SANDIA REPORT

SAND2005-1219
Unlimited Release
Printed February 2005

Knowledge Discovery and Data Mining (KDDM) Survey Report

Dr. Leon Chapman, Dr. Rossitza A. Homan, Jim N. Treadwell, Mark T. Elmore,
Dr. Travis L. Bauer, Laurence R. Phillips, Shannon V. Spires, Danyelle N. Jordan

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

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.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

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 ordering: http://www.osti.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

SAND2005-1219
Unlimited Release
Printed February 2005

Knowledge Discovery and Data Mining (KDDM) Survey Report

Dr. Leon D. Chapman
Sandia National Laboratories

Dr. Rossitza A. Homan
Sandia National Laboratories

Jim N. Treadwell
Oak Ridge National Laboratory

Mark T. Elmore
Oak Ridge National Laboratory

Dr. Travis L. Bauer,
Sandia National Laboratories

Laurence R. Phillips
Sandia National Laboratories

Shannon V. Spires
Sandia National Laboratories

Danyelle N. Jordan
Sandia National Laboratories

Abstract

The large number of government and industry activities supporting the Unit of Action (UA), with attendant documents, reports and briefings, can overwhelm decision-makers with an overabundance of information that hampers the ability to make quick decisions often resulting in a form of gridlock. In particular, the large and rapidly increasing amounts of data and data formats stored on UA Advanced Collaborative Environment (ACE) servers has led to the realization that it has become impractical and even impossible to perform manual analysis leading to timely decisions. UA Program Management (PM UA) has recognized the need to implement a Decision Support System (DSS) on UA ACE. The objective of this document is to research the commercial Knowledge Discovery and Data Mining (KDDM) market and publish the results in a survey. Furthermore, a ranking mechanism based on UA ACE-specific criteria has been developed and applied to a representative set of commercially available KDDM solutions. In addition, an overview of four R&D areas identified as critical to the implementation of DSS on ACE is provided. Finally, a comprehensive database containing detailed information on surveyed KDDM tools has been developed and is available upon customer request.

Table of Contents

List of Tables

List of Figures

1.0 Executive Summary

The Unit of Action (UA) Program Management Office and Lead System Integrator (LSI) have developed a comprehensive Advanced Collaborative Environment (ACE) to create, store, access, manipulate, and/or exchange data digitally. The UA ACE is an information/collaboration environment providing controlled access to all information, both released and in work, and workflows relevant to PM UA. The large and rapidly increasing amounts of data and data formats stored on UA ACE servers has convinced the UA Program Management (PM) of the urgent need to implement appropriate Decision Support System (DSS) on ACE. PM UA has solicited the assistance of the Future Force Integrated Support Team (FIST) composed of Sandia National Laboratories (SNL) as lead laboratory and Oak Ridge National Laboratory (ORNL) to provide an overview of Knowledge Discovery and Data Mining (KDDM) technologies and commercial solutions available today. Another key objective of UA ACE is to implement multiple levels of access control and security. PM UA has asked FIST to report on DOE experience in secure data exchange environments.

1.1 Review Objectives

The objectives of this document are to:

- Provide a glimpse into the evolving world of knowledge discovery
- Provide insight on a limited set of commercially available KDDM tools with a perspective on tools that are currently available
- Review and classify these tools according to their potential applicability to the UA ACE system
- Explore the applicability of the Advanced Decision Support System (ADSS), Intelligent Agent, and Cognitive Model technologies in development at the National Laboratories
- Discuss DOE methods of secure data exchange.

The main goal of this survey is to provide relevant information to the Program Office and LSI that could help determine potential KDDM candidate tools that could contribute to a DSS system on UA ACE.

1.2 What is KDDM?

The field of KDDM can be summarized as the development of methods and techniques for automating the process of making sense of data. The basic problem addressed by the KDDM process is one of mapping voluminous, low-level data into high-level, compact and information rich data forms. At the core of the process is the application of data-mining methods for pattern discovery and extraction.

The traditional method of turning data into knowledge relies on manual analysis and interpretation. However, manual analysis of a data is slow, expensive, and subjective; and as the volume of data increases, this type of manual data analysis becomes impractical. The area of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data

pattern processing. The term *Data Mining* has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. The phrase *Knowledge Discovery and Data Mining* emphasizes that knowledge is the end product of a data-driven discovery process. KDDM refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.

Data mining is the application of specific algorithms for extracting patterns from data. KDDM includes the additional processes of data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining to derive useful knowledge from the data. Application of data-mining in a vacuum, without being driven by a strategic KDDM objective, can lead to the discovery of meaningless and invalid patterns.

KDDM has evolved from research fields of machine learning, pattern recognition, databases, statistics, Artificial Intelligence (AI), knowledge acquisition for expert systems, data visualization, and high-performance computing. The unifying goal is extracting high-level knowledge from low-level data in large data sets. The data-mining component of KDDM currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data-mining step of the KDDM process. KDDM differs from pattern recognition or machine learning in that these fields provide some of the data-mining methods that are used in the data-mining step of the KDDM process, but KDDM focuses on the overall process of knowledge discovery from data, including how results can be interpreted and visualized.[**1**.].

Potential Benefits

Incorporation of a KDDM tools in a Decision Support System, such as the FIST ADSS on UA ACE will enhance information acquisition and assimilation through advanced enabling technologies. This will offer a capability to correlate, analyze, and display selected UA related problems and will offer alternatives for solutions. The ability to conduct data mining and sorting for problem solving will offer a cost effective model for problem displays and solutions. This can improve the technology investment and R&D process as follows:

- Provide program information to UA, military services & defense industrial partners
- Facilitate enhanced technology program investment portfolio strategies
- Identify technology gaps and project overlaps
- Aid management in road mapping progress toward specific goals
- Provide required security level data protection
- Provide rapid updating of data sets from diverse partner sources

The results of this survey can be used as an initial resource when developing a KDDM technology roadmap for UA ACE.

1.3 Survey Study Methodology and Resources

Leveraging past work in KDDM toolset evaluation, the team polled approximately 100 potential KDDM tools. A high-level, **Level 1** questionnaire was developed and sent to the companies representing this set of tools. The purpose of this questionnaire was to provide initial filtering based on tool type, attributes, and capabilities with regard to:

- Supported Operating Systems
- Web-based capabilities
- Supported Database packages

Based on evaluation of all responses a subset of 60 tools was selected to move forward to the detail evaluation phase in the survey. This initial down-select was based on inputs provided by each company of their ability to meet our stated needs. In effect, the first cut was made by the software vendors themselves. A second, more detailed **Level 2** questionnaire was sent to the corresponding companies. This questionnaire was based on criteria provided by Jim McNicol, Unit of Action Advanced Collaborative Environment Senior Technical Advisor and focused on:

- System Scalability
- System's Development Code
- System Training & Support Options
- Future Development Plans

As can be determined by the tenor of the questions, this survey provides a general overview of the functionality of commercially available KDDM tools. In this survey, we do not attempt to determine optimum KDDM tools for ACE, since choosing such optimum requires a detailed evaluation of the ACE requirements.

1.4 Document Organization

Section One, the Executive Summary, defines the objectives and potential benefits of this survey, providing brief overview of KDDM and stating methodology and resources used throughout the survey. **Section Two** is an overview of the Advanced Collaborative Environment with possible KDDM Needs identified. **Section Three** is a ranking of the KDDM tools by tool type.

Sections four though seven depart from the survey approach and are provided as point papers on the work being performed at the National Laboratories and the significant potential the work may provide to ACE. Specifically, **Section Four** provides an overview of the FIST Advanced Decision Support tool that integrates KDDM component toolsets into an integrated and easily expandable system. **Section Five** is an overview of the work performed in Intelligent Agents at Oak Ridge National Laboratory and elsewhere. **Section Six** provides insight to advanced technology development being performed at Sandia National Laboratories in the areas of Augmented Cognition and Cognitive Modeling. **Section Seven** describes work done in the area of secure data exchange throughout the DOE Complex.

2.0 ACE Overview

The UA Advanced Collaborative Environment (ACE) provides a framework to facilitate and coordinate the use of M&S by both Government-authorized users and the contractor team during UA design and development. ACE enables the systems engineering, design, development, test, production and support of the integrated UA systems-of-systems network and platforms; and allows UA engineers and program managers to conduct preliminary and critical design reviews in a secure medium from hundreds of sites across the country. ACE is a primary medium for support of all of the UA program decisions and milestone reviews. The ACE environment is illustrated in Figure 1: ACE Environment.

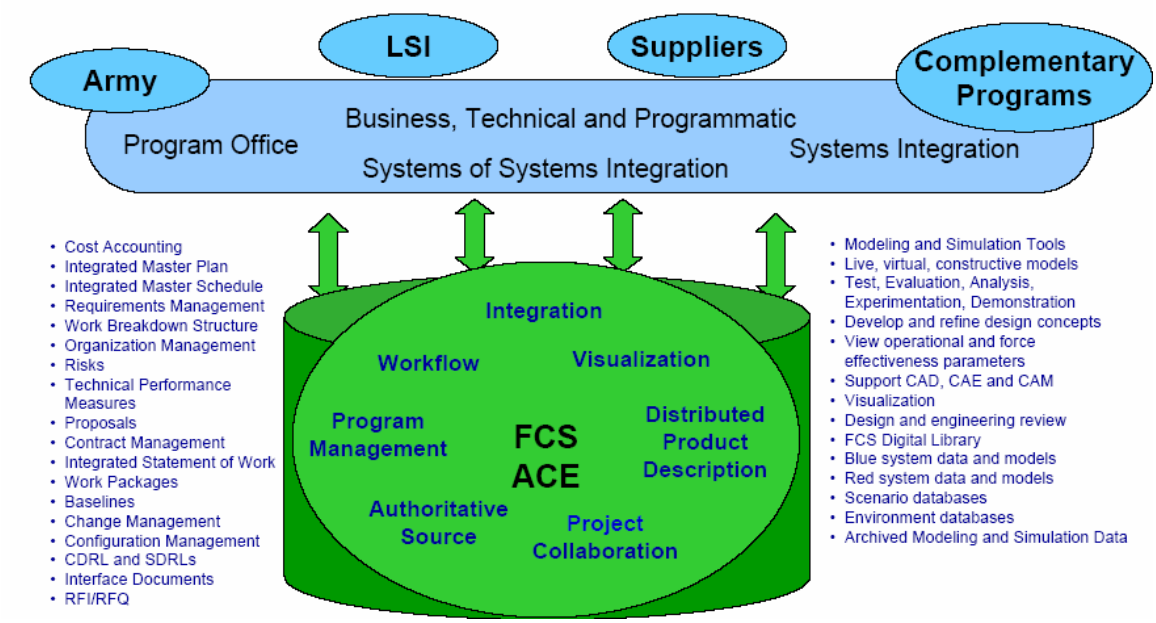


Figure 1: ACE Environment

ACE is comprised of five functional areas:

1. Distributed Program Data Management,
2. Programmatic Management,
3. Project Collaboration,
4. Workflow,
5. Visualization.

A brief description of each functional area is given below.

2.1 Distributed Program Data Management

Distributed program data management is implemented using a Distributed Product Description (DPD), which enables access to all sources of authoritative program data resident in multiple data sources, including: requirements, specifications, SW, product, modeling & simulation, business management, training, logistics, test, and manufacturing.

The DPD is a central architectural element. The DPD maintains the system design information for alternative PM UA platform designs and provides this information as needed for M&S analyses. The strong inter-networking capabilities of the UA concept, along with the variety of

innovative, coordinated operations to be conducted by the UA, place unusually stringent requirements upon the design of the UA DPD. In particular, the DPD must maintain coordinated system design (structural) and behavioral (performance) views, must be able to incrementally reflect changed performance parameters in response to design changes, and must address the performance impacts on coordinated UA operations due to changes in any one of the UA platform designs (including the effects of combat damage or component failures). To respond to these requirements, an interface-centric approach has been advanced as the basis for the DPD logical design. This innovation is expected to provide significant advantages for representing an UA compared to more traditional design approaches, which have typically emphasized either a product structure viewpoint or an M&S-oriented, performance characteristics viewpoint.

2.2 Programmatic Management

Programmatic Management is implemented to allow for the management of all programmatic data (cost, schedule, performance, risk) as well as providing a single view into information, allowing users to develop earned value metrics and risk mitigation plans.

2.3 Project Collaboration

Project collaboration is implemented using an application for sharing and teaming among all ACE constituents that provides access control, discussion forums, resource management and project reporting.

2.4 Workflow

Workflow is implemented using automated process management ensuring consistency across the program, process improvement, paperless review, data pushed to users, integration of multiple data sources through the DPD, comment and change tracking, and management-to-deadline decision tracking.

2.5 Engineering Visualization

Visualization allows for a web-centric capability providing 3D and 2D visual collaboration, mock-up, prototyping, review and study as part of the product development process.

2.6 Simple Analysis of Potential ACE KDDM Needs

In order to begin to identify and comment on KDDM toolset capabilities that might be of benefit to the ACE, a rudimentary evaluation of needs was developed and is displayed below in Table 1. No attempt was made to determine a second level of need across the ACE enterprise. However, it should be noted that the DPD Management Functional Area is assumed to represent the most challenge to KDDM. In order to take this survey to the next level of evaluation, the specific needs of these Functional Areas will need to be identified.

Table 1: Basic ACE Need Evaluation

Functional Area	Need
Distributed Program Data Management	Data Mining Text Mining (All Project Documents Clustering & Visualization
Programmatic Management	Data Mining (Project Trends) Clustering & Visualization
Project Collaboration	Visualization
Workflow	Data Mining (Project Trends) Clustering & Visualization
Engineering Visualization	Text Mining (Specifications) Clustering & Visualization

3.0 Commercial KDDM Tools

The ACE has data ranging from free form text to numerical databases. This diversity of data types necessitates a discussion on the differences between:

- Information Access & Retrieval
- Data Mining
- Text Mining

In addition, the following section describes:

- Data Clustering
- Data Visualization

3.1 KDDM Tool Descriptions

Information Access and Retrieval

Information access and retrieval, as the title implies, is the ability to find documents relevant to a user's needs at the time of need. In the ACE, document repositories appear to have proliferated to the point that information access may be a pressing need. It is important to note that while the information access and retrieval system may return a document, no new information will have been created. This is not the intent of an information access and retrieval system. While some of the tools identified in this study may have information access capabilities, the tools were not evaluated for that specific capability.

Data Mining

Data mining, also known as knowledge discovery in databases, is the application of specific algorithms for identifying data patterns within large databases. Thus, data mining techniques and tools are appropriate for looking at large datasets to determine relevant trends for use in problem solving and decision making. Data mining is designed to create new information from the lower levels of data. In the industry, data mining is specific to large volumes of numerical data that exist within the organization's realm of interest that can be evaluated to find meaningful trends. Thus a data mining effort within ACE might consist of evaluation modeling & simulation runs to determine meaningful statistical information regarding some element of the data set.

Text Mining

Text mining, also known as intelligent text analysis and knowledge discovery in text is the application of advanced computational methods to create information from unstructured textual data. Text mining may hold the most immediate promise to users of the ACE. Advances in text mining have been occurring since the late 1990's as fundamental work in the mathematics of languages was combined with the emerging power of the computer. Advances continue to be made

Data Clustering

Data clustering is a common technique for data analysis. Clustering consists of partitioning data into subsets (clusters), so that the data in each subset share some common trait [2]. There are a number of clustering techniques in use today. An evaluation of the ACE information repository clustering needs will determine the techniques best used. It is anticipated that multiple techniques would be required.

Data Visualization

Data visualization is the visual representation of the results of the data and text mining with clustering techniques applied. While many tools have some data visualization capabilities, it is believed that tools exist that provide greater flexibility to the user.

Extensible Markup Language (XML) Conversion Tools

In addition to the Data/Text Mining and Data Clustering and Visualization tools, a series of toolsets providing XML conversion capabilities was also examined. XML provides a flexible data structure designed to facilitate the sharing of structured text and information across multiple software packages.

The following sections present results from the analysis of individual tools based on the responses to the survey questionnaires and the developed UA ACE-specific ranking mechanism. A database of the responses and a complete response sheet from each company is available upon request.

Ranking Methodology

The tools were ranked using a two level criteria approach. The high level (**Level 1**) criteria were:

- Compliant with the Windows Client and Server Operating Systems
- Are a Web-Based Solution
- Are Solaris Server OS compliant
- Are HP-Unix Server compliant
- Are IBM AIX 5-2 or OS/400 Server OS compliant
- Work with any open database compliant data source
- Can provide output in HTML or XML

The companies provided a Yes/No Response to each. Each Yes response was given a value of one (1) and each No response a value of zero (0). The values were added up and then divided by the total possible score to determine the rank. The Level 1 criteria were not weighted.

After applying the Level 1 criteria to filter the initial pool of tools, another ranking was performed using more detailed, lower level (**Level 2**) criteria including:

- What is your compatibility with core ACE software packages? (9)
- Is your software scalable? (8)
- What are your data input and output processes? (7)
- What are your supported data types? (6)
- What are your Data Access Speeds and what testing supports your claims? (5)
- What is your Development Code? (4)
- What are your access control schemes? (3)
- What are your Help System options? (2)
- What are your Training options? (1)

Each **Criterion** was assigned a numerical weight indicated in parentheses in the Criteria List above. Therefore, Configuration with ACE tools was given the highest weight at nine (9) and the Training System criteria was given the lowest value at one (1).

A value of High, Medium or Low was given to each written response to each criteria question. These response values were then correlated with a numerical value; High = 1; Medium = 2; and Low = 1. Note: Not all companies responded to the Level 2 questionnaire.

The Response value was multiplied against the Criterion weight, summed across all criteria and then divided by the highest possible score of 135, where all criteria have a Response of High (3). This resulted in the rank percentage shown in the **Overall** column of all synopsis tables.

This method was chosen to provide the greatest flexibility for PM UA to revise Criteria weight and Response scores.

Each of the following toolset sections is comprised of

1. A table identifying all of the toolsets evaluated in the section, a brief description of tool provided by the company and the published price (when available).
2. A table of the top-ranking tools Level One and Level Two scores
3. A synopsis table of the Level Two responses
4. Detailed Table of each Level Two Criteria

Therefore each section can easily be reviewed at the summary or detailed response level depending on the level of interest.

3.2 Data Mining Toolsets

Twenty-one data mining toolsets were evaluated during this project. The following table provides a short description and price for the tool.

Table 2: Data Mining Tool Descriptions.

Product	Description	Price
Alice d'Isoft	Discovers hidden trends and relationships in data. First to integrate an OLAP engine and advanced dynamic data aggregation functions. Uses interactive decisions trees.	Contact Vendor
Alyuda Forecaster	Easy business and financial forecasting and data analysis. Wizard-like interface that guides you through easy forecasting. Three different interfaces for all skill levels.	$249.00
Analyst's Notebook	Brings clarity to complex investigations and intelligence analysis. Turns large amounts of unrelated data into actionable intelligence. Proven to save both time and resources. Generates analysis manually or automatically form structured data.	$3,464.00
Cart	Uses a decision tree tool to search relationships and patterns between data in large databases. Manageable for technical and non-technical users. Application include: telecommunication, banking, transportation insurance, health care, education, etc.	$9,299.00
Clementine	Turns data into better business results. Data mining tools to develop predictive models using business expertise and deploy them into business operations to improve decision making. Maximum return in limited amount of time. Clementine makes data mining a business process by focusing data mining technology solving on specific business problems.	$90,000.00
Data Desk 6	Interactive graphical tools to explore and understand your data. Explores any set of data, from a few hundred to a few million. Fast computations that do several analyses at the same time.	$750.00

Product	Description	Price
DataDetective	Predicts trends and detects patterns. Flexible data interface. The format is completely optimized for data mining. C++ core engine. An extremely fast mining engine.	$20,000.00
DB2 Intelligent Miner for Data	Identifies and extracts high value business intelligence from enterprise data assets. Focuses on large scale data mining. Application programming interface that enables the development of customized, industry specific mining application. Long run mining operations.	$74,950.00
Enterprise Miner	Easy to use integrated data mining tools. Exploit and explore corporate data for tactical business advantages. Streamlines the entire data mining process from data access model assessment by supporting all necessary tasks within a single integrated solution. Combines descriptive models and algorithms. Well suited for data mining in large organizations.	Contact Vendor
GhostMiner Developer	Contains a data module, a set of models for training and testing the accuracy of predictions and post analysis tools for evaluation and visualization of the results. Object-oriented design for easy extensions of models. Requires some knowledge of data mining.	$2,000.00
IDOL Server	Platform for understanding the meaning and significance of information. Quickly processes digital information automatically and communicate with multiple applications without the need for manual processing or meta-data. Open architecture and is entirely data agnostic and scalable.	Contact Vendor
Insightful Miner	Highly scalable data analysis workbench that gives new and skilled analysts the ability to deploy predictive intelligence throughout the enterprise. Analyze large datasets. Accelerated discovery. Full featured support for the full data analysis life cycle.	$17,850.00
ModelMAX	Uses data mining for descriptive and predictive modeling. Handles all mathematical and statistical steps necessary for data mining. Button Repetitive model that shortens data mining cycle by 50%. No programming experience, easy to learn.	Contact Vendor
NeuroSolutions	Many different levels available depending on what you need. Support Vector Machine to separate data into designated areas. Probing capabilities. Produces lowest possible error due to Genetic Optimization. 2 separate wizards.	$2,495.00
PolyAnalyst 4.6	Incorporates the latest achievements in data mining and knowledge discovery that can analyze structured and unstructured data. Solves complicated business problems that will help make more informed decisions. Universal data mining tool: ranging fro data importing, cleaning, reporting, modeling, scoring, visualization, and manipulation. Uses decision trees and clustering.	$24,150.00
RoboSuite	Access to any application of data on the web with the ability to integrate into any environment. Has full programming capabilities, but requires very little or no programming experience. Main goal to exploit the fact many applications already have an HTML-based web interface to their functionality and data.	$50,000.00
Starlight	Data and text mining functionality. Designed to capture and graphically portray relationships among multiple pieces of information of various types. Includes a geographic information system. XML conversion.	$15,000.00
SuperQuery Office Edition	Data mining tool designed for Microsoft Excel and Access data files. Discover important facts hidden within business and scientific data. Analyzes sales, financial, customers, inventory data and accounting. Easy to use instructions.	$149.95

Product	Description	Price
Verity K2 Developer	Exploit and manage highly relevant, high value content. Build intelligent applications that learn form end-user actions. Accommodate your release cycle with flexible integration options. Supports multiple languages.	Contact Vendor
WebQL	Data mining tool designed for the World Wide Web. Extracts data from unstructured data sources. Reformats data into structured formats. Designed for business professionals looking for market trends, studying consumer behavior, or doing scientific research.	$18,400.00
XpertRule Miner	Full data mining process addressed. Uses ActiveX Technology. Extensive data information, visualization and reporting features. For both skilled and unskilled data miners. Supports the discovery of association both "case" and "transaction" data.	$3,225.00

For each tool we computed an integer score value and an overall percentage based on the Level 1 and Level 2 criteria. A threshold value of greater than or equal to 60% in both Levels was chosen for determination of the top ranking toolsets. The top ranking toolsets (over 60% in both Levels) are as follows:

Table 3: Data Mining Level 1 and Level 2 Ranking Results.

Tool Name	Level 1 Criteria Results	Level 2 Criteria Results
IDOL Server	100%	92%
Verity K2 Developer	100%	75%
Enterprise Miner	100%	71%
RoboSuite	100%	66%

Of the twenty-one toolsets, the following percentages apply to the first round of questions. It is important to note that the survey team is not able to verify any of the company's claims at this time.

- 100% are compliant with the Windows Client and Server Operating Systems
- 38% are a Web-Based Solution
- 48% are Solaris Server OS compliant
- 33% are HP-Unix Server compliant
- 33% are IBM AIX 5-2 or OS/400 Server OS compliant
- 90% work with any open database compliant data source
- 76% can provide output in HTML or XML

Five tools meet all criteria across the spectrum of the compliance categories above. They include:

- IDOL Server from Autonomy
- Clementine from SPSS
- Enterprise Miner from SAS
- RoboSuite from Kapow
- Verity K2 Developer from Verity

The remaining tools ranked as follows:

- Seven tools meet between five and seven criteria
- Nine tools meet less than five criteria

Of the twenty-one initial queries sent out, only eight provided responses.

The following synopsis provides a subjective overall score based on the relative rankings each tool received in the criteria provided by PM UA ACE for the Level 2 questions.

Table 4: Data Mining Level Two Response Synopsis.

Data Mining Tools Synopsis										
Tool Name	Configuration with ACE Core Tools	Scalability	Data I/O	Supported Data Types	Access Speeds	Development Code	Access Control	Help System	Training System	Overall
IDOL Server	High	High	High	High	High	Med	Med	Med	Low	92%
Verity K2 Developer	Med	Low	High	High	High	Med	Med	Med	High	75%
Enterprise Miner	Med	High	Med	Med	Med	Med	Low	High	Low	71%
RoboSuite	Med	Med	Med	High	Low	Med	Low	High	Low	66%
WebQL	Med	Low	Med	Med	Med	Med	Low	Med	Low	58%
Starlight	Low	Med	Med	Med	Low	Med	Low	Low	Low	52%
GhostMiner Developer	Low	Med	Low	Med	Med	Med	Low	Med	Low	52%
Neuro Solutions	Low	High	Low	Low	Low	Med	Low	Med	Low	50%

Configurations with ACE Core Tools

Table 5: Data Mining Configurations with ACE Core Tools

Product	Configure with ACE core tools	ACE Relevance
NeuroSolutions	No	Low
IDOL Server	Windchill - We have an Omni Fetch that can be configured to access proprietary repositories. Convera Retrieval Ware - Yes through Federated Search Oracle Database Systems - We have a Fetch specifically configured for Oracle. DOORS - We have an Omni Fetch that can be configured to access proprietary repositories. ClearCase - We have an Omni Fetch that can be configured to access proprietary repositories.	High
GhostMiner Developer	Not applicable - GhostMiner does not need any external software to work. We are not aware of any conflict between configurations the above mentioned software.	Low
Verity K2 Developer	Windchill - No integrations, but we would index the data directly from the underlying database; Convera Retrieval Ware - No K2 Enterprise would replace this product; Oracle Database Systems - Yes, we have many implementations indexing data from Oracle. We have a database gateway that allows Verity K2 to index Oracle data. We have hundreds of other references for Oracle. DOORS - No integrations, but we would index the data directly from the underlying database; ClearCase -	Med

Product	Configure with ACE core tools	ACE Relevance
Enterprise Miner	SAS is ODBC and OLE DB compliant, making SAS flexible and allowing our solutions to access any ODBC or OLE DB compliant data source. SAS can connect to Oracle directly.	Med
WebQL	We have both a Windows and Linux version of WebQL and support various APIs. WebQL supports interfaces for programmatically controlling query execution for C++, Java, .NET and ActiveX programmers. The ActiveX API supports a variety of Windows based environments, including Visual Basic, Delphi, and Microsoft Office. In addition, a SOAP interface is available for accessing WebQL Server from any programming language. In addition, WebQL will allow you output and input through ODBC. WebQL has not been specifically configured to work with the mentioned systems (exception: WebQL is configured to work with Oracle Database Systems) but we will assist with any implementation.	Med
Starlight	No, however we have an interface utility that allows automated data flows from tool to tool The next version of Starlight is being developed with APIs.	Low
RoboSuite	- Kapow RoboSuite can work with any system that has a web front end, has an API, accepts XML or has a database that we can access. - Support JSR 168 Portal standard plus direct interfaces with BEA WebLogic, IBM Websphere, Microsoft .Net and generic HTTP web servers. Support all Content Management Systems, Packaged Applications like SAP, PeopleSoft and Siebel, EAI vendors, etc.	Med

Software Scalability

Table 6: Data Mining Software Scalability.

