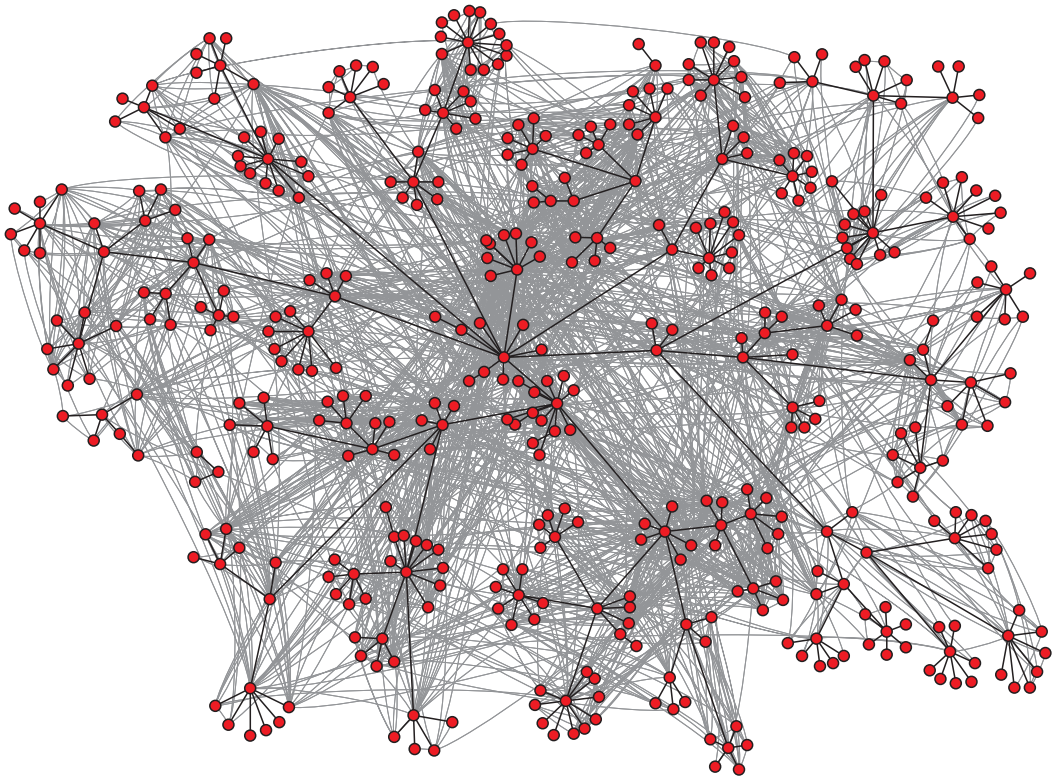
Data Sciences Technology for Homeland Security Information Management and Knowledge Discovery



Report of the DHS Workshop on Data Sciences
September 22-23, 2004

Jointly released by Sandia National Laboratories and Lawrence Livermore National Laboratory

T. Kolda, D. Brown, J. Corones, T. Critchlow, T. Eliassi-Rad, L. Getoor, B. Hendrickson, V. Kumar, D. Lambert, C. Matarazzo, K. McCurley, M. Merrill, N. Samatova, D. Speck, R. Srikant, J. Thomas, M. Wertheimer, P. C. Wong

SANDIA REPORT
SAND2004-6648
Unlimited Release
Printed January 2005

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

LLNL REPORT
UCRL-TR-208926
Unlimited Release
Printed January 2005

Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy at the University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online order http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

Data Sciences Technology for Homeland Security Information Management and Knowledge Discovery

Authors

Tamara Kolda, Sandia National Laboratories

David Brown, Lawrence Livermore National Laboratory

James Corones, Krell Institute

Terence Critchlow, Lawrence Livermore National Laboratory

Tina Eliassi-Rad, Lawrence Livermore National Laboratory

Lise Getoor, University of Maryland

Bruce Hendrickson, Sandia National Laboratories

Vipin Kumar, University of Minnesota

Diane Lambert, Bell Laboratories, Lucent

Celeste Matarazzo, Lawrence Livermore National Laboratory

Kevin McCurley, IBM Almaden Research Center

Michael Merrill, National Security Agency

Nagiza Samatova, Oak Ridge National Laboratory

Douglas Speck, Lawrence Livermore National Laboratory

Ramakrishnan Srikant, IBM Almaden Research Center

Jim Thomas, Pacific Northwest National Laboratory

Michael Wertheimer, Innovative Solutions, RABA Technologies, LLC

Pak Chung Wong, Pacific Northwest National Laboratory

Preface & Acknowledgements

On September 22-23, 2004, the Department of Homeland Security (DHS) sponsored a workshop on data sciences in Alexandria, Virginia. This was the third in a series of workshops to identify the DHS Advanced Scientific Computing (ASC) Research and Development (R&D) Program. The focus of this workshop was to identify mathematics and computer science R&D areas that will support future DHS operational requirements. This report summarizes the findings of this workshop.

The organizing committee gratefully acknowledges the assistance provided by the Krell Institute and by the staff of Institute for Scientific Computing Research (ISCR) at LLNL in preparing for and running the workshop.

Credit for picture on the front page: HP Labs' email communication (light grey lines) mapped onto organization hierarchy (black lines). Figure 3 from Lada Adamic and Eytan Adar, "How to search a social network," preprint submitted to *Social Networks*, October 26, 2004. Used with permission of the authors.

Table of Contents

Acronyms and Abbreviations

ADVISE	Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement, a thrust area in the TVTA portfolio
ASC	Advanced Scientific Computing, a program in the TVTA portfolio
BKC	Biodefense Knowledge Center, a program in the TVTA portfolio
DHS	U.S. Department of Homeland Security
IA	Information Analysis, part of the DHS Information Analysis and Infrastructure Protection Directorate
ISC	Information Sharing and Collaboration, a cross-directorate DHS program
IDS	Institute for Discrete Sciences, a program in the TVTA portfolio
NBACC	National Biodefense Analysis and Countermeasures Center
NVAC	National Visualization and Analytics Center, a program in the TVTA portfolio
R&D	Research and Development
RN	Radiological and Nuclear
RTAS	Regional Threat Assessment System
S&T	Science and Technology, a DHS directorate
SIGMOD	Special Interest Group on Management of Data
TVIS	Threat Vulnerability Information System
TVTA	Threat and Vulnerability, Testing and Assessment, an S&T portfolio
VLDB	Very Large Database

Workshop Participants

Steven Ashby
Deputy Associate Director
Computing Applications and Research,
Lawrence Livermore National Laboratory
and DHS ASC Thrust Area Lead

Mark Bradley
Science and Technology Directorate
U.S. Department of Homeland Security

David Brown[1]
Computing Applications and Research
Lawrence Livermore National Laboratory

Eric Yisroel Brumer
Science and Technology Directorate
U.S. Department of Homeland Security

John Conroy
Institute for Defense Analyses Center
for Computing Sciences

James Corones[1]
President
Krell Institute

Terence Critchlow[5]
Lawrence Livermore National Laboratory

George Cybenko[3]
Dorothy and Walter Gramm
Professor of Engineering
Dartmouth College

Susan Davidson[3]
Weiss Professor
Department of Computer & Information Science
University of Pennsylvania.

Stephen Dennis[3]
Deputy Director of Research & Engineering
Information Sharing and Collaboration
Program Office
U.S. Department of Homeland Security

Cynthia Dwork[3]
Microsoft Research

Tina Eliassi-Rad[1]
Lawrence Livermore National Laboratory

Auroop R. Ganguly
Oak Ridge National Laboratory

Johannes Gehrke[3]
Assistant Professor
Department of Computer Science
Cornell University

Lise Getoor[1]
Assistant Professor
Department of Computer Science
University of Maryland

Bruce Hendrickson[2]
Distinguished Member of Technical Staff
Sandia National Laboratories

John Hoyt
Program Manager,
Office of Research and Development
Science and Technology Directorate
U.S. Department of Homeland Security

David Jensen[3]
Associate Professor
Department of Computer Science
University of Massachusetts

Cliff Joslyn
Los Alamos National Laboratory

Joseph Kielman[3]
TVTA Portfolio Manager,
Science and Technology Directorate,
U.S. Department of Homeland Security

[1] Member of Organizing and Writing Committees
[2] Session Chair and Member of Writing Committee
[3] Speaker
[4] Chair of Organizing and Writing Committees
[5] Member of Writing Committee

Jon Kleinberg[3]
Associate Professor
Department of Computer Science
Cornell University

Tamara Kolda[4]
Principal Member of Technical Staff
Sandia National Laboratories

Vipin Kumar[1]
Director,
Army High Performance Computing Research Center
University of Minnesota

Diane Lambert[2]
Bell Laboratories, Lucent

Teresa Lustig
Science and Technology Directorate
U.S. Department of Homeland Security

Celeste Matarazzo[1]
NAIC Division Leader
Lawrence Livermore National Laboratory

Andrew McCallum[3]
Associate Professor
Department of Computer Science
University of Massachusetts

Kevin McCurley[1]
IBM Almaden Research Center

Michael Merrill[1]
Engineering Leader
National Security Agency

Noël Nachtigal[3]
Principal Member of Technical Staff
Sandia National Laboratories

Frank Olken
Computer Scientist
Scientific Data Management
Research Group
Lawrence Berkeley National Laboratory

Thomas Potok
Group Leader
Applied Software Engineering
Research Group
Oak Ridge National Laboratory

Doron Rotem
Lawrence Berkeley National Laboratory

Nagiza Samatova[1]
Oak Ridge National Laboratory

Peter Sand[3]
Director of Privacy Technology
Privacy Office
U.S. Department of Homeland Security

David Shepherd
Strategic Analysis, Inc
Technical Advisor for Knowledge Discovery and Dissemination
TVTA Portfolio
Science and Technology Directorate
U.S. Department of Homeland Security

Tom Slezak[3]
Lawrence Livermore National Laboratory

Burton Smith
Chief Scientist
Cray Inc.

Douglas Speck[1]
Lawrence Livermore National Laboratory
ASC Program Liaison
Science and Technology Directorate
Department of Homeland Security

Ramakrishnan Srikant[2]
IBM Almaden Research Center

Latanya Sweeney[3]
Assistant Professor
Department of Computer Science
Carnegie Mellon University

Jim Thomas[1]
Pacific Northwest National Laboratory

John van Rosendale
Program Manager
DOE Office of Advanced Scientific Computing Research

Karen Verspoor
Computational Sciences Division
Los Alamos National Laboratory

Michael Wertheimer[2]
Director, Innovative Solutions
RABA Technologies, LLC

Pak Chung Wong[1]
Chief Scientist
Pacific Northwest National Laboratory

Agenda

Wednesday, September 22, 2004
Opening Remarks • Joseph Kielman, DHS TVTA • Steven Ashby, Lawrence Livermore National Laboratory
DHS Panel • Stephen Dennis, DHS TVTA • Pete Sand, DHS Privacy Technology Office • Noël Nachtigal, Sandia National Laboratories
Session 1: Applications *Chair: Mike Wertheimer, RABA Technologies* • Tom Slezak, Lawrence Livermore National Laboratory • Everett Wheelock, Lawrence Livermore National Laboratory[6]
Session 2: Security and Privacy *Chair: Ramakrishnan Srikant, IBM-ARC* • Latanya Sweetney, Carnegie Mellon University • Cynthia Dwork, Microsoft Research
Thursday, September 23, 2004
Session 3: Relational Data and Graphs *Chair: Diane Lambert, Bell Labs, Lucent* • George Cybenko, Dartmouth College • David Jensen, University of Massachusetts • Jon Kleinberg, Cornell University
Session 4: Data Analysis, Modeling and Inference *Chair: Pak Chung Wong, Pacific Northwest Laboratory* • Johannes Gehrke, Cornell University • Andrew McCallum, University of Massachusetts
Session 5: Data Integration *Chair: Bruce Hendrickson, Sandia National Laboratories* • Susan Davidson, University of Pennsylvania

Workshop presentations are available for download at the workshop Web site at http://www.ascworkshop.info/sep_2004/.

[6] Dr. Wheelock's presentation presented by Dr. Slezak

I. Executive Summary

The Department of Homeland Security (DHS) has vast amounts of data available, but its ultimate value cannot be realized without powerful technologies for knowledge discovery to enable better decision making by analysts. Past evidence has shown that terrorist activities leave detectable footprints, but these footprints generally have not been discovered until the opportunity for maximum benefit has passed. The challenge faced by the DHS is to discover the money transfers, border crossings, and other activities in advance of an attack and use that information to identify potential threats and vulnerabilities.

