberts (Rutgers), Laura McNamara (Sandia), Gene Spafford (Purdue), Sara Kiesler (CMU) and Michael Ledeen (AEI). Finally we thank Margaret MacDonald and Paula MacDonald for their editorial and administrative support, respectively.

References

Bogan, Chris and Dampier, David. 2005. Preparing for Large-Scale Investigations with Case Domain Modeling. Digital Forensic Research Workshop. August 17-

19, 2005, New Orleans, LA. http://www.dfrws.org/2005/proceedings/bogen_domainmodel.pdf

Carrier, Brian. 2006. CERIAS TR 2006-06. A Hypothesis-based approach to digital forensic investigations. PhD Thesis. Purdue University, West Lafayette, IN 47907-2086.

Carrier, Brian. and Spafford, Eugene H. 2005. Automated digital evidence target definition using outlier analysis and existing evidence. Digital Forensic Research Workshop. August 17-19, 2005, New Orleans, LA. http://www.dfrws.org/2005/proceedings/carrier_targetdefn.pdf

Majchzrak, Ann, Malhotra, Arvind, Stamps, Jeffrey, and Lipnack, Jessica. 2004. Can Absence Make a Team Grow Stronger. *Harvard Business Review.* May 1, 2004.

Reust, Jessica 2005. (ed). DFRWS 2005 Final Report. Digital Forensic Research Workshop. August 17-19, 2005, New Orleans, LA. http://www.dfrws.org/2005/DFRWS2005FinalReport.pdf

Sun, Ying, Kantor, Paul. To appear. Cross-evaluation: A New Model for Information System Evaluation. *Journal of the American Society for Information Science and Technology*.