



Knowledge Exploration, Analysis, and Discovery Workshop

By

Mark Maybury and Penny Chase

Presented At

The Digital Forensic Research Conference

DFRWS 2006 USA Lafayette, IN (Aug 14th - 16th)

DFRWS is dedicated to the sharing of knowledge and ideas about digital forensics research. Ever since it organized the first open workshop devoted to digital forensics in 2001, DFRWS continues to bring academics and practitioners together in an informal environment. As a non-profit, volunteer organization, DFRWS sponsors technical working groups, annual conferences and challenges to help drive the direction of research and development.

<http://dfrws.org>

KNEAD



Knowledge Exploration, Analysis, and Discovery (KNEAD) Challenge Workshop

**Mark Maybury
Penny Chase
The MITRE Corporation**

**DFRWS
15 August 2006**

KNEAD



Workshop Objective

To identify tools and methods to enable groups of interdisciplinary forensic analysts to organize, access, and “mine” maximally relevant information from large volumes of continuously changing multimedia, multilingual, and multicultural data



Workshop Approach

- **Bring together a cross-disciplinary group of experts from academia, industry, and government**
- **Two-day meeting at MITRE McLean, November 15 and 16, 2006**
 - **First day**
 - **Micro-problem**
 - **Macro-problem**
 - **Participants gave presentations on what contributions their disciplines bring to the problem**
 - **Second day**
 - **Focused on key issues that emerged during first day**
 - **Developed recommendations**
- **“Virtual Workshop”**
 - **Results were refined over several months**
 - **Small groups worked on different topics, collaborating via email, telecons, and face-to-face meetings**



Participants

- **Brian Carrier (Purdue)**
- ***Brant Cheikes (MITRE)***
- **Jeremy Christianson (IRS)**
- ***Chris Elsaesser (MITRE)***
- **LeeEllen Friedland (MITRE)**
- **Susan Fussell (CMU)**
- **Jessica Glick Turnley
(Galisteo Consulting Group)**
- **Paul Kantor (Rutgers)**
- ***Sara Kiesler (CMU)***
- ***Michael Ledeen (AEI)***
- ***Laura McNamara (Sandia)***
- **Sue Lee (JHU APL)**
- **Flo Reeder (MITRE)**
- **Fred Roberts (Rutgers)**
- ***Eugene Spafford (Purdue)***
- ***Frank Stech (MITRE)***
- **Sarah Taylor (LMCO)**

Participated in workshop follow-up



Micro and Macro Problems



- **Micro-problem**
 - Hands-on experiment
 - Small, interdisciplinary groups
 - Problem: what can you tell us about a single disk drive?
- **Macro-problem**
 - Thought experiment
 - Problem: what happens when you try to scale to 100s or 1000s of data collections?



Data Characteristics

- **Massive**
 - Volume
 - Complexity
- **Multimedia**
 - Data, Text, audio, image, etc.
 - Structured, unstructured, semi-structured
- **Multilingual**
 - Data is not just in English
 - A single document could be in multiple languages
- **Multicultural**
 - Multiple cultural backgrounds and cognitive styles of the data's users and creators
- **Multiscale**
 - Document to drive to computer to network
- **Streaming**
 - Non disk resident
 - Might require real-time analysis
- **Heterogeneous Purposes**
 - Investigative, tactical, strategic
- **Analysis Techniques Evolve**
 - Analysis results change
- **Denial & Deception**
 - Intentionally hidden or distorted data

Data and Information Discovery



- **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

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



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.

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)



Collaboration



hunter

- chase moving targets
- specialized tools to extend range and effectiveness



gatherer

- collect stationary objects
- known, fixed locations
- known times



explorer

- map unknown territory
- react opportunistically
- navigation/transportation



Evaluation

- **Requirements**
 - Results must be valid, reliable, and objective
 - Metrics should be simple to specify and straightforward to measure
 - Replicable and ideally automatable to support evaluation of large data sets
 - Independent of (natural) language, theory, and development paradigm
 - Process must be cost-effective across resource dimensions (time, cost, data, human)
 - Results must be useful to the consumer of the evaluation (users, developers, program managers)
- **Leverage prior work**
 - EAGLES (Expert Advisory Group for Language Engineering Standards) task-based approach
 - Task-based cross-evaluation (initially developed for DARPA AntWorld project and refined for AQUAINT)

Summary Findings and Recommendations



- **Solutions must address the need to scale, reduce noise, process heterogeneous sources, support multidisciplinary analysis, and manage uncertainty**
- **Research must be driven by realistic data and analysts/operators**
- **Research and analysis must be iterative and rapid**
- **Small experiments are necessary to converge on progress**
- **Both unclassified and classified/sensitive data sets are needed to effectively evaluate performance of tools and methods**

Summary Findings and Recommendations



- **Employ multidisciplinary research teams (including ethnographers, psychologists, computer science, domain experts)**
- **A “jump start” demonstration would accelerate progress**
- **Augment existing programs to advance KNEAD-specific gaps**
- **Areas for further research**
 - **Scaleable forensics**
 - **Contextual and cultural processing to enhance signal from noise (in an evolving haystack)**
 - **Collaborative, multiperspective analysis (awareness, annotation, discovery, and debate)**
 - **Exploring forensic hypotheses under uncertainty**
 - **Tailorable analytic environments**