The data to be analyzed by DHS comes from many sources ranging from news feeds, to raw sensors, to intelligence reports, and more. The amount of data is staggering; some estimates place the number of entities to be processed at 1015. The uses for the data are varied as well, including entity tracking over space and time, identifying complex and evolving relationships between entities, and identifying organization structure, to name a few. Because they are ideal for representing relationship and linkage information, semantic graphs have emerged as a key technology for fusing and organizing DHS data. A semantic graph organizes relational data by using nodes to represent entities and edges to connect related entities. Hidden relationships in the data are then uncovered by examining the structure and properties of the semantic graph.

The DHS has three primary missions: to prevent terrorist attacks within the United States, to reduce America's vulnerability to terrorism, and to minimize the damage from potential attacks and natural disasters. The Directorate of Science and Technology (S&T) serves as the primary research and development arm of Homeland Security, and the Threat and Vulnerability, Testing and Assessment (TVTA) portfolio within S&T creates advanced modeling, information and analysis capabilities to evaluate extensive amounts of data and information from diverse sources.

Within TVTA, the Advanced Scientific Computing (ASC) program serves as a research and development arm to develop innovative computational technologies for deployment in next-generation homeland security applications.

The ASC program sponsored a workshop on September 22-23, 2004 in Alexandria, Virginia, with the purpose of identifying and elucidating the future R&D needs of the DHS in the data sciences. The workshop brought together approximately 50 invited participants, representing the DHS, the U.S. Department of Energy (DOE) and its national laboratories, academia, and industry. This report summarizes the findings of the workshop.

The R&D needs identified by the workshop focus on semantic graph technologies. Semantic graphs are at the core of the Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE) thrust area in TVTA. This thrust area will provide a common platform that supports scalable knowledge management across multiple missions. It will have an integrated suite of tools for the analyst, including visualization and query interfaces, as well as methods for ingesting data and integrating disparate data sources.

Although semantic graphs are not a new concept, the DHS faces unique challenges due to the scale of the data and the complex knowledge discovery needs inherent in the homeland security mission. The development of an R&D program encompassing the following five areas will greatly advance the DHS's capabilities in fulfilling its mission to prevent terrorist attacks and reduce the nation's vulnerability to terrorism:

1. **Architecture and management of databases for large-scale semantic graphs,** including issues associated with distributed databases, ingestion of large volumes of data from both structured and unstructured sources, and integration of data from different sources with different representations.

2. **Scalable algorithms and interfaces for information retrieval and analysis on semantic graphs,** such as scalable algorithms for the discovery of complex relationships between nodes, an efficient query language for semantic graphs, scalable visualization tools and intuitive user interfaces, and the integration of simulation results.

3. **Models for detection and prediction on semantic graphs**, including detection of missing or incorrect data; statistical prediction of attributes, links and subgraph patterns; identification of anomalous nodes or relationships; and models that incorporate temporal or spatial effects.

4. **Models for discovering and detecting processes on graphs**, such as statistical and machine learning models for determining organization structure, finding portions of the graph that are undergoing abrupt changes, and using processes to aid in analysis.

5. **Algorithms to provably ensure privacy and security**, including the development of policies that account for multiple levels of trust and access, new anonymization methods with provable privacy guarantees, and development of models that allow for trade-offs between privacy and national security.

Supporting the highest caliber research is critical to the success of this R&D program. To attract the best researchers, the DHS should provide strong support for open research, release test data for competitive analysis, and establish programs for exchanges between researchers and analysts, for postdoctoral fellowships, and for summer research institutes to focus on critical research problems.

II. Introduction

The Department of Homeland Security (DHS) has three primary missions: to prevent terrorist attacks within the United States, to reduce America's vulnerability to terrorism, and to minimize the damage from potential attacks and natural disasters. The Directorate of Science and Technology (S&T) serves as the primary research and development arm of Homeland Security. The Threat and Vulnerability, Testing and Assessment (TVTA) portfolio within S&T creates advanced modeling, information and analysis capabilities that are used to enhance S&T's ability to evaluate extensive amounts of data and information from diverse sources. Within TVTA, the Advanced Scientific Computing (ASC) program provides computing expertise and capabilities for homeland security; its mission is to develop innovative computational technologies for deployment in next-generation homeland security applications. This report outlines research and development (R&D) objectives on semantic graph technologies for the ASC program, in support of the missions of DHS. In this chapter, each of the DHS components mentioned above is described and an outline of the remainder of the document is provided.

To support its primary missions, the DHS leverages resources within federal, state, and local governments, coordinating the transition of multiple agencies and programs into a single, integrated agency focused on protecting the American people and their homeland. More than 87,000 different governmental jurisdictions at the federal, state, and local level have homeland security responsibilities. The comprehensive national strategy seeks to develop a complementary system connecting all levels of government without duplicating effort.[7]

The S&T directorate serves as the primary R&D arm of DHS, using our nation's scientific and technological resources to provide federal, state, and local officials with the technology and capabilities to protect the homeland. The focus is on catastrophic terrorism—threats to the security of our homeland that could result in large-scale loss of life and major economic impact. S&T's work is designed to counter those threats with evolutionary improvements to current technological capabilities and development of revolutionary technological capabilities. It unifies and coordinates much of the federal government's efforts to develop and implement scientific and technological countermeasures, including channeling the intellectual energy and extensive capacity of scientific institutions, such as the national laboratories and academic institutions.[7]

The TVTA portfolio within S&T creates advanced modeling, information and analysis capabilities, which includes advancing the nation's capabilities in weapons of mass destruction intelligence analysis; developing capabilities in advanced supercomputing; creating advanced systems capable of merging terrorist threat data with infrastructure vulnerability data to improve warning and response; and integrating analytic, scientific and technological resources in performing net assessments of capabilities versus known or projected threats. [8,9]

The TVTA program uses a strategy of multi-year investments that infuse new capabilities into the DHS mission directorates on a regular basis based on strategic five-year road maps.

[7] From the DHS Web site at http://www.dhs.gov

[8] From the statement of Under Secretary Dr. Charles E. McQuery, DHS S&T Directorate, before the House Select Committee on Homeland Security Subcommittee on Cybersecurity, Science, and Research and Development on May 21, 2003.

[9] From the ORAU Web site at: http://www.orau.gov/dhsed/04abstracts.htm

A spiral development process ensures early use and feedback by intended users and operators of all technologies developed within the program. Successively, more complete and refined prototypes lead to operational pilots and fully operational systems for the department organizations.[8]

Within the TVTA, the mission of the ASC program is to develop innovative computational technologies for deployment in next-generation homeland security applications. A strategic planning process conducted by the S&T Directorate in the Spring of 2004 identified four strategic focus areas in advanced scientific computing that will be required to meet DHS needs:Integrated simulation analysis capabilities, scalable information management and knowledge discovery, scalable discrete mathematics, and high-performance computing resources.

To meet these needs, ASC has recommended the formation of the Institute for Discrete Sciences (IDS), a virtual center that will engage academia as well as the national laboratories. The initial focus of the IDS will be on TVTA programs.[10]

As is discussed further in Chapter I, knowledge management is a major component of TVTA's responsibilities. In particular, the Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE) thrust area is focused on providing a common platform that supports scalable knowledge management across multiple missions. ADVISE is being developed in response to the needs of the DHS Information Sharing and Collaboration (ISC) program. The goal of the ISC program is to coordinate and facilitate efforts for enabling information sharing throughout DHS and with its customers and partners (especially the federal, state, and local governments). Further, other parts of S&T also have data sciences needs that TVTA can help support, including the Radiological and Nuclear Countermeasures (RN) portfolio.

To fully understand the five-year R&D data sciences needs of TVTA, the ASC program sponsored a workshop on data sciences on September 22-23, 2004 in Alexandria, Virginia. The workshop brought together approximately 50 participants, representing DHS, the U.S. Department of Energy (DOE) and its national laboratories, academia, and industry. The research goals identified by this team of experts are discussed in Chapter I. Further recommendations critical to bringing excellence to the R&D on data sciences technologies are presented in Chapter V.

[10] From "Advanced Scientific Computing for Homeland Security" presented by Steve Ashby, August 5, 2004.

III. Data Sciences Needs in the DHS S&T Directorate

The mission of the TVTA portfolio, within the DHS S&T directorate is to detect elusive indicators of threat, assess adversary capabilities, understand adversary motives and behaviors, and relate threats with vulnerabilities to evaluate risk. Analysts will not be able to accomplish this task without the aid of more sophisticated, automatic tools for data analysis.

Knowledge management is an important component of the TVTA portfolio, and the ADVISE program is focused on providing a common platform that supports scalable knowledge management across multiple missions. The first section of this chapter (Section III.A) discusses ADVISE as well as a project that is built on top of it: the Biodefense Knowledge Center (BKC). Further, this section describes the Information Sharing and Collaboration (ISC) office, which DHS created to facilitate the sharing of data with all stakeholders. The ISC coordinates the data that will be fed into systems such as ADVISE.

TVTA is not the only portfolio in S&T facing data sciences challenges. The second section of this chapter (Section III.B) discusses the needs of the Radiological and Nuclear Countermeasures portfolio that is doing real-time sensor data processing to, for example, detect radiation signatures of nuclear materials. The recommendations of this report will not focus on the data sciences of this type of data processing; however, the consideration of how to integrate this data with auxiliary threat information is discussed.

For example, a radiation detector at a Canadian border crossing may pick up an anomalous reading that might be too ambiguous to trigger an alarm, but the incorporation of additional data (e.g., the driver is associated with a group known to be collecting nuclear materials or the same anomalous reading appears every week from the same driver and truck) would greatly improve the threat detection ability of these systems.

Figure 1: Knowledge Management plays a critical role in the TVTA's mission. It is used for management and planning for threat preparation, anticipation, prevention, detection, and restoration.[11]

[11] From the workshop presentation of Dr. Joseph Kielman, DHS TVTA portfolio manager.

A. Threat and Vulnerability, Testing and Assessment Portfolio

The TVTA portfolio has a broad charter to help prepare the nation against terrorism by assessing threats and vulnerabilities. In his opening remarks at the workshop, Dr. Joseph Kielman, Portfolio Manager for TVTA, summarized the TVTA mission:

> *"Through science and technology, develop capabilities that enable the creation, application, and dissemination of knowledge to prepare for, anticipate, prevent, and detect terrorist activities and, if necessary, restore the nation's operational capabilities."*

The TVTA portfolio is sponsoring research and activities in the following three areas: knowledge management technology, social and behavioral sciences, and intelligence and specialized information (e.g., WMD and nuclear capability assessments).

Knowledge management plays a critical role in TVTA's mission. It is used in management and planning for threat preparation, anticipation, prevention, detection, and restoration. See Figure 1.

TVTA's goal is to develop a scalable, all-source knowledge management system to discover terrorist capabilities and threats. This system must process and integrate multiple types of data, ranging from news feeds, to raw sensors, to intelligence reports, and more. Multiple types of analyses must be supported. The system must also have the capability to share the analysis and the associated data with appropriate with federal, state, and local officials. The cycle that is envisioned is illustrated in Figure 2.

Within the domain of Knowledge Management Technology, TVTA is investing in fundamental technologies, new capabilities, and integration activities, as well as several pilot programs. ADVISE, discussed in Section III.A.1, is a thrust area that has been developed to support the full range of information fusion needs of the DHS. The BKC is being built on top of the infrastructure developed by ADVISE, as discussed in Section III.A.2. The ISC, on the other hand, is a sort of "human information fusion" program that brings together experts from across the DHS and elsewhere to facilitate information sharing; this is discussed in Section III.A.3.