Product	Scalable	ACE Relevance
NeuroSolutions	Yes. There is no limit to the size of neural networks you can implement. There is no limit to the size of data file you feed into NeuroSolutions or into a neural network DLL	High
IDOL Server	We are powering some of the largest systems in the world. We are powering systems with over 10 Terabytes of Data.	High
GhostMiner Developer	This is desktop Windows application; Scalability in terms of size: Maximum number of columns - 40 000, Maximum number of records - 500 000, Maximum number of cells - 20 000 000	Med
Verity K2 Developer	Indexing scalability Indexing is resource intensive. It takes many CPU and I/O cycles, allocation of large amounts of memory is required to process the documents	Low
Enterprise Miner	Absolutely. SAS has no restriction on data size, depends on your hardware.	High
WebQL	Amount of data (size) Answer: Limitations are set by your hardware and bandwidth.	Low
Starlight	A lot depends on the speed of the server and the number of concurrent users. We have been recommending 4-6 clients per server. Starlight visualizations can accommodate record sets number in the thousands to tens of thousands of records.	Med
RoboSuite	Kapow RoboSuite scales linearly with the # of CPU's deployed. Thus you can start economically with just one CPU for a few users and then add more CPU's as the # of users increase. - There is no limit to the data size that can be handled by Kapow RoboSuite.	Med

Data Input and Output Processes

Table 7: Data Mining Data I/O Processes.

Product	Data input and output processes	ACE Relevance
NeuroSolutions	None	Low
IDOL Server	IDOL Server Fetches are configured to automatically aggregate and index information from the designated repositories at specified times. The information is automatically retrieved when user submits a query or a query is submitted based on a users profile or the agents they have created. In the case of a profile or agent query this information is automatically pushed to the user.	High
GhostMiner Developer	Data preparation, feature selections, model validation, data sampling	Low
Verity K2 Developer	Verity provides a web based spider that can automatically index data from any of the Verity K2 supported data types.	High
Enterprise Miner	Processes can be automated via batch.	Med
WebQL	We can support any number of processes which include, but are not limited to, ftp, html post, etc. for our Client Center Apps. WebQL can retrieve data from any number of disparate data sources dependent on the structure of the developed application.	Med
Starlight	- Near real-time updating - Data export as CSV, xml and bmp-screen shots - Our data manipulation utility, XEE (XML Engineering Environment) facilitates data preparation	Med
RoboSuite	- Kapow RoboSuite can be called from a command line tool, and we have API's for .Net and Java. In addition, we support web services, XML output, and interfaces for BEA Weblogic, IBM Websphere and Microsoft .NET.	Med

Supported Data Structures or Types

Table 8: Data Mining Supported Data Structures or Types.

Product	Supported Data Types	ACE Relevance
NeuroSolutions	ASCII, Binary and Bitmap images	Low
IDOL Server	We support over 350 Data Types. I have included a Connector PDF (Connector.pdf) detailing these.	High
GhostMiner Developer	Numerical and categorical (binary) data of the following formats: ASCII text files (including CSV files); Excel spreadsheets; Any database conforming to ODBC standard including MS Access, MS SQL Server and Oracle; Any database conforming to OLE DB standard including MS SQL Server, Informix, Oracle and many more.	Med
Verity K2 Developer	K2 Enterprise can index structured and unstructured repositories into a single collection. Files are stored in their native formats on many types of platforms and storage media. K2 Enterprise can detect, access and index over 280 of the most popular file types	High
Enterprise Miner	supports any relational database; please refer to the following link: http://www.sas.com/technologies/dw/etl/access/index.html	Med
WebQL	We support all standard data types (xml, csv, xls, html, pdf, doc, tsv, images, databases, etc.)	Med
Starlight	- xml	Med
RoboSuite	- There is no limitation here.	High

Data Access Speeds & Testing

The ranking for this section is qualitative only. Two questions were asked: 1) Does there appear to be a reasonable number of test results? and 2) Did someone else provide or perform the test?

Table 9: Data Mining Software Data Access Speeds.

Product	Data Access Speeds and Testing	ACE Relevance
NeuroSolutions	There have not been any independent speed tests done to our knowledge.	Low
IDOL Server	At the bottom of the Scalability and Performance Whitepaper there are results to test that where performed	High
GhostMiner Developer	Data import performance was tested on dataset created by duplicating rows/columns from reference file iris.txt. Performance tests for importing dataset with large number of columns through ODBC couldn't be performed due to MS SQL limitation - maximum number of columns in a single table is limited to 1024. ODBC tests were performed using MS SQL running locally. GM has limitation of ~ 20,000,000 cells per dataset. During import most time is spent on initialization of internal structures/memory allocation/GUI preparation (no significant difference between Text File/ODBC imports).	Med
Verity K2 Developer	This equates to 800 documents per second across 8 CPUs.	High
Enterprise Miner	Again, the speed of the solution depends on your hardware and network configurations. SAS has a department, referred to as the Enterprise Excellence Center, who focuses on hardware sizing and configuration if that would be needed.	Med
WebQL	The QL2 Client Center consists of a cluster of parallel servers that currently processes 25GB of data per day from the web. In the last 17weeks, the QL2 Client Center has processed over 3 Terabytes of data.	Med
Starlight	- No independent tests have been performed. Data access is good with near real-time updating.	Low
RoboSuite	- Our Robot middleware typically adds sub-second throughput. Very large data sets will increase throughput time in a linear fashion.	Low

Development Code

Table 10: Data Mining Software Development Code.

Product	Development Code	ACE Relevance
NeuroSolutions	Visual C++ 6.0	Med
IDOL Server	C++ & C - Backend Java, JSP, .Net - Can all be used on the front end. We have several out of the box and supported integrations for the interface.	Med
GhostMiner Developer	Borland C++	Med
Verity K2 Developer	Access to the K2 search engine is through a browser. Verity also provides a set of published APIs that allow for easy interface to the Verity K2 search engine. Many companies OEM K2 into their product to provide search. Fat Wire, Documentum, FileNet, Cold Fusion are just a few of over 200 companies that OEM Verity's search software. Verity provides APIs in the following formats, JAVA, COM, ASP, .NET and Perl.	Med
Enterprise Miner	Enterprise Miner supplies complete scoring in SAS, C, Java and PMML. SAS is developed in C and C++ and the client piece of Enterprise Miner 5.1 is java. SAS itself is a 4th generation language (4GL) that is "flexible and extensible with an easy-to-learn syntax and hundreds of language elements	Med

Product	Development Code	ACE Relevance
	and functions that support programming everything from data extraction, formatting and cleansing to data analysis, reporting and information delivery"	
WebQL	WebQL is developed using C/C++. The language used for query development is similar to Oracle's implementation of SQL	Med
Starlight	- C++	Med
RoboSuite	- Java	Med

Access Control Schemes

Table 11: Data Mining Software Access Controls.

Product	Access Control Schemes	ACE Relevance
NeuroSolutions	OLE Automation	Low
IDOL Server	Through our IAS (Intellectual Asset Protections System) we will honor the existing security architecture. If you have LDAP or Active Directory we can map right into them. If you are controlling Access at the application or document level we can index the ACL at ingestion which we call mapped security. The other option is before we display a document to a user we will verify they have access this is unmapped security.	Med
GhostMiner Developer	No Answer Provided	Low
Verity K2 Developer	Security Once you identify the information in your repositories and unify it with Verity Gateways, into Verity Collections, you need to ensure it remains secure. Security in Verity K2 Enterprise is flexible and easy to use. It doesn't require you to implement a new security system for your enterprise information. Rather, it enforces your native applications' security mechanisms, such as LDAP, NT or UNIX logins	Med
Enterprise Miner	No Answer Provided	Low
WebQL	We need clarification on this question. I can say that the QL2 Client Center is extremely secure and requires biometric security clearance to access the data center.	Low
Starlight	- USERNAME:PASSWORD - Groups, permissions	Low
RoboSuite	- Since we provide middleware this is not relevant as access control is done elsewhere.	Low

Help System Configuration

Table 12: Data Mining Software Help System.

Product	Help System Configuration	ACE Relevance
NeuroSolutions	We offer email and phone support during normal business hours. The software comes with a complete help file.	Med
IDOL Server	12/7 Combined Telephone Support and Web-Based with escalation to on-site support if necessary.	Med
GhostMiner Developer	12/5 - People based (email/phone)	Med
Verity K2 Developer	Verity provides normal business hours phone support with purchase of K2 Enterprise. Post sales support can be provided on an as needed basis. 24 hour support can be provided for additional cost.	Med

Product	Help System Configuration	ACE Relevance
Enterprise Miner	The SAS Tech Support is 24/7 and is web-based and people-based.	High
WebQL	Technical Support at QL2 is available via phone during normal business hours and via email 24/7. Technical support is people based and correspondence is on a personal level due to the application specific issues that are usually addressed. Full help files are available for WebQL via the WebQL Studio Interface or via download files.	Med
Starlight	- People based	Low
RoboSuite	- Have world wide support during normal business hours in each local time zone. There is 24/7 support via email.	High

Software Training Options

Table 13: Data Mining Software Training Options.

Product	Training Options	ACE Relevance
NeuroSolutions	We offer a 5-day training course in Florida twice per year	Low
IDOL Server	We have a basics training course for one week and an advanced training course which last an additional week.	Low
GhostMiner Developer	2 days course - basic training (allows for using a software by non-experienced user)	Low
Verity K2 Developer	All courses are online and can be done via the web from your office.	High
Enterprise Miner	Training is required, but recommended. - Predictive Modeling using SAS Enterprise Miner Software is the first course to offer. 3-days onsite is $2,850/day plus expenses for up to 20 students. Click on the title in the curriculum page for details. - Text Mining Using SAS Software has the Predictive Modeling course as a prerequisite. It is 1-day. An on-site course is $2,850/day plus expenses for up to 20 students. The same instructor can probably do both courses for a 4-day stretch of training. ** Our Training Dept. will need about 4 weeks to schedule and arrange these courses for your group **	Low
WebQL	We have two training options. OPTION 1 : Training at QL2 Facilities (2 day introductory and 2 day advanced training sessions) Introductory training sessions will give the student a strong basis to further explore query development using WebQL. Advanced training is geared toward individuals who have some background in SQL, PERL regular expressions, and web page layout. Advanced training is usually geared towards customized application development for the client. OPTION 2: Training at the Clients Location; similar to objectives outlined for option 1 except training occurs at the Clients Location. The client must provide the facility and equipment necessary for the training sessions. There is no minimum training necessary in certain cases and we have had clients use WebQL without any formal training with success.	Low
Starlight	- Two days minimum training. A week with some initial support recommended.	Low
RoboSuite	- Training is done on-site hands-on during project implementation. Typical training takes about 1 week.	Low

3.3 Text Mining Toolsets

Thirteen text mining toolsets were evaluated during this project. The following table provides a short description and price for the tool.

Table 14: Text Mining Tool Descriptions.

Text Mining Tools		
Product	**Description**	**Price**
ClearForest	Reads vast amounts of text. Pinpoints relevant information. Uncovers relationships. Provides visual, interactive analytical and executive summaries. Manages information overload. Turbo-charged search.	$200,000.00
docyoument	Discovers knowledge from tons of unstructured textual data to get short and useful information for later operational, marketing or general strategic decision making. Uses text-classification algorithms that will classify all news messages with a small distance to these samples to sort them into those folders. Provide the possibility to get a short summary of a document or even all documents in a folder.	Contact Vendor
dtSearch Text Retrieval Engine	Provides access to dtSearch indexed and unindexed search options and hit highlighted file display features. Includes extensive support for existing fields in documents, as well as support for adding on-the-fly classification information and other fields to documents during indexing. Supports SQL and other COM or non-file data.	$2,500.00
Enkata Enterprise Insight Suite	Addresses problems of business processing modeling, unstructured data analysis, cost allocation, and project portfolio management to deliver unprecedented time to value. Enables you to model, analyze, and optimize interdependent business processes and systems across the enterprise. Has the ability to perform text classification on free-form text documents. Its framework is designed to interpret the content of these text documents and assign a probability score that represents the likelihood of the document belonging to a specific category.	Contact Vendor
InFact	The knowledge access solution that moves intelligence analysts from key word search to event discovery. Its powerful analysis of text content includes new search operators that move beyond competitive knowledge access solutions. Offers human-like understanding of text, dramatically reducing the time knowledge workers spend gathering and analyzing information, vastly improving knowledge discovery, transfer, and decision making. Produces easy-to-read concept maps that quickly summarize intelligence from vast amounts of information across many documents.	$250,000.00
LexiQuest	Analyzes text with a high degree of accuracy-and about 4,000 times faster than you can. Ability to process more than one gigabyte of text or approximately 250,000 pages, per hour. More accurate than other text mining solutions because it is based on natural language processing technologies. Uncovers concepts contained in large collections of text and displays them in a color -coded graphical map so that analysts and business users can clearly see relationships among them. Processes text in many common formats, including plain text, HTML, XML, PDF, and Microsoft Office documents formats.	Contact Vendor
PowerDrill	An advanced information exploration and retrieval application that allows analysts and researchers to drill quickly and deeply into written information and uncover important patterns and relationships. Uses sophisticated linguistic technologies to extract the individuals, actions, and objects described in free-form written text. Quickly drills into written information and categorizes all actions, actors, and objects. Replaces	$50,000.00

Text Mining Tools		
Product	**Description**	**Price**
	slow and expensive manual analysis and tagging of written documents.	
Predictive Text Analytics	Predictive Text Analytics fro SPSS Inc. enables you to combine structured and unstructured data to draw more accurate conclusions about future events and actions. By knowing not only what happened, but why, you can make better decisions for the future.	Contact Vendor
Readware Information Processor	Intelligently analyzes the text it reads. Transforms each text into a compact mathematical format that is stored as a signature record in a readware collection while it is indexing it. Allows users to interrogate and compare the information in readware collections in ways that are simply not possible with statistical, information-theoretic, NLP or so called AI models. Capable of analyzing and indexing tens of thousands of pages an hour.	$250,000.00
SemioMap	Its graphical interface gives users the ability to scan text collections-no matter how big-in a matter of minutes to understand the overall content of the documents. Ability to quickly find out more about a specific topic, seeing how and where concepts are related and the strength of these relationships. Advances search capabilities lets you find documents using document attributes such as text, title, author, abstract, creation date, URL, etc.	$25,000.00
SmartDiscovery	Uses the most comprehensive set of advanced text analysis tools on the market, including search, entity extraction, fact finding, categorization, and visualization. Allows users to quickly find and employ the precise, relevant information needed to get their jobs done more effectively. Allows you to identify the relationships and links between people, organizations, and other entities inside unstructured data sets.	Contact Vendor
VisualText	Ideal tool for quickly developing accurate and fast information extraction, natural language processing, and text analysis systems for the most complex needs. Enables you to build analyzers that can be maintained and enhanced by non-programmers and non-linguists. Automatically generates new rules and layers them into the analyzer framework. Ideal for text analysis applications to combat terrorism, narcotics, espionage, and nuclear proliferation. Ability to find important nuggets of information in voluminous texts.	$50,000.00
XML Miner	A system and class library for mining data and text expressed in XML, extracting knowledge and re-using that knowledge in products and applications in the form of fuzzy logic expert system rules. Predicts numeric values, categorize and classify data, infer the relevance and topics in text, and mines the structure of XML documents. Integrates text mining seamlessly so that blocks of embedded text can be handled at the same time as numeric and categorical data.	$4,499.00

The top ranking toolsets (over 60% in both Levels) are as follows:

Table 15: Text Mining Level 1 and Level 2 Ranking Results.

Tool Name	Level 1 Criteria Results	Level 2 Criteria Results
LexiQuest	75%	87%
ClearForest	63%	78%
dtSearch Text Retrieval Engine	63%	73%
Readware Information Processor	88%	70%
PowerDrill	75%	63%

Of the thirteen toolsets, the following percentages apply to the first round of questions. It is important to note that the survey team is not able to verify any of the company's claims at this time.

- 100% are compliant with the Windows Client Operating System
- 85% Windows Server Operating System
- 92% are a Web-Based Solution
- 69% are Solaris Server OS compliant
- 31% are HP-Unix Server compliant
- 31% are IBM AIX 5-2 or OS/400 Server OS compliant
- 92% provide an XML Configurable Custom Dictionary
- 100% can provide output in HTML or XML

Three tools meet all criteria across the spectrum of the compliance categories above. They include:

- Docyoument from Media Style
- XML Miner from Scientio, LLC
- Predictive Text Analytics from SPSS

The remaining tools breakdown as follows:

- Nine tools meet between five and seven criteria
- One tool meets less than five criteria

Of the thirteen initial queries sent out, six provided responses.

The following synopsis provides a subjective overall score based on the relative rankings each tool received in the criteria provided by PM UA ACE for the Level 2 questions.

Table 16: Text Mining Level Two Response Synopsis.

Text Mining Tools Synopsis										
Tool	Configuration with ACE Core Tools	Scalability	Data I/O	Supported Data Types	Access Speeds	Development Code	Access Control Schemes	Help System	Training System	Overall
LexiQuest	High	High	Med	High	High	Med	Med	Med	Med	87%
ClearForest	High	Med	Med	High	Med	Med	Med	Med	Med	78%
dtSearch Text Retrieval Engine	Low	High	Med	High	High	Med	Med	Med	Low	73%
Readware Information Processor	Med	Med	Med	Med	High	Med	Med	Med	Med	70%
VisualText	Med	Med	Med	Med	Med	Med	Med	Med	Med	67%
PowerDrill	Med	Low	Med	Med	Med	Med	High	Med	Med	63%

Configurations with ACE Core Tools

Table 17: Text Mining Configurations with ACE Core Tools.

Product	Configure with ACE core tools	ACE Relevance
Readware Information Processor	Readware has been used with Oracle and with many content management applications. Readware has an API and accepts HTTP-style commands over a TCP socket, so it is very easy to query a Readware collection or "feed" files or records to the Readware Analyst for indexing and classification.	Med
dtSearch Text Retrieval Engine	Convera is a competitor of ours; it is unlikely that you could use the dtSearch Engine on their indexes and vice versa. A large portion of our developer case studies mention developing with SQL, XML, and other database formats.	Low
ClearForest	Windchill (Yes), Convera Retrieval War (Yes), Oracle Database Systems (Yes), DOORS (Yes), ClearCase (Yes). ClearForest has been architected to provide an open API, to allow applications to interface with the ClearForest outputs, via web services.	High
PowerDrill	RES supports Oracle out of the box. More detail and analysis would be required to better understand the other tools and how you would want them to work with RES. RES is very extensible and offers the ability to be customized to a particular environment via the creation of custom pre and postprocessors, therefore working with other tools is seldom a problem.	Med
VisualText	Analyzers use ODBC calls to work with databases including Oracle. Analyzers compile to DLL libraries on Windows and .a archive libraries on Unix/Linux, which can be called from any application via a programmer's API.	Med
LexiQuest	1. SPSS Inc.'s text mining technology works with the drivers available for configuring to a number of search engines. Creating a driver for a new search engine, such as Convera, is a fairly simple process. One of SPSS Inc.'s customers, Peugeot, has embedded Convera into their text mining environment and the driver for that configuration was developed in about an hour. 2. SPSS Inc.'s text mining technology works with an ODBC driver for connecting to Oracle or any other database.	High

Software Scalability

Table 18: Text Mining Software Scalability.

Product	Scalable	ACE Relevance
Readware Information Processor	There is no theoretical limit to the numbers of documents or records that can be indexed in a single Readware installation Statistics show that a single Readware query server, in a cluster of seven P5 servers, processes up to 250 queries per hour.	Med
dtSearch Text Retrieval Engine	Version 7.x of the dtSearch Engine will be able to hold a quarter of a terabyte or more in a single index. And there is still no limit on the number of indexes that you can build and simultaneously search.	High
ClearForest	ClearForest has been used in environments that store petabytes of input data. The ClearForest solution is scalable to allow multiple ClearTags CPUs, running in parallel, to handle these large data sources.	Med
PowerDrill	Attensity PowerDrill is a web-enabled text query tool which operates against an enterprise database repository and scales for large user communities.	Low
VisualText	Analyzers can process both large (multiple MB) and small text inputs, as well as large and small numbers of inputs. Because analyzers focus on intensive and deep analysis, typical speed for an accurate Analyzer is about 0.1 seconds per 1000 characters of text.	Med
LexiQuest	1. SPSS Inc.'s text mining technology can process on average 1GB of open text per hour on a standard PC (2Ghz, 512 MB RAM). Using Grid Computing technology, SPSS Inc.'s text mining technology can process 60GB of open text in 3 hours. 2. There is no specific limit on the size of the raw data being accessed or with the number of concepts extracted from that data. Plans for a 64 bit architecture in January 2005 will improve scalability to an even higher degree.	High

Data Input and Output Processes

Table 19: Text Mining Data I/O Processes.

Product	Data input and output processes	ACE Relevance
Readware Information Processor	Any automation that can inter-operate using TCP/IP standards and protocols. In addition customers have created ASPs, JavaBeans, DLLs and all manner of other programmatic automations encapsulating., exporting and/or integrating Readware functionality and/or exposing Readware analysis, filtering or classification results. We also have console tools that can be automated using shell-scripts and batch files that may be preferred by some IT shops.	Med
dtSearch Text Retrieval Engine	In addition to supporting all of the file types above, the dtSearch Engine includes a data source API for non-file data. The dtSearch Engine also includes multiple options for exporting search results, in a wide variety of Web-based and other formats.	Med
ClearForest	ClearForest provides documented APIs for all data input and output applications. Once data has been inserted into a ClearForest folder, it will automatically by ingested by ClearTags, and the tagged data will be automatically sent to either a database or application. The ClearForest APIs are used to create interfaces with other commercial or government	Med

Product	Data input and output processes	ACE Relevance
	applications.	
PowerDrill	Attensity RES supports an API where users can submit text and receive extractions as output in asynchronous fashion. However, RES is most often utilized in automated batch mode where text is collected and entities and events extracted and output in hands-off, automated fashion.	Med
VisualText	Analyzers may accept text via files and buffers. Stream input can easily be wrapped around Analyzers. Analyzers produce output to files, buffers, and streams, and combinations of these. Additional input and output is readily available via ODBC connectivity to databases.	Med
LexiQuest	1. Command mode and API are available for managing input and output. Files containing information on extracted concepts and relationships are automatically generated.	Med

Supported Data Structures or Types

Table 20: Text Mining Supported Data Structures or Types.

Product	Supported Data Types	ACE Relevance
Readware Information Processor	It appears this software package is capable of reading most of the common formats. However, the response to the question was not clear so it has been omitted. The full text can be read in the Appendix.	Med
dtSearch Text Retrieval Engine	dtSearch supports all of the following document types: "Office" (word processor, database, spreadsheet, presentation, etc.), emails, HTML, PDF, XML, SQL, ZIP, CSV, RTF, ANSI, Unicode files, and more.	High
ClearForest	We support all data types relating to textual base files such as: Word, PDF, PowerPoint, Excel, .RTF, .TXT, Adobe, HTML, XML, Any ASCII text file Clearforest also provides a data access layer, which allows the user to handle proprietary formats and inputs from data bases.	High
PowerDrill	Document types out-of-the box include PDF, Word, RTF, ASCII, .txt, html, and email. Additional document types are supported via the inclusion of custom converters.	Med
VisualText	Analyzers can in principle accept arbitrary binary input files. However, the primary data input types are human readable texts, such as XML, HTML, plain text, and RTF. Conversion to text is readily available for formats such as PDF and PS.	Med
LexiQuest	1. The ODBC driver allows SPSS Inc.'s text mining technology to read data from any database. It can also access raw data from pdf, html, xml, txt, bibliographic format, and MS Office.	High

Data Access Speeds & Testing

The ranking for this section is qualitative only. Two questions were asked: 1) Does there appear to be a reasonable number of test results? and 2) Did someone else provide or perform the test?

Table 21: Text Mining Software Data Access Speeds.

Product	Data Access Speeds and Testing	ACE Relevance
Readware Information Processor	ADAC of Germany tested the throughput of the Readware Query processor. They found that the throughput of the query processor on an 800 MHz PIII with 512MB of main memory was 83 queries per minute. These tests were performed by the lead programmer at ADAC of Germany.	High
dtSearch Text Retrieval Engine	Our case studies such as http://www.dtsearch.com/CS_Premirus_CC.html, cite search times of less than a second, through hundreds of gigabytes or more of data.	High
ClearForest	Information Extraction speed for the recommended configuration with ClearForest products Version 6 is as follows: Performance (KB/min). These results were obtained for creation of XML files on P4 2.8GHz single CPU HT PC with 2GB RAM on Window 2003 Server. Basic, 800 Intelligence, 240 Business, 180 Biological, 190 Patents, 1400	Med
PowerDrill	Benchmark data regarding RES ingest speeds have been performed internally only, and our estimates indicate that RES is capable of ingesting and processing 1-2GB of raw text per 24 hour period per CPU. These tests were done on a Windows platform with Pentium 1+GHz CPUs. If the processing load exceeds 1-2GB per 24 hour period, multiple CPUs can be employed as the software is scalable.	Med
VisualText	Compiled analyzer speed is typically on the order of 0.1 seconds per 1000 characters. Analyzers may also interact directly with databases via ODBC-based interfaces	Med
LexiQuest	Pfizer is a company that uses SPSS Inc.'s technology for bioinformatics research. Tests are done on accuracy of extraction and speed. Pfizer was highly satisfied with the results they were able to achieve. Accuracy ranged from 85% to 90% in extracting relevant gene patterns with a speed of 60GB (15 million documents) in 3 hours.	High

Development Code

Table 22: Text Mining Software Development Code.

Product	Development Code	ACE Relevance
Readware Information Processor	ANSI-C with some C++ classes	Med
dtSearch Text Retrieval Engine	The dtSearch Text Retrieval Engine comes in two versions: one for Win & .NET and one for Linux. The dtSearch Engine for Win & .NET supports Delphi, Java, C++, C++.NET, C#, VB.NET, ASP.NET & more. The dtSearch Engine for Linux supports C++ and Java.	Med
ClearForest	The ClearLab Development Environment uses DIAL4, an OO Based language, to customize ClearTags and the Discovery Modules.	Med
PowerDrill	C++ for the RES engine in conjunction with Java.	Med
VisualText	TAI's NLP++ programming language is used for the development	Med

Product	Development Code	ACE Relevance
	of Analyzers. For optimization and deployment, the definition of an Analyzer, i.e., its passes, rules, and code are compiled to C++ code. Under the hood, VisualText, runtime libraries, and Analyzers are written in C++: Microsoft Visual C++ and Gnu C++ are available for Analyzers, while VisualText itself is built in MS Visual C++ with MFC and CodeJock's XTreme library. VisualText runs on Windows platforms, while Analyzers may run on any platform with Gnu C++. A bridge for calling Analyzers from MS .NET code is available as well.	
LexiQuest	SPSS Inc.'s text mining technology is written in C and C++. The user interfaces are written in Java 1.	Med

Access Control Schemes

Table 23: Text Mining Software Access Controls.

Product	Access Control Schemes	ACE Relevance
Readware Information Processor	We have no specific access control schemes rather we support whatever access control scheme is imposed upon us.	Med
dtSearch Text Retrieval Engine	Please see http://support.dtsearch.com/faq/dts0127.htm and http://support.dtsearch.com/faq/dts0179.htm for sample answers to that question.	Med
ClearForest	ClearForest does not have its own security functionality. We typically use the client's access control schemes for database and system security.	Med
PowerDrill	Attensity RES requires user login and authentication. Multiple user accounts are common in an RES environment and each may be granted roles which provide different levels of access controls for varying tasks. More detail is available upon request.	High
VisualText	As an Analyzer is a component, rather than an end-to-end solution, this question does not apply.	Med
LexiQuest	Access control schemes can be based on MS Identification.	Med

Help System Configuration

Table 24: Text Mining Software Help System.

Product	Help System Configuration	ACE Relevance
Readware Information Processor	24/7 people-based via email describes us best. We are experimenting with wiki tools and other collaborative environments and will probably move the developer's kit manual to that environment, once we decide. Until then we encourage email and offer telephone, and web-based conferencing where and whenever practical.	Med
dtSearch Text Retrieval Engine	In addition to providing extensive online and offline documentation, dtSearch Corp. provides developer technical support by phone, by email to tech@dtsearch.com, as well as through a developer user's group.	Med
ClearForest	We offer telephone, web-based and people-based technical support. Our standard hours for support our 8 AM - 6PM, Monday -	Med

Product	Help System Configuration	ACE Relevance
	Friday. We offer additional 24/7 premium support for customers requiring additional coverage.	
PowerDrill	Support is available via telephone and pager as well as online bug reporting capabilities. Standard support is 5x8 hours. Additional support options such as 24x7 are available upon request.	Med
VisualText	Our support is telephone-based and email-based. Visits to customer and developer sites can be arranged as well. Help is provided on a seven-day-per-week basis.	Med
LexiQuest	1. SPSS Inc. offered technical support for all it's software by the following means: phone, email, and web-based. 2. 24/7 support can be set up based on the complexity of each customer usage scenario.	Med

Software Training Options

Table 25: Text Mining Software Training Options.