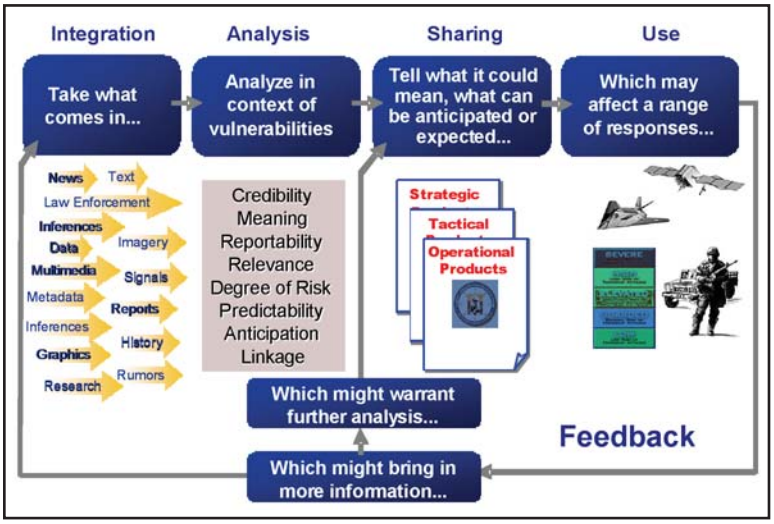


Figure 2: The TVTA information cycle includes integration of data from various sources, analysis of the data, sharing with federal, state, and local officials, and using the data to predict threat levels and so on.[11]

1. Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement

Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE) is a system that is under "spiral" development (meaning that it is being deployed simultaneously with development) and will provide a common platform that supports scalable knowledge management across multiple missions; see Figure 3 for an illustration that shows the overall architecture.

The system includes tools for ingesting and canonicalizing massive quantities of information from many different sources, as shown at the bottom of the figure. Some of the data comes from other databases, as indicated by the green cylinders. Other data comes from free-form text document sources that must be processed to discover the entities and their relationships. Automatic tools for event extraction are used for some reports but are not yet very good. ("Louisiana" is a system for interfacing to various extraction tools.) Manual extraction (aided by the "SPUD" tool) is still necessary for critical documents.

At ADVISE's core, semantic graphs are used to organize the data entities and their relationships. (The graph system is called "Nebraska.")

A semantic graph organizes relational data by using nodes to represent entities and edges to connect related entities. Hidden relationships in the data are uncovered by examining the structure and properties of the semantic graph. Privacy and support policies are enforced by a security infrastructure. Several interfaces for browsing, querying, and viewing the results of queries are under development, including IN-SPIRE and Starlight, from the DHS NationalVisualization and Analytics Center (NVAC).

The key to fusing disparate data from many sources in ADVISE is the exploitation of "pre-computed" relationship information by storing the data in a semantic graph. All nodes are related by the links between them on the graph. For example, Figure 4 shows a simple semantic graph that links people (black nodes), workplaces (red nodes), and towns (blue nodes). The different link (or edge) types indicate different relationship types.

For example, the fact that Person 13 and Person 15 have a green link between them indicates that they are friends with one another, while the orange link from Workplace 19 to Town 22 indicates that Workplace 19 is located in Town 22. In this example, the links are all bidirectional, but directed links can also be used.

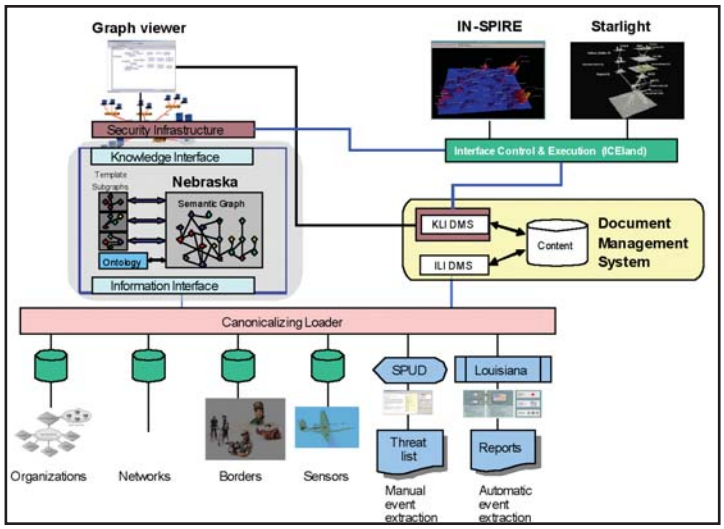


Figure 3: The ADVISE architecture incorporates many different data sources and stores the data in a semantic graph. The sources of the data can be retrieved from the document management system. The data can also be visualized using tools such as IN-SPIRE and Starlight. [11]

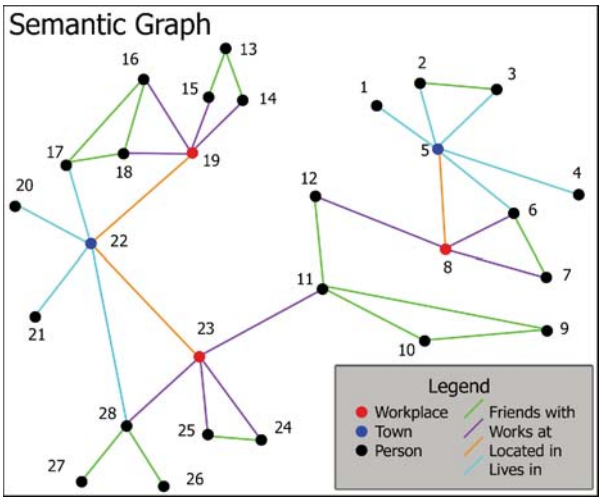


Figure 4: An example of a semantic graph with three nodes types and four edge types.

Confidences (or uncertainties) are attributes of both the nodes and edges. Studying such graphs can help in understanding the relationships between entities (e.g., what's the shortest path between Persons 16 and 26?) and in making intelligent hypotheses (e.g., Persons 15 and 14 are linked by a common workplace and a common friend, so we may hypothesize that there is a good chance that they should also be connected by a "Friends with" link).

Advanced, long-range R&D is needed to support ADVISE, including research on automatic processing of text documents, semantic graph representation, querying on semantic graphs, automatic knowledge discovery on semantic graphs, and methods for ensuring privacy and security. These topics are discussed in more detail and generality in Chapter I.

Several systems are built on top of the ADVISE architecture (see Figure 5), including the Threat Vulnerability Information System (TVIS) for the Information Analysis (IA) organization, the Regional Threat Analysis System (RTAS) for Border and Transportation Security (BTS), and the Biodefense Knowledge Center (BKC) for the National Biodefense Analysis and Countermeasures Center (NBACC).

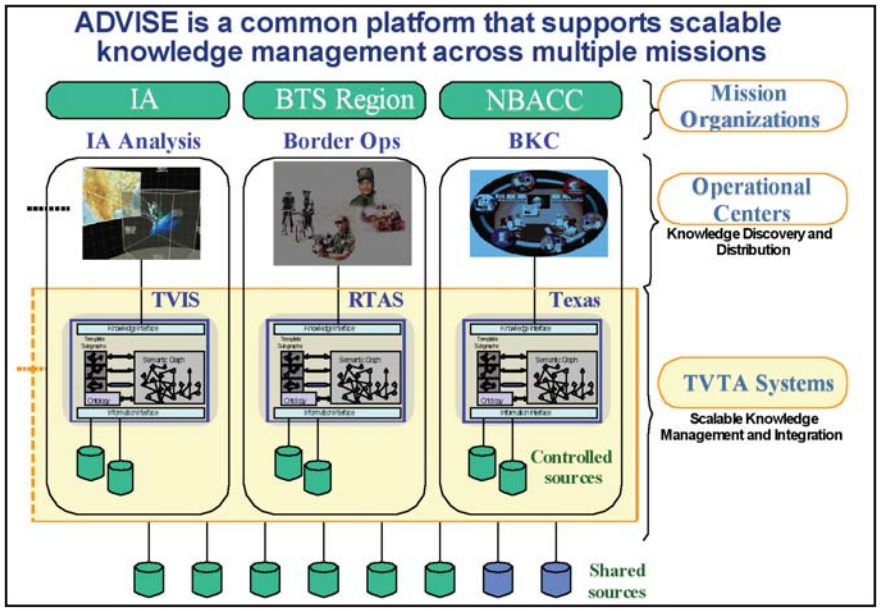


Figure 5: ADVISE will be the underlying architecture for applications in Information Analysis, Border Operations, and the Biodefense Knowledge Center. [11]

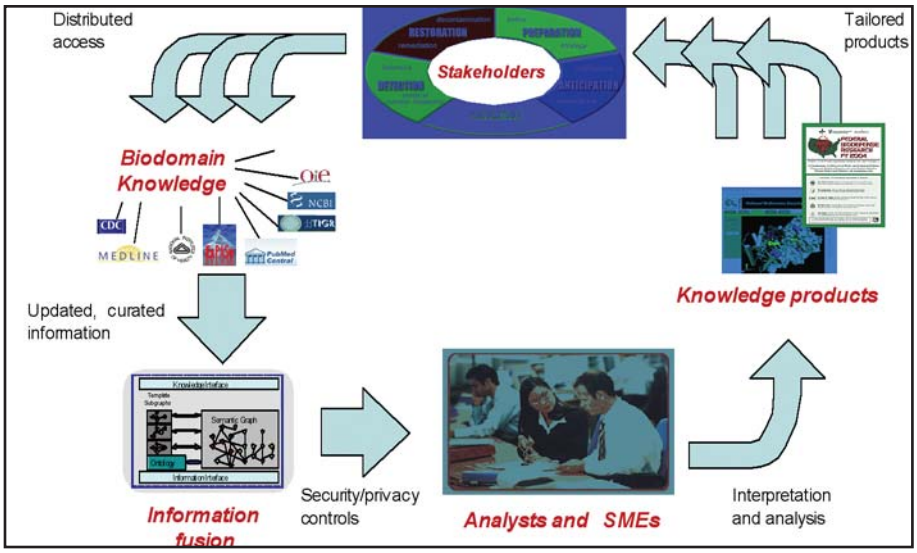


Figure 6: The BKC will connect analysts, researchers, and stakeholders within a trusted information network.[12]

2. Biodefense Knowledge Center

The Biodefense Knowledge Center (BKC) provides an interesting example of a specific use of the ADVISE system. The goal of the BKC is to create an overarching architecture that integrates disparate components in order to anticipate, prepare for, prevent, detect, respond to, and attribute biological threats. The BKC system will address the need for a trusted information sharing and analysis system for biodefense stakeholders. Current architectures for biodefense are limited by a number of concerns ranging from a lack of appropriate security measures to a lack of scalability to a lack of shareability.

The BKC will provide access across different missions and stakeholders and provide critical links to subject matter experts and curated, continually updated information; see Figure 6.

For the BKC, the semantic graph (see Figure 7) will contain information for a variety of data sources, ranging from intelligence data to basic biological information (such as genomic data) to data on virus outbreaks and vaccine stockpiles.

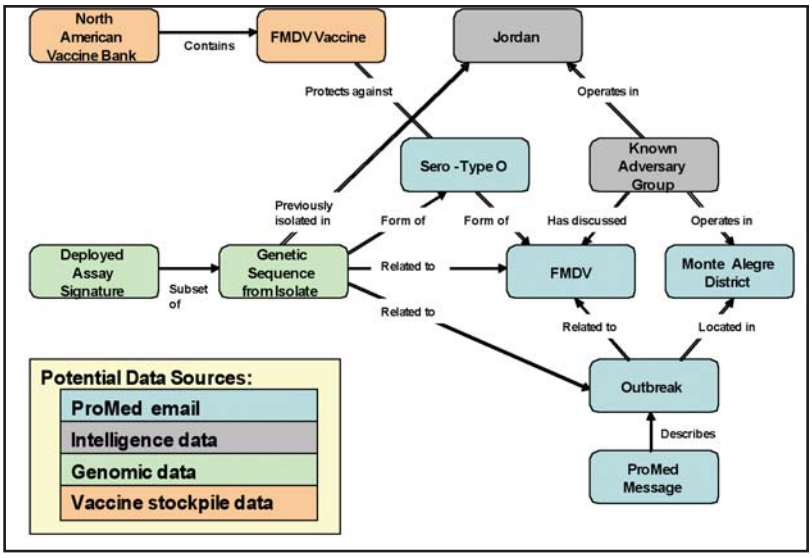


Figure 7: The BKC will fuse data from disparate sources to answer questions regarding biological threats.[12]

[12] From the workshop presentation of Tom Slezak, Lawrence Livermore National Laboratory

3. Information Sharing and Collaboration

The needs of the Information Sharing and Collaboration (ISC) Office provide a broader perspective on the needs of the DHS.

The ISC was created to address the stove piping of data sources across DHS. The mission of the ISC Program is to coordinate and facilitate efforts to enable information sharing throughout the DHS and with its customers and partners, especially the federal, state, and local governments as well as the private sector. The goal is to achieve an interoperable system-of-systems (including systems such as ADVISE) across the DHS enterprise that will facilitate the sharing of information with all stakeholders in a timely and effective manner appropriate to the mission needs of DHS and its partners.

The desired end state is an integrated system where any employee of DHS can log in from any location and have full access (as appropriate to his or her access permissions). The aggregated whole will be independent of any particular facility or system. All federal, state, local, and private sector security entities will be able to share and collaborate in real time with distributed data warehouses that will provide full support for analysis and action. The ISC blueprint incorporates a "privacy and policy" piece that is a key to combining privacy and policy down to the lowest levels of information. Policy-based filtering will provide role-based access to data integrated into the system.

All data will be collected into *one* shared analytic space. The data itself will be distributed across multiple collections (1300 distributed sources), but all data will potentially be available to all analysts. The anticipated data sources are enormous in size and number and come from multiple agencies.

The data will be from both structured and unstructured sources. Entity and event extraction will be required to capture and expose the "who, what, and when" information from the data sources. An illustration of the different data sources and the shared analytic space is shown in Figure 8.

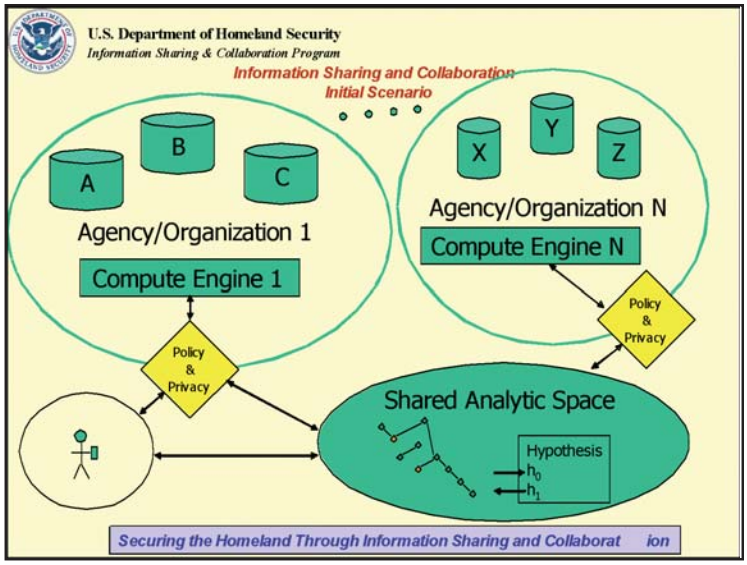


Figure 8: The DHS Information Sharing and Collaboration office envisions a shared analytic space that integrates data from multiple domains.[13]

The ISC has identified the following technical goals for its program, which are very similar to the goals for ADVISE:

- High-speed information extraction of entities, attributes, and relations, including tracking entities over space and time and identification of "protected" entities
- Detection of novel and relevant events
- Detection of meaningful groups against models of behavior that are "of interest"
- Graph algorithms for information retrieval on large-scale, distributed semantic data
- Automated pattern learning such as learning expert examples of threat and non-threat patterns

[13] From the workshop presentation of Stephen Dennis, DHS ISC program deputy director for research and engineering.

B. An example of the needs of other S&T portfolios: Radiological and Nuclear Countermeasures

DHS collects radiological and nuclear (RN) data and associated metadata at border checkpoints for real-time decision making to decide whether or not a vehicle crossing into the U.S. is carrying illegal RN materials. For example, Figure 9 shows a truck passing through a new DHS radiation portal. The collection, processing, use, and storage of this data give a sense of the diverse needs of the DHS.

For each vehicle, RN data is recorded by commercial devices and is stored in many different non-standard formats. For example, the data can be in the form of radiation counts, captured full or partial energy spectra, or identified isotopes. The data may be stored as simple binary information, formatted text, unformatted text (which is human-readable, but difficult to automatically parse), "blobs" (binary large objects, e.g., Zip file archives), or a set of multiple structured files (e.g., Microsoft Access database files). RN data also has been obtained at testing facilities under controlled conditions for analysis and studies. However, RN data by itself may be insufficient to determine whether or not a vehicle is carrying illegal RN materials.

Metadata on each vehicle is recorded as well, including information about the vehicle (manufacturer, model, color, and weight), the cargo (type, quantity, origin, and destination), personal information about the vehicle's occupants (names, identification information), and environmental data (date and time, weather, and background measurements). This metadata is entered in various ways such as free-form text or filled-in templates that are completed by the customs agent and/or electronic information from other sources. Still images or streaming video of the vehicles may also be available.

As each vehicle crosses, the DHS agent must make a real-time decision because commerce at ports of entry, in general, cannot be interrupted or delayed beyond the few seconds normally allowed to process a vehicle. The RN readings are often difficult for non-experts to interpret, so additional information (i.e., historical RN readings as well as the metadata) could be useful in making determinations on what to do. Some vehicles are stopped for further inspection due to suspect RN data, but those vehicles cannot be held for much more than an hour.

Figure 9: A truck passes through a radiation portal at a port in Newark, NJ. These portals collect data that is used to determine whether the vehicle is carrying illegal RN materials.[14]

In fact, all the data is stored long-term for *a posteriori* data analysis, detector studies, and algorithm development and validation. The amount of data is enormous. There are gigabytes to terabytes of data per database because the data is collected continuously at border points. Unfortunately, at this stage, even straightforward operations on these data are tedious.

[14] Photo Credit: Gerald L. Nino. From "Scanning for Nukes," May 2004, at http://www.cbp.gov/xp/CustomsToday/2004/May/nukeScanner.xml.

For example, a query such as "get all information from March 2nd" is difficult (or even impossible) to execute. On the other hand, answers are typically needed in a short amount of time (less than 1 hour) due to the impact that stopping a vehicle on the border has on transportation and commerce.

Data stored by non-DHS agencies (such as data obtained at a truck weighing station) could be useful in making decisions. However, it is currently difficult to obtain such data because it is owned by state agencies. By and large, the agencies involved are cooperative, but there are no tools available to streamline integration and no resources available to provide these tools.

Finally, the current privacy and security controls to historical data are rudimentary and a barrier to the analysis that needs to be performed. Access is typically granted at the database level via, e.g., password access to the host computer. There are no provisions to filter the data or enable selective access. Because the data contains personal information (e.g., information about each driver) and trade secrets (shipment manifests), it cannot be easily shared. The ability to support variable access permissions depending on the user's need to know is important.

IV. Research and Development Goals for Information Management and Knowledge Discovery

Semantic graphs play an important role in the knowledge management capabilities of the TVTA portfolio within the DHS S&T directorate, as discussed in Chapter I. Advanced, long-range R&D on semantic graph technologies is needed to support the data sciences needs of the TVTA, especially for thrust areas such as ADVISE (see Section III.A.1). The volume of data that must be analyzed is enormous, and the TVTA is responsible for providing the tools for evaluating that data. Current technologies are inadequate for the task. The technologies that exist must be reengineered in order to handle the massive amounts of data that need to be evaluated, and new technologies must be developed to aid in advanced knowledge discovery.

Advances in the following R&D areas will provide great improvements to the DHS's capabilities in fulfilling its mission to prevent terrorist attacks and reduce the nation's vulnerability to terrorism:

A. **Architecture and management of databases for large-scale semantic graphs,** including issues associated with distributed databases, ingestion of large volumes of data from both structured and unstructured sources, and integration of data from different sources with different representations. See Section IV.A, below.

B. **Scalable algorithms and interfaces for information retrieval and analysis on semantic graphs,** such as scalable algorithms for the discovery of complex relationships between nodes, an efficient query language for semantic graphs, scalable visualization tools and intuitive user interfaces, and the integration of simulation results. See Section IV.B, below.

C. **Models for detection and prediction on semantic graphs**, including detection of missing or incorrect data, statistical prediction of attributes and links, identification of anomalous nodes or relationships, and models that incorporate temporal or spatial effects. See Section IV.C, below.

D. **Models for discovering and detecting processes on graphs**, such as statistical and machine learning models for determining organization structure, finding portions of the graph that are undergoing abrupt changes, and using processes to aid in analysis. See Section IV.D, below.

E. **Algorithms to provably ensure privacy and security**, including the development of policies that account for multiple levels of trust and access, new anonymization methods with provable privacy guarantees, and development of models that allow for trade-offs between privacy and national security. See Section IV.E, below.

Note that these focus areas do not present a complete picture of the R&D needs of the DHS. For example, little emphasis is placed on the processing of multimedia data although such processing will be required in numerous contexts ranging from the "US Visit" program data to aerial surveillance photos and more. Thus, while the inclusion of a particular focus area is an indication that the area is of critical importance to the DHS, its absence should *not* be construed as indicating that it is unimportant.

Critical measures to bring excellence to the R&D process are outlined in Chapter V.

A. Architecture and management of databases for large-scale semantic graphs

The amount of data to be processed by the DHS is staggering. Some estimates put the total number of entities to be stored at 10^{15} or higher. The Information Sharing and Collaboration (ISC) office (see Section III.A.3) envisions four primary functions that its systems must provide to customers. First, users must be able to **read** source data, either through interactive queries or through an application programming interface (API), using the source structure provided by the original supplier. Second, users must be able to **search** and retrieve information from all integrated repositories. Third, there must be a capability to **import** sources by re-hosting supplier's data within the DHS information space. Fourth, users must be able to conduct **link** analysis on structured information gleaned from sources, even if sources themselves are unstructured. These four requirements depend on the architecture and management of the databases.

Given the unprecedented amount of data that is to be stored, designing a suitable architecture and management system for the databases is of major importance. R&D is needed in the following areas, ordered by importance, and described in detail in the subsections that follow.

1. **Methods for ingesting large-scale data streams,** including filtering and entity, relationship, and event extraction. The automatic tools of today are insufficient, achieving low levels of precision-recall on complex tasks such as co-reference determination and event extraction. Algorithm developments that improve the level of accuracy on these tasks are desperately needed.

2. **The identification of appropriate databases for storing and querying** on large-scale, distributed semantic graphs. There are actually two issues here. The first is which type of database organization is most appropriate for representing the semantic graph. The second is how to best coordinate a collection of distributed databases for access by graph queries.

3. **Methods for integrating databases that have different schemas and ontologies,** as well as methods for automatically identifying subsets of databases that should be integrated.

1. Ingesting large-scale data streams

Data ingestion is the process of taking raw data and converting it to a format appropriate for storage in a database. In his presentation, Stephen Dennis, DHS ISC Deputy Director of Research and Engineering, said that the ISC expects to process one billion structured and one million unstructured text-formatted messages per hour. All of those who were already working with DHS identified the problem of automatic text processing as being particularly difficult. Processing such data is still labor intensive, as evidenced by the "manual event extraction" component in the architecture of ADVISE (see Figure 3). The development of more effective automatic tools that combine filtering with entity, relationship, and event extraction is critical for populating the DHS's databases.

Figure 10 illustrates the process that is needed to go from raw unstructured data to usable information. The first step (not shown) is to translate various types of documents (e.g., email, text messages, Word files, PDF files) into text documents, noting that it is critical to maintain any useful structured information such as document author, date, title, etc. Next, irrelevant documents are filtered to greatly reduce the number of documents that are sent for further processing and help to reduce the overall size of the database. Then the documents must be processed to discover information: what entities are named, how they are related, and what events have taken place. The next steps can be separate or combined, depending on the techniques that are employed. An example of entity and relationship extraction is shown in Figure 11.

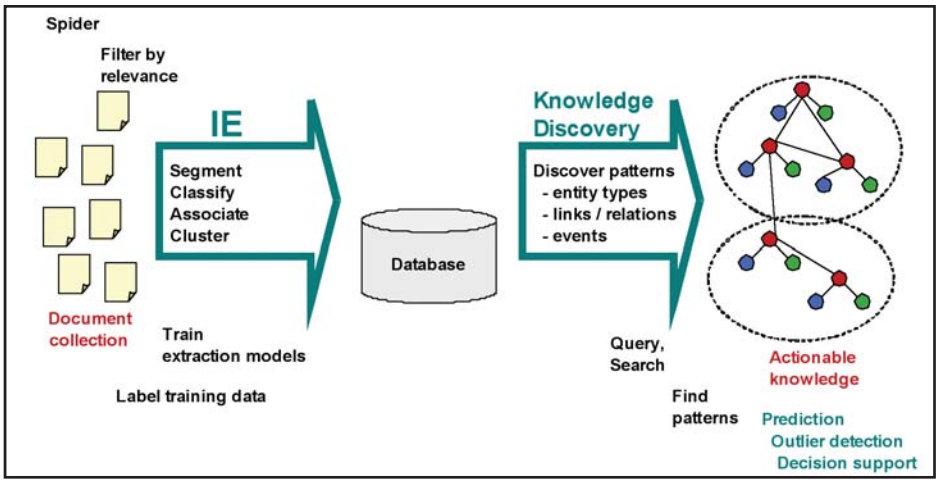


Figure 10: The process of going from raw data to a semantic graph includes filtering, information extraction (IE), and knowledge discovery.[15]

Prof. Andrew McCallum of the University of Massachusetts cited the performance of the state-of-the-art tools to be as follows.