Product	Training Options	ACE Relevance
Readware Information Processor	We offer on-site installation and operations training in one, three and five day packages. In addition to these basic courses on the technical use and operation of a Readware installation, we offer one and two week course on the principles and best practices for building Readware topics, filters and classifiers using Readware Knowledge-Types and information cultures.	Med
dtSearch Text Retrieval Engine	We sell the dtSearch Engine as an "out of the box" developer product. We do not provide training, although our developers are certainly available for technical questions.	Low
ClearForest	We offer formal training at our facility as well as at the customer's site. We recommend the following classes: ClearResearch Analyst Training - 1 Day ClearTags Administration Training -2 Days ClearLab Development Environment Training - 3 Days	Med
PowerDrill	Generally speaking a 1 week training course is required to become proficient with the Attensity product set. This includes 1 day of training on PowerDrill and 4 days of training on RES and the Knowledge Engineering process.	Med
VisualText	TAI provides training tailored to the needs of the customer, focusing on the particular customer application. Minimal training consists of following the documentation and tutorials provided in the VisualText Help. Self-training is available via our well-reviewed Help documentation, which includes a set of Tutorials to be read and executed. Sample analyzers are provided both for study purposes and for use as initial Analyzers for customization. A general Analyzer and a resume analyzer are available for free download at http://www.textanalysis.com/Apps/apps.html In addition; VisualText installation comes with other examples, such as an Analyzer that connects to a database, and a business-events analyzer.	Med
LexiQuest	1. The minimum training we offer is 2 days, the most is 7 days. Knowledge transfer sessions can exceed that timeframe depending on the complexity of the defined challenge. 2. Onsite Training - SPSS Inc. will send a trainer to your facility, or reserve one of our public training facilities, to train any number of users on customized	Med

Product	Training Options	ACE Relevance
	content relevant to their environment. 3. Public Training - users can attend a Public Training course at any SPSS facility nationwide. 4. SPSS Consulting - SPSS Inc. will send a consultant to your facility to address a specific challenge. Knowledge transfer will take place as the consultant will teach users as the solution is being developed.	

3.4 Clustering Tools

Eight clustering toolsets were evaluated during this project. The following table provides a short description and price for the tool.

Table 26: Clustering Tool Descriptions.

Clustering Tools		
Product	**Description**	**Price**
Clustan Graphics 6	Offers hierarchical agglomerative cluster analysis, k-means analysis, focal point clustering, outlier analysis and proximity analysis. Ability to construct a cluster model from a hierarchical cluster analysis and then classify any number of new cased by reference to it. With the Clustan Wizard you're clustered in three clicks. Data mining in a particular strength which can cluster 200,000 cases or more hierarchically and run k-means on a million cases using a PC.	$375.00
Clustering Engine	Automatically organizes search or database query results into meaningful hierarchical folders completely on-the-fly, out-of-the-box. Automatically clusters search results into categories that are intelligently selected from the words and phrases contained in the search results themselves. Categories are always up-to-date and as fresh as your content. Human level accuracy and a familiar intuitive folders-style interface.	$50,000.00
Insight Discoverer Clusterer	Innovative solution for structuring previously unstructured information. Classifies and regroups documents, based on their semantic similarities, into coherent classes - the clusters. Offers excellent visibility on large sets of documents and on complex domains. The server stores the results of the clustering in an interactive application, ready to be navigated by the user.	$36,500.00
StarProbe	A star schema based data mining system that works smoothly with most common database systems and incorporates statistics, machine learning, and data warehousing. Contains neural clustering, which is based on Kohonen feature maps, creates grid-shaped partitions. Clusters bundles of similar objects into clusters.	$12,000.00
TextAnalyst	A unique software tool for semantic analysis, navigation, and search of unstructured texts. Helps to quickly summarize, efficiently navigate, and cluster documents in your text base. Breaking links representing weak relations in the original Semantic Network enables clustering of the text base.	$1,290.00
VisuaLinks	Used to discover patterns, trends, associations and hidden networks in any number and type of data sources. VisuaLinks presents data graphically uncovering underlying relationships and patterns. It's designed to expose clusters, or networks, of related information. Searches your data looking for particular types of data that are related by particular types of associations.	$2,800.00
WordStat v4.0	A text analysis module specifically designed to study textual information. Includes numerous exploratory data analysis and graphical tools that may be used to explore the relationship between the content of the documents and information stored in categorical or numeric variables. Relationships among	$1,095.00

Clustering Tools		
Product	**Description**	**Price**
	words or categories as well as document similarity may be identified using hierarchical clustering and multidimensional scaling analysis.	
Text Miner	A suite of tools for discovering and extracting knowledge from text documents. Text documents can be clustered automatically into groups, classified into predefined categories and used in conjunction with structured data to build predictive models. Text clustering algorithms group documents into common themes and topics based on their content. Cluster summaries are easy to interpret in the context of the original text documents.	Contact Vendor

The top ranking toolsets (over 60% in both Phases) are as follows:

Table 27: Clustering Tools Level 1 and Level 2 Ranking Results.

Tool Name	Level 1 Criteria Results	Level 2 Criteria Results
VisuaLink	89%	74%
StarProbe	89%	64%
Clustering Engine	67%	87%

Of the eight toolsets, the following percentages apply to the first round of questions. It is important to note that the survey team is not able to verify any of the company's claims at this time.

- 100% are compliant with the Windows Client Operating System
- 88% Windows Server Operating System
- 50% are a Web-Based Solution
- 50% are Solaris Server OS compliant
- 38% are HP-Unix Server compliant
- 25% are IBM AIX 5-2 or OS/400 Server OS compliant
- 100% provide Hierarchical Cluster Analysis
- 63% can provide customizable output

None of the tools meet all criteria across the spectrum of the compliance categories above. The tools breakdown as follows:

- Two tools meet eight of the nine criteria
- Four tools meet between five and seven criteria
- Two tools meets less than five criteria

Of the eight initial queries sent out, four provided responses.

The following synopsis provides a subjective overall score based on the relative rankings each tool received in the criteria provided by PM UA ACE for the Level 2 questions.

Table 28: Clustering Level Two Response Synopsis.

Clustering Tools Synopsis										
Tool	Configuration with ACE Core Tools	Scalability	Data I/O	Supported Data Types	Access Speeds	Development Code	Access Control Schemes	Help System	Training System	Overall
Clustering Engine	High	High	Med	High	High	Med	Med	Low	High	87%
VisuaLinks	High	Med	Med	High	Low	Med	Med	Med	Med	74%
Text Miner	Med	High	Med	Med	Med	Med	Low	High	Low	71%
StarProbe	Med	High	Med	Low	Med	Med	Low	Low	Low	64%

Configurations with ACE Core Tools

Table 29: Clustering Configurations with ACE Core Tools.

Product	Configure with ACE core tools	ACE Relevance
VisuaLinks	Windchill- no - Convera Retrieval Ware-yes - Oracle Database Systems- Yes - DOORS- Yes, through a jdbc driver. - ClearCase- no	High
Clustering Engine	Vivisimo software has been configured to work with Convera & Oracle Database System. Although we do not have direct experience with Windchill from PTC, DOORS from Telelogic or IBM's ClearCase, We can work with any product having a web interface.	High
StarProbe	StarProbe is primarily designed to work with relational DBMSs. It works with any database systems with ODBC and/or JDBC support. This includes Oracle, DB2, SQL Server, MS Access, Infomix, MySQL, and many others.	Med
Text Miner	SAS is ODBC and OLE DB compliant, making SAS flexible and allowing our solutions to access any ODBC or OLE DB compliant data source. SAS can connect to Oracle directly.	Med

Software Scalability

Table 30: Clustering Software Scalability.

Product	Scalable	ACE Relevance
VisuaLinks	Yes, the tool is scalable. We are deployed in many locations one of which is 12 databases each with upwards of 30 million records with hundreds of user.	Med
Clustering Engine	Our Clustering Engine supports search indexes of well over 30 million documents. Our Content Integrator software for metasearch scales to 200 data sources. There is no limitation on the number of simultaneous users.	High
StarProbe	Current maximum logical size for a single dataset is 2 billion records. This may be further limited by the capacity of single disk drives. T	High
Text Miner	Absolutely. SAS has no restriction on data size, depends on your hardware.	High

Data Input and Output Processes

Table 31: Clustering Data I/O Processes.

Product	Data input and output processes	ACE Relevance
VisuaLinks	VisuaLinks is able to perform queries to and from database's. Dig processes unstructured documents.	Med
Clustering Engine	It is unclear exactly what you are asking, however our software can fetch content automatically where ever it is and output it in RSS, HTML or XML.	Med
StarProbe	Input from databases is automated interactively. However, input from data files requires user to specify input data format. Outputs are all in proprietary formats and can be re-opened from StarProbe. There are several export facilities. For example, charts and graphics can be made to image files (in GIF format). Textual outputs can be saved as files. In addition, predictive models and clustering outcomes can be applied directly to database tables.	Med
Text Miner	Processes can be automated via batch.	Med

Supported Data Structures or Types

Table 32: Clustering Supported Data Structures or Types.

Product	Supported Data Types	ACE Relevance
VisuaLinks	Any relational data source, emails, document repositories, or websites.	High
Clustering Engine	All standard MS office types, PDF, compressed files, HTML, XML, DB files, etc.	High
StarProbe	There are three generic data types: Character strings, 64-bit integer and 64-bit IEEE-754 floating point designed for scientific data. Taxonomic data structures (or hierarchical drill down structures) are supported using star schema.	Low
Text Miner	supports any relational database; please refer to the following link: http://www.sas.com/technologies/dw/etl/access/index.html	Med

Data Access Speeds & Testing

The ranking for this section is qualitative only. Two questions were asked: 1) Does there appear to be a reasonable number of test results? and 2) Did someone else provide or perform the test?

Table 33: Clustering Software Data Access Speeds.

Product	Data Access Speeds and Testing	ACE Relevance
VisuaLinks	We have people who have tested our product based on certain criteria from the Gartner group, the JIVA project which has been certified in Intel networks, IRS and many more. These tests have been marked as classified. A copy may be obtained by other classified individuals.	Low
Clustering Engine	Clustering: 150ms to cluster 200 results 400ms to cluster 500 results Content Integrator: 300ms to metasearch 12 sources and cluster 200 results Note: Vivisimo can cluster a maximum of 1000 search results, and metasearch over 200 data sources.	High
StarProbe	We have no independent test records. The tests were performed on Sony Vaio Laptop with Pentium4 1.6GHz, 512MB memory, Windows	Med

Product	Data Access Speeds and Testing	ACE Relevance
	XP, Sun Microsystems JRE 1.4.2. The dataset used has 1 million records with 13 numeric and non-numeric fields. Pie charts - 2 seconds, Scatterplots - 6 seconds, Decision trees (depth = 20 levels) - 150 seconds, Hotspot analysis (depth = 2 levels) - 20 seconds, Clustering (1 epoch training for 2 fields) - 7 seconds.	
Text Miner	Again, the speed of the solution depends on your hardware and network configurations. SAS has a department, referred to as the Enterprise Excellence Center, who focuses on hardware sizing and configuration.	Med

Development Code

Table 34: Clustering Software Development Code.

Product	Development Code	ACE Relevance
VisuaLinks	VisuaLinks is developed completely in JAVA. DIG is developed in .NET	Med
Clustering Engine	Vivisimo software products are written in: C, C++, ASP, C#, JSP, Java or Perl. Though for most applications, no dvpt at the API level is required, just configuration through the web interface extensively based on XML and XSL.	Med
StarProbe	100% written in Java 1.1.8 and also can be run on any later versions of Java runtime. This means that StarProbe can be deployed on a wide variety of workstations. Note that some systems (such as Microsoft) do not support recent Java versions.	Med
Text Miner	Enterprise Miner supplies complete scoring in SAS, C, Java and PMML. SAS is developed in C and C++ and the client piece of Enterprise Miner 5.1 is java. SAS itself is a 4th generation language (4GL) that is "flexible and extensible with an easy-to-learn syntax and hundreds of language elements and functions that support programming everything from data extraction, formatting and cleansing to data analysis, reporting and information delivery"	Med

Access Control Schemes

Table 35: Clustering Software Access Controls.

Product	Access Control Schemes	ACE Relevance
VisuaLinks	User name/password with permission attributes to read/write sources.	Med
Clustering Engine	We can mimic any scheme of an underlying application. Provide a username/password repository. Support user groups.	Med
StarProbe	Current version is for workstations and therefore does not employ any access controls. However, there is license verification and users have to enter database connection information if they want to access database data.	Low
Text Miner	No Answer Provided	Low

Help System Configuration

Table 36: Clustering Software Help System.

Product	Help System Configuration	ACE Relevance
VisuaLinks	. We have Web-based, on our support site, which is accessible 24/7. The help consist of Knowledge-Based Articles, FAQ, and a "Help" guide which is also integrated into our software. We have People-based help, via phone and email, from the hours of 8:00 - 6:00 (EST)	Med
Clustering Engine	No Answer Provided	Low
StarProbe	Currently we can provide email (and telephone) based support only. Since we are located in Sydney, Australia, we cannot provide people-based support for the time being. It might be possible in future when we have partnership in US.	Low
Text Miner	The SAS Tech Support is 24/7 and is web-based and people-based.	High

Software Training Options

Table 37: Clustering Software Training Options.

Product	Training Options	ACE Relevance
VisuaLinks	Our training is broken into 10 days, each class is strongly recommended. VisuaLinks 2 day Intro to VisuaLinks Training. 1 day Advanced VisuaLinks Training. 2 day Modeling VisuaLinks Training. DIG 1 day End User DIG Training 2 day Administrator DIG Training.	Med
Clustering Engine	We provide web based and on-site training for the administrator of the technology. Most clients require less than four hours of training in order to have the software fully installed and functioning. No end user training is required.	High
StarProbe	We provide technical support only. Once users understand data mining and what they want to do, the rest is very simple and easy. Problems and questions can be asked as part of on-line technical support.	Low
Text Miner	Training is required, but recommended. - Predictive Modeling using SAS Enterprise Miner Software is the first course to offer. 3-days onsite is $2,850/day plus expenses for up to 20 students. Click on the title in the curriculum page for details. - Text Mining Using SAS Software has the Predictive Modeling course as a prerequisite. It is 1-day. An on-site course is $2,850/day plus expenses for up to 20 students. The same instructor can probably do both courses for a 4-day stretch of training. ** Our Training Dept. will need about 4 weeks to schedule and arrange these courses for your group **	Low

3.5 Visualization Tools

Ten clustering toolsets were evaluated during this project. The following table provides a short description and price for the tool.

Table 38: Visualization Tool Descriptions.

Visualization Tools		
Product	**Description**	**Price**
AnswerTree	Contains four powerful algorithms, the widest choice of decision trees available. Displays models visually to allow you to easily see the groups that matter. The diagrams display a snapshot of the segments, patterns, and relationships in the data that enable the user to make confident decisions. Built with scalability in mind, therefore the user can work with their data more efficiently.	$1,495.00
CoMotion	Provides sophisticated visualization components to create interactive, analytic, collaborative environments that bridge the gap between intelligence and knowledge management. Provides full access to data and a clear visual environment to explore it. Performs routine and exploratory analysis. Move data from one visualization to another.	Contact Vendor
Honeycomb Analyzer	Transforms data from a database into an information map. All data are presented in a treemap. Filters allow end-users to eliminate irrelevant data-elements. Graphical icons highlight key attributes of the data	$75,000.00
Insightful S-PLUS	4,200 data analysis functions that include the most comprehensive set of robust and modern methods available anywhere. The user can import their data, select statistical functions and display results. Easy to examine and visually explore data, run functions one step at a time and visually compare models for fit.	$2,400.00
IN-SPIRE	A discovery tool that integrates information visualization with interaction and query capabilities. Quickly and automatically conveys the gist of large sets of unformatted text documents such as technical reports, web data, newswire feeds, and message traffic. Allows the user to easily identify trends, anomalies, and relationships in huge volumes of text.	$500.00
Miner 3D Enterprise	Empowers the user to understand trends and relationships in their data. An all-visual intuitive tool, that helps the user to work effectively without extensive training. Integrated model builders automatically create charts on currently available data. Data points are visualized as graphic objects with properties capable of carrying information. Users have complete control over almost every part of the visualization space.	$1,195.00
MineSet 3.1.1	Reveals the hidden value in your data warehouse with tools for both data mining and data visualization. Allows business users to enjoy visual interpretation of complex data mining algorithms. Scalability that handles massive amounts of data. Contains visualization tools unique to the industry.	Contact Vendor
PowerAnalyzer	A robust visualization solution designed for the rapid deployment of IT and corporate data in real time through dashboards and custom applications. Accelerates implementations. Customize to unique requirements. Ensures security, performance, and scalability. Transforms data into immediate, accurate, and understandable information.	$50,000.00
Statistica	Provides the most comprehensive array of data analysis, data management, data visualization, and data mining procedures. Techniques include: predictive modeling, clustering, classification, and exploratory techniques. Offers the speed and capacity to handle datasets/designs of	$795.00

Visualization Tools		
Product	**Description**	**Price**
	practically unlimited size and unusual comprehensiveness of its procedures.	
Thinkmap SDK	Enables organizations to incorporate data-driven visualization technology into their enterprise Web application. Allows users to make sense of complex information in ways traditional interfaces are incapable of. Composed of a number of loosely coupled components that can be quickly reconfigured to fulfill many different visualization tasks. Template includes Spider, Hierarchy, Clustering and Chronology.	Contact Vendor

Only one tool, Statistica, met the criteria of exceeding over 60% in both Phases.

Table 39: Visualization Level 1 and Level 2 Ranking Results.

Tool Name	Phase 1 Results	Phase 2 Results
Statistica	67%	82%

Of the ten toolsets, the following percentages apply to the first round of questions. It is important to note that the survey team is not able to verify any of the company's claims at this time.

- 50% are compliant with the Windows Client Operating System
- 50% Windows Server Operating System
- 100% are a Web-Based Solution
- 100% are Solaris Server OS compliant
- 80% are HP-Unix Server compliant
- 60% are IBM AIX 5-2 or OS/400 Server OS compliant
- 100% provide Hierarchical Cluster Analysis
- 80% can provide customizable output

Three tools meet all criteria across the spectrum of the compliance categories above. The remaining tools breakdown as follows:

- Two tools meet eight of the nine criteria
- Four tools meet between five and seven criteria
- 1 tools meets less than five criteria

Phase 2 Questions included

- Is your software scalable?
- What are your Data Access Speeds and what testing supports your claims?
- What is your Development Code?
- What is your compatibility with core ACE software packages?
- What are your supported data types?
- What are your access control schemes
- Describe your Help desk and Training options?
- What is your data input and output processes?

Of the ten initial queries sent out, three provided responses.

The following synopsis provides a subjective overall score based on the relative rankings each tool received in the criteria provided by PM UA ACE for the Level 2 questions.

Table 40: Visualization Level Two Response Synopsis.

Visualization Tools Synopsis										
Tool	Configuration with ACE Core Tools	Scalability	Data I/O	Supported Data Types	Access Speeds	Development Code	Access Control Schemes	Help System	Training System	Overall
Statistica	High	High	High	Med	Low	Med	Med	High	Med	82%
MineSet 3.1.1	High	High	Med	High	Low	Med	Low	High	Med	79%
Miner 3D Enterprise	Low	Med	Low	Med	Low	Med	Med	High	Low	52%

Configurations with ACE Core Tools

Table 41: Visualization Configurations with ACE Core Tools.

Product	Configure with ACE core tools	ACE Relevance
Statistica	We are used in a wide variety of applications by our customers who use Windchill, Oracle, and ClearCase. With our open systems you could easily use Covera Retrieval Ware or Doors from Hologram to access our technology from Com, ODBC, OLE, SOAP or other standards. Some of our customers include Caterpillar, John Deer which are large PTC WindChill uses and we are Oracle Partners.	High
MineSet 3.1.1	MineSet provides native support for Import/Export for over 27 different flat file and statistical formats via StatTransfer from Circle Systems, Inc. This wide range of import/export formats enables transfer of data between applications such as Windchill, RetrievalWare and DOORS. MineSet supports direct connection to Oracle, Sybase and Informix running on any major platform and connectivity to ODBC-compliant data sources including SQL Server and DB2. MineSet provides an API and plug-in interface for accessing external analytic algorithms and functions plus API's to other OLAP tool vendors.	High
Miner 3D Enterprise	From the listed environments we directly support ORACLE. If your systems can be set up to deliver XML, CSV or TXT output, then even better - we can load data very easily. Also, if the other platforms are accessible via Microsoft ADO or ODBC, or via a data pipelining tools (SciTegic), then we also can access it.	Low

Software Scalability

Table 42: Visualization Software Scalability.

Product	Scalable	ACE Relevance
Statistica	Yes STATISTICA features a multithreading and distributed processing architecture delivers unmatched performance (offered in the Client-Server version) including super-computer-like parallel processing technology that optionally scales to multiple server computers that can work in parallel to rapidly process computationally intensive projects. F	High

Product	Scalable	ACE Relevance
MineSet 3.1.1	MineSet™ is architected as a multi-threaded application, which enables unequalled analytical performance.	High
Miner 3D Enterprise	Miner3D is perfectly scalable in the mean of amount of data as well as in quality of scenes. Miner3D is adaptable software, it continually monitors the computer's performance and when it detects a slower response it automatically downgrades the quality of visualization, so then the interactivity is always good.	Med

Data Input and Output Processes

Table 43: Visualization Data I/O Processes.

Product	Data input and output processes	ACE Relevance
Statistica	Our tools offer a wide range of options for automation of data input and output from a wide range of data input and output formats.	High
MineSet 3.1.1	Running MineSet in batch mode allows the software to perform operations without bringing up any visualization. Batch mode can be particularly useful in projects requiring lengthy computations that need to be done frequently. For instance, the computations can be run at night so the data will be ready the next morning. Batch mode operation is controlled/configured by scripts.	Med
Miner 3D Enterprise	No Answer Provided	Low

Supported Data Structures or Types

Table 44: Visualization Supported Data Structures or Types.

Product	Supported Data Types	ACE Relevance
Statistica	We support all Windows supported data types and we are fully OLE compliant.	Med
MineSet 3.1.1	1-2-3, Access, ASCII - Delimited, ASCII - Fixed Format, dBASE, Excel, Epi Info, FoxPro, Gauss, JMP, LIMDEP, Matlab, MiniTab, Mineset, OSIRIS, Paradox, Quattro Pro, SAS data file, SAS Transport, S-PLUS, SPSS Data, SPSS Portable, Stata, Statistica, SYSTAT MineSet supports direct connection to Oracle, Sybase and Informix running on any major platform and connectivity to ODBC-compliant data sources including SQL Server and DB2.	High
Miner 3D Enterprise	Currently we support common data types - integers, real numbers, scientific format, currencies, date, time, text strings, and links, all in a wide range of formats. From Version 5 (planned to be released in November 2004) we will support also images, pictures and textures. With chemistry-specific software libraries we will support also chemical structures, 2D drawings of molecules. Similarly we can support also other specific areas… In future we plan to support also video, sounds and speech input.	Med

Data Access Speeds & Testing

The ranking for this section is qualitative only. Two questions were asked: 1) Does there appear to be a reasonable number of test results? and 2) Did someone else provide or perform the test?

Table 45: Visualization Software Data Access Speeds.

Product	Data Access Speeds and Testing	ACE Relevance
Statistica	We have benchmarks of comparison to some of our competitors for some of our modules.	Low
MineSet 3.1.1	There is no formal industry standard for benchmarking Data-Mining and analysis tools. As the practical performance of an installation of MineSet is dependant on data type and available hardware resource.	Low
Miner 3D Enterprise	Unfortunately, I don't have access to any independent tests results. We can share with you an externally-researched study that was focused mostly on issues like ease-of-use, deployment and support.	Low

Development Code

Table 46: Visualization Software Development Code.

Product	Development Code	ACE Relevance
Statistica	C++ for Com Objects with Java WebSTAT applets	Med
MineSet 3.1.1	C++/Java	Med
Miner 3D Enterprise	The source code is written in C/C++. We also use OpenGL libraries for graphics.	Med

Access Control Schemes

Table 47: Visualization Software Access Controls.

Product	Access Control Schemes	ACE Relevance
Statistica	We use the Windows Access Control which is quite extensive but can be extended with other Access Control Methods.	Med
MineSet 3.1.1	MineSet client and server are licensed via a FLexLM license file. The license could be configured to be floating or node-locked and either open or, limited to a specific user.	Low
Miner 3D Enterprise	Miner3D is not a comprehensive application environment, but rather a user interface software technology. Applications built on/with Miner3D require integration with target operating systems, database systems that provide access control schemes.	Med

Help System Configuration

Table 48: Visualization Software Help System.

Product	Help System Configuration	ACE Relevance
Statistica	We offer service for the North America 9 to 6 CST Monday thru Friday. Our STATISTICA Help is also available online and is extremely extensive.	High
MineSet 3.1.1	MineSet ships with a comprehensive online tutorial, user guide, and reference and interface manuals. Purple Insight provides a full range of online and/or telephone based support packages.	High
Miner 3D Enterprise	By now it is a combination of online help with web and email support. We plan however to open a US-office (we are Europe-based company) and within the next few months we should have a sales office providing	High

Product	Help System Configuration	ACE Relevance
	also telephone support and consultancy services for North American customers.	

Software Training Options

Table 49: Visualization Software Training Options.

Product	Training Options	ACE Relevance
Statistica	StatSoft offers both introductory and advanced training courses in major cities in the United States and overseas as well as on site. StatSoft's training classes offer: Practical hands-on experience with the program, An introduction to real-world example applications, Energetic, helpful, knowledgeable instructors, Comprehensive take-home course manual, Personal attention, small class size, Interactive, class-paced learning.	Med
MineSet 3.1.1	The training options are: self teach, class room training or bespoke knowledge transfer: Self Teach - The step by step tutorial that ships with MineSet can give an overview of the main MineSet features in less than a day. The tutorial includes a 'further exploration' section that takes user through more advanced features, the time required depends on the users requirements. Class Room Training Course/Knowledge Transfer - Purple Insight offers consulting, knowledge transfer and training through Purple Insight consultants or partners. Training can be delivered as a standard 3 day training class or as knowledge transfer with the content tailored to the customer's requirement. Training can be delivered at the customer site or Purple Insight can provide facilities.	Med
Miner 3D Enterprise	The training for developers is not necessarily to be too extensive, because we use standards and commonly accepted methods and technologies. Usually a couple of days of email support were sufficient for effective deployment. Training for end-users depends on your application and thus you should provide it internally with your own people.	Low

3.6 XML Conversion Tools

Eight clustering toolsets were evaluated during this project. The following table provides a short description and price for the tool.

Table 50: XML Conversion Tool Descriptions.