- Named entity recognition (i.e., identifying a person, location, or organization) had a recall-precision rate of between 80-95%.

- Binary relationship extraction (i.e., determining Location 1 is "contained in" Location 2 or Person 1 is a "member of" Organization 1) had a recall-precision rate of between 60-80%, depending on the type of relationship.

These accuracies should be much closer to 100%. Inaccuracies at this early stage of processing should be as minimal as possible.

Another closely related topic is the development of methods for assigning levels of confidence. For example, data from controlled sources such as intelligence feeds may be viewed as more reliable than uncontrolled sources. For example, data from CNN's Web site is generally viewed as more reliable than data from a college student's blog (i.e., a Web log that contains periodic, reverse chronologically ordered posts on a common Web page.). Furthermore, the processing of the data may introduce uncertainties since relationships may be detected incorrectly.

Disambiguation and duplicate detection may also happen during the ingestion phases and have a large impact on later processing. For example, would two reports of the same event be stored as two events or one event with two reports? And what if those two reports are actually the *same* report, e.g., the same news story appearing in two newspapers?

All the tools that are developed for ingestion must be able to scale to the needs of DHS. Some phases of the processing, such as entity extraction, can be done in an embarrassingly parallel fashion because they only involve an individual information unit, i.e., one document. More sophisticated levels of analysis, such as duplicate detection, require the processing of multiple sources of data simultaneously, making issues of scaling more challenging.

[15] From the workshop presentation of Prof. Andrew McCallum, University of Massachusetts.

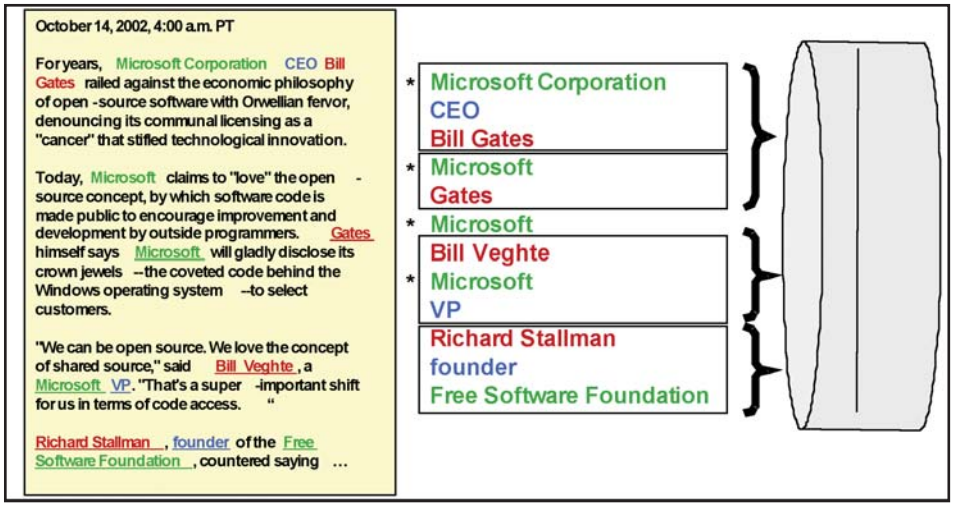


Figure 11: Information extraction is used to identify entities (.e.g, "Bill Gates"), determine the entity type (e.g., "name"), and discover relationships (e.g., "Bills Gates is the CEO of Microsoft."[15]

2. Appropriate databases for storing and querying large-scale, distributed semantic graphs

Once the data is ingested, the issue becomes one of storage. Stephen Dennis, DHS ISC Deputy Director of Research and Engineering, cited the importance of the identification of new architectures for efficient management of distributed knowledge and information. There are actually two major issues here. The first question is which type of database is most appropriate for storing semantic graphs—possible choices include relational, object-oriented, or vertical. The second issue is how to best implement a distributed semantic graph database.

Relational databases are popular because of their efficiency if the proper indexing is used. However, relational databases are not appropriate for every situation, and early evidence suggests that they may not be the most appropriate device for efficiently storing semantic graphs. Recall that a semantic graph looks something like what is seen in Figure 12. There is a collection of nodes connected by a collection of links. Each node and link has a type (e.g., a "Paper" node or an "Author" node) and possibly, though not always, some additional attributes.

An ontology specifies the rules for the types of nodes that are allowed and the types of links that are allowed between different specific node types. The ontology for a semantic graph is analogous to the database schema that defines the structure of a database.

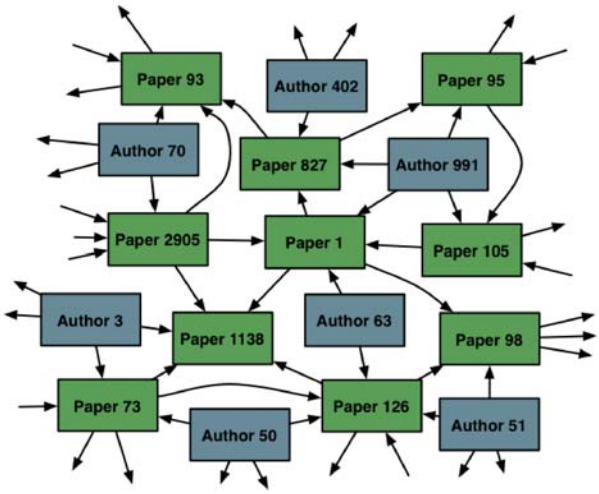


Figure 12: A sample semantic graph that shows papers, the authors of the papers, and the citations between papers.[16]

[16] From the workshop presentation of Prof. David Jensen, University of Massachusetts.

One of the major challenges of database (or semantic graph) integration is how to merge the schemas (or ontologies). Figure 13 shows an example of an ontology for a citation database. The ontology says that there are four types of nodes (Journal, Paper, Author, and Domain) and five types of links (Published in, Authored, Citation, Co-Authored, and Affiliation). Furthermore, the ontology specifies the rules on which types of links can go between which types of nodes. An "Authored" link can only go from an "Author" node to a "Paper" node.

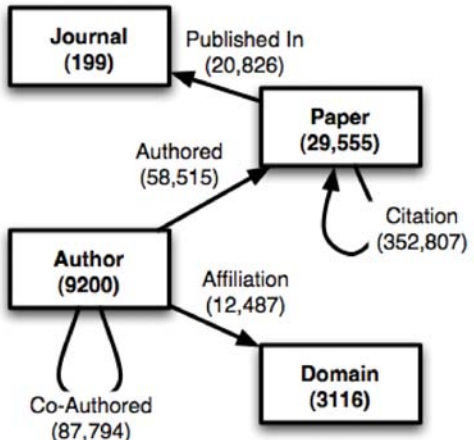


Figure 13: An example of an ontology that specifies the rules for the types of nodes and the links that are allowed between them.[16]

In his presentation, Prof. David Jensen of the University of Massachusetts observed that traditional relational databases are optimized for row-wise access with a fixed schema, whereas vertical databases are optimized for column-wise access with no fixed schema and so are potentially better suited for storing semantic graphs. In his group's experiments, switching from a relational to a vertical database resulted in a 20 times improvement in query speed *in addition to* simplifying code development. A better understanding is needed of when (partial) schema information can be exploited and when a representation which makes minimal ontological commitment is preferred.

Whichever type of database is used, there will be far too much data to store on a single computer system. The question is how should the data best be distributed across multiple systems in order to support efficient ingestion and queries.

In answering the above questions, the following related concerns must also be taken into account.

- The hardware for the distributed systems.
- Tracking the sources of all data and propagating changes when data is changed or deemed incorrect.
- Restricting access to the data to authorized users. The practice of controlling data through compartmentalization and classification is well known. Another common process is to restrict access to data to the organization that collects it. For example, data collected by one branch of DHS, such as immigration, may not be readily shared with another branch due to privacy and, more generally, data control issues. Several DHS personnel cited a need for data storehouses that enforce policies on accessing the data. (See Section IV.E for more on security and privacy issues.)
- Creating an audit trail by tracking all accesses to the data. (In fact, the size of the audit data could become larger than the original data itself.)
- The development of efficient methods for "retiring data." Research must solve the problem of maintaining awareness and factoring in the applicability of information by its period of germaneness.
- Archiving data, especially since the original sources may be altered or become unavailable. As an example, see the Internet Archive project at http://www.archive.org.
- The ability to retrieve historical data such as a snapshot-in-time. The Internet Archive has, for example, the "Wayback Machine," which enables users to pull up past versions of any Internet Web page (http://www.archive.org/web/web.php).

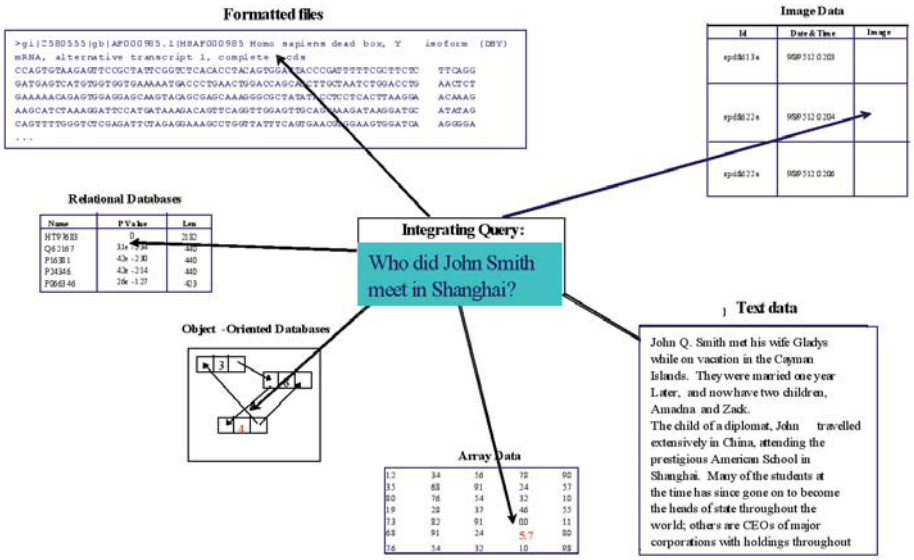


Figure 14: The diversity of data that might be needed to answer a single query includes formatted files, raw text files, image data, and various other databases.[17]

3. Data integration

Data integration is an important problem to DHS. Much of the data needed to address the DHS mission is distributed across multiple agencies (e.g., CDC, FBI, CIA, local law enforcement agencies, open source public data sets), is in multiple formats (stored in flat files, XML, HTML pages, and different database management systems using different schemas), and is multi-modal (e.g., unstructured text, video, speech, call records); see Figure 14. The data is distributed at multiple levels of granularity (e.g., hourly vs. weekly, city vs.state, gene vs. disease), confidence (some more trustworthy than others), classification, and privacy sensitivity (e.g., citizen vs. non-citizen data, open source vs. private data).

Most of the DHS knowledge discovery applications require the fusion of multiple different kinds of data. Some data is already in structured databases, but other data consists of unstructured text, images, transactions, relationships, biological strings, spectra, etc. Diversity of data is not unique to DHS, and special purpose solutions for data integration can be found in a variety of commercial and academic settings. However, the breadth of DHS data types, the dispersed nature of the data repositories, the highly dynamic nature of the data, and the mixture of classification and privacy concerns make the DHS data integration problems particularly difficult.

There are three general techniques to performing data integration:

- Moving the data to a central location (such as data warehouses or data marts)
- Accessing distributed data sources dynamically (such as in federated databases or a multi-database infrastructure)
- Peer-to-peer data management.

Each of these approaches has benefits and shortcomings. For example, the centralized data management approach can lead to very large databases, which can be difficult to maintain in the face of continually evolving data. Similarly, distributed data approaches have issues with security and reliable access to the remote sites. As a result, hybrid systems are evolving that combine desired characteristics from each of these basic approaches.

[17] From the workshop presentation of Prof. Susan Davidson, University of Pennsylvania.

Virtually every database or semantic graph has its own schema or ontology, and the development of methods for integrating data with different ontologies is an important need. The ability to learn schema mappings and transfer mappings from one related domain to another is essential to achieve the scaling factors required by DHS.

Most current data integration systems do not fully support multi-modal data integration, e.g., a single query that should combine information from a relational database, a collection of free text documents, and a set of images. While some primitive functionality exists, the capabilities of these systems lag far behind those of tools developed specifically for a single modality, and this capability gap must be reduced.

Since much of the information that is being processed may be incorrect in some way or another, integration provides potential to better understand the certainty of the information. For example, confidence in a "fact" may become higher if that fact appears in multiple databases. Of course, such a simple rule can be misleading if all databases have used the same source as a basis for the fact. Thus, it is important to track the provenance of the data, i.e., where it originated, and use that information during the integration process so that the chance of over-valuing information from a single source is reduced.

Another issue in integration is disambiguation, i.e., deciding if two entities are the same. Most techniques that exist today are fairly limited in their abilities because they do not take all the relational information into account. The development of new techniques in this regard will help both in the integration of existing databases as well as in the integration of data into an existing database.

B. Scalable algorithms and interfaces for information retrieval and analysis on semantic graphs

Semantic graphs are graphical structures that display relationships between entities. In particular, a semantic graph consists of nodes (i.e., entities) and links (i.e., relationships). See Figure 4 and Figure 12 for examples. Each node and link is categorized to be of a particular type and may additionally be characterized by different sets of attributes. For example, a "Person" node could have a "Last Name" and "Zip Code" attributes. Many real-world semantic graphs that DHS would use have the following properties.

- They are large-scale with billions of nodes and links and thousands of node and link types.
- They may have "influential nodes," i.e., nodes that are closely connected to many other nodes.
- They encode (sometimes unknown) community structures.
- They are stored in physically distributed systems, mainly due to their massive size.
- They are noisy and incomplete.
- They may have weights on the nodes and links that measure uncertainty (or alternatively confidence) associated with the facts and relations.
- They are dynamic, changing as new facts and relations become known and old facts and relations become irrelevant.

Developing *scalable* algorithms and interfaces for information retrieval on semantic graph data is important for supporting DHS architectures such as ADVISE (see Section III.A.1) that are based on semantic graphs. In this case, scalable means that the algorithms can be applied to graphs with billions of nodes and links. In his workshop presentation, Prof. David Jensen of the University

of Massachusetts said, "Existing commercial tools are well-developed, widely deployed, and nearly useless for analyzing semantic graphs." Thus, development of tools for analyzing semantic graphs is critical. The most important R&D areas are the following, listed in order of importance.

1. **Scalable algorithms for relationship analysis,** such as solving shortest path problems. Relationship analysis algorithms are fundamental tools for analysts but still require substantial computational time because of the difficulty of adapting these algorithms to large-scale semantic graphs.

2. **Development of scalable and intuitive user interfaces** is critical for allowing the analyst to use the data to "connect the dots." This means not only making it easy for the user to express queries, but presenting results in a manner that can be easily understood.

3. **Connecting modeling and simulation** to use the data obtained from the semantic graphs.

1. Scalable algorithms for relationship analysis

Developing scalable algorithms for relationship analysis on semantic graphs was identified as an important need by Dr. Joseph Kielman, DHS TVTA portfolio manager, in his opening remarks as well as in the presentation on ADVISE. Graph algorithms need to be extended to work on large-scale graphs where even algorithms that execute in linear time in the number of nodes may be too expensive. Many traditional graph algorithms are not only not linear, they are NP-hard (e.g., subgraph isomorphism, maximum and maximal cliques).

In many cases, algorithms that give approximate answers need to be devised. Entirely novel graph algorithms need to be developed that are appropriate for the types of analysis that will be done on semantic graphs, such as algorithms that find the web of connections between two people.

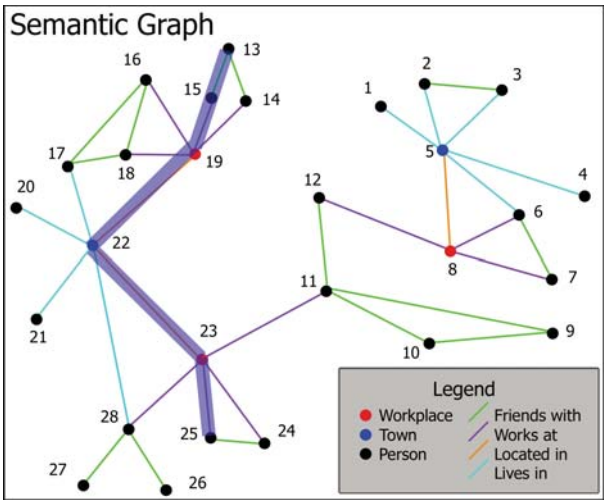


Figure 15: The shortest path between two people is one way of understanding their relationship. The figure shows the shortest path from Person 13 to Person 25.

Consider the shortest path problem. The goal is to find the shortest path between two nodes. Figure 15 shows an example of the shortest path between Person 13 and Person 25. The shortest path between two entities can be used to try and understand their relationship. The shortest path problem is well studied on smaller graphs without semantic structure.

However, much work remains to be done on adapting shortest path algorithms to semantic graphs. For example, it may be possible to exploit ontological information. Alternatively, different types of nodes and edge types should be treated differently. For example, it may be that the analyst wishes to ignore all "Located in" edges in computing the shortest path shown in Figure 15, which would substantially change the solution to the problem. Or, different edge and nodes types may be weighted differently to reflect the relative importance of each type of connection. It is important to note that even the shortest path algorithms that exist in the (non-semantic) graph case do not yet scale adequately to very large graphs.

Shortest paths are just one way to analyze a relationship, but shortest path and its ilk are not the most interesting or relevant methods to understand the broader connections between

entities. The problem with shortest path is obvious: Two individuals may be related by a single commonality (e.g., a common acquaintance) but have no deeper relationship. One answer is to look at all short paths (where "short" is defined by a pre-specified maximum length) between two nodes, which has been identified as a specific goal of the ADVISE program.

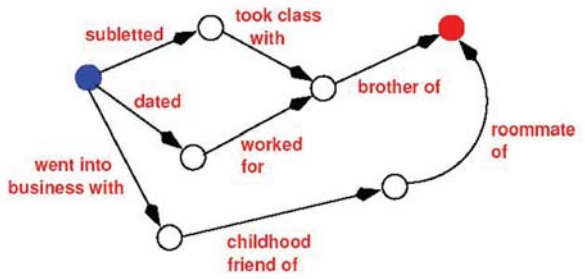


Figure 16: A connection subgraph is an alternative way of showing the connection between two people by showing multiple paths.[18]

Researchers at IBM Almaden Research Center have proposed a more general approach, called the "connection subgraph."[19] Here the idea is to find a graph with some maximum number of vertices (as opposed to edges) that shows the various relationships between two nodes. Figure 16 shows an example of what might be returned by a connection subgraph; the red and blue nodes are not directly connected but are connected by three relatively short and significant paths.

2. Scalable and intuitive user interfaces for querying and browsing semantic graphs

Although semantic graphs are potentially a powerful tool for aiding in decision analysis, the analysts must be able to express their queries and view the results in some sort of intuitive and meaningful fashion. Workshop participants noted that effective query interfaces—both manual and automatic—are crucial for achieving customer adoption. Deep integration of researchers into the customer environment was cited as one way to develop a customer-friendly interface.

Though the development of an infrastructure for querying on semantic graphs is still in its infancy and needs further investment, Prof. David Jensen of the University of Massachusetts discussed the idea of a visual language for querying graphs and cited his group's work on QGraph as an example. Figure 17 shows an example of a visual query on a graph.

Several workshop participants noted that visualization of results, particularly results deriving from graphs, is problematic. Often a query amounts to the selection of one or more subgraphs. However, when the graphs are massive, the subgraphs that are returned by a query can themselves be extraordinarily large, making it difficult to know how to best refine further analysis. The proper tools for exploring relatively large graphs, such as those returned from a query, are critical for DHS analysts.

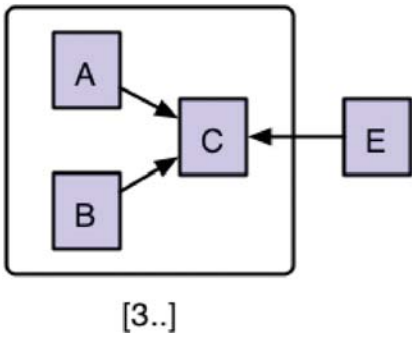


Figure 17: Visual queries may be the optimal way for analysts to work with semantic graphs. This figure shows an example of a visual query on a graph.[16]

Any query infrastructure that is designed, be it visual or otherwise, must be efficient to be adopted by users. This means that it should translate the queries into efficient operations and include built-in support for common queries such as shortest path. Query response speed is extremely important for customer adoption.

[18] From the workshop presentation of Prof. Jon Kleinberg, Cornell University.

[19] See, for example, Faloutsos, McCurley, and Tomkins, Connection subgraphs on IBM WebFountain Data, 2004.

3. Connecting modeling and simulation

Information from models and simulations can also be quite useful to an analyst. For example, a simulation of the effect of a chemical release in an air duct may help in planning for evacuations. The data from modeling and simulations, however, is generally stored in table form rather than as a semantic graph. For example, the raw RN sensor data discussed in Section III.B is more appropriately represented in standard databases. There are several scenarios in which it may be necessary to integrate relational data from a semantic graph and non-relational data from modeling and simulation.

- Modeling and simulation results can be used in the analysis of patterns. For example, it might be worthwhile to compare the results of the RN scan of a truck at a border crossing with known RN scans or prior history. The results of these analyses need to be combined with the relational data to provide a complete picture.

- Modeling and simulation may use data from the semantic graph as input data. For instance, if certain chemicals are purchased by known terrorists, simulations may help determine the level of threat posed by the purchased chemicals. For example, the analysts may simulate a chemical weapon in a subway system.

4. Crosscutting issues: scalability, accounting for measures of uncertainty, and including temporal or spatial phenomena

In addressing the challenges outlined above, several other issues should be considered.

- **Scalability –** Scaling to the enormous data sets of interest to DHS.

- **Uncertainty –** Methods for handling uncertainty should be incorporated into the algorithms and visualization tools for the results should have intuitive mechanisms for displaying confidence. Furthermore, sources of variability and bias (e.g., sample selection) in the data should be accounted for.

- **Temporal and spatial information cannot be ignored.** Such information can be used in higher-level analysis; see Section IV.D.

- **Compression of information –** In a semantic graph, what is the trade off between the richness of information and the ingestion of a large number of facts? It was estimated that the number of facts that needs to be ingested is of the order of 1015. Methods to reliably estimate the intrinsic amount of information in a graph are needed, especially if those methods can help to compress or prune the large-scale graphs that DHS is working with.

C. Models for detection and prediction on semantic graphs

In the previous section, the focus was on basic scalable algorithms and tools for semantic graphs. However, straightforward queries alone are not sufficient to analyze large-scale graphs. For example, relationship analysis only helps when the analysts know which entities to focus on. In this section, the emphasis is on tools that enable more automatic knowledge discovery, enabling detection and prediction. In order of importance, the focus areas are listed below and discussed in greater detail in the following subsections.

1. **Scalable algorithms for structure identification in graphs** to understand the underlying structure such as important hubs, interesting communities, and so on.

2. **Methods for prediction on semantic graphs,** including identifying missing or incorrect information and estimating unknown attributes and links.

3. **Consider crosscutting issues such as scalability,** accounting for measures of uncertainty, and including temporal or spatial phenomena.

1. Structure-identifying algorithms for semantic graphs

Identifying structure in large-scale graphs is a daunting task but can be used for both analysis and for accelerating queries.