XML Conversion Tools		
Product	**Description**	**Price**
ClearTags	Powerful research tools that present a single-screen view of complex inter-relationships as well as Web-based monitoring and visualizations enabling users to gain new insights from news and research content. Bridges the gap between text and business intelligence. Tags and extracts information form inside text that can then be incorporated into any business intelligence system or publishing database. Automatically categorizes documents and structures entities contained within the text. Generates metadata in a highly customizable XML format, which can populate datamarts, used for business intelligence in a third-party analytic tool or within ClearForest Analytics.	Contact Vendor
ECS Engine	An extensible, modular technology that can be adapted to meet the content conversion needs of any organization. All content goes through four processes in order to uncover the documents structure and generate valid XML that can then be transformed to meet specific customer needs. The XML file is designed for ease of use in XSL Transformation scripts. Can be integrated with both traditional and XML-enabled applications.	Contact Vendor
KeyView Software Developer Kits	Gives your applications the ability to interact with the widest range of intellectual capital and languages possible - up to 295 file formats in 70 languages. Converts the most file formats possible to valid XML or Web-ready HTML. KeyView Export dynamically converts documents to well-formed, valid XML using a predefined Verity Document Type Definition (DTD). The XML output can be displayed in standard browsers using cascading style sheets (CSS) or extensible stylesheet language (XSL).	Contact Vendor
RLO-Xtractor	Automatically extracts multimedia elements from your PowerPoint presentation for easy re-use. Every content element (text box, graphic, video clip, etc.) becomes an abstracted object that can be referenced by XML, resulting in truly Re-usable Learning Objects (RLOs) that can be used independently in presentations, sales material or testing applications. Single step translation of PowerPoint presentations into XML-tagged objects and files. Open XML specifications for flexible import controls.	Contact Vendor
W2XML	A Word to XML conversion software. Converts DOC, RTF, HTM, and more to well-formed XML. Completely scalable and allows for custom XML exporting by allowing you to create and apply custom XSLTs to modify the standard output. The software comes with an XSLT that allows you to output Docbook-compliant XML.	$259.95
xDoc XML Converter	Converts PDF files to XML. Converts WordPerfect into XML, PDF, or HTML. Has a point-and-click interface to make the process of transforming content from legacy formats into meaningful XML simpler. Able to manage your data conversion project without writing custom code or manually converting documents. Contains a powerful rules engine that is used to extract content from existing sources.	$2,495.00
xmlspy 2004	Contains robust, intelligent XML editing features of Text View, including code completion, syntax coloring and built-in XML validation and wellformedness checker. Has a built-in Authentic View to allow developers to create customized views and data input forms. A powerful XSLT Debugger allows a developer to troubleshoot problematic XSLT stylesheets node-by-node,	$1,248.74

XML Conversion Tools		
Product	**Description**	**Price**
	viewing node sets, testing Xpath expressions, inspecting variables, and setting breakpoints. Encompasses the entire XML development life cycle, starting with application and data modeling in WSDL and XML Schema, all the way to XML transformation, storage, and syndication.	
X-Style	Automatically converts any document form any pre-existing format into XML. Ability to realize completely automated conversions from textual formats, HTML and Microsoft Word documents into XML. An effective way to remove both obstacles from the path leading to XML integration. Word styles are mapped into XML elements and attributes according to a set of rules encoded in the customization phase.	Contact Vendor

Only one tool, KeyView, met the criteria of exceeding over 60% in both Phases.

Table 51: XML Conversion Level 1 and Level 2 Ranking Results.

Tool Name	Phase 1 Results	Phase 2 Results
KeyView	100%	66%

Of the eight toolsets, the following percentages apply to the first round of questions. It is important to note that the survey team is not able to verify any of the company's claims at this time.

- 70% are compliant with the Windows Client Operating System
- 70% Windows Server Operating System
- 70% are a Web-Based Solution
- 20% are Solaris Server OS compliant
- 20% are HP-Unix Server compliant
- 20% are IBM AIX 5-2 or OS/400 Server OS compliant
- 80% import MS Office
- 60% provide an Automated Publishing process
- 80% have Batch-processing capabilities
- 60% are Customizable to some extent

Two tools meet all criteria across the spectrum of the compliance categories above. The remaining tools breakdown as follows:

- Four tools meet between five and seven criteria
- Two tools meets less than five criteria

Of the ten initial queries sent out, two provided responses.

The following synopsis provides a subjective overall score based on the relative rankings each tool received in the criteria provided by PM UA ACE for the Level 2 questions.

Table 52: XML Conversion Level Two Response Synopsis.

Visualization Tools Synopsis										
Tool	Configuration with ACE Core Tools	Scalability	Data I/O	Supported Data Types	Access Speeds	Development Code	Access Control Schemes	Help System	Training System	Overall
KeyView Software	Med	Med	Med	High	Low	Med	Med	Low	Med	66%
X-Style	Low	Med	Med	Low	Low	Med	Low	Med	Med	50%

Configurations with ACE Core Tools

Table 53: XML Conversion Configurations with ACE Core Tools.

Product	Configure with ACE core tools	ACE Relevance
KeyView Software Developer Kits	No customizations of the KeyView Export SDK APIs are required to operate in different application environments. Oracle is a user of the Keyview filters. That means they OEM it into their product. Verity has hundreds of vendors that OEM the Verity filters like Oracle, Lotus, Documentum, etc.	Med
X-Style	Our tool does not access any third party software. In terms of integration, configuring it to work in the context of a web application is trivial, as it can be accessed either via java API or via HTTP (web service).	Low

Software Scalability

Table 54: XML Conversion Software Scalability.

Product	Scalable	ACE Relevance
KeyView Software Developer Kits	The KeyView Export SDK is a thread safe toolkit that supports scalable deployments. Multiple applications of Keyview can be implemented in parallel to support scalable deployments.	Med
X-Style	The tool is best suited to deal with large amount of data. As it presupposes manual configuration for targeting a specific DTD, it is worth only if the documents to be processed are in a relevant number.	Med

Data Input and Output Processes

Table 55: XML Conversion Data I/O Processes.

Product	Data input and output processes	ACE Relevance
KeyView Software Developer Kits	Data input and output automation is controlled through the calling application. What that means is that the calling application will pass the data to Keyview/export through our API set.	Med
X-Style	Via script, Via web input. Different forms of data input can be envisaged.	Med

Supported Data Structures or Types

Table 56: XML Conversion Supported Data Structures or Types.

Product	Supported Data Types	ACE Relevance
KeyView Software Developer Kits	CONTAINER FORMATS Microsoft Outlook (MSG) 97, 2000, 2002 (XP), 2003 DISPLAY FORMATS Adobe Portable Document Format (PDF) 1.1 (Acrobat 2.0) to 1.5 (Acrobat 6.0) GRAPHICS FORMATS AutoCAD Drawing format (DWG) 13, 14, and 2000 - extracts text only AutoCAD Drawing format (DXF) 13, 14, and 2000 - extracts text only Encapsulated PostScript (EPS) (raster only) TIFF header only Enhanced Metafile (EMF) no specific version Graphic Interchange Format (GIF) 87, 89 JPEG File Interchange Format no specific version Lotus AMIDraw Graphics (SDW) no specific version Lotus Pic (PIC) no specific version Macintosh Raster (PICT/PCT) 2 MacPaint (PNTG) no specific version Microsoft Windows Bitmap (BMP) no specific version PC PaintBrush (PCX) 3 Portable Network Graphics (PNG) no specific version SGI RGB Image (RGB) no specific version Sun Raster Image (RS) no specific version Tagged Image File (TIFF) 5 Truevision Targa (TGA) 2 Windows Animated Cursor (ANI) no specific version Windows Metafile (WMF) 3 WordPerfect Graphics 1 (WPG) 1 WordPerfect Graphics 2 (WPG) 2, 7 MULTIMEDIA FORMATS MPEG-1 Audio layer 3 (MP3) ID3 versions 1 and 2 - metadata only PRESENTATION FORMATS Applix Presents (AG) 4.0, 4.2, 4.3, 4.4 Corel Presentations (SHW) 6, 7, 8, 10, 2000, 2002, 11 Lotus Freelance Graphics (PRE) 2, 96, 97, 98, Millennium Edition R9, 9.8 Lotus Freelance Graphics 2 (PRE) 2 Microsoft PowerPoint for Windows (PPT) 95 through 2003 Microsoft PowerPoint for PC (PPT) 4 Microsoft PowerPoint for Macintosh (PPT) 98 Microsoft Project (MPP) 98, 8, 2000, 2002 (XP) - metadata only Microsoft Visio (VSD) 5, 6 (2000), 2002 (XP), 2003 - metadata only Microsoft Visio XML format (VDX) 2003 - text only OpenOffice (SXI, SXP) 1, 1.1 - text only StarOffice (SXI, SXP) 6, 7 SPREADSHEET FORMATS Applix Spreadsheets (AS) 4.2, 4.3, 4.4 Comma Separated Values (CSV) no specific version Corel Quattro Pro (QPW, WB3) 6, 7, 8, 10, 2000, 2002, 11 Lotus 1-2-3 (123) 96, 97, Millennium Edition R9, 9.8 Lotus 1-2-3 (WK4) 2, 3, 4, 5 Lotus 1-2-3 Charts (123) 2, 3, 4, 5 Microsoft Excel for Windows (XLS) 2.2, through 2003 Microsoft Excel for Windows XML format 2003 - text only Microsoft Excel for Macintosh (XLS) 98 Microsoft Excel Charts (XLS) 2, 3, 4, 5, 6, 7 Microsoft Works Spreadsheet (S30,S40) 1, 2, 3, 4 OpenOffice (SXC) 1, 1.1 - text only StarOffice (SXC) 6, 7 - text only WORD PROCESSING & TEXT FORMATS Microsoft Word all versions ANSI (TXT) all versions ASCII (TXT) all versions HTML 2.0, 3.2, 4.0 IBM DCA/RFT (Revisable Form Text) (DC) SC23-0758-1 Rich Text Format (RTF) 1 through 1.7 Unicode Text 3, 4 XHTML 1.0 Generic XML 1.0 - text only	High
X-Style	HTML/PDF/WORD/PURE TEXT	Low

Data Access Speeds & Testing

The ranking for this section is qualitative only. Two questions were asked: 1) Does there appear to be a reasonable number of test results? and 2) Did someone else provide or perform the test?

Table 57: XML Conversion Software Data Access Speeds.

Product	Data Access Speeds and Testing	ACE Relevance
KeyView Software Developer Kits	We don't have any third party tests.	Low
X-Style	We do not any independent testing. In any case performance depends on the amount of linguistic processing to be performed in order to match the target DTD.	Low

Development Code

Table 58: XML Conversion Software Development Code.

Product	Development Code	ACE Relevance
KeyView Software Developer Kits	The KeyView Export SDK is written in C and provides APIs for development in C, Java, and COM.	Med
X-Style	The system is 100% Java. CELI has full ownership on the source code.	Med

Access Control Schemes

No Answers Provided

Help System Configuration

Table 59: XML Conversion Software Help System.

Product	Help System Configuration	ACE Relevance
KeyView Software Developer Kits	N/A	Low
X-Style	People based. We follow our customers for all the duration of the software setup on a project base.	Med

Software Training Options

Table 60: XML Conversion Software Training Options.

Product	Training Options	
KeyView Software Developer Kits	Full documentation and sample code is provided in the product. Minimum training required is experience with the C, Java, or COM API's and the KeyView Export SDK documentation.	Med
X-Style	In order to use the system, no training is required, as it performs automatic XML conversion.	Med

4.0 FIST Advanced Decision Support System (ADSS)

4.1 Purpose

Decisions related to spiraling forward technologies to the warfighter are complex. There is a myriad of technologies and strategies, a large number of shareholders with related information, and capabilities are spread across U.S. Army, other DoD laboratories, other government and private Federally Funded Research and Development Centers (FFRDCs), and academia. The Advanced Decision Support System (ADSS) offers a mechanism for bringing together data that is distributed in location, diverse in content, and disparate in format. ADSS provides access to appropriate data, mechanisms for coalescing the data, tools for refining the data, and visualizations for human understanding of the data. ADSS provides effective means for gaining spiral technology investment insights and enhancing spiral technologies and R&D decision-making support.

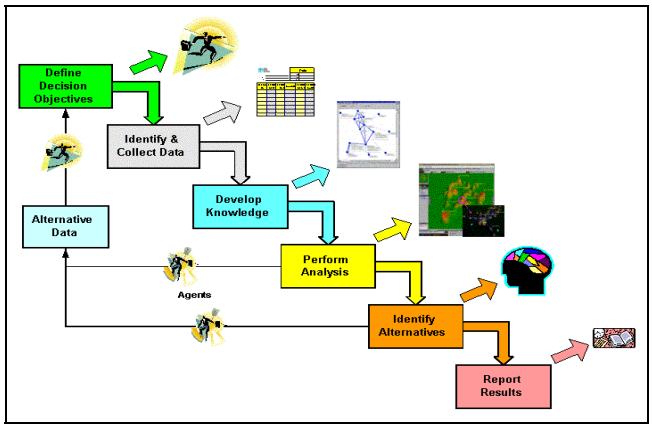


Figure 2: ADSS Process

Planning for and investing in spiral technology research and development (R&D) for ultimate transition to the warfighter throughout the product life cycle is becoming increasingly complex. There is a need to evaluate the strategies and impacts of spiral technologies prior to selection for funding and potential integration into the Army's Unit of Action (UA) program (formerly called the Future Combat System). The numerous technology investments by government, academia, and industry often overlap. One goal of ADSS is to help maximize spiral technologies and research and development (R&D) investments into the UA program while minimizing duplication with appropriate insights to the rapidly increasing volume of technology and programmatic information.

A major goal of ADSS is to assist decision-makers by providing the right information to the right person at the right time. ADSS is a transformational system leveraging advances in textual processing of non-homogenous datasets, advanced information display, and robust data mining using government off the shelf (GOTS), commercial off the shelf (COTS) and developmental software to enable robust information transfer to the system users. ADSS strives to improve program management through increased information visibility for spiral technology investments in the UA and enhance information acquisition and assimilation through advanced enabling technologies to include intelligent agents, data mining, 3D data visualization, and augmented cognition. See the ADSS Process Figure.

4.2 Benefits

ADSS provides effective means for gaining spiral technology investment insights and enhancing spiral technologies and R&D decision-making support. ADSS provides enhanced progress and program management visibility and facilitates program management across a diverse portfolio of spiral technologies and R&D programs. ADSS is currently being developed under the Program Manager, Unit of Action, Technologies (PM UA-T) office and is being used to aid in accessing and understanding diverse data relevant to the UA acquisition process. ADSS will improve the spiral technologies and R&D process by providing relevant program information to PM UA-T, facilitating enhanced spiral technology program investment portfolio strategies, and identifying spiral technology gaps and project overlaps. It will also provide visual navigation of technology views to aid management in road mapping progress toward specific goals and provide rapid updating of datasets from diverse partner sources using intelligent agents. ADSS is being developed as a tailored system to aid in making timely decisions with the right balance of information.

Integration of semantic text processing, tailored program interfaces and graphical display of textual data is in the early stages of development and poses a number of challenging issues. Current advances in augmented cognition will enable ADSS to minimize the cognitive gap between available data and human cognitive capacity. ADSS is concurrently developing architecture, to enable inclusion and integration of a family of data manipulation/analysis tools for use in managing large sets of projects potentially benefiting a number of agencies or companies. ADSS goals include enhanced human machine interface and reduced time and effort required to identify spiral technology gaps and duplication of efforts.

4.3 Architecture

There are many good GOTS and COTS tools available on the market today that perform semantic processing and visualization. It has been the goal of the ADSS team to identify an initial set of tools that provide advanced capabilities with potential for flexible integration, which are free for government use.

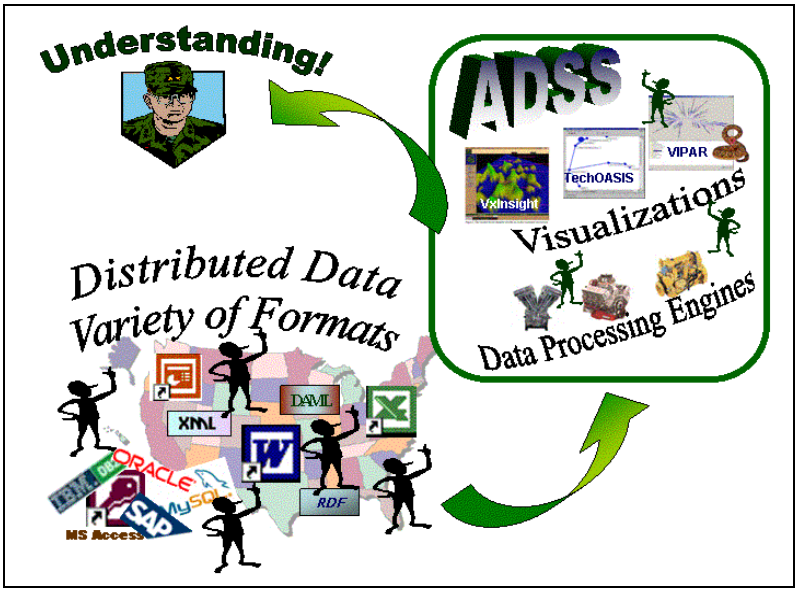


Figure 3: ADSS Architecture

The initial ADSS toolset is comprised of Oak Ridge National Laboratory's *Virtual Information Processing Agent Research (VIPAR)* system, the Sandia National Laboratories' *VxInsight* system and *Funding Analysis* tool, and Search Technologies *TechOASIS* system. Each of those tools can deal with massive volumes of distributed data and knowledge that exists in a variety of formats to include Word documents, PowerPoint files, PDF files, Excel files, and text files. See the ADSS Architecture figure above.

These documents are stored on a variety of platforms across a number of locations and organizations. ADSS was created to aid in the solution of access and understanding of all relevant data in a timely manner. The ADSS team is designing and implementing a flexible system architecture that integrates those tools along with intelligent agents for rapid data updating and conversion to meta-data for analysis. A primary ADSS goal is to provide an architecture that empowers the user by providing different cross sections of distributed datasets of interest.

For example, current ADSS tools are able to cluster data on a word, phrase, paragraph, and document level. In addition, varieties of advanced visualizations are available to aid analysts and decision makers. A demonstration version of ADSS has been completed using combined datasets that span the UA Operational Requirements Document (ORDs), the Army Science and Technology (S&T) Objectives (STOs), the TRADOC Force Operating Capabilities (FOCs), the DoDs S&T projects, and others. Current ADSS analysis allows insights into these diverse datasets that allows analysis to compare requirements (ORDs & FOCs) against technologies (STOs & S&Ts) in visualizations to help quantify gaps and overlaps of technology developments.

4.4 Interoperability

ADSS is a system-of-systems, strongly parallel to the Future Force system-of-systems approach. To the current component systems, *VIPAR*, *VxInsight*, and *TechOASIS*, other desired analysis tools may be easily added to the ADSS toolbox. While each component system, in a short-sighted view, may be seen as a competitor to the others, the goal of ADSS is to remove the barriers and allow the systems to interoperate. ADSS moves the environment from one of competition to one of collaboration and complementation. This interoperation will bring a more powerful system to the ADSS user.

By enabling the interoperation of discreet systems, we believe that ADSS will be an important component of an extensible and expandable architecture for the ACE.

The following three sections provide an overview of on-going FIST R&D efforts that can be leveraged, further developed, and tailored to provide state of the art DSS system for ACE. Those three functional areas have been identified as critical to potential DSS application on ACE and information has been specifically requested by UA ACE.

5.0 Intelligent Agents

5.1 Description of Technology

Software Intelligent Agents are increasingly becoming a desired solution for KDDM type applications due to their ability to process huge amounts of raw data. We are finding that existing technologies are facing difficulties due to limitations in scalability, mobility, and security. Agent technology provides a number of advantages in these areas, mainly through much stronger messaging and coordination models that allow cooperating agents to collaborate while distributed across multiple processing platforms. The demonstration of such capabilities in an ACE environment does not currently exist, although a number of strong agent-based systems have been deployed in related areas. The decision maker will be able to leverage intelligent agents in his information systems to assist him in filtering through the vast amount of information so that he only focuses on the most pertinent items to assist in his decision making process.

Software agents differ from conventional software in that they are long-lived, semi-autonomous, proactive, and adaptive. Cooperation among a community of agents can provide a synergy that is greater than the sum of the individual agents. Agent approaches to systems gain popularity from several of their observed advantages and characteristics but mainly from autonomy, adaptability, sociability, and mobility.

5.2 Background on Agent Technology

Agent Technology has been around for a little more than 10 years now. They have been used in many fields ranging from web intelligence to personal assistants. The main characteristics of an agent that allow them this freedom are:

1. Varying levels of autonomy
2. Degree of intelligence
3. Reliable and robust peer to peer communication protocols.

This section is a look at how the agent community has evolved in the past few years and a look at some of the major contributions to the agent community.

Early Years

In one of the earlier works in agent technology, Maes et al demonstrated how agents could be used in a framework of collaborating agent interfaces [**3**]. While a particular agent may not have any prior knowledge about something, there may exist a number of agents belonging to other users who do. Instead of each agent re-learning what other agents have already learned through experience, agents can simply ask for help in such cases.

As web browsers and the internet were becoming more and more of a household thing, an agent work in assisting users with their browsing was done in his work by Henry Lieberman [**4**]. The agent automates a browsing strategy consisting of a best-first search augmented by heuristics inferring user interest from browsing behavior.

The Remembrance Agent (RA) is a program which augments human memory by displaying a list of documents which might be relevant to the user's current context [**5**]. Unlike most information retrieval systems, the RA runs continuously without user intervention. Its unobtrusive interface allows a user to pursue or ignore the RA's suggestions as desired.

One of the first truly mobile agent systems, the Agent Tcl is described in [**6**]. Agent Tcl addresses the weaknesses of transportable agent systems by running on standard hardware, supporting multiple languages and transport mechanisms, provide transparent migration and communication. The paper describes the architecture and implementation of the Agent Tcl system.

Foner, L., investigates approaches and applications that will serve as a test bed for evaluating solutions to the problem of scaling agent systems that stand to gain from distribution over the network [**7**]. Most approaches scale poorly because they assume the feasibility of knowing about all or most of any agents peers.

Agent Technology Making Headway

Substantial work was being done in the mobile agent community and Bellavista [**8**], describes how a Mobile Agent based system could represent an interesting alternative to traditional tools built upon the client/server model. The paper describes an MA-based management system with security and interoperability as the two main design objectives.

As part of his Masters Thesis [**9**], Minar describes a system of distributed agents called Straum. The system is based on the idea of an *ecology of distributed agents* as a paradigm for building distributed software. Computers run servers that are local environments of computation. Applications are built out of agents that live in these servers. Mobile agents move to servers to use local resources and servers support agent query services to allow agents to discover each other and communicate information over the network.

Lesser et al describes a rationale, architecture, and implementation of a next generation information gathering system – a system that integrates several areas of Artificial Intelligence (AI) research under a single umbrella [**10**]. They used an information gathering agent, BIG, that plans to gather information to support a decision process, reasons about the resource trade-offs of different possible gathering approaches, extracts information from both unstructured and structured documents, and uses the extracted information to refine its search and processing activities.

Chia et al investigate coordination issues in a distributed job-shop scheduling system in which agents schedule potentially contentious activities asynchronously in parallel [**11**]. The paper formally describe two types of agent behaviors, poaching and distraction, arising from the asynchronous nature of distributed systems that decrease scheduling effectiveness, and present experimental results from a distributed airport resource management system demonstrating a significant improvement in scheduling performance when coordination mechanisms are used to prevent such behaviors.

The paper by Daniel et al presents a concise and accessible presentation of the issues associated with agent-based organizations [**12**]. The paper discusses the importance of organization design in agent-based systems and the relevance of traditional organization-design theories to software agents. It describes the need for organizational diversity at two levels:

1. In the structure of organizations
2. In the functional behavior (and structure) of the agents themselves.

Let's Browse [**13**], is an experiment in building an agent to assist a group of people in browsing, by suggesting new material likely to be of common interest. It is built as an extension to the single user Web browsing agent Letizia. Let's Browse features automatic detection of the presence of users, automated "channel surfing" browsing, and dynamic display of the user profiles and explanation of recommendations

In their research at the MIT Media Laboratory [**14**], Minar et al built systems that use mobile software agents to manage complex real-world networks. In this paper they describe a strategy for using a collection of cooperating mobile agents to solve routing problems for dynamic, peer-to-peer networks.

Singh et al were able to build a referral system for expertise location [**15**]. They developed an approach that:

1. Combines techniques from information retrieval, multi agent learning and adaptive user modeling to refine the social network according to the user's needs, and
2. Allows agents to unobtrusively exchange explicit profile information and adds coverage regarding a user's expertise.

The AGLETS mobile agent system is best described in their book "Programming and Deploying Java Mobile Agents with Aglets" by Lange et al [**16**]. The Aglets Workbench, developed at IBM's research labs in Japan, is aimed at producing stand-alone mobile agents. The complete package offers a graphical environment for building mobile agent applications in Java, an agent server, and the specification for an Agent Transfer Protocol (ATP). The experimental work discussed later has been achieved through use of the Aglets Workbench.

COLLAGE is a learning system, that endows agents with the capability to lean how to choose the most appropriate coordination strategy form a set of available coordination strategies [**17**]. COLLAGE relies on meta-level information about agents' problem solving situations to guide them towards a suitable choice for a coordination strategy.

Trust is important wherever agents must interact [**18**]. It becomes more important especially in electronic communities, where the agents assist and represent principal entities, such as people and businesses. This paper proposes a social mechanism of reputation management, which aims at avoiding interaction with undesirable participants. The approach taken towards reputation management leads to a decentralized society in which agents help each other weed out undesirable players.

The Multi-Agent System Simulator (MASS) is a discrete, event-based multi-agent simulator, providing environment, messaging, and execution and sensor services [**19**]. The Intelligent Home domain environment is also covered, as an example of how this system is used in practice. Horling et al in the paper show how (MASS) provides a concrete, re-runnable, well-defined environment to test multi-agent coordination/negotiation.

The Impulse research project at the MIT Media Lab examines what happens when the rich experience of the physical world is augmented with the low search costs and information resources available through the Internet. This paper presents a subset and implementation of one aspect of the Impulse vision: a scenario demonstrating a mobile device which uses location-aware queries to digitally augment and explore the physical world [**20**].

Morris et al examine seller strategies for dynamic pricing in an auction driven market place [**21**]. Specifically, the paper focuses on the airline industry, a field experienced in demand forecasting and dynamic pricing capabilities. It describes relevant factors when a seller dynamically evaluates incoming bids on a finite number of goods. Two adaptive pricing strategies and evaluation using a market simulator are presented.

Recent Developments

Woolf et al describe how to use software agents to build a web-based Education Marketplace that matches student requests to available and appropriate resources [**22**]. The paper discusses the open learning environment where a learner has choices; it describes how the Internet might replace the existing education monopoly and help dissolve the cottage industry of education in which a teacher handcrafts materials fixed by space and time.

An agent framework for evaluating the coordination and adaptive qualities of multi-agent systems is described by Vincent et al [**23**]. The framework "Java Agent Framework (JAF)" also allows to build different types of agents rapidly, and to facilitate the addition of new technology. There are also ongoing efforts to standardize intelligent agent technology, such as the Foundation for Intelligent Physical Agents (FIPA).

Singh et al in this article consider the problem of service location [**24**]. It describes, an approach that places the intelligence on the endpoints, enabling the users to locate desirable services based on trustworthy, personalized recommendations of their peers. The task is not only to locate a particular service, but also to locate a service that is rated highly by one's friends and associates.

A multi-agent approach for interoperation of business process in e-commerce is described by Xing et al [**25**]. The approach consists of a behavior model, a meta model that provides a language representing various trading entities and an execution architecture that supports persistent and dynamic (re)execution.

Currently service description and composition use simplistic approaches and do not accommodate interactions between consumers and providers. Cheng et al present richer representations that enable capturing more of the semantics of web services than in currently possible [**26**]. Also the paper describes algorithms that detect irregularities during the composition and execution of web services.

Within the Virtual Information Processing Agent Research (VIPAR)1 project at the Oak Ridge National Lab, CSE division, Potok et al have developed a process using Internet ontologies and intelligent software agents to perform automatic HTML to XML conversion for Internet newspapers [**27**]. The VIPAR software has the ability for intelligent agents to use a flexible RDF ontology to transform HTML documents to XML tagged documents.

The challenge to organize/classify and comprehend immense amounts of information is vitally important to the scientific, business, and defense/security communities. The VIPAR system is a multi-agent system that demonstrates the ability to self-organize newspaper articles in a manner comparable to humans. The VIPAR system demonstrates the important ability where agents use a flexible RDF ontology to monitor/manage Internet-based newspaper information [**28**]. Moreover, VIPAR extends this capability by dynamically adding/clustering new information entering the system. The VIPAR system includes thirteen information agents that manage thirteen different newspaper sites. Results from the project show that VIPAR can organize information in a way comparable to human organized information and validates the agent approach taken.

In this article Horling et al describe how different types of multi-agent organizations can be used to address the challenges posed by a large-scale distributed sensor network environment [**29**]. The high-level architecture is given in some detail, and empirical data is provided showing the various effects that organizational characteristics have on the system's performance.

Challenges for Software Agent Design

A major technical issue in designing intelligent software agents is that they need to be capable of anticipating and reasoning out information requirements of teammates involved in a highly dynamic environment, like the U.S. Department of Homeland Security.

Critical requirements for intelligent agents are the ability to capture the knowledge that an agent needs to reason, ability of the agent to manipulate the knowledge, the ability of the agent to learn as new information is presented, and the ability of the agent to communicate with other agents to share knowledge. Agent designers must have an accurate and complete representation of the knowledge that will be used and generated by the system at the time the agents are designed.

6.0 Cognitive Modeling

6.1 Basic Description of the Modeling Capability

While there are many behavioral and cognitive modeling capabilities available (e.g., ACT-R, SOAR, iGEN), we are aware of none that utilize an architecture based on electrophysiological (EEG [30]) understanding of human cognition as a basis upon which well-understood psychological phenomena are built. In addition, none of the cognitive simulations, of which we know, include entities that are aware of their spatial surroundings, nor do any behavioral simulations create a variable population of unique individuals. By way of comparison, Sandia's Cognitive Framework is an architecture based on electrophysiological behavior that underlies semantic and contextual processing in the human cognitive system (Forsythe & Xavier, 2002). This framework currently includes visual perception abilities (in virtual environments) and an understanding of the physical environment including object permanence (Wagner, et al., 2004). Because the Cognitive Framework itself is a generic representation of human cognition, we are able to populate that framework with data from unique individuals, creating models of as many unique individuals as needed.

Sandia National Laboratories' Cognitive Systems group has the capability to accurately model an individual human's decision-making and context recognition in a given domain (Forsythe, et al., 2003; Forsythe, et al., 2002; Forsythe & Xavier, 2002; Jordan et al., 2002; Skocypec & Herrmann, 2004). The basic diagram of our framework is shown below. The core components of this framework are the semantic memory (i.e., memory for basic concepts such as individual people, technologies, places, items in the environment, emotional states of one's self and others), contextual memory (collections of concepts constituting different situations), episodic memory (described below), and an evidence accumulation approach to the recognition of contexts (e.g.., presence in a fast food restaurant or interacting with someone who is becoming increasingly hostile), but more complex versions include association of context with explicit emotional processes and the inclusion of an episodic memory. Additional components are currently being added and will be discussed below. This framework has been populated with knowledge from individual experts (also called *models* for various applications and validated with regard to its faithful representation of the knowledge and cognitive processes of a given individual (Forsythe et al., 2003; Jordan et al., 2002).

Episodic memory is separate and distinct from contextual memory. Contextual memory comprises memory for generic events, such as going to a fast-food restaurant or watching a movie. Episodic memory, however, is characterized by memory for specific events –such as the time Ronald McDonald walked into the restaurant while you were eating. In the cognitive framework, episodic memory is based on *event indexing*, which provides an empirically-grounded model for how people parse their day-to-day experiences (Swan & Radvansky, 1998).

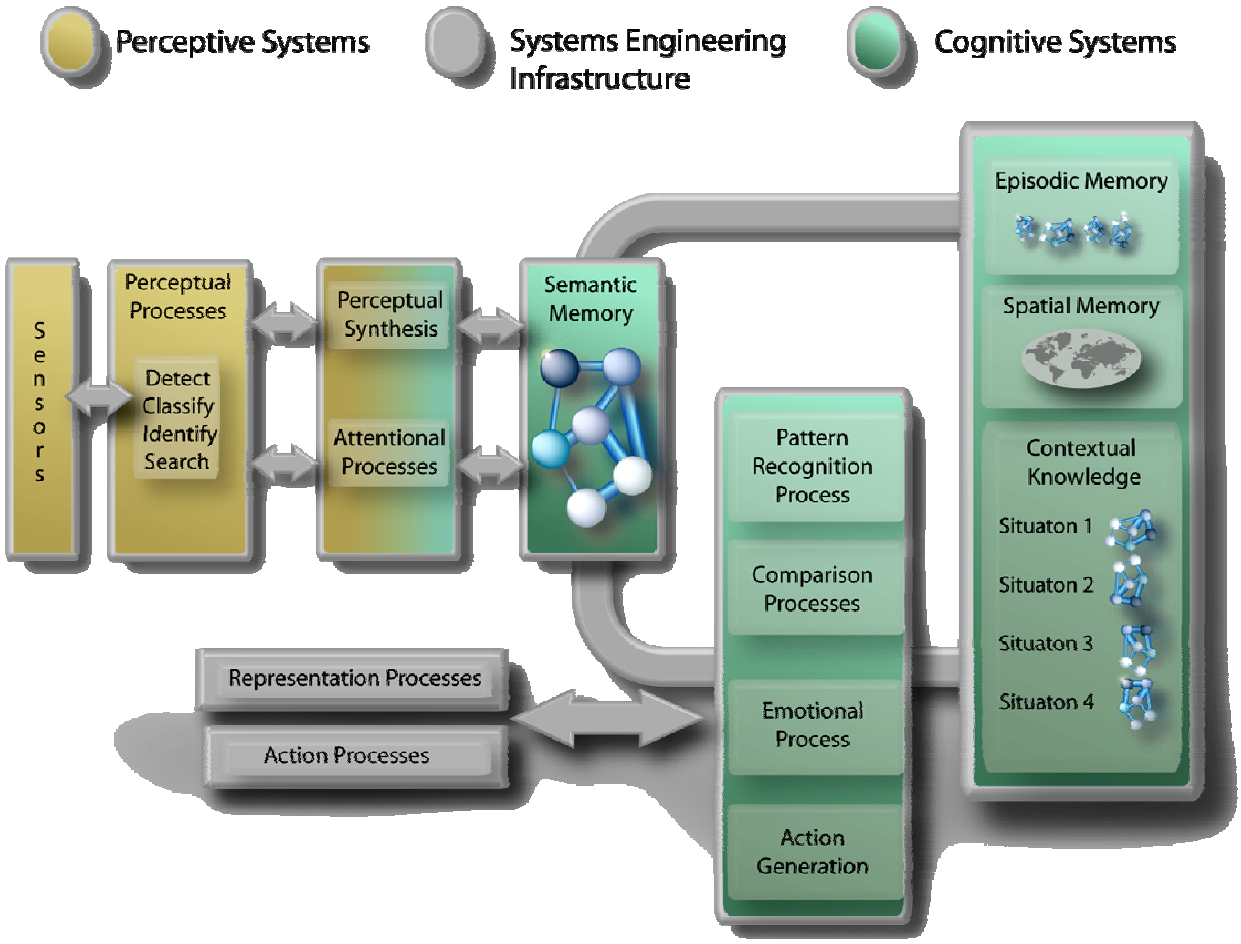


Figure 4: Diagram of Existing Cognitive Model Framework.

Its six dimensions (time, place, objects/actors, emotions, intention, and causality) provide meta-tags for the representation of and distinction between episodes. Experience is stored as a continuous series of data events with transitions from one episode to another marked by changes in one or more of the event-indexing dimensions. The magnitude of the transition (e.g. change in a single versus multiple dimensions) establishes the hierarchical structure of experience. Similarly, continuity on one or more dimensions provides the basis for threads that cut across events and mark the beginning and end of distinguishable episodes.

Modeling Expert Decision Making

We have build models of multiple experts by populating the above cognitive model framework through in person interviews with actual experts. The goal in one project was to model how certain experts would look at data collected from individuals working at Sandia and determine which individuals might possibly be an insider threat. The data collected included (but was not limited to) building access records and computer network access records. This data was not data from actual individuals, but was realistic data.

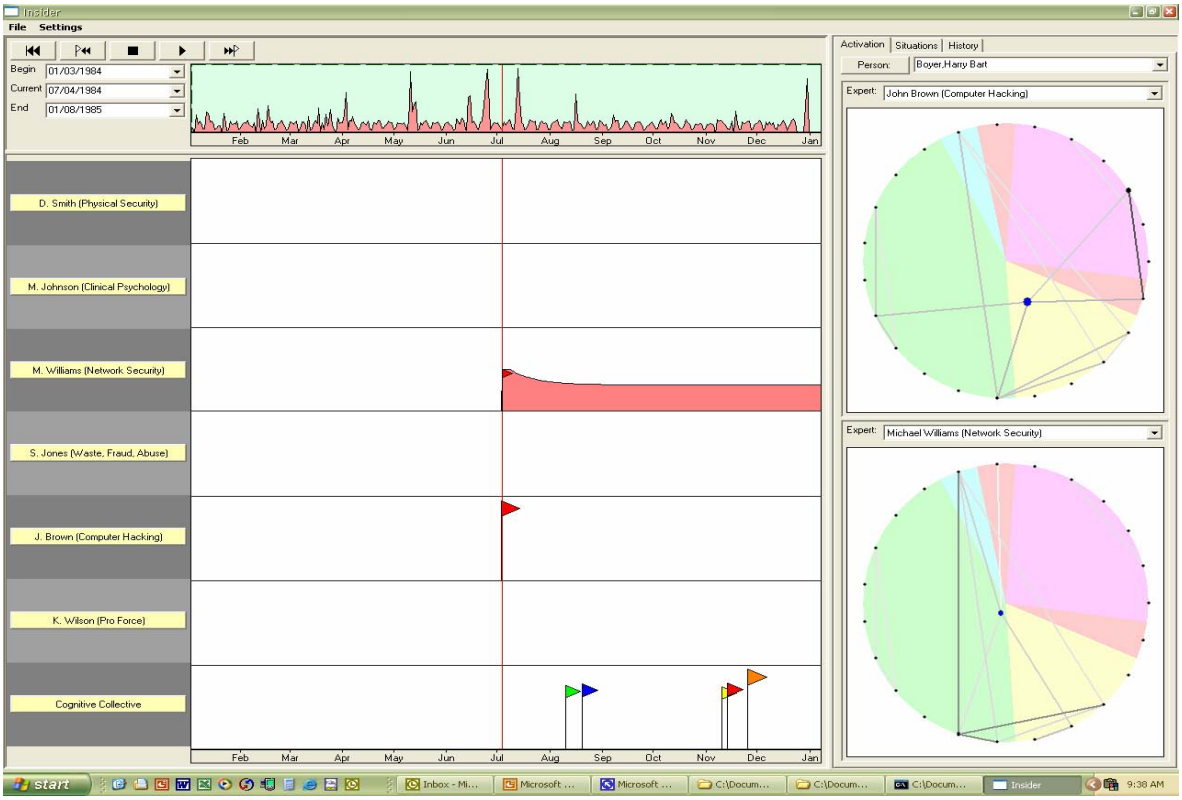


Figure 5: Insider Threat Application

We discovered that our expert models agreed with their human counterparts 90% of the time. In other words, the models could look at the data and make the same decision as the actual expert with great regularity.

6.2 Automated Knowledge Elicitation through Text Analysis

One of the major roadblocks in creating many individual models (i.e., the framework populated with information from a given individual) has been the *knowledge elicitation bottleneck*. The software framework for representing knowledge is the same across individuals and application domains, enabling us to use the same software framework for multiple domains. However, the data that populate that framework is different for each individual. In the past, we have built models through a laborious interview process in which key concepts, contexts, and the relationships between the two were identified for the individual of interest. This method, while effective, is very labor intensive.

Another roadblock is that we may not have access to a particular individual we wish to model. In this case we can build a large number of plausible individuals. This approach prevents the use of traditional knowledge elicitation methods.

During FY04 the Cognitive Systems group at Sandia developed the capability to automatically create a semantic memory of an individual using text generated by that individual (Skocypec & Herrmann, 2004). These automatically generated semantic spaces can contain anywhere from 2000 to 16,000 or more concepts and the semantic relationships between these concepts (which can number over 2 million relationships), enabling more complex models to be built, which will

help enable more complex behavior of the avatars. By way of comparison, traditional interview-based knowledge elicitation methods yield at most around 200 such concepts and 1500 relationships because of the burden traditional knowledge elicitation places on the knowledge engineer and on the person being modeled. Validation of this automated method is currently underway and should be complete around the end of FY05.

Recall that each model comprises both a semantic memory and a contextual memory. Automatic generation of contexts is in its early stages of development at Sandia. Generally, using text-based sources, this process involves identifying key concepts that occur together across clusters of documents – that is, this process identifies unique *patterns* of concept usage. Therefore, while an individual concept might occur in multiple contexts, the other concepts that occur in each context occur differentially such that the pattern of concepts is diagnostic of a context. For example, in an analysis of open-source documents and statements from Osama bin Laden, we found that he tended to use the concept *house* in two different contexts – one in which he talks about the "holy house" or the Kasbah, and the other in which he talks about the "house of idiots" or the "White House."

While these contexts were derived through human analysis of the OBL text, we have been able to automatically derive some gross contexts from user generated text. For example, one of the team members produced a corpus of text that, when analyzed, yielded seven contexts - two of which the system called "Cognitive Models" and "Sandia Cognitive." These contexts share several terms like *cognitive*, and *Sandia*. However, upon inspection, we find that the first context clearly is concerned with the technical aspects of the program, and as such includes terms like *phase, collective, cues,* and *data*. The second context, on the other hand, is concerned with the more programmatic aspects of the work, containing terms such as *staff, program, university*, and the name of a contractor to Sandia, *Orion*. Interestingly, there are several terms that contribute positive evidence to one context while contributing negative evidence to the other such as *data, research, computer,* and *experts*. These two contexts are presented in the Appendix.

Notice that there are some odd terms that show up as concepts in each of the contexts. Terms like *returns*, *interact*, *set*, and *key* probably do not constitute real concepts in our team member's cognitive model. The syntactic parsing components we are currently developing will help to eliminate these, as well as identify concepts represented by phrases, such as *cognitive model* and *tray icon*, both of which would be legitimate concepts for our team member. However, it may be that terms that appear odd on the surface are actually diagnostic of contexts. We plan to have the syntactic parsing capacity in place before work begins on the development we are proposing here.

Finally, Sandia is in the early stages of developing the capability to automatically generate these models using original-language texts (e.g., French, Italian) and in non-Roman scripts (e.g., Arabic, Farsi), reducing the problem with loss of subtle cultural information through translation. We anticipate that this capability will be ready by the middle of FY05.

7.0 DOE Secure Data Exchange

This section describes elements of the Department of Energy's Nuclear Weapons Complex ("the complex") information security practice and process. Because of the sensitivity of the data in question, information about much of what is actually done is not publicly available. In general, no business shares deep information about how it protects its information. In particular, "lessons learned" can be considered to be descriptions of vulnerabilities, even if only of past systems and practices. This may allow inference about process and technology in use that can expose vulnerabilities of existing systems. Material presented here has been approved for general release.

The complex does not use unified security architecture. In some cases cooperative practices have been agreed upon between sites that have specific communication requirements, in particular high-speed, high-volume interchange. In special cases specialized hardware and internally developed software is used to protect very sensitive transmissions.

For general business purposes, the complex uses Commercial Off-the-Shelf (COTS)[31] technology and adheres to public standards for e-mail, general file exchange, web traffic and control, and the like. COTS encryption technology, such as Entrust, is commonly used.

Individual sites use dedicated computer security staff to ensure the safety of the information at that site. These personnel watch for and respond to intrusion attempts, e-mail worms and viruses, Trojan software, and so forth. In addition, computer operations staff maintains and updates firewalls, patch individual user systems, maintain password processes, and other security elements. Security practices vary from site to site because individual sites have different functions, use different computers, have different business strategies, and are managed by different entities. It has been observed that this site-to-site variation prevents an attack that's effective at one site from being successfully applied complex-wide.

7.1 Need-To-Know (NTK)

There are two universally-applied rules for granting access to classified material.

1. The accessor must be cleared to the classification level of the document and
2. The accessor must be required in his or her professional capacity to know—i.e., must have a need to know—the information in the document.

The phrase "need to know" refers to the second of the two rules.

Y-12 uses a rule-based need-to-know process to control in-house document access within the Y-12 Electronic Data Management System (EDMS). The Sandia NTK/DACS (Need-to-Know/Data Access Control System) is part of a Need-to-Know (NTK) infrastructure that has been put in place to control access to data stored in information systems on the Sandia Classified Network. Both systems were developed at their respective sites. This is an example of two sites using different processes and technology to achieve similar ends.

The function of the NTK/DACS is to maintain a data structure that records what roles have access to what documents and what persons are assigned to each role. The NTK then grants access to an individual requesting access to information when that individual has been assigned at least one role that has a need to know the document information. In practice the NTK function is carried out by a software construct called the NTK Engine (NTKE).

Currently the NTKE is used by several different areas at Sandia. The NTK/DACS has been designed so that data controlled or owned by non-Sandia entities can be included. Currently, two different Sandia areas store data owned by other sites. NTK is used to control access to that data according to the site owner's direction and determination of NTK.

An important aspect of the NTK/DACS and the NTKE is that a general framework is being used to control access to several different independent information sets. That means that the cost of developing the framework is amortized across a large user population and doesn't need to be re-expended to accommodate new need-to-know access control requirements in other areas.

Possibly the most significant "lesson learned" from the NTK effort is that it was expensive to develop and deploy and now requires a dedicated staff of several people who fix bugs, add functionality, assist new customers, and manage the NTK software. The cost is fairly high, but the system source and experts are in-house if needed for any reason.

7.2 Cryptography

Data being transferred from one location to another must be protected if the transmission medium allows access by unauthorized parties. Most data transfer today is electronic, for which the essential protection mechanism is encryption. Encryption is the alteration of a message into a form unreadable (or nearly so) by any means but which can easily be converted back into its original meaningful form.

Many encryption mechanisms are used by elements of the complex to transmit data internally and externally. Some are freely available COTS, some are export-controlled COTS, and some are purpose-built in-house solutions for special requirements.

Special-purpose encryption/decryption hardware has been developed in-house to meet special requirements for speed, uniformity of application, and reliability under stress. Sandia National Laboratories are one of few business entities allowed to pursue cryptographic research because of their role in engineering nuclear weapon components and operations.

7.3 Automated classification

Work on automated classification has been done within the DOE. While not directly concerned with data exchange, it is necessary to know the level of classification of a document prior to any exchange, so that whether to allow the exchange can be determined. All documents being accessed by an NTK-type access control system (see above description) must have already been assigned classification level and type.

An automated classification system appropriate for use in "live" access control does not exist. This would require a system capable of understanding written language. Classification is performed today by humans specifically trained to decide whether information is classified or not. These "classifiers" assess each document proposed for release and assign it a level and type of classification.

A useful automated classifier eases the burden of determining the classification level of a given document. An automated process is especially useful for "triage": separating a large set of documents into three groups: Almost certainly classified, almost certainly unclassified, and possibly classified. Human classifiers can then scan the "almost certainly" groups to verify the automatic classifier output and devote full attention to the "possibly classified" group. The primary benefit of using an automated classifier is speed; a secondary benefit is uniformity.

7.4 Aggregation

An area of particular concern is *aggregation*, which refers to the concept that there are two pieces of information, Item A and Item B, such that A is unclassified and B is unclassified, but A and B seen together—i.e., in aggregation—is classified.

This is particularly difficult to deal with, because if either of these is ever published, the other must be withheld from publication even though it is not itself classified. In addition, because of the large number of potential combinations, it may not be possible to say *a priori* which combinations are of concern.

7.5 Agent-based Data Access

Sandia has performed research in agent-based data access which would potentially solve two problems:

1) Provide the ability for various stakeholders to obtain and change data from federated data repositories such that only specific allowed data is made available to specific stakeholders. Data could then remain under the control of its providers without the need for creating duplicated, "sanitized" databases for publication, because the sanitization takes place on demand.

2) Provide m-of-n authentication over sensitive operations such as adding a new stakeholder to an access list. Such authentication requires at least m of a group of n administrators to agree on sensitive changes, thus reducing the threat posed by a single malicious administrator.

This work has been conducted purely as research over the past several years and has not been funded at the level necessary to field a working system, although prototypes have been demonstrated.

7.6 DOE Secure Data Exchange Summary

The DOE nuclear weapons complex uses a variety of COTS and locally-developed systems to securely exchange data. Details of practice and process are ordinarily closely held to preserve operational security. Specific interchange pathways between sites are protected by security mechanisms developed for that pathway. Other pathways might use the same security solution if the original users have found it to be reliable, convenient, and trustworthy. Site-specific need-to-know (NTK) systems are used to determine whether a requester is allowed access to requested

information when it's in a system that contains sensitive information. At all sites, dedicated computer security personnel use a combination of COTS and local software to provide firewall services, encryption, authentication, intrusion detection, and virtual private networks for off-site employees.

Knowledge Discovery and Data Management Appendices

8.0 Appendix 1 - Example Commercial Survey Response Sheets

8.1 Data Mining Tools

Level One Questions		
Product	**Vendor**	**URL**
Price	**Price Explanation**	
Description		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes/No	Yes/No	Yes/No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes/No	Yes/No	Yes/No
Open Database Compliant Data Source		**Output in HTML**
Yes/No		Yes/No
Level Two Questions		
Scalable		
Data Access Speeds and Testing		
Development Code		
Configuration with Core ACE Tools		
Supported Data Types		
Access Control Schemes		
Help Systems Configuration		
Training Options		
Data Input and Output Processes		
Current Developmental Plans		
Strategic Plan		
Compare / Contrast with Similar Systems		
Toolset Strengths		

8.2 Text Mining Tools

Level 1 Questions		
Product	**Vendor**	**URL**
Price	**Price Explanation**	
Description		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes/No	Yes/No	Yes/No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes/No	Yes/No	Yes/No
XML Configurable	**Output in HTML**	
Yes/No	Yes/No	
Level 2 Questions – Same as Data Mining Questions		

8.3 Data Clustering Tools

Level 1 Questions		
Product	**Vendor**	**URL**
Price	**Price Explanation**	
Description		
Web-Based System	**Custom Output**	**Client Side Requirement**
Yes/No	Yes/No	Yes/No
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes/No	Yes/No	Yes/No
HP-Unix Server	**IBM AIX 5-2 or OS/400 Server OS**	
Yes/No	Yes/No	
Level 2 Questions – Same as Data Mining Questions		

8.4 Visualization Tools

Level 1 Questions		
Product	**Vendor**	**URL**
Price	**Price Explanation**	
Description		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes/No	Yes/No	Yes/No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes/No	Yes/No	Yes/No
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes/No	Yes/No	Yes/No
Level 2 Questions – Same as Data Mining Questions		

8.5 XML Conversion Tools

Level 1 Questions		
Product	**Vendor**	**URL**
Price	**Price Explanation**	
Description		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes/No	Yes/No	Yes/No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes/No	Yes/No	Yes/No
Import MS Office		**Automated Publishing Process**
Yes/No		Yes/No
Batch Processing		**Customizable**
Yes/No		Yes/No
Level 2 Questions – Same as Data Mining Questions		

9.0 Appendix 2 - Listing of Tools and Companies Participating in the Survey

The following tables provide a complete listing of all companies evaluated. The first table contains those companies that participated in the entire survey. The second table contains the companies that did not respond to the first questionnaire

Type Legend

DM – Data Mining Tool
TM – Text Mining Tool
VIS – Visualization Tool
CL – Clustering Tool
XML – XML Conversion Tool
l

Phase	Type	Tool	Company/Year	URL
1,2	DM	Alice d'Isoft	Isoft/2001	www.alice-soft.com
1,2	DM	Alyuda Forecaster	Alyuda Research Company/2003	www.alyuda.com
1,2	DM	Analyst's Notebook	Investigative Analysis Software/1997	www.i2inc.com
1,2	VIS	AnswerTree	SPSS Inc./2000	www.spss.com
1,2	DM	Cart	Salford Systems/2000	www.salford-systems.com
1,2	DM	ClearForest	ClearForest Corporation	www.clearforest.com
1,2	XML	ClearTags	ClearForest Corp.	www.clearforest.com
1,2	DM	Clementine	SPSS/2004	www.spss.com/clementine
1,2	CL	ClustanGraphics 6	Clustan Ltd.	www.clustan.com
1,2	CL	Clustering Engine	Vivisimo, Inc.	www.vivisimo.com
1,2	VIS	CoMotion	Maya Viz	www.mayaviz.com
1,2	DM	Data Desk 6	Data Description Inc./1998	www.datadesk.com
1,2	DM	DataDetective	Sentient Information Systems/1991	www.sentient.nl/
1,2	DM	DB2 Intelligent Miner for Data	IBM/2000	www-306.ibm/software/data/iminer/fordata/
1,2	TM	docyoument	Media Style	www.media-style.com
1,2	TM	dtSearch Text Retrieval Engine	dtSearch Corp./1991	www.dtsearch.com
1,2	XML	ECS Engine	Exegenix Canada Inc.	www.exegenix.com
1,2	TM	Enkata Enterprise Insight Suite	Enkata Technologies/1999	www.enkata.com
1,2	DM	Enterprise Miner	SAS/2003	www.sas.com/index/html
1,2	DM	GhostMiner Developer	Fujitsu/2002	www.fqspl.com/pl/ghostminer/
1,2	VIS	Honeycomb Analyzer	The Hive Group/1991	www.hivegroup.com
1,2	DM	IDOL Server	Autonomy/1997	www.autonomy.com

Phase	Type	Tool	Company/Year	URL
1,2	TM	InFact	Insightful Corporation/1988	www.insightful.com
1,2	CL	Insight Discoverer Clusterer	Temis	www.temis-group.com
1,2	DM	Insightful Miner	Insightful/2003	www.insightful.com
1,2	VIS	Insightful S-PLUS	Insightful Corporation/1995	www.insightful.com
1,2	VIS	IN-SPIRE	Pacific Northwest National Laboratory	www.in-spire.pnl.gov/index.html
1,2	XML	KeyView Software Developer Kits	Verity	www.verity.com
1,2	TM	LexiQuest	SPSS Inc./2001	www.spss.com
1,2	VIS	Miner 3D Enterprise	Dimension 5/1995	http://miner3d.com
1,2	VIS	MineSet 3.1.1	Purple Insight Ltd./2003	www.purpleinsight.com
1,2	DM	ModelMAX	ASA/2003	www.asacorp/products/mmxover.jsp
1,2	DM	NeuroSolutions	NeuroDimensions Inc./1995	www.neurosolutions.com
1,2	DM	PolyAnalyst 4.6	Megaputer/2003	www.megaputer.com
1,2	VIS	PowerAnalyzer	Informatica Corporation/2000	www.informatica.com
1,2	TM	PowerDrill	Attensity/2000	www.attensity.com
1,2	TM	Predictive Text Analytics	SPSS	http://www.spss.com/predictive_text_analytics/
1,2	TM	Readware Information Processor	Readware, Inc./1988	www.readware.com
1,2	XML	RLO-Xtractor	Multimedia Design Corporation	www.mmdesigncorp.com
1,2	DM	RoboSuite	Kapow/2003	www.kapowtech.com
1,2	TM	SemioMap	Entrieva, Inc./1997	www.entrieva.com
1,2	TM	SmartDiscovery	Inxight Software, Inc./1997	www.inxight.com
1,2	DM	Starlight	Pacific Northwest National Laboratory/2003	http://starlight.pnl.gov
1,2	CL	StarProbe	Rosella Dependable Technology	www.roselladb.com
1,2	VIS	Statistica	StatSoft, Inc./1990	www.statsoftinc.com
1,2	DM	SuperQuery Office Edition	AZMY Thinkware Inc./1997	www.azmy.com
1,2	CL	Text Miner	SAS Institute Inc.	www.sas.com
1,2	CL	TextAnalyst	Megaputer	www.megaputer.com
1,2	VIS	Thinkmap SDK	Thinkmap, Inc./1997	www.thinkmap.com
1,2	DM	Verity K2 Developer	Verity/1998	www.verity.com
1,2	CL	VisuaLinks	Visual Analytics Inc.	www.visualanalytics.com
1,2	TM	VisualText	Text Analysis International, Inc/1998	www.textanalysis.com

Phase	Type	Tool	Company/Year	URL
1,2	XML	W2XML	DocSoft, Inc.	www.docsoft.com
1,2	DM	WebQL	QL2 Software/2000	www.ql2.com
1,2	CL	WordStat v4.0	Provalis Research	www.simstat.com
1,2	XML	xDoc XML Converter	CambridgeDocs LLC	www.cambridgedocs.com
1,2	TM	XML Miner	Scientio, LLC/2003	www.metadatamining.com
1,2	XML	xmlspy 2004	Altova	www.xmlspy.com/download_spy_enterprise.html
1,2	DM	XpertRule Miner	Attar Software/2002	www.attar.com
1,2	XML	X-Style	CELI S.R.L.	www.celi.it