Consider that Google's strength lies in identifying the structure inherent in the Web. The PageRank algorithm, developed by Google founders Sergei Brin and Larry Page, uses eigenanalysis to estimate a Web page's importance by analyzing the pages that point to that page, and the pages that point to those pages, and so on. (This is a powerful tool but can and has been exploited by something called "link spam.") Prof. Jon Kleinberg, Cornell University, is well known for developing an even more powerful concept of "authorities" and "hubs." These techniques can be used to answer questions such as "Which nodes are the most influential?" This can focus the analyst's attention on important nodes in the graph.

Clustering and partitioning can also be used for detecting structure in semantic graphs, which can in turn be used for identifying communities and for decisions on how to partition the graph for parallel computing. Consider, for example, the diagram in Figure 18 that shows the email and organization structure of HP Labs. The illustration shows that the email exchange largely abides by the organizational structure. Clustering and partitioning can be used to analyze such a graph.

Clustering and partitioning techniques have been developed in the field of scientific computing where graph partitioning is used to determine the distribution of finite-element matrices for operations such as parallel matrix-vector multiply. Clustering techniques can potentially aid in the identification of communities of interest within a large-scale semantic graph, again focusing the analyst's attention.

In both cases, the techniques that have been developed are only for directed graphs. They need to be extended to semantic graphs with different node and edge types. Furthermore, both rely on large-scale eigenvalue computations, so these linear algebra techniques need to be extended to massive sizes. Other approaches based on link-based clustering are also applicable.

2. Methods for prediction on semantic graphs, including identifying missing or incorrect information and estimating unknown attributes

The DHS knows that its data is error prone and incomplete. The errors are due to both errors in the original data and errors introduced during ingestion. Some of the errors in the data may even be due to subversive information that is intended to mislead. Gaps in the data are due to the fact that not everything is in data format. For example, two individuals may communicate frequently via untraceable means.

Being able to identify missing or incorrect information and to make predictions about the future (e.g., these two groups will likely begin collaborating because there are several key connections between them) are extremely valuable tools. These tools could also be useful for discovering anomalous nodes, relationships, or subgraphs.

There are some emerging techniques being developed for modeling attributes on semantic graphs, including relational Bayesian networks, relational Markov networks, and relational dependency networks. For example, Prof. David Jensen of the University of Massachusetts discussed his group's work on using relational models to predict which co-authors in physics will likely write papers together. Prof. Lise Getoor's group at the University of Maryland has also done work on link-based classification and other link-based prediction tasks.

Furthermore, Jensen and others are developing methods for discovering statistical biases in relational learning, which will aid in improving machine learning techniques and detecting unusual behavior in the presence of background noise.

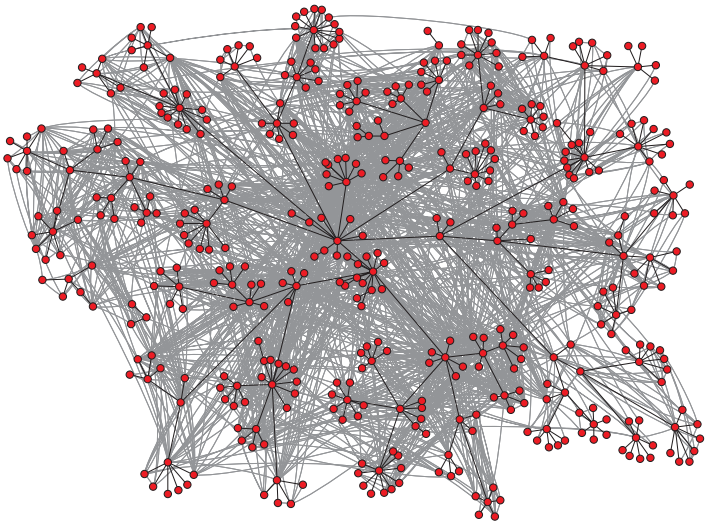


Figure 18: HP Labs' email communication (light grey lines) mapped onto organization hierarchy (black lines). [20]

3. Crosscutting issues: scalability, accounting for measures of uncertainty, and including temporal or spatial phenomena

In all the above areas of interest, there are several crosscutting issues. Once again, a primary challenge is scalability. Techniques must be developed to scale these methods to the enormous DHS data sets. For example, how can the eigenanlysis that is so critical for identifying underlying structure be extended to graphs with billions of nodes?

Most data have accompanying measures of uncertainty. All of the algorithms above should be adapted to incorporate this information. Finally, temporal and spatial information cannot be ignored and has a bearing on testing processes, as discussed in the next section.

D. Models for discovering and detecting processes on graphs

Modeling and detection of processes and structure is an important and novel area to consider, which may help in answering some of the most difficult and complex questions such as:

- Can patterns in the financial transactions of terrorists be detected and exploited?
- What is the structure of power in a group of terrorists?
- What were the main topics of intercepted terrorist messages over the past five years?
- Can a group that is purposely trying to deceive by swapping cell phones with innocents be tracked? In other words, can such changes be tracked over time?

Researchers are just beginning to understand and exploit this new domain, so algorithm development and understanding is key. However, even if researchers are successful in developing algorithms to do the more complex analysis outlined above, the methods will be useless without a query architecture to support the analyst in using the advanced tools.

[20] Shown on front cover. Figure 3 from Lada Adamic and Eytan Adar, "How to search a social network," preprint submitted to Social Networks, October 26, 2004.

For example, the SQL query architecture returns answers to straightforward queries. It can answer the question, "State the number of people with passports from Italy who passed through customs at the JFK airport in January, 2004," but not "Identify individuals whose passports may have been forged that passed through customs at the JFK airport in January, 2004."

Analysts may also be interested in queries that require finding subgraphs that match any of a library of patterns. For example, the query might be "Identify any suspicious group of individuals that passed through customs at JFK in January, 2004." The answer might include things like, "Fifteen men between ages 24 and 44, all employees of the same chemical processing plant and all traveling without their families, flew into JFK during the time period in question."

1. Algorithms to detect processes on semantic graphs

Modeling of data allows information to be processed more efficiently and also allows recognition of the structure of information embedded within data. For example, spectral decomposition of an audio stream can allow extraction of different features, such as background noise or the intermingled voices of different people. The recognition that spectral decomposition of audio can separate these features is a crucial observation, and every data source holds the potential for such modeling to advance the state of the art in moving up the knowledge hierarchy.

A variety of mathematical models can be applied to data and information in order to form higher-level structures for the synthesis and extraction of higher-level understanding. Successful examples can be found in hidden Markov models, Bayesian networks, statistical machine learning, and conditional random fields.

The mathematical abstraction of models for data is crucial for the conception and analysis of algorithms for extraction and synthesis of information and knowledge. For example, the Viterbi algorithm can be used to predict the most likely state of a hidden Markov model, and the abstraction of events on relational data forms the basis for extraction of association rules.

Algorithms such as Hidden Markov models can automatically learn how to detect such behavior and offer an alternative to rule-based systems. Such systems have typically been used on non-relational data and so need to be adapted to relational data and semantic graphs. Recent work on probabilistic relational models and statistical relational learning are approaches that extend statistical modes to relational and semantic graph data.

Much research has been done on studying the structure and processes on the World Wide Web, and this understanding has been turned into design principles. For example, decentralized peer-to-peer systems such as Gnutella[21] have been developed based on random-graph models of small-world models. Studies of email traffic have also been used to understand the influence and communication structure at various companies. Once the influence structure is understood, plans can be made to, for example, disrupt a network; see Figure 19.

2. Identification of subgraphs that are undergoing abrupt changes or bursts in activity

Clearly DHS data will be dynamic, so algorithms that exploit the transient structure of the graph will need to designed. For example, discovering portions of the graph that are particularly active is one possible application. For example, analysts may be interested in detecting an intense email exchange between suspected terrorists.

One promising line of research was discussed by Prof. Jon Kleinberg of Cornell University. There is technology to discover "word bursts" in text. Such technology has already been adapted by Web sites such as Daypop[22] to analyze what Webloggers are writing about. Kleinberg's own studies have turned up interesting trends in the titles of papers at the SIGMOD and VLDB database conferences.

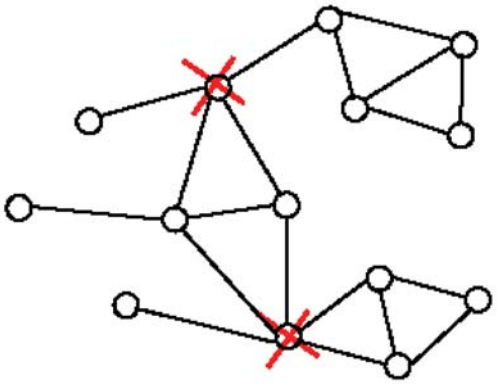


Figure 19: Analysis can help in determining how to disrupt a network.[18]

For example, terms like "data," "base," "schema" and "large" were popular in the late 70s. "Object-oriented" and "parallel" were popular in the late 80s and early 90s, and more recently terms like "warehouse," "indexing," and "xml" are prevalent. The technology that has been developed so far is for documents and not graphs but is an example of the type of work that could be adopted.

Eventually, all of these algorithms will needs to scale to large sizes. According the Stephen Dennis, DHS ISC Deputy Director of Research and Engineering, there will be over three million relationships to analyze *per hour,* and it will be important to detect both threat and non-threat patterns.

3. Using processes and structures for improved analysis

In his workshop presentation, Prof. George Cybenko of Dartmouth College contended that much of the work of analysts is in trying to express processes and structures in terms of the limited query capabilities on data. Instead, if processes and structures were automatically identified in the data, then they could be used in interacting with the database; see Figure 20.

Exposing and making the process models explicit will lead to shareable, large-scale, dynamically updatable data analysis capabilities.

In addition to developing algorithms to detect processes, interfaces will need to be developed so that these processes can be easily understood and used by analysts. For example, a possible process (i.e., a hypothesis) might be discovered by an algorithm and then shown to an analyst. The analyst should have the opportunity to not only study the process but also label it and save it for future use (i.e., find more processes like this one). The analyst should also have the capability to track changes in the process over time.

E. Algorithms to provably ensure privacy and security

Peter Sand, Director of Privacy Technology of DHS Privacy Office, spoke at the workshop about privacy and the DHS. Precursory information about a terrorist event is difficult to discover because the terrorists act in secrecy as much as possible. Clues of potential terrorist activity (movement of money, information, materials, and people) are hidden in the vast amount of data available to DHS. However, discovering these clues requires clever investigation of large amounts of data that may have no apparent significance and that may also be potentially protected by privacy law and policy. Although some practitioners may view privacy as an obstacle or barrier to using the data, Sand proposes an alternative view that privacy should be viewed instead as an attribute of the data that must be considered along with other attributes.

Privacy is a politically and emotionally charged term and means different things to different people. One view of privacy is that it corresponds to the control of personal or institutional space. From a homeland security perspective, the objective is to protect privacy without impeding the flow of information needed to identify threats and vulnerabilities to the nation's security.

[21] http://www.gnutella.com/
[22] http://www.daypop.com/

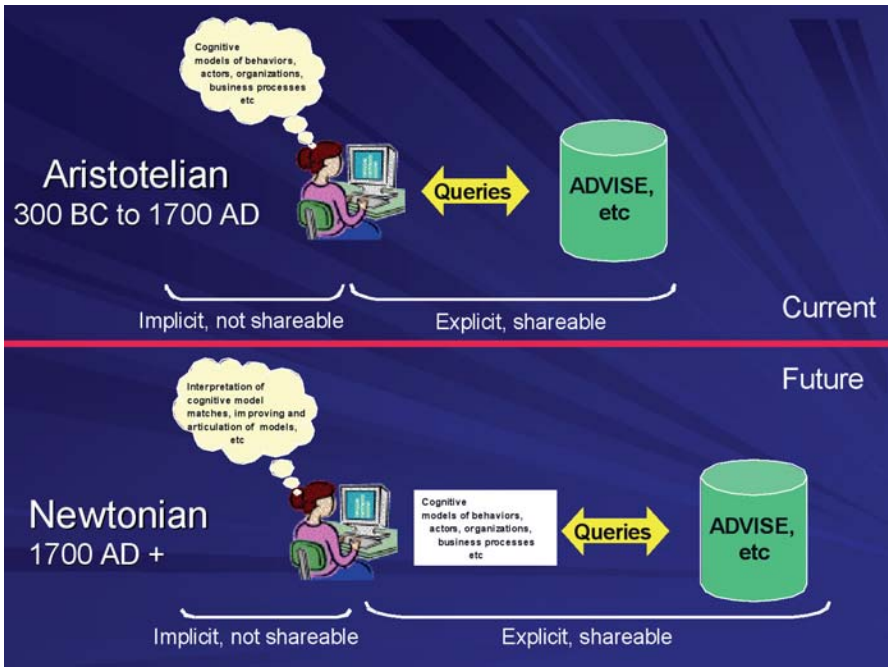


Figure 20: Aristotelian queries require that the analyst's natural cognitive processes be converted into the query language whereas Newtonian queries would use cognitive models for representing the queries.[23]

As Mr. Sand advocated, security safeguards and privacy protections should be considered an integral part of any information system. The following are the R&D needs on the topic of privacy and security, in order of importance:

1. **An understanding of how privacy and security can be integrated into semantic graph techniques.** For example, if one link of a subgraph is protected but is part of the answer to a query, is it safe to return the remainder of the subgraph to the analyst?