Phase	Tool	Company	URL
1	NeuroXL Classifier	AnalyzerXL LLC.	www.neuroxl.com
1	Optix	Mindwrap, Inc./1988	www.mindwrap.com
1	Partek Discover	Partek Incorporated	www.partek.com
1	See5/C5.0	RuleQuest/2000	www.rulequest.com
1	TeraText Product Suite	TeraText Solutions/1993	www.teratext.com
1	The Visualizer Workstation	Computer Science Innovations Inc.	www.csi-inc.com
1	TreeAge Pro	TreeAge Software, Inc./2004	www.treeage.com
1	txtkit	Schoenerwissen/OfCD/1998	www.txtkit.sw.ofcd.com
1	Weka 3	The University of Waikato	www.cs.waikato.ac.nz/ml/weka
1	Xcise	Brosis Innovations, Inc./1996	www. brosisii.com
1	X-ICE	Turn-Key Systems	www.turnkey.com.au
1	XMLCapture Suite	XMLCities, Inc.	www.xmlcities.com
1	Advanced Information Extractor	Poorva, Inc./2001	http://poorva.com
1	Advisor ToolKit	Computer Science Innovations Inc.	www.csi-inc.com
1	Affinium	Unica Corporation/2001	www.unica.com
1	Analytica Decision Engine (ADE)	Lumina Decision Systems	www.lumina.com
1	Application Foundation	Business Objects/2001	www.businessobjects.com
1	Authorware 7.0	MacroMedia, Inc.	www.macromedia.com
1	Cluto	University of Minnesota	www-users.cs.umn.edu/~karypis/cluto
1	Cognos PowerPlay	Cognos/2000	www.cognos.com

Phase	Tool	Company	URL
1	Cubist 1.13	RuleQuest/2000	www.rulequest.com
1	DataScope Professional Suite	Cygron/2000	www.cygron.com
1	DataX	Zaptron Systems, Inc.	www.zaptron.com
1	DB/TextWorks	Inmagic, Inc./2003	www.inmagic.com
1	DB2 Intelligent Miner Visualization	IBM/2000	www-306.ibm.com/software/data/iminer/visualization/index/html
1	DolphinSearch	DolphinSearch/2001	www.dolphinsearch.com
1	EnSight	CEI/1994	www.ensight.com
1	iData Analyzer	Information Acumen Corp./1998	www.infoacumen.com
1	Insightful Miner 3	Insightful Corporation/1995	www.insightful.com
1	KnowledgeMiner	Script Software Intl.	www.knowledgeminer.net
1	KnowledgeSTUDIO	ANGOSS/1998	www.angoss.com
1	LNKnet Pattern Classification Software	Lincoln Laboratory	www.ll.mit.edu//IST/lnknet/index.html
1	NeuralWorks Predict	NeuralWare	www.neuralware.com

10.0 Appendix 3 – Commercial Survey Response Sheets

10.1 Data Mining Tools

Product	Vendor	URL
Alice d'Isoft	Isoft/2001	www.alice-soft.com
Price	**Price Explanation**	
$0.00	Contact Vendor	
Description		
Discovers hidden trends and relationships in data. First to integrate an OLAP engine and advanced dynamic data aggregation functions. Uses interactive decisions trees.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No
Open Database Compliant Data Source	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
Alyuda Forecaster	Alyuda Research Company/2003	www.alyuda.com
Price	**Price Explanation**	
$249.00		
Description		
Easy business and financial forecasting and data analysis. Wizard-like interface that guides you through easy forecasting. Three different interfaces for all skill levels.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Open Database Compliant Data Source	**Output in HTML**	
No	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
Analyst's Notebook	Investigative Analysis Software/1997	www.i2inc.com

Price	Price Explanation
$3,464.00	

Description

Brings clarity to complex investigations and intelligence analysis. Turns large amounts of unrelated data into actionable intelligence. Proven to save both time and resources. Generates analysis manually or automatically form structured data.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
Cart	Salford Systems/2000	www.salford-systems.com
Price	**Price Explanation**	
$9,299.00		
Description		
Uses a decision tree tool to search relationships and patterns between data in large databases. Manageable for technical and non-technical users. Application include:telecommunication, banking, transportation insurance, health care, education, etc.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
Open Database Compliant Data Source	**Output in HTML**	
Yes	No	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
Clementine	SPSS/2004	www.spss.com/clementine
Price	**Price Explanation**	
$90,000.00	$70,000-$90,000 due to specific requirements	

Description

Turns data into better business results. Data mining tools to develop predictive models using business expertise and deploy them into business operations to improve decision making. Maximum return in limited amount of time. Clementine makes data mining a business process by focusing data mining technology solving on specific business problems.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
Yes	Yes	Yes

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
Data Desk 6	Data Description Inc./1998	www.datadesk.com
Price	**Price Explanation**	
$750.00	$650-$750 due to specifications	
Description		
Interactive graphical tools to explore and understand your data. Explores any set of data, from a few hundred to a few million. Fast computations that do several analysis's at the same time.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Open Database Compliant Data Source	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
DataDetective	Sentient Information Systems/1991	www.sentient.nl/
Price	**Price Explanation**	
$20,000.00	$20,000 Euro	
Description		
Predicts trends and detects patterns. Flexible data interface. The format is completely optimized for data mining. C++ core engine. An extremely fast mining engine.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Open Database Compliant Data Source	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
DB2 Intelligent Miner for Data	IBM/2000	http://www-306.ibm.com/software/data/iminer/fordata/

Price	Price Explanation
$74,950.00	

Description

Identifies and extracts high value business intelligence from enterprise data assets. Focuses on large scale data mining. Application programming interface that enables the development of customized, industry specific mining application. Long run mining operations.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
Yes	No	Yes

Open Database Compliant Data Source	Output in HTML
Yes	No

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
Enterprise Miner	SAS/2003	http://www.sas.com/technologies/analytics/datamining/miner/

Price	Price Explanation
$0.00	Contact Vendor, pricing varies due to specification

Description

Easy to use integrated data mining tools. Exploit an explore corporate data for tactical business advantages. Streamlines the entire data mining process from data access model assessment by supporting all necessary tasks within a single integrated solution. Combines descriptive models and algorithms. Well suited for data mining in large organizations.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
Yes	Yes	Yes

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

Absolutely. SAS has no restriction on data size, depends on your hardware.

Data Access Speeds and Testing

Again, the speed of the solution depends on your hardware and network configurations. SAS has a department, referred to as the Enterprise Excellence Center, who focuses on hardware sizing and configuration if that, would be needed. Here is a brief description: The Sizing and Configuration Program is a program which will provide SAS Sales and their customers a recommendation of the hardware required for integrating a distinct SAS solution into the customer's infrastructure. The Enterprise Excellence Center will assist with assessing the client's current operations and in collaboration with our platform partners to determine a customer's computing needs. The resulting recommendations will become part of our technical marketing reference collateral. The Sizing and Configuration Program requires that SAS interacts directly with the customer and their IT organization to understand what is needed, the relative use of SAS products and their IT's hardware vendor platform and computing standards.

Development Code

Enterprise Miner supplies complete scoring in SAS, C, Java and PMML. SAS is developed in C and C++ and the client piece of Enterprise Miner 5.1 is java. SAS itself is a 4th generation language (4GL) that is "flexible and extensible with an easy-to-learn syntax and hundreds of language elements and functions that support programming everything from data extraction, formatting and cleansing to data analysis, reporting and information delivery" (see http://www.sas.com/technologies/bi/appdev/base/). Enterprise Miner 5.1 can be extended easily with custom tools using just SAS code and XML.

Configuration with Core ACE Tools

SAS is ODBC and OLE DB compliant, making SAS flexible and allowing our solutions to access any ODBC or OLE DB compliant data source. SAS can connect to Oracle directly. http://www.sas.com/technologies/dw/etl/access/relational.html

Supported Data Types
supports any relational database; please refer to the following link: http://www.sas.com/technologies/dw/etl/access/index.html
Access Control Schemes
No Answer Provided
Help Systems Configuration
The SAS Tech Support is 24/7 and is web-based and people-based.
Training Options
Training is required, but recommended. - Predictive Modeling using SAS Enterprise Miner Software is the first course to offer. 3-days onsite is $2,850/day plus expenses for up to 20 students. Click on the title in the curriculum page for details. - Text Mining Using SAS Software has the Predictive Modeling course as a prerequisite. It is 1-day. An on-site course is $2,850/day plus expenses for up to 20 students. The same instructor can probably do both courses for a 4-day stretch of training. ** Our Training Dept. will need about 4 weeks to schedule and arrange these courses for your group **
Data Input and Output Processes
Processes can be automated via batch.
Current Developmental Plans
There is continuous development on this solution. SAS invests about 26% of total revenue into our R&D group, this is nearly twice the average investment of large software companies. This investment helps us be able to have continuous developments on all of our solutions. Customers are able to receive the latest versions as long as they are up to date on their maintenance.
Strategic Plan
The SAS Public Sector group was formed in order to focus on Federal, State, and Local government. We have had a lot of success in the government industry and several references. We work as a contractor to understand your business problems, determine a solution, and provide services via our Pilot Program. The pilot program is a way to determine the business requirements and implement/customize the solution to the organizational needs. Also, knowledge transfer via an on-site contractor and training is a very important part of the pilot program as well. We have had our clients' meet with established references to review over their solution and discuss how SAS has been beneficial. Also, we are definitely open to non-disclosure agreements if in the case that we would need to utilize your data for a demo or within the pilot program.
Compare / Contrast with Similar Systems
Please refer to the competitive overview document in my email, this provides a comparison between SAS Enterprise Miner & Text Miner to SPSS Clementine, Angoss, and IBM Intelligent Miner. I have included some general bullet points on the SAS solution below: - SAS provides flexible software that supports all steps necessary to address the business problems at hand in a single, integrated solution - Easy-to-use GUI interface helps both business analysts and statisticians - SAS provides all the components of data mining: data access, exploring, modifying, modeling and model assessment - Full control of model creation; no black box - Only solution that fully integrates text mining and data mining for more productivity - Our solution allows analysts to build more models faster which enables more collaboration between analysts - 40% of market share. . .have included an article that explains further
Toolset Strengths
Please refer to the attached documents.

Product	Vendor	URL
GhostMiner Developer	Fujitsu/2002	http://www.fqspl.com.pl/ghostminer/gm_developer.asp

Price	Price Explanation
$2,000.00	Per annual license

Description

Contains a data module, a set of models for training and testing the accuracy of predictions and post analysis tools for evaluation and visualization of the results. Object-oriented design for easy extensions of models. Requires some knowledge of data mining.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

Open Database Compliant Data Source	Output in HTML
Yes	No

Scalable

Scalability in terms of operating systems - No; This is desktop Windows application; Scalability in terms of size: Maximum number of columns - 40 000, Maximum number of records - 500 000, Maximum number of sells - 20 000 000

Data Access Speeds and Testing

Data import performance was tested on dataset created by duplicating rows/columns from reference file iris.txt. Performance tests for importing dataset with large number of columns through ODBC couldn't be performed due to MS SQL limitation - maximum number of columns in a single table is limited to 1024. ODBC tests were performed using MS SQL running locally. GM has limitation of ~ 20 000 000 cells per dataset. During import most time is spent on initialization of internal structures/memory allocation/GUI preparation (no significant difference between Text File/ODBC import).

Development Code

Borland C++

Configuration with Core ACE Tools

Not applicable - GhostMiner does not need any external software to work. We are not aware of any conflict between configurations the above mentioned software.

Supported Data Types

Numerical and categorical (binary) data of the following formats: ASCII text files (including CSV files); Excel spreadsheets; Any database conforming to ODBC standard including MS Access, MS SQL Server and Oracle; Any database conforming to OLE DB standard including MS SQL Server, Informix, Oracle and many more.

Access Control Schemes

No Answer Provided

Help Systems Configuration

12/5 - People based (email/phone)

Training Options

2 days course - basic training (allows for using a software by non-experienced user)

Data Input and Output Processes
Data preparation, feature selections, model validation, data sampling
Current Developmental Plans
1. PMML format for import/export of models (projects) 2. Models optimization 3. Web version of the software 4. New models for regression and association rules
Strategic Plan
Non-disclosure agreement is required
Compare / Contrast with Similar Systems
GhostMiner is unique data mining software from Fujitsu that not only supports common databases (or spreadsheets) and mature machine learning algorithms, but also assists with data preparation and selection, model validation, multimodels like committees or k-classifiers, and visualization. All of this and more is available in one package - a large range of data preparation techniques, a broad scope of selection of features methods and a choice of data mining algorithms and visualization techniques are integrated. This means that only one data format (project) is needed, and so trying out and comparing different approaches becomes extremely easy. The package also comes with an intuitive interface, which should make it easier to use even for non-technical user.
Toolset Strengths
No Answer Provided

Product	Vendor	URL
IDOL Server	Autonomy/1997	www.autonomy.com
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description

Platform for understanding the meaning and significance of information.Quickly processes digital information automatically and communicate with multiple applications without the need for manual processing or meta-data. Open architecture and is entirely data agnostic and scalable.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
Yes	Yes	Yes

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

While there is not a single measure of Scalabililty I have provided a detailed whitepaper on Scalability and Performance. IDOL is Multi Tiered, Modular, Distributed Linear, Provides Caching, High Performance, Cross Platform, Replicable for load balancing and redundancy, Reliable and provides central controls for administration and monitoring. We are powering some of the largest systems in the world. We are powering systems with over 10 Terabytes of Data.

Data Access Speeds and Testing

At the bottom of the Scalability and Performance Whitepaper there are results to test that where performed

Development Code

C++ & C - Backend Java, JSP, .Net - Can all be used on the front end. We have several out of the box and supported integrations for the interface.

Configuration with Core ACE Tools

Windchill - We have a Omni Fetch that can be configured to access proprietary repositories. Convera Retrieval Ware - Yes through Federated Search Oracle Database Systems - We have a Fetch specifically configured for Oracle. DOORS - We have a Omni Fetch that can be configured to access proprietary repositories. ClearCase - We have a Omni Fetch that can be configured to access proprietary repositories.

Supported Data Types

We support over 350 Data Types. I have included a Connector PDF (Connector.pdf) detailing these.

Access Control Schemes

Through our IAS (Intellictual Asset Protections System) we will honor the existing security architecture. If you have LDAP or Active Directory we can map right into them. If you are controlling Access at the application or document level we can index the ACL at ingestion which we call mapped security. The other option is before we display a document to a user we will verify they have access this is unmapped security. I have attached a Security Whitepaper which has more details. In addition to the existing security you have we also can have user and

group authentication through user name and password and control access and user privledges at the application/portal level.
Help Systems Configuration
12/7 Combined Telephone Support and Web-Based with escalation to on-site support if neccesarry.
Training Options
We have a basics training course for one week and an advanced training course which last an additional week.
Data Input and Output Processes
The Fetches are configured to automatically aggregate and index information from the designated repositories at specified times. The information is automatically retrieved when user submits a query or a query is submitted based on a users profile or the agents they have created. In the case of a profile or agent query this information is automatically pushed to the user.
Current Developmental Plans
We will need an NDA in place to discuss this.
Strategic Plan
Yes we will need an NDA.
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
I have attached our Technology Whitepaper (Autonomy-Technology) which details our approach.

Product	Vendor	URL
Insightful Miner	Insightful/2003	www.insightful.com
Price	**Price Explanation**	
$17,850.00		

Description

Highly scalable data analysis workbench that gives new and skilled analysts the ability to deploy predictive intelligence throughout the enterprise. Analyze large datasets. Accelerated discovery. Full featured support for the full data analysis life cycle.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No
Open Database Compliant Data Source	**Output in HTML**	
Yes	Yes	

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
ModelMAX	ASA/2003	http://www.asacorp.com/products/mmxover.jsp?&tc=true

Price	Price Explanation
$0.00	Contact Vendor, pricing varies to specifications

Description

Uses data mining for descriptive and predictive modeling. Handles all mathematical and statistical steps necessary for data mining. Button Repetitve model that shortens data mining cycle by 50%. No programming experience, easy to learn.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
NeuroSolutions	NeuroDimensions Inc./1995	www.neurosolutions.com
Price	**Price Explanation**	
$2,495.00	$195-$2,495 due to specifications	

Description

Many different levels available depending on what you need. Support Vector Machine to separate data into designated areas. Probing capabilities. Produces lowest possible error due to Genetic Optimization. 2 separate wizards.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

Open Database Compliant Data Source	Output in HTML
No	No

Scalable

Yes. There is no limit to the size of neural networks you can implement. With the Custom Solution Wizard, it is very economical to deploy a large number of neural network solutions, since there are no runtime fees associated with the neural network DLLs that you generate. There is no limit to the size of data file you feed into NeuroSolutions or into a neural network DLL generated by the Custom Solution Wizard.

Data Access Speeds and Testing

There have not been any independent speed tests done to our knowledge.

Development Code

Visual C++ 6.0

Configuration with Core ACE Tools

No

Supported Data Types

ASCII, Binary and Bitmap images

Access Control Schemes

OLE Automation

Help Systems Configuration

We offer email and phone support during normal business hours. The software comes with a complete help file.

Training Options

We offer a 5-day training course in Florida twice per year

Data Input and Output Processes

None

Current Developmental Plans

During the 1st quarter of 2005, we plan to release NeuroSolutions 5.0, which will include new learning algorithms.

Strategic Plan

Our strategic plan is currently in the works. Once we have this plan we would have no problem sharing it provided that we had a NDA.

Compare / Contrast with Similar Systems
Our tool is generally more flexible than our competitors.
Toolset Strengths
No Answer Provided

Product	Vendor	URL
PolyAnalyst 4.6	Megaputer/2003	www.megaputer.com
Price	**Price Explanation**	
$24,150.00	$24,150 for professional license	

Description

Incorporates the latest achievements in data mining and knowledge discovery that can analyze structured and unstructured data. Solves complicated business problems that will help make more informed decisions. Universal data mining tool: ranging fro data importing, cleaning, reporting, modeling, scoring, visualization, and manipulation. Uses decision trees and clustering.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
RoboSuite	Kapow/2003	www.kapowtech.com
Price	**Price Explanation**	
$50,000.00		

Description

Access to any application of data on the web with the ability to integrate into any environment. Has full programming capabilities, but requires very little or no programming experience. Main goal to exploit the fact many applications already have an HTML-based web interface to their functionality and data.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

- Kapow RoboSuite scales linearly with the # of CPU's deployed. Thus you can start economically with just one CPU for a few users and then add more CPU's as the # of users increases. - There is no limit to the data size that can be handled by Kapow RoboSuite.

Data Access Speeds and Testing

- Our Robot middleware typically adds sub-second throughput. Very large data sets will increase throughput time in a linear fashion.

Development Code

- Java

Configuration with Core ACE Tools

- Kapow RoboSuite can work with any system that has a web front end, has an API, accepts XML or has a database that we can access. - Support JSR 168 Portal standard plus direct interfaces with BEA WebLogic, IBM Websphere, Microsoft .Net and generic HTTP webservers. Support all Content Management Systems, Packaged Applications like SAP, PeopleSoft and Siebel, EAI vendors, etc.

Supported Data Types

- There is no limitation here.

Access Control Schemes

- Since we provide middleware this is not relevant as access control is done elsewhere.

Help Systems Configuration

- Have world wide support during normal business hours in each local time zone. There is 24/7 support via email.

Training Options

- Training is done on-site hands-on during project implementation. Typical training takes about 1 week.

Data Input and Output Processes

- Kapow RoboSuite can be called from a command line tool, and we have API's for .Net and Java. In addition, we support web services, XML output, and interfaces for BEA Weblogic, IBM Websphere and Microsoft .NET.
Current Developmental Plans
- We have a major release one time per year and a service pack release about every 3 months in between. Next major release 6.0 is due out late next year.
Strategic Plan
- We will provide this under NDA.
Compare / Contrast with Similar Systems
- The other systems are typically scripting tools with little or no Graphic User Interface.
Toolset Strengths
See attachments

Product	Vendor	URL
Starlight	Pacific Northwest National Laboratory/2003	http://starlight.pnl.gov

Price	Price Explanation
$15,000.00	

Description

Data and text mining functionality. Designed to capture and graphically portray relationships amoung multiple pieces of information of various types. Includes a geographic information system. XML conversion.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

- Starlight is a client server application that operates over a network. We haven't determined ideal or maximum client to server ratios. A lot depends on the speed of the server and the number of concurrent users. We have been recommending 4-6 clients per server. Starlight requires PCs using Windows 2000 or later. Starlight will run on a laptop, however "screen real estate" becomes an issue. Systems running Starlight require a good to high-end video card. There are license and maintenance costs associated with Starlight. - Starlight visualizations can accommodate recordsets number in the thousands to tens of thousands of records. Visualizations lose effectiveness as the number of records and complexity of relationships increases. - The upcoming version of Starlight will sit atop a repository that can accommodate millions of records and allows rapid sub setting to create sets appropriate for visualization.

Data Access Speeds and Testing

- No independent tests have been performed. Data access is good with near real-time updating.

Development Code

- C++

Configuration with Core ACE Tools

No, however we have an interface utility that allows automated data flows from tool to tool The next version of Starlight is being developed with APIs.

Supported Data Types

- xml

Access Control Schemes

- USERNAME:PASSWORD - Groups, permissions

Help Systems Configuration

- People based

Training Options

- Two days minimum training. A week with some initial support recommended.

Data Input and Output Processes
- Near real-time updating - Data export as csv, xml and bmp-screen shots - Our data manipulation utility, XEE (XML Engineering Environment) facilitates data preparation
Current Developmental Plans
- Starlight continues to evolve. A new integer release of Starlight is due out before the end of 2004. Dot releases are planned every six months
Strategic Plan
- Starlight has been developed at PNNL, a government laboratory operated by Battelle Memorial Institute, a non-profit R&D organization. We continue to evolve Starlight with government and Battelle funding.
Compare / Contrast with Similar Systems
- Starlight integrates capabilities into a single tool. ESRI GIS, Free text analysis and clustering, network representations. Directed graphs. The strength of Starlight is the ability to apply multiple models to a dataset.
Toolset Strengths
No Answer Provided

Product	Vendor	URL
SuperQuery Office Edition	AZMY Thinkware Inc./1997	www.azmy.com

Price	Price Explanation
$149.95	

Description

Data mining tool designed for Microsoft Excel and Access data files. Discover important facts hidden within business and scientific data. Analyzes sales, financial, customers, inventory data and accounting. Easy to use instructions.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

Open Database Compliant Data Source	Output in HTML
Yes	No

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
Verity K2 Developer	Verity/1998	www.verity.com
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description

Exploit and manage highly relevant, high value content.Build intelligent applications that learn form end-user actions. Accommodate your release cycle with flexible integration options. Supports multiple languages.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
Yes	Yes	Yes

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable

Indexing scalability Indexing is resource intensive. It takes many CPU and I/O cycles, allocation of large amounts of memory is required to process the documents and the LAN can get saturated. In order to be able to index all the data sources in a timely manner, a distributed indexing operation must be invoked. The following Figure shows an amplification of the Verity Spidering process.

Data Access Speeds and Testing

DIA - WISE At the Defense Intelligence Agency we upgraded a real time message handling system called SAFE from one search engine to the Verity K2 Enterprise search engine. The new system is called WISE. Here are the results: Here is actual user feedback from Arms Transfer Desk on their "WISE Evaluation" form: Once we began testing the system with complex searches related to our work, we were amazed to see how fast the search engine is. During the evaluation phase I ran simultaneous searches in WISE and WebSAFE (the legacy search system). The WISE searches were completed within a few seconds or a few minutes. The same searches in WebSAFE sometimes took as long as 30 minutes to an hour. Here are the systems statistics: running on SunFire 6800 with 8 x 900MHz CPUs, across 55M documents, XML-marked gov't messages averaging 20KB size, (Times in seconds). - worst case (word "the" present in most docs) 7.6s - slower searches (using a word present in 1% of docs) 3.4s - typical case (using a word present in 4000 docs) 2.1s - best case (using a word returning 50 docs) 1.9s - moderate search (nested and/or over ten words returning 5K docs) 5.2s - very complex search (1KB Query with boolean, proximity, and wildcards, returning 700K docs) 58s Indexing speed is 100 docs per CPU per second. This equates to 800 documents per second across 8 CPUs.

Development Code

Access to the K2 search engine is through a browser. Verity also provides a set of published APIs that allow for easy interface to the Verity K2 search engine. Many companies OEM K2 into their product to provide search. Fat Wire, Documentum, FileNet, Cold Fusion are just a few of over 200 companies that OEM Verity's search software. Verity provides APIs in the following formats, JAVA, COM, ASP, .NET and Perl.

Configuration with Core ACE Tools
Windchill - No integrations, but we would index the data directly from the underlying database; Convera Retrieval Ware - No K2 Enterprise would replace this product; Oracle Database Systems - Yes, we have many implementations indexing data from Oracle. We have a database gateway that allows Verity K2 to index Oracle data. We have hundreds of other references for Oracle. DOORS - No integrations, but we would index the data directly from the underlying database; ClearCase - Yes, Concord communications has an integration with Verity and clearcase. Oracle - The DIA WISE system explained in the "What data access speeds and testing has been performed on your software" section uses Oracle to store the data. All data is indexed from the Oracle DB. ClearCase - Concord Communications.
Supported Data Types
K2 Enterprise can index structured and unstructured repositories into a single collection. Files are stored in their native formats on many types of platforms and storage media. The file server represents one repository, while the Web server represents another. Verity K2 Enterprise allows analysts to search all types of repositories for specific information simultaneously. In a single search from a single point of access, analysts can locate documents from the following repository types using Verity Gateways: Databases (ODBC/JDBC), Lotus Notes, Documentum, Livelink, FileNet, Microsoft Exchange, Web servers (HTTP & HTTPS), File systems (NTFS and UNIX), Multimedia, Others using the Gateway Development Kit (PCDocs, InfoRouter, Plumtree, etc.). File Formats Repositories store document files in hundreds of native formats. Verity K2 Enterprise can detect, access and index over 280 of the most popular file types, including: Word processing files, such as Microsoft Word, Lotus Word Pro and Corel WordPerfect. Spreadsheet documents, such as Lotus 1-2-3, Corel QuattroPro and Microsoft Excel. Presentation files, such as Corel Presentations, Microsoft PowerPoint and Lotus Freehand. Adobe Acrobat PDF, Applix. HTML and XML. The Verity spider accesses information simultaneously on multiple Internet and intranet sites and allows Verity K2 Enterprise to explore all CGI-compliant Web servers. It supports proxy and firewall authentication, HTTPS/SSL and various login methods, including CGI, cookies and forms.
Access Control Schemes
Security Once you identify the information in your repositories and unify it with Verity Gateways, into Verity Collections, you need to ensure it remains secure. Security in Verity K2 Enterprise is flexible and easy to use. It doesn't require you to implement a new security system for your enterprise information. Rather, it enforces your native applications' security mechanisms, such as LDAP, NT or UNIX logins, leveraging your existing infrastructure and reducing administration overhead. K2 Enterprise provides two levels of security: collection-level and document-level. It also offers results list filtering so that specific users are not aware of certain documents. Collection-level Security Collection-level security limits the Verity Collections a user can search. To search a collection, a user must be a member of a particular user group. Network administrators create these user groups regularly when they set up native LDAP, NT or UNIX security models. When Verity K2 Enterprise is installed, administrators simply assign each collection to one or more user groups. In this way, K2 Enterprise knows exactly which users have permission to search each Verity Collection. For example, if you index all your human resources information into a single collection, your administrator must identify which LDAP, NT or UNIX user groups can access this collection. This way, if an employee queries for enterprise-wide salary information, K2 Enterprise searches only the collections this employee can access. Since the employee is excluded from the human resources user group, his search doesn't access the human resources collection,

and no salary information appears in his search results. In addition to providing flexible restrictions to sensitive material, collection-level security accelerates search performance by limiting the number of collections that queries are run against. This is helpful when an enterprise organizes its information in a large number of collections. Obtaining Access to Secure Collections To obtain access to a secure collection, you must authenticate to Verity K2 Enterprise. When you authenticate, you provide K2 Enterprise with the login information for your native security model. K2 Enterprise then uses this information to verify your user group. For example, if you are using an NT Domain security model, you provide K2 Enterprise with a valid NT user name, password and domain name. K2 Enterprise stores this information, authenticates you with NT, and obtains your NT group information. Only then will K2 Enterprise grant you access to your authorized collections. Verity K2 Ticket Server When you authenticate to a security model by providing valid login information, you receive a "ticket" from the Verity K2 Ticket Server. This centralized authentication service stores information in memory for users who have been authenticated to LDAP, NT or UNIX, or secure repositories. Once users end their session by logging off, their credentials are deleted from memory. Optionally, the Ticket Server can save credentials to an encrypted store to retain them from session to session. The Verity K2 Ticket Server doesn't search any repositories, but monitors search and category viewing requests on a TCP/IP port. As each request is made, the Verity K2 Ticket Server only gives users access to the collections for which they have the correct ticket. This integrates your native security model into the Verity K2 Enterprise system. The Verity K2 Ticket Server can handle authentications from multiple native security models. For example, if you provide K2 Enterprise with credentials to authenticate you to your NT Domain security model, and then use K2 Enterprise to search a collection, you are able to search documents behind NT security, but only those you have permission to read. However, a collection can include information from two separate repositories, each with its own native security model. In this case, if the collection you are searching also contains documents from a Microsoft Exchange repository, K2 Enterprise prompts you to authenticate to Microsoft Exchange. If you don't provide a valid Exchange user name and password, the Verity K2 Ticket Server denies you a ticket to the Exchange portion of the collection, and K2 Enterprise only searches documents from the NT repository. However, if you successfully log into Microsoft Exchange, the Verity K2 Ticket Server gives you a ticket to Exchange in addition to your NT Domain ticket, stores your login credentials, and grants you access to Exchange documents that you have permission to read. The Verity K2 Ticket Server stores both your NT and Exchange tickets until you end your browser session. This allows you to access different NT and Exchange information without continuously logging in. Document-level Security Verity K2 Enterprise uses authentication and the Verity K2 Ticket Server to achieve collection-level security. However, you may also want to limit certain documents within collections to specific users. In this case, you use document-level security to control whether a document appears in a results list, and whether a user can retrieve it. Document-level security uses Verity Gateways to determine a user's access rights for individual documents. Each gateway respects and enforces the document repository's existing security model. Since Verity Gateways support databases, Lotus Notes, Documentum, Microsoft Exchange, Web servers and file systems, K2 Enterprise can examine access rights for multiple models and use them to provide document-level security. No Results Filtering Verity K2 Enterprise offers different methods for document-level security. The first is "no results list filtering," in which you configure K2 Enterprise to display all documents in a results list or category, regardless of user access rights. If a user doesn't have access rights to view a document, he can see its results list information, such as its title and summary, but he can't retrieve it. This method is useful when you want users to be aware of documents, but unable to view the details

contained within them. Results Filtering The second method of document-level security is called "results list filtering," in which Verity K2 Enterprise checks each document for access rights before it displays a results list to the user. Filtered results lists and categories only show documents that a user can retrieve. Results list filtering is useful when you don't want particular users to be aware of certain documents within a secure collection. After all, a query result sometimes provides as much information as the entire contents of a document. This feature is particularly useful for those organizations concerned about covert channels. Access Control List Verity K2 Enterprise uses the access control list (ACL) to regulate security at the document level. To enhance performance, an ACL for each document is cached within Verity Collections. When users submit queries or view categories, K2 Enterprise uses the cached information to determine whether the user can access a document, instead of examining the access rights of each document in its remote repository. This approach limits calls to repositories to determine access privileges, which dramatically increases the speed with which K2 Enterprise returns results and significantly decreases the load on each repository. In some enterprises, document access rights change rapidly, requiring K2 Enterprise to re-index collections as quickly as the changes occur. During the re-indexing process, each collection updates its ACL as well. If access rights change more quickly than an enterprise wants to re-index, K2 Enterprise gives you the option to turn off ACL caching. In this case, when generating a results list, you can elect to check remote repositories for access rights instead. This slows the search and retrieval process, but ensures K2 Enterprise observes up-to-date user access rights. K2 Enterprise provides the flexibility to use cached ACLs or to check repository access rights when it generates a results list. However, when a user selects a document from a results list for viewing, K2 Enterprise always checks the repository for access rights. It does not use cached ACLs to determine whether a user can open a document for viewing. This ensures that even if a document's access rights change immediately after a collection is indexed, K2 Enterprise always applies the most current security measures before it displays a document to the user. Anonymous Access Along with collection- and document-level security, enterprises can use anonymously accessible collections and documents. This method is useful for organizations that expose information to external users. External documents are stored in a non-secure collection and are configured in LDAP, NT or UNIX to be anonymously available. If a user signs on to Verity K2 Enterprise without first logging into the native security model, K2 Enterprise only gives access to these external documents. Although this method is transparent to the end user, it allows you to designate specific documents as external information. Single Sign-on to Verity K2 Enterprise Verity K2 Enterprise enables single sign-on, in which a user provides one user name and password to access all repositories. K2 Enterprise can be configured to reuse a single user's ID and password for authenticating to all repositories. With single sign-on, users don't have to remember extra passwords, and network administrators don't have to maintain them.

Help Systems Configuration

Verity provides normal business hours phone support with purchase of K2 Enterprise. Post sales support can be provided on an as needed basis. 24 hour support can be provided for additional cost.

Training Options

All courses are online and can be done via the web from your office.

Data Input and Output Processes

Verity provides a web based spider that can automatically index data from any of the Verity K2 supported data types. Please see the "What are your supported data types" section for a complete description of the Verity spider.

Current Developmental Plans
Verity typically provides two major releases each year. The releases provide enhancements to the Verity K2 product line. Currently we are in the process of adding new entities to the entity extraction module (from 14 to 30 out of the box entities) and implementation templates to simplify the implementation of the product.
Strategic Plan
Yes we do. A non-disclosure agreement is required.
Compare / Contrast with Similar Systems
K2 Enterprise is the only product on the Market that indexes, extracts entities, provides full text search, exports data to HTML or XML, creates taxonomies using 5 methods and has a collaborative taxonomy tool with version control and access control. All of these features are completely integrated into the K2 product. Implementation risk is reduced because the integration is already done. Additionally Verity is a profitable company with 6 strait years of profitability. We have over 200 million dollars in the bank. Contrast that to our competitors like Convera (Retrieval Ware) who has lost money for 3 strait years, just announce their most recent quarter as a 7 million dollar loss. They only have 19 million dollars left in the bank. That's less than 1 year of viability. Please go to the Yahoo finance site to get the specifics about Verity and Convera financials by going to the following sites: Verity - http://finance.yahoo.com/q/cf?s=VRTY&annual Convera - http://finance.yahoo.com/q/cf?s=CNVR&annual Additionally Convera announced that they are now going to compete directly with Google and Yahoo by indexing the internet. Not sure where they are going in the intranet portal market. Please see the Convera earnings release at http://biz.yahoo.com/bw/040818/185639_1.html for a complete description of their most recent quarter and their strategic plan.
Toolset Strengths
See Attachment_1

Product	Vendor	URL
WebQL	QL2 Software/2000	www.ql2.com
Price	**Price Explanation**	
$18,400.00	$18,400 first year, $15,800 subsequent years	
Description		
Data mining tool designed for the World Wide Web. Extracts data from unstructured data sources. Reformats data into structured formats. Designed for business professionals looking for market trends, studying consumer behavior, or doing scientific research.		

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	No

Open Database Compliant Data Source	Output in HTML
Yes	Yes

Scalable
- Tools must be economically deployable in a wide range of sizes and configurations. Answer: Yes, our pricing model allows for different levels of licensing depending on the firepower needed and the size of the implementation. Any application developed can be run on more powerful versions of WebQL and is only a matter of licensing to scale up. I will include a price list of current annual licensing fees, licensing service add-on options, and custom development fees, - Amount of data (size) Answer: Limitations are set by your hardware and bandwidth.
Data Access Speeds and Testing
We are most interested in independent tests and an explanation of the test criteria Answer: We have a real world implementation with our secured QL2 Client Center where we host developed applications for our clients. The QL2 Client Center consists of a cluster of parallel servers that currently processes 25GB of data per day from the web. In the last 17weeks, the QL2 Client Center has processed over 3 Terabytes of data.
Development Code
WebQL is developed using C/C++. The language used for query development is similar to Oracle's implementation of SQL
Configuration with Core ACE Tools
We have both a Windows and Linux version of WebQL and support various APIs. WebQL supports interfaces for programmatically controlling query execution for C++, Java, .NET and ActiveX programmers. The ActiveX API supports a variety of Windows based environments, including Visual Basic, Delphi, and Microsoft Office. In addition, a SOAP interface is available for accessing WebQL Server from any programming language. In addition, WebQL will allow you output and input through ODBC. WebQL has not been specifically configured to work with the mentioned systems (exception: WebQL is configured to work with Oracle Database Systems) but we will assist with any implementation that may be necessary if a viable application is presented that require these systems. We also support all standard input and output file formats (xml, csv, xls, etc.). Due to the nature of our business and relationships with our clients and business partners, we are unable to disclose information that may be deemed to be confidential.
Supported Data Types
We support all standard data types (xml, csv, xls, html, pdf, doc, tsv, images, databases, etc.)
Access Control Schemes
We need clarification on this question. I can say that the QL2 Client Center is extremely secure and requires biometric security clearance to access the data center.
Help Systems Configuration
Technical Support at QL2 is available via phone during normal business hours and via email 24/7. Technical support is people based and correspondence is on a personal level due to the application specific issues that are usually addressed. Full help files are available for WebQL via the WebQL Studio Interface or via download files.
Training Options
We have two training options. OPTION 1 : Training at QL2 Facilities (2 day introductory and 2 day advanced training sessions) Introductory training sessions will give the student a strong basis to further explore query development using WebQL. Advanced training is geared toward individuals who have some background in SQL, PERL regular expressions, and web page layout. Advanced training is usually geared towards customized application development for the client. OPTION 2: Training at the Clients Location; similar to objectives outlined for option 1 except training occurs at the Clients Location. The client must provide the facility and equipment necessary for the training sessions. There is no minimum training necessary in certain cases and we have had clients use WebQL without any formal training with success.

Data Input and Output Processes
We can support any number of process which include, but are not limited to, ftp, html post, etc. for our Client Center Apps. WebQL can retrieve data from any number of disparate data sources dependent on the structure of the developed application.
Current Developmental Plans
NDA required
Strategic Plan
NDA required.
Compare / Contrast with Similar Systems
Please forward us a list of competitor's products so we can compare features. I can state that query and application development using WebQL Technology can greatly reduce development and maintenance costs for your applications.
Toolset Strengths
Please see attached appendices.

Product	Vendor	URL
XpertRule Miner	Attar Software/2002	www.attar.com
Price	**Price Explanation**	
$3,225.00		
Description		
Full data mining process addressed. Uses ActiveX Technology. Extensive data information, visualization and reporting features. For both skilled and unskilled data miners. Supports the discovery of association both "case" and "transaction" data.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
No	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Open Database Compliant Data Source	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		

Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

10.2 Text Mining Tools

Product	Vendor	URL
Docyoument	Media Style	www.media-style.com
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description

Discovers knowledge from tons of unstructured textual data to get short and pregnant information for later operational, marketing or general strategic decision making. Uses text-classification algorithms that will classify all news messages with a small distance to these samples to sort them into those folders. Provide the possibility to get a short summary of a document or even all documents in a folder.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes

XML Configurable	Output in HTML
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
dtSearch Text Retrieval Engine	dtSearch Corp./1991	www.dtsearch.com

Price	Price Explanation
$2,500.00	

Description

Provides access to dtSearch indexed and unindexed search options and hit highlighted file display features. Includes extensive support for existing fields in documents, as well as support for adding on-the-fly classification information and other fields to documents during indexing. Supports SQL and other COM or non-file data.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

XML Configurable	Output in HTML
Yes	Yes

Scalable

The current released version of the dtSearch Text Retrieval Engine can index from 4-8 GB of text in a single index, and there is no limit on the number of indexes that you can build and simultaneously search. Version 7.x of the dtSearch Engine will be able to hold a quarter of a terabyte or more in a single index. And there is still no limit on the number of indexes that you can build and simultaneously search.

Data Access Speeds and Testing

We have over a hundred third-party developer case studies at http://www.dtsearch.com/casestudies.html. Many of these third-party case studies, such as http://www.dtsearch.com/CS_Premirus_CC.html, cite search times of less than a second, through hundreds of gigabytes or more of data. Our own testing confirms such results. But we encourage you to browse our case studies section for third-party affirmation, or try the dtSearch Engine yourself (downloads are attached to the email message).

Development Code

The dtSearch Text Retrieval Engine comes in two versions: one for Win & .NET and one for Linux. The dtSearch Engine for Win & .NET supports Delphi, Java, C++, C++.NET, C#, VB.NET, ASP.NET & more. The dtSearch Engine for Linux supports C++ and Java.

Configuration with Core ACE Tools
Convera is a competitor of ours; it is unlikely that you could use the dtSearch Engine on their indexes and vice versa. A large portion of our developer case studies mention developing with SQL, XML, and other database formats. Example 1: For a sample SQL case study, please see http://www.dtsearch.com/CS_Premirus_CC.html (as well the above-mentioned journal article). Example 2: A sample XML usage would be Battelle's use of the dtSearch Engine in its XML-based Starlight Information Visualization System, developed at the Pacific Northwest National Laboratory. But for most other database-type programs, integration would be easy. Please see, for example, http://support.dtsearch.com/faq/dts0111.htm and http://support.dtsearch.com/faq/dts0183.htm, describing integration with a wide range of database types. There is a third-party journal article that directly compares a somewhat earlier release of the dtSearch Engine with some other database search engines. Here are some excerpts from this article: With regard to indexing performance, "dtSearch ran considerably faster than MS-SQL ... dtSearch took from 3.5 [to] 4 hours, whereas MS-SQL took from 20 [to] 28 hours." With regard to search performance, of 14 test queries, "dtSearch was faster in all of the queries except two" and "documents returned by dtSearch are always equal to, or a superset of, the documents returned by MS-SQL." Further, "There were no false positives" with dtSearch. A chart comparison of dtSearch with MS-SQL, Oracle, and IBM DB2 also found that only dtSearch had all 17 search features listed. "Based on the above benchmarks, we elected to use the dtSearch engine for text indexing and searching of data." If you wish to review the full article, it is posted at http://www.dtsearch.com/CS_JAMIA.pdf
Supported Data Types
dtSearch supports all of the following document types: "Office" (word processor, database, spreadsheet, presentation, etc.), emails, HTML, PDF, XML, SQL, ZIP, CSV, RTF, ANSI, Unicode files, and more. Please see http://support.dtsearch.com/faq/dts0103.htm for a full list of support file types, and for additional information on our handling of "binary" data.
Access Control Schemes
Please see http://support.dtsearch.com/faq/dts0127.htm and http://support.dtsearch.com/faq/dts0179.htm for sample answers to that question.
Help Systems Configuration
In addition to providing extensive online and offline documentation, dtSearch Corp. provides developer technical support by phone, by email to tech@dtsearch.com, as well as through a developer user's group. For the latter, please see http://www.dtsearch.com/dt_subscribe.html#other. (The developer's forum is the second on the list.)
Training Options
We sell the dtSearch Engine as an "out of the box" developer product. We do not provide training, although our developers are certainly available for technical questions. For a list of technical FAQ articles which you may find helpful, please see http://support.dtsearch.com/faq/default.htm. We also maintain on our Web site a long list of third-party companies with development experience in the dtSearch Engine, many of which maintain security clearance. Please see http://www.dtsearch.com/customDevelop.html for details. I would be happy to put you in touch with some appropriate options in this regard, if that would be helpful.

Data Input and Output Processes
In addition to supporting all of the file types above, the dtSearch Engine includes a data source API for non-file data. The dtSearch Engine also includes multiple options for exporting search results, in a wide variety of Web-based and other formats.
Current Developmental Plans
Please see above for some information on Version 7.x. For additional information on the Version 6.5 dtSearch Engine beta, please see http://www.dtsearch.com/beta.html
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
Please feel free to browse our case studies section; many of these describe benchmarking dtSearch against a large number of other tools, and choosing the dtSearch Engine. Many of our case studies also describe customers that have undertaken a "competitive upgrade" to the dtSearch Engine. The above-mentioned third-party journal article also compares the dtSearch Engine with various database-specific indexing products. You may also find our publication reviews page useful. Please see http://www.dtsearch.com/dtreviews.html
Toolset Strengths
The dtSearch Engine tends to excel in indexing and searching a very large quantity of data. We encourage you to test it out in full on any databases that you may have. Attached is download information to try a fully-functional copy of the dtSearch Engine for Win & .NET. (Please let us know if you would like the dtSearch Engine for Linux downloads too.) Also attached is our most recent product line press release, as it contains a good overview of the dtSearch product line's over two dozen search options and other features. For more information, please see our online "features map" at http://www.dtsearch.com/PLF_Features_2.html Two white papers that you might find useful are also attached. One of these is on XML-based distributed searching. The other is on different methods for fielded data / full-text searching options using the dtSearch Engine. Additional white paper and other PDF document downloads are available at http://www.dtsearch.com/PLF_F_download.html

Product	Vendor	URL
Enkata Enterprise Insight Suite	Enkata Technologies/1999	www.enkata.com
Price	**Price Explanation**	
$0.00	Contact Vendor	
Description		
Addresses problems of business processing modeling, unstructured data analysis, cost allocation, and project portfolio management to deliver unprecedented time to value. Enables you to model, analyze, and optimize interdependent business processes and systems across the enterprise. Has the ability to perform text classification on free-form text documents. Its framework is designed to interpret the content of these text documents and assign a probability score that represents the likelihood of the document belonging to a specific category.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No

XML Configurable	Output in HTML
Yes	Yes
Scalable	
No Answer Provided	
Data Access Speeds and Testing	
No Answer Provided	
Development Code	
No Answer Provided	
Configuration with Core ACE Tools	
No Answer Provided	
Supported Data Types	
No Answer Provided	
Access Control Schemes	
No Answer Provided	
Help Systems Configuration	
No Answer Provided	
Training Options	
No Answer Provided	
Data Input and Output Processes	
No Answer Provided	
Current Developmental Plans	
No Answer Provided	
Strategic Plan	
No Answer Provided	
Compare / Contrast with Similar Systems	
No Answer Provided	
Toolset Strengths	
No Answer Provided	

Product	Vendor	URL
InFact	Insightful Corporation/1988	www.insightful.com
Price	**Price Explanation**	
$250,000.00		
Description		
The knowledge access solution that moves intelligence analysts from key word search to event discovery. Its powerful analysis of text content includes new search operators that move beyond competitive knowledge access solutions. Offers human-like understanding of text, dramatically reducing the time knowledge workers spend gathering and analyzing information, vastly improving knowledge discovery, transfer, and decision making. Produces easy-to-read concept maps that quickly summarize intelligence from vast amounts of information across many documents.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No

XML Configurable	Output in HTML
Yes	Yes
Scalable	
No Answer Provided	
Data Access Speeds and Testing	
No Answer Provided	
Development Code	
No Answer Provided	
Configuration with Core ACE Tools	
No Answer Provided	
Supported Data Types	
No Answer Provided	
Access Control Schemes	
No Answer Provided	
Help Systems Configuration	
No Answer Provided	
Training Options	
No Answer Provided	
Data Input and Output Processes	
No Answer Provided	
Current Developmental Plans	
No Answer Provided	
Strategic Plan	
No Answer Provided	
Compare / Contrast with Similar Systems	
No Answer Provided	
Toolset Strengths	
No Answer Provided	

Product	Vendor	URL
LexiQuest	SPSS Inc./2001	www.spss.com
Price	**Price Explanation**	
$0.00	Contact Vendor	
Description		
Analyzes text with a high degree of accuracy-and about 4,000 times faster than you can. Ability to process more than one gigabyte of text, or approximately 250,000 pages, per hour. More accurate than other text mining solutions because it is based on natural language processing technologies. Uncovers concepts contained in large collections of text and displays them in a color -coded graphical map so that analysts and business users can clearly see relationships among them. Processes text in many common formats, including plain text, HTML, XML, PDF, and Microsoft Office documents formats.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No

XML Configurable	Output in HTML
Yes	Yes

Scalable
1. SPSS Inc.'s textmining technology can process on average 1GB of open text per hour on a standard PC (2Ghz, 512 MB RAM). Using Grid Computing technology, SPSS Inc.'s textmining technology can process 60GB of open text in 3 hours. 2. There is no specific limit on the size of the raw data being accessed or with the number of concepts extracted from that data. Plans for a 64 bit architecture in January 2005 will improve scalability to an even higher degree.
Data Access Speeds and Testing
1. SPSS Inc. has several customers that have performed their own tests. Most notably, Pfizer is a company that uses SPSS Inc.'s technology for bioinformatics research. Tests are done on accuracy of extraction and speed. Pfizer was highly satisfied with the results they were able to achieve. Accuracy ranged from 85% to 90% in extracting relevant gene patterns with a speed of 60GB (15 million documents) in 3 hours.
Development Code
1. SPSS Inc.'s textmining technology is written in C and C++. The user interfaces are written in Java 1.
Configuration with Core ACE Tools
1. SPSS Inc.'s textmining technology works with the drivers available for configuring to a number of search engines. Creating a driver for a new search engine, such as Convera, is a fairly simple process. One of SPSS Inc.'s customers, Peugeot, has embedded Convera into their textmining environment and the driver for that configuration was developed in about an hour. 1. Many of SPSS Inc.'s customers have unique environments. SPSS Inc. has successfully configured our textmining technology to work in each of these environments. Our open and scalable applications reduce complexity and the time needed for integration. 2. SPSS Inc.'s textmining technology works with an ODBC driver for connecting to Oracle or any other database. 3. One important point to note is that because SPSS Inc.'s textmining technology is natively integrated with one and other, and all belong and are developed by SPSS Inc., configuring them to work within an organization's specific environment is rendered relatively simple compared with offerings comprised of several components from different vendors.
Supported Data Types
1. The ODBC driver allows SPSS Inc.'s textmining technology to read data from any database. It can also access raw data from pdf, html, xml, txt, bibliographic format, and MS Office.
Access Control Schemes
1. Access control schemes can be based on MS Identification.
Help Systems Configuration
1. SPSS Inc. offered technical support for all it's software by the following means: phone, email, and web-based. 2. 24/7 support can be set up based on the complexity of each customer usage scenario. Any MS Office, pdf, html, xml.
Training Options
1. The minimum training we offer is 2 days, the most is 7 days. Knowledge transfer sessions can exceed that timeframe depending on the complexity of the defined challenge. 2. Onsite Training - SPSS Inc. will send a trainer to your facility, or reserve one of our public training facilities, to train any number of users on customized content relevant to their environment. 3. Public Training - users can attend a Public Training course at any SPSS facility nationwide. 4. SPSS Consulting - SPSS Inc. will send a consultant to your facility to address a specific challenge. Knowledge transfer will take place as the consultant will teach users as the solution is being developed.

Data Input and Output Processes
1. Command mode and API are available for managing input and output. Files containing information on extracted concepts and relationships are automatically generated.
Current Developmental Plans
1. New language integration to include Arabic, Chinese, Somali, and Hindi 2. Tighter integration between textmining and data mining (structured data) workbench 3. Additional easy to use UI for configuration of dictionary files 4. 64 bit architecture
Strategic Plan
1. SPSS Inc.'s strategic plan, with special emphasis on textmining within the Public Sector (to include Intelligence, DOD, and DHS) can be shared once a non-disclosure agreement is in place.
Compare / Contrast with Similar Systems
1. SPSS Inc.'s textmining technology is the only application currently available that allows users to combine the power of data mining (structured data) and text mining in an environment that is open, scalable, and easily configurable by non-experts. 2. Our textmining offerings are automatically deployed…unlike some of our competitors who provide only desktop solutions. 3. Most of our competitors have an offering at the "named entity extraction" level, meaning they can pull out concepts from documents and type each concept into a category. 4. A few competitors are at the "fact extraction" level, meaning they can pull out events and relationships between concepts that occur in the text. 5. SPSS Inc.'s text mining technology can handle the previous two points, but goes a step further. Our powerful extraction and pattern matching capabilities present users with the most relevant concepts, relationships, and events that occur in a set of unstructured text data at a speed unsurpassed in the industry (1GB per hour and up to 60GB in 3 hours). 6. Users can easily create new categories and patterns to find the relationships and events in the text that are most relevant to their given scenario. 7. We also allow users to easily build statistical models through an integrated UI that turns extracted information into actionable intelligence. Without an easy to integrate combination of data mining and text mining, users will get stuck mid level in terms of data analysis - unable to reach the goal of using their information for predicting future events.
Toolset Strengths
1. SPSS Inc. has over 20 years of experience in computational linguistics and currently have over 500 satisfied customers using our textmining technology. We have developed dictionaries for several domains, including Public Sector, bioinformatics, and call centers. In addition to these dictionaries, we have open and user-configurable dictionaries which can be tuned and customized to a specific organization's needs. All this without having the customer rely on SPSS Inc. for assistance. This allows for proprietary and classified information to remain confidential within the organization. 2. We can currently work with English, French, Spanish, Italian, German, Dutch, and Japanese. Development plans are underway for the incorporation of Arabic, Chinese, Hindi, and Somali. 3. An additional strength of our textmining technology comes from the strong integration among the tools that work together to provide customers with an end-to-end solution in predicting future events and taking intelligent action. Our applications are open, scalable, easy to configure, and give the users the power not only to understand the meaning held within their text, but most importantly, to use that information directly in making decisions that affect their organization. SPSS Inc's unparalleled synergy of text and data mining gives organizations a way to predict most accurately, and therefore have the most effect on the events that shape the future landscape.

Product	Vendor	URL
PowerDrill	Attensity/2000	www.attensity.com
Price	**Price Explanation**	
$50,000.00		

Description

An advanced information exploration and retrieval application that allows analysts and researchers to drill quickly and deeply into written information and uncover important patterns and relationships. Uses sophisticated linguistic technologies to extract the individuals, actions, and objects described in free-form written text. Quickly drills into written information and categorizes all actions, actors, and objects. Replaces slow and expensive manual analysis and tagging of written documents.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No
XML Configurable	**Output in HTML**	
Yes	Yes	

Scalable

Attensity PowerDrill is a web-enabled text query tool which operates against an enterprise database repository and scales for large user communities. Attensity Relational Extraction Server (RES) is a scalable text extraction engine that scales to take advantage of large machine architectures. Architectures featuring multiple CPUs on one machine (Symmetrical MultiProcessor or SMP) are supported as well as architectures featuring multiple physical machines.

Data Access Speeds and Testing

Benchmark data regarding RES ingest speeds have been performed internally only, and our estimates indicate that RES is capable of ingesting and processing 1-2GB of raw text per 24 hour period per CPU. These tests were done on a Windows platform with Pentium 1+GHz CPUs. If the processing load exceeds 1-2GB per 24 hour period, multiple CPUs can be employed as the software is scalable.

Development Code

C++ for the RES engine in conjunction with Java.

Configuration with Core ACE Tools

RES supports Oracle out of the box. More detail and analysis would be required to better understand the other tools and how you would want them to work with RES. RES is very extensible and offers the ability to be customized to a particular environment via the creation of custom pre and postprocessors, therefore working with other tools is seldom a problem.

Supported Data Types

Document types out-of-the box include PDF, Word, RTF, ascii, .txt, html, and email. Additional document types are supported via the inclusion of custom converters. For example, we are supplying a Word Perfect (.wpd) converter for a current Government customer.