2. **Methods that can automatically account for the difference between entities whose privacy should be protected and those that are not protected,** as well as different situations where the trade offs on security and privacy might be allowed to change, i.e., less privacy may be assured when threat levels have been elevated or when a search warrant has been obtained on a particular individual or group of individuals.

3. **Anonymization methods with provable privacy guarantees,** even in the presence of auxiliary data. Many times "anonymous" data is released which, when combined with other data, reveals private information about an individual because the methods do not use a definition of anonymity that accounts for auxiliary data.

1. An understanding of privacy and security in the context of semantic graphs

Semantic graphs present complex data access issues. It is conceivable that each node and each edge in the graph may be assigned a security classification. There is the question of how query results should be presented when some of the data is restricted. If some information is not used because it is protected, the impact on the reliability of the result may be detrimental. Even presenting negative answers and summary answers may reveal restricted information.

[23] From the workshop presentation of Prof. George Cybenko, Dartmouth College.

For example, consider again Figure 21. If an analyst looks for the shortest path between Persons 15 and 16, but all the "Works at" (purple) edges are restricted information, what should the query system return? If nodes 15, 16, and 19 are returned with the paths deleted, it indicates that there *is* a relationship and that somehow those nodes are involved with the relationship. Even a "cannot answer the query due to security concerns" implicitly indicates that there is a connection and that it involves restricted information, which may still be more than what should be revealed.

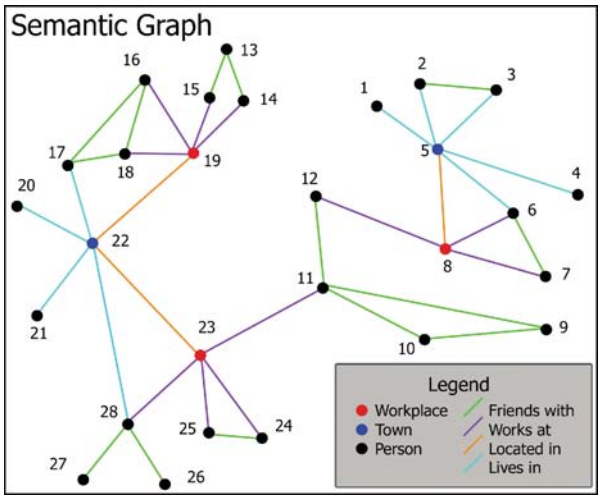


Figure 21: If some edges of classified or other protected information, is there any safe way to present aggregate results to uncleared users?

This problem is of critical importance since the goal of the DHS ISC includes having multiple analysts with multiple levels of authorization and need-to-know accessing the same massive collection of DHS data.

2. Privacy and security policies that account for multiple trust levels

Privacy policies are typically expressed as rules on collection, retention, usage, and dissemination. For example, P3P (Platform for Privacy Preferences) is a W3C standard for Web sites to express their privacy policies in a machine-readable format. The current state-of-the-art allows enforcement of privacy policies for structured data with a single trust boundary, e.g., analyze queries to the database, and either block the query if it is in violation of the privacy policy, allow the query to proceed if it is conformant, or modify the query to reflect user opt-ins or opt-outs such that the modified query will be conformant.

In general, access to data is binary—an analyst either has or does not have access. Such an access policy does not account for important differences, such as the following.

- **Multiple levels of trust.** While it may not be acceptable to have data available directly to analysts, it may be acceptable in some scenarios to allow a computer to access the data. For example, names and personal information about truck drivers crossing at the border may be restricted from an analyst's view, but it might be allowable for the computer to access such information for computing statistics (e.g., how many different/distinct drivers crossed at a particular border point).

- **Access restrictions on data may temporarily be lifted in certain emergency events.** For example, it may be allowable to make private passenger data available to government in the event of a suspected aircraft hijacking.

- **Reversible anonymization is a powerful technique** that allows human beings to look at information about a person and only learn the identity of the person if the information fits a profile of suspicious behavior. The challenge is to ensure that the information that is suppressed does not affect the decision of whether or not the behavior is suspicious.

3. Anonymization methods with provable statistical privacy guarantees

A primary issue with respect to privacy and security is that the release of data that should not be identifiable with individuals often is identifiable *when it is combined with data from other sources*.

For example, Prof. Latanya Sweeney, Carnegie Mellon University, discussed the example of the release of sanitized medical records which, when combined with publicly available voter registration information, did reveal private medical information about individual patients. Sweeney discussed other examples as well, including the fact that pixilation can actually improve facial recognition; see Figure 22.

Figure 22: Pixelation of images has been shown to *improve* facial recognition.[24]

There are several approaches to anonymizing data. One can release only partial information, for example, just the first three digits of a zip code. Or, one can use data perturbation. In either case, the goal is to protect privacy in individual records while still allowing the building of accurate data models at the aggregate level (leveraging the fact that one cares about privacy at the individual level, whereas data models are built at the aggregate level).

The second approach focuses on privacy (sovereignty) of organizational data and uses a secure multi-party computation approach to build data models across organizations while revealing minimal information apart from the data model (i.e., the model is built without the organizations sharing the underlying data).

One of the main failings of the techniques that have been developed up until now is that they fail to use rigorous definitions of privacy that account for auxiliary information.

For example, if a person's medical records are released without a name but did include a street address, gender, and age, it is fairly obvious that the information can be combined with other sources of information (e.g., credit reports, voter records, etc.) to reveal private information.

Subtle variations of this problem crop up repeatedly, including the release of census data. This is termed re-identification, i.e., the process of linking anonymous data to the actual identity of an individual. Prof. Sweeney has also shown that nearly 9 out of every 10 people in the U.S. can be uniquely identified by their date of birth, gender, and zip code; see Figure 23.

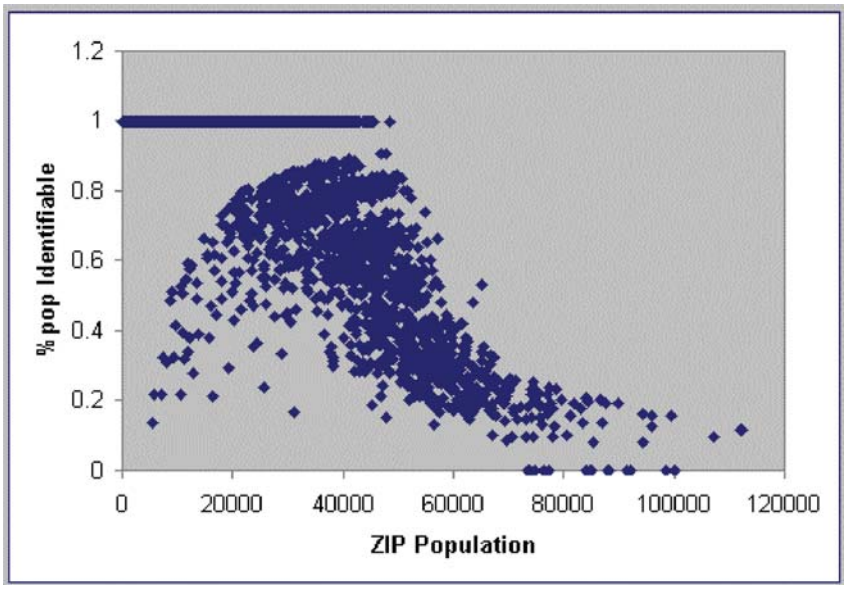


Figure 23: The date of birth, gender, and 5-digit zip uniquely identifies 87.1% of the US population.[24]

[24] From the workshop presentation of Prof. Latanya Sweeney, Carnegie Mellon University.

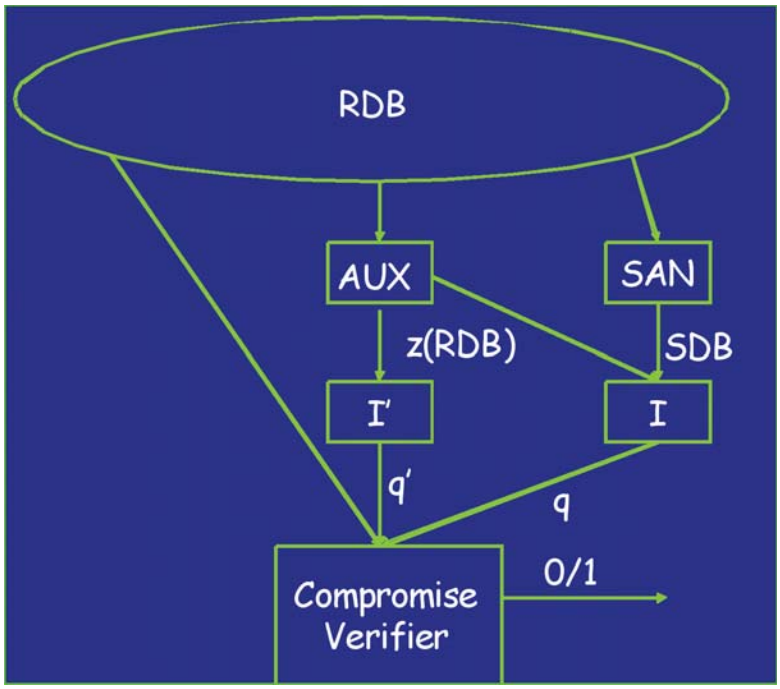


Figure 24: Abstraction of sanitization of data. Auxiliary information (AUX) must not aid in the revelation of any santizied (SAN) information.[25]

Using proper definitions to define privacy is critical to properly implement privacy policies and practices. Dr. Cynthia Dwork, Microsoft Research, explained the complexities of adequately defining privacy and stressed the importance of doing it correctly; see Figure 24 for a description of the type of model that should be used for properly testing the sanitization of data. The idea is that a sanitized database (SAN) should not reveal any more information about a given database (RDB) than can already be determined with known auxillary information (AUX).

[25] From the workshop presentation of Dr. Cynthia Dwork, Microsoft Research.

V. Critical Components of the R&D Process

We have reviewed the many research challenges faced by homeland security applications, not the least of which is the unprecedented scale and complexity of the data in homeland security applications. Fast progress will only be possible if many top researchers from diverse communities, including discrete mathematics, sociology, computer science, and statistics, can be persuaded to enthusiastically join the effort. The following measures are critical components to ensure excellence in R&D on data sciences technology for DHS.

- **Strong support for open research is vital.** Open research in knowledge discovery algorithms encourages wide scrutiny from the technical community, identifies and remedies technical errors, and makes the capabilities and limitations of algorithms known widely and known with confidence. Examples of open algorithms include Linux, Internet protocols, and public key cryptography.

- **Release of test data for competitive analysis.** Realistic test data, including metadata and analysis objectives, are the raw material needed to spur advances in algorithms, analysis and visualization. Such data could be released for competitions sponsored by various conferences such as KDD or the Statistical Computing and Graphics sections of the American Statistical Association.

- **Visiting Scientist and Analyst Programs.** Deeper immersion in homeland security problems and close, sustained contact with DHS analysts will focus the R&D to best support the DHS. Visiting scientist positions could range from 10 week summer internships to year-long sabbaticals. Conversely, sending analysts to spend time at the various research locations would give them a sense of what's new on the horizon and give them the chance to help shape research.

- **Post-doctoral fellowship positions.** Enticing the best new scientists to join the DHS R&D programs will require attractive post-doctoral positions, such as named fellowships with attractive stipends.

- **Summer institutes.** Focused workshops lasting from one week to three months, especially during the summer, offer a venue for bringing the best academic and lab researchers together in a forum where they can focus exclusively on a single problem.

Measures such as these will assure the success of the DHS R&D program on semantic graphs.