Access Control Schemes
Attensity RES requires user login and authentication. Multiple user accounts are common in an RES environment and each may be granted roles which provide different levels of access controls for varying tasks. More detail is available upon request. Attensity PowerDrill also requires the user/analyst to log on before querying and exploring any text, therefore different datasets can be protected via username and password schemes.
Help Systems Configuration
Support is available via telephone and pager as well as online bug reporting capabilities. Standard support is 5x8 hours. Additional support options such as 24x7 are available upon request.
Training Options
Generally speaking a 1 week training course is required to become proficient with the Attensity product set. This includes 1 day of training on PowerDrill and 4 days of training on RES and the Knowledge Engineering process.
Data Input and Output Processes
Attensity RES supports an API where users can submit text and receive extractions as output in asynchronous fashion. However, RES is most often utilized in automated batch mode where text is collected and entities and events extracted and output in hands-off, automated fashion.
Current Developmental Plans
The current release of RES is 3.0.2. The next near-term release will be 3.1 and will include enhanced support for creating PowerDrill datasets housed in an enterprise database such as Oracle. Otherwise, there are interface improvements and general minor bug fixes included in 3.1. The next release subsequent to 3.1 will be 4.0 and will be a major release with many new enhancements. The feature list for 4.0 has yet to be decided upon however and is therefore not releasable at this time. PowerDrill's current release level is 1.6. The next release will rename the product to Attensity Discover 2.0. This release, due out in November of 04, will offer a web-enabled architecture, scalability for large concurrent user populations, and a number of ease of use enhancements to the interface.
Strategic Plan
Yes - this information would be made available to the government and it's integrators via Non Disclosure Agreements (NDAs).
Compare / Contrast with Similar Systems
Attensity differs from it's competitors with regard to the deep linguistic parse which we perform on the input text. This patented approach represents a fundamental breakthrough in converting unstructured text into structured tables with a high degree of accuracy. Rather than using statistical and probabilistic algorithms to extrapolate representations of meaning from word content and proximity, Attensity's technology understands English language by using computational linguistics to parse sentences into fundamental linguistic elements, and then analyzes these elements using sophisticated algorithms. This approach results in the following benefits: - Unprecedented accuracy allowing the extraction of events and entities from within documents, not just categorization of documents themselves; - Identification of an event's attributes, i.e. what participated in an event, and how; - High raw text throughput exceeding 1MB/minute on 1GHz Intel CPU; - Features designed to handle noisy text: misspellings, poor grammar and bad punctuation.
Toolset Strengths
No Answer Provided

Product	Vendor	URL
Readware Information Processor	Readware, Inc./1988	www.readware.com

Price	Price Explanation
$250,000.00	$99 to $250,000

Description

Intelligently analyzes the text it reads. Transforms each text into a compact mathematical format that is stored as a signature record in a readware collection while it is indexing it. Allows users to interrogate and compare the information in readware collections in ways that are simply not possible with statistical, information-theoretic, NLP or so called AI models. Capable of analyzing and indexing tens of thousands of pages an hour.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	Yes	Yes

XML Configurable	Output in HTML
Yes	Yes

Scalable

Yes. Readware is economically deployable in a wide range of sizes and configurations. There is no theoretical limit to the numbers of documents or records that can be indexed in a single Readware installation by Readware's finite balanced-tree indexing methods. The Readware IpServers, optimized for web and intranet application, are scalable by users and by indexed documents. The site with the most users (average 1.5 million/day) using Readware as a search engine for 10,000 pages of highly dynamic content is on-line at http://www.adac.de. Statistics show that a single Readware query server, in a cluster of seven P5 servers, processes up to 250 queries per hour. ADAC also runs a separate Readware installation on their Intranet where they have hundreds of different collections of information that need updating and re-indexing on a daily basis. The ADAC intranet has more than one million documents in the collections, though it is all centrally organized and managed from a single pair of services for querying and indexing their information. For very large installations, the Readware Analyst, the indexes and collections (supporting metadata) and the Readware Query processor can all be distributed (and still inter-operate) across heterogeneous networks on the same of different hardware platforms.

Data Access Speeds and Testing

Management Information Technologies, Inc. (MITi) the corporation that has produced Readware technology, has participated in two NIST examinations of text-retrieval effectiveness. The results of these examinations have been published. The latest TREC paper is available at: http://trec.nist.gov/pubs/trec8/papers/READWARE.pdf Some customers have issued criteria on which performance was tested. ADAC of Germany, for example, tested the throughput of the Readware Query processor as that was an important factor in their speedy system. They found that the throughput of the query processor on an 800 mhz PIII with 512MB of main memory was 83 queries per minute. These tests were performed by the lead programmer at ADAC. Another customer required that we be able to classify a query (answer the question: To what category or classification does this query pertain?) in one-half second or less given 100,000 categories from classification headings in yellow pages. In another case we had to be able to identify cursing or name-calling in instant messages, and highlight and annotate the pertinent parts of message, while being delivered.

Development Code
ANSI-C with some C++ classes
Configuration with Core ACE Tools
Readware has been used with Oracle and with many content management applications. Readware has an API and accepts HTTP-style commands over a TCP socket, so it is very easy to query a Readware collection or "feed" files or records to the Readware Analyst for indexing and classification. The results of Readware's analysis and classification (the metadata) for any document in the collection, can be queried and utilized independently by automated processes, so configuration usually present no technical difficulty. In Germany, the newspaper publisher Springer uses Oracle for sorting news articles and the relevant data: byline, publication, page number, date and section, and some other data. Recently they wanted to see if there were better ways to support editors and writers to use and refer to stories in their archives. Of course the front office "content management and personal access" system was based more on XML and used web-type devices, while the back-end office was a typical Oracle IT shop. We showed that Readware could index up to 2,800,000 records in the archives in a few days and that the incremental indexing could easily handle the roughly 150,000 to 600,000 new records each day. We also showed that that indexing included extensive topic analysis and classification and how these services could support their editors and historians. While our interface initially did not support their DTDs, or use parameters like daterange, we were able to read their DTDs, agree on a formal specification of special search parameters, and provide an HTML-form search and retrieval interface for a pair of Readware servers on their Sun enterprise systems with four processors and 2GB of RAM. It took about two man-weeks to create an interface that automatically read, indexed and classified the XML-output generated by Oracle for their content management and personalization system.
Supported Data Types
When you say data type I infer that you mean as in a spreadsheet where one has number, dollar, decimal, text, etc. So on this understanding, Readware is capable of the following data types - out-of-the-box - so to speak. First, everyone should understand that Readware is mainly for unifying access to unstructured text-based information, in addition to or in conjunction with other software. Text is the main type. Text is further parsed and recognized as: Constant (names, nouns, places, things, acronyms). These may be processed phonetically (Jon, John, Joan are similar) or literally, and, Word (an inflected, conjugated, plural, singular term or verbal stem), Concept (a member of group of related words), Super-concept (a member of group of related concepts), Idiom (noiseWord and word/constant ([this way], [that way]). Readware recognizes numbers, and it also treats the words one, two, three, "two thousand and four" as the numbers 1,2,3 and 2004 repectively. The number_places, ones, tens, hundreds, and thousands are indexed. Readware handles Dollars and Euro's as an instance of a topic called money. And that brings us to the point that any data type that can be defined can be specified as a Readware Knowledge-Type by administrators/users of the system using plain language and a few text files. Several hundred knowledge-types are delivered with Readware. These are delivered as Topics, Issues, Probes and Categories. A Readware Topic is composed of one or more queries that specifically identifies and usually also disambiguates some data. For example, if I wanted a topic called stocks and bonds, I could create it by declaring it and making at least one query. I could just query for the keywords 'stocks' (boolean) and 'bonds' but it turns out that would only skim the surface and it would include irrelevant information like Barry Bonds the baseball player. A Readware topics lets you specify in plain language in a readable text file that you want to look for corporate stocks and not livestock, or stock of a store and that a T-Bill is acceptable as a bond. Barry Bonds could be specifically

excluded or (easier, better) the category of sports could be excluded, having the run-time effect of excluding any sports related documents from being analyzed by the query at all.

Access Control Schemes

We have no specific access control schemes rather we support whatever access control scheme is imposed upon us. For most TCP-standard access control for fetching documents on a network we utilize the public domain "curl" functions. We support NTLM in Microsoft environments. FYI: The Query server requires read-only access to files outside its own operations and collection logs. The analyst server requires administrator style rights on most machines and is usually run only by administrators from within secure environments.

Help Systems Configuration

24/7 people-based via email describes us best. We are experimenting with wiki tools and other collaborative environments and will probably move the developer's kit manual to that environment, once we decide. Until then we encourage email and offer telephone, and web-based conferencing where and whenever practical.

Training Options

There is no minimum training requirement. Readware is configured in a straight-forward way and there are no interdependencies outside the network and/or http server. A developer's kit manual and release notes contain all the information necessary to install and operate the software. We offer on-site installation and operations training in one, three and five day packages. The one day course requires a very smart systems engineer that has intimate knowledge of installing software applications on the platform. The engineer will learn how to install Readware on that platform. He will learn the parameters of the Readware initialization file and the work of creating, incrementing, controlling, and maintaining Readware collections and services. The three day course covers the above in fine detail for people that know basic computer operations and how to install software on their machine. These people could then install Readware servers on other machines at the end of three days. The five day course is for developers or larger-enterprises that intend to integrate Readware into other systems. This course covers everything in the three-day course in addition to installation on clusters, using multiple Query processors in the same or different machines. How to distribute the index and signature database and other intricate technical configuration issues will be covered in this course. In addition to these basic courses on the technical use and operation of a Readware installation, we offer one and two week course on the principles and best practices for building Readware topics, filters and classifiers using Readware Knowledge-Types and information cultures.

Data Input and Output Processes

Any automation that can inter-operate using TCP/Ip standards and protocals. In addition customers have created ASPs, JavaBeans, DLLs and all manner of other programmatic automations encapsulating., exporting and/or integrating Readware functionality and/or exposing Readware analysis, filtering or classification results. We also have console tools that can be automated using shell-scripts and batch files that may be preferred by some IT shops.

Current Developmental Plans

Our plans are currently focused on improving the GUI tools to create and institute topics and classifiers/filters and incrementally adding new knowledge to topics and classifiers we deliver to customers. Otherwise we plan to develop the Readware services to the needs and specifications of our customers.

Strategic Plan
As the original developers of Readware technology our strategic plan is simple enough, however a non-disclosure agreement would suffice to reveal the details of our perception of our technological position, our progress to date and our future plans.
Compare / Contrast with Similar Systems
The Readware Ip Servers compare favorably to Verity's indexing and enterprise classification system, Autonomy's offerings and Microsoft's indexing products for enterprises.
Toolset Strengths
In addition to being scalable the multi-threaded Readware IpServers are reliable and readily adaptable to unstructured information identification, classification, browsing, navigating and querying in a low cost easily hosted and administered software package.

Product	Vendor	URL
SemioMap	Entrieva, Inc./1997	www.entrieva.com
Price	**Price Explanation**	
$25,000.00		
Description		
It's graphical interface gives users the ability to scan text collections-no matter how big-in a matter of minutes to understand the overall content of the documents. Ability to quickly find out more about a specific topic, seeing how and where concepts are related and the strength of these relationships. Advances search capabilities lets you find documents using document attributes such as text, title, author, abstract, creation date, URL, etc.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No
XML Configurable	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		

Data Input and Output Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
SmartDiscovery	Inxight Software, Inc./1997	www.inxight.com
Price	**Price Explanation**	
$0.00	Contact Vendor	
Description		
Uses the most comprehensive set of advanced text analysis tools on the market, including search, entity extraction, fact finding, categorization, and visualization. Allows users to quickly find and employ the precise, relevant information needed to get their jobs done more effectively. Allows you to identify the relationships and links between people, organizations, and other entities inside unstructured data sets.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No
XML Configurable	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		

Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
VisualText	Text Analysis International, Inc/1998	www.textanalysis.com
Price	**Price Explanation**	
$50,000.00		

Description
Ideal tool for quickly developing accurate and fast information extraction, natural language processing, and text analysis systems for the most complex needs. Enables you to build analyzers that can be maintained and enhanced by non-programmers and non-linguists. Automatically generates new rules and layers them into the analyzer framework. Ideal for text analysis applications to combat terrorism, narcotics, espionage, and nuclear proliferation. Ability to find important nuggets of information in voluminous texts.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

XML Configurable	Output in HTML
No	Yes

Scalable
Analyzers can process both large (multiple MB) and small text inputs, as well as large and small numbers of inputs. Because analyzers focus on intensive and deep analysis, typical speed for an accurate Analyzer is about 0.1 seconds per 1000 characters of text. TAI is committed to providing economically feasible deployments on both the small and the large scale, to ensure affordable deployment. Costs are agreed to based on the business value of the application.
Data Access Speeds and Testing
Compiled analyzer speed is typically on the order of 0.1 seconds per 1000 characters. Analyzers may also interact directly with databases via ODBC-based interfaces Independent testing of VisualText has been performed to ensure commercial-grade functionality and consistency with documentation.
Development Code
TAI's NLP++ programming language is used for the development of Analyzers. For optimization and deployment, the definition of an Analyzer, i.e., its passes, rules, and code are compiled to C++ code. Under the hood, VisualText, runtime libraries, and Analyzers are written in C++: Microsoft Visual C++ and Gnu C++ are available for Analyzers, while VisualText itself is built in MS Visual C++ with MFC and CodeJock's XTreme library. VisualText runs on Windows platforms, while Analyzers may run on any platform with Gnu C++. A bridge for calling Analyzers from MS .NET code is available as well.

Configuration with Core ACE Tools
Analyzers use ODBC calls to work with databases including Oracle. Analyzers compile to DLL libraries on Windows and .a archive libraries on Unix/Linux, which can be called from any application via a programmer's API. The API may invoke text analyzers with files, buffers, and streams. The API also enables calls to update and access the Conceptual Grammar knowledge base associated with an Analyzer.
Supported Data Types
Analyzers can in principle accept arbitrary binary input files. However, the primary data input types are human readable texts, such as XML, HTML, plain text, and RTF. Conversion to text is readily available for formats such as PDF and PS.
Access Control Schemes
As an Analyzer is a componet, rather than an end-to-end solution, this question does not apply.
Help Systems Configuration
Our support is telephone-based and email-based. Visits to customer and developer sites can be arranged as well. Help is provided on a seven-day-per-week basis.
Training Options
TAI provides training tailored to the needs of the customer, focusing on the particular customer application. Minimal training consists of following the documentation and tutorials provided in the VisualText Help. Self-training is available via our well-reviewed Help documentation, which includes a set of Tutorials to be read and executed. Sample analyzers are provided both for study purposes and for use as initial Analyzers for customization. A general Analyzer and a resume analyzer are available for free download at http://www.textanalysis.com/Apps/apps.html In addition, VisualText installation comes with other examples, such as an Analyzer that connects to a database, and a business-events analyzer.
Data Input and Output Processes
Analyzers may accept text via files and buffers. Stream input can easily be wrapped around an Analyzers. Analyzers produce output to files, buffers, and streams, and combinations of these. Additional input and output is readily available via ODBC connectivity to databases.
Current Developmental Plans
Current plans include the development of vanilla applications utilizing the power of VisualText. A world-class part-of-speech (POS) tagger is currently under development, to be followed by a Categorizer, Summarizer, and more advanced version of TAIParse (our combined tagger, chunker, and parser). Following the development of general purpose Analyzers, starter Analyzers will be developed for particular domains and applications, including business, financial, and medical.
Strategic Plan
As we are a relatively small startup company, our plans include ramping up operations, revenue flow, and financing. We have recently completed a full development and sales cycle with IBM, and expect to continue to expand and broaden that and similar relationships.
Compare / Contrast with Similar Systems
VisualText is the first commercial-grade IDE for natural language processing. While competitors provide substantial off-the-shelf analysis capabilities, such capabilities are typically hard-wired or black-box. By contrast, our offering is unique in enabling developers to customize, prototype, enhance, and maintain their own text analysis capability to an unprecedented degree. VisualText further provides a framework for integrating multiple and diverse text analysis capabilities within a single system.
Toolset Strengths
No Answer Provided

Product	Vendor	URL
XML Miner	Scientio, LLC/2003	www.metadatamining.com
Price	**Price Explanation**	
$4,499.00		

Description

A system and class library for mining data and text expressed in XML, extracting knowledge and re-using that knowledge in products and applications in the form of fuzzy logic expert system rules. Predicts numeric values, categorize and classify data, infer the relevance and topics in text, and mines the structure of XML documents. Integrates text mining seamlessly so that blocks of embedded text can be handled at the same time as numeric and categorical data.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes

XML Configurable	Output in HTML
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
Predictive Text Analytics	SPSS	http://www.spss.com/predictive_text_analytics/
Price	**Price Explanation**	
$0.00	Contact Vendor	
Description		
Predictive Text Analytics fro SPSS Inc. enables you to combine structured and unstructured data to draw more accurate conclusions about future events and actions. By knowing not only what happened, but why, you can make better decisions for the future.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
XML Configurable	**Output in HTML**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
ClearForest	ClearForest Corporation	www.clearforest.com
Price	**Price Explanation**	
$200,000.00		

Description

Reads vast amounts of text. Pinpoints relevant information. Uncovers relationships. Provides visual, interactive analytical and executive summaries. Manages information overload. Turbo-charged search.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
XML Configurable	**Output in HTML**	
Yes	Yes	

Scalable

The ClearForest solution can be deployed in a stand-alone workstation configuration which provides entity and relationship extraction and visualization/analysis application. ClearForest can also be deployed in a client/server configuration with multiple CPUs, used in parallel, to process large data amounts. ClearForest has been used in environments that store petabytes of input data. The ClearForest solution is scalable to allow multiple ClearTags CPUs, running in parallel, to handle these large data sources.

Data Access Speeds and Testing

Information Extraction speed for the recommended configuration with ClearForest products Version 6 is as follows: Performance (KB/min). These results were obtained for creation of XML files on P4 2.8GHz single CPU HT PC with 2GB RAM on Window 2003 Server. Basic, 800 Intelligence, 240 Business, 180 Biological, 190 Patents, 1400

Development Code

The ClearLab Development Environment uses DIAL4, an OO Based language, to customize ClearTags and the Discovery Modules.

Configuration with Core ACE Tools

Windchill (Yes), Convera Retrieval War (Yes), Oracle Database Systems (Yes), DOORS (Yes), ClearCase (Yes). ClearForest has been architected to provide an open API, to allow applications to interface with the ClearForest outputs, via web services. Our entity tagging and relationship extraction tool has been integrated with the following types of applications: Visualization Applications (I2 Analyst Notebook, Visual Links, …), Databases (Oracle, SQL Server, Sybase, DB2…), Portals, Custom Applications, Government Developed Applications (Starlight)

Supported Data Types

We support all data types relating to textual base files such as: Word, PDF, Powerpoint, Excel, .RTF, .TXT, Adobe, HTML, XML, Any ASCII text file Clearforest also provides a data access layer, which allows the user to handle proprietary formats and inputs from data bases.

Access Control Schemes

ClearForest does not have its own security functionality. We typically use the client's access control schemes for database and system security.

Help Systems Configuration
We offer telephone, web-based and people-based technical support. Our standard hours for support our 8 AM - 6PM, Monday - Friday. We offer additional 24/7 premium support for customers requiring additional coverage.
Training Options
We offer formal training at our facility as well as at the customer's site. We recommend the following classes: ClearResearch Analyst Training - 1 Day ClearTags Administration Training -2 Days ClearLab Development Environment Training - 3 Days
Data Input and Output Processes
ClearForest provides documented APIs for all data input and output applications. Once data has been inserted into a ClearForest folder, it will automatically by ingested by ClearTags, and the tagged data will be automatically sent to either a database or application. The ClearForest APIs are used to create interfaces with other commercial or government applications.
Current Developmental Plans
The ClearForest Solution version currently shipping is Release 6.1. We typically release two major releases and two minor or point releases per year.
Strategic Plan
Details of our strategic product plan can be shared with the government under a signed Non-Disclosure Agreement.
Compare / Contrast with Similar Systems
ClearForest's biggest differentiators from other entity and relationship extraction tools is: - Ease of Use - Best Return on Investment (as documented by the Joint Forces Command Evaluation of December 2003) - The ability to customize extraction rules through a development environment - The use of Domain Specific extraction rules called Discovery Modules
Toolset Strengths
A white paper and product brochures discussing how entities and relationships are extracted, using ClearForest, are attached for your review. Also shown on the next page is the ClearForest Solution Architecture.

10.3 Data Clustering Tools

Product	Vendor	URL
ClustanGraphics 6	Clustan Ltd.	www.clustan.com
Price	**Price Explanation**	
$375.00		
Description		
Offers hierarchical agglomerative cluster analysis, k-means analysis, focalpoint clustering, outlier analysis and proximity analysis. Ability to construct a cluster model from a hierarchical cluster analysis and then classify any number of new cased by reference to it. With the Clustan Wizard your clustered in three clicks. Data mining in a particular strength which can cluster 200,000 cases or more hierarchically and run k-means on a million cases using a PC.		
Web-Based System	**Custom Output**	**Client Side Requirement**
No	No	Yes
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	Yes	No
HP-Unix Server	**IBM AIX 5-2 or OS/400 Server OS**	
No	No	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
Clustering Engine	Vivisimo, Inc.	www.vivisimo.com

Price	Price Explanation
$50,000.00	

Description

Automatically organizes search or database query results into meaningful hierarchical folders completely on-the-fly, out-of-the-box. Automatically clusters search results into categories that are intelligently selected from the words and phrases contained in the search results themselves. Categories are always up-to-date and as fresh as your content. Human level accuracy, and a familiar intuitive folders-style interface.

Web-Based System	Custom Output	Client Side Requirement
Yes	Yes	No

Windows Client OS	Windows Server OS	Solaris Server OS
Yes	Yes	Yes

HP-Unix Server	IBM AIX 5-2 or OS/400 Server OS
No	No

Scalable

Our Clustering Engine supports search indexes of well over 30 million documents. Our Content Integrator software for metasearch scales to 200 data sources. There is no limitation on the number of simultaneous users.

Data Access Speeds and Testing

In terms of our Clustering Engine and Content Integrator software, the data access speed has been tested by Vivisimo, and confirmed by each member of our client base. Following are the rates at which are Clustering Engine and Content Integrator process and displays search results: Clustering: 150ms to cluster 200 results 400ms to cluster 500 results Content Integrator: 300ms to metasearch 12 sources and cluster 200 results Note: Vivisimo can cluster a maximum of 1000 search results, and metasearch over 200 data sources.

Development Code

Vivisimo software products are written in: C, C++, ASP, C#, JSP, Java or Perl. Though for most applications, no dvpt at the API level is required, just configuration through the web interface extensively based on XML and XSL.

Configuration with Core ACE Tools

Vivisimo software has been configured to work with Convera & Oracle Database System. Although we do not have direct experience with Windchill from PTC, DOORS from Telelogic or IBM's ClearCase, We can work with any product having a web interface. Vivisimo has successfully deployed its technologies atop of Convera search, Oracle Databases, other web based data management products and combinations there of. A specific case of integrating multiple web based data systems and Convera search would be our implementation at Johnson & Johnson where we meta search over twenty sources, several of which use Convera.

Supported Data Types

All standard MS office types, PDF, compressed files, HTML, XML, DB files, etc.

Access Control Schemes

We can mimic any scheme of an underlying application. Provide a username/password repository. Support user groups.

Help Systems Configuration
No Answer Provided
Training Options
We provide web based and on-site training for the administrator of the technology. Most clients require less than four hours of training in order to have the software fully installed and functioning. No end user training is required.
Data Input and Output Processes
It is unclear exactly what you are asking, however our software can fetch content automatically where ever it is and output it in RSS, HTML or XML.
Current Developmental Plans
Our near term development plans include extending the Clustering Engine beyond the current languages we currently support (English, European, Arabic and Korean) to all the world's major languages (Japanese, Chinese, etc.) as well as extending the capabilities of Vivisimo's enterprise crawler in terms of scale, access control, and others.
Strategic Plan
In terms of company direction, yes we do have a strategic plan and would be open to discussing it with the government so long as a non-disclosure agreement is in place.
Compare / Contrast with Similar Systems
The greatest advantages to Vivisimo's technologies over our competitor's offerings are that it is effective in helping users find information with out the expense of manual methods of categorizing information, our intuitive interface requires no end user training, the system is flexible in terms of interoperability with other search technologies and with minimal total cost of ownership we provide the highest RIO in the industry. All products are completely web based.
Toolset Strengths
For further details regarding technical aspects of the Vivisimo product and the product in general visit the following web pages and use the menu items on the left to navigate through the documents.

Product	Vendor	URL
Insight Discoverer Clusterer	Temis	www.temis-group.com
Price	**Price Explanation**	
$36,500.00		
Description		
Innovative solution for structuring previously unstructured information. Classifies and regroups documents, based on their semantic similarities, into coherent classes - the clusters. Offers excellent visibility on large sets of documents and on complex domains. The server stores the results of the clustering in an interactive application, ready to be navigated by the user.		
Web-Based System	**Custom Output**	**Client Side Requirement**
No	No	No
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	Yes	Yes
HP-Unix Server	**IBM AIX 5-2 or OS/400 Server OS**	
Yes	Yes	
Scalable		
No Answer Provided		

Data Access Speeds and Testing
No Answer Provided
Development Code
No Answer Provided
Configuration with Core ACE Tools
No Answer Provided
Supported Data Types
No Answer Provided
Access Control Schemes
No Answer Provided
Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
StarProbe	Rosella Dependable Technology	www.roselladb.com
Price	**Price Explanation**	
$12,000.00		

Description
A star schema based data mining system that works smoothly with most common database systems and incorporates statistics, machine learning, and data warehousing. Contains neural clustering, which is based on Kohonen feature maps, creates grid-shaped partitions. Clusters bundles of similar objects into clusters.

Web-Based System	Custom Output	Client Side Requirement
Yes	Yes	No
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	Yes	Yes

HP-Unix Server	IBM AIX 5-2 or OS/400 Server OS
Yes	Yes

Scalable
Current maximum logical size for a single dataset is 2 billion records. This may be further limited by the capacity of single disk drives. This is because that current version does not allow a dataset split into multiple disks. Generally this size is sufficient for large private enterprises and census data. In addition, StarProbe can be deployed on any Java-enabled workstations, ranging from small laptops/PCs to large Unix super computers.

Data Access Speeds and Testing
We have no independent test records. Data mining tools and techniques we offer are quite different from others. Testing side by side with other packages may not be possible. However, the following test results may indicate the performance of StarProbe in general. The tests were performed on Sony Vaio Laptop with Pentium4 1.6GHz, 512MB memory, Windows XP, Sun microsystems JRE 1.4.2. The dataset used has 1 million records with 13 numeric and non-numeric fields. Pie charts - 2 seconds, Scatterplots - 6 seconds, Decision trees (depth = 20 levels) - 150 seconds, Hotspot analysis (depth = 2 levels) - 20 seconds, Clustering (1 epoch training for 2 fields) - 7 seconds. It is noted that this does not include disk I/Os since the size of main-memory was sufficiently large to cache entire data by the operating system. The first two tests can be indication how fast StarProbe can perform compared to other systems. The last three tests may indicate maximum data size that can be tackled in practical time. Most performance problems lie with connections between StarProbe and database servers. Reading and updating database can take very significant time depending on the performance of DBMS and networks. This problem is not specific to StarProbe. It applies to all other systems.
Development Code
100% written in Java 1.1.8 and also can be run on any later versions of Java runtime. This means that StarProbe can be deployed on a wide variety of workstations. Note that some systems (such as Microsoft) do not support recent Java versions.
Configuration with Core ACE Tools
StarProbe is primarily designed to work with relational DBMSs. It works with any database systems with ODBC and/or JDBC support. Definitely, Oracle is specially supported. It is noted that Oracle has an unusual numeric data type and StarProbe handles it in a special way. However, StarProbe does not work directly with other above systems. It is assumed that other systems will deposit data into database systems. StarProbe directly works with data in database systems. Starprobe uses ODBC and JDBC connection to get data and update database. It can work with any database that supports ODBC and/or JDBC. This includes Oracle, DB2, SQL Server, MS Access, Infomix, MySQL, and many others.
Supported Data Types
There are three generic data types: Character strings, 64-bit integer and 64-bit IEEE-754 floating point designed for scientific data. Taxonomic data structures (or hierarchical drill down structures) are supported using star schema. (Note that this is what StarProbe is named after!)
Access Control Schemes
Current version is for workstations and therefore does not employ any access controls. However, there is license verification and users have to enter database connection information if they want to access database data.
Help Systems Configuration
Currently we can provide email (and telephone) based support only. Since we are located in Sydney, Australia, we cannot provide people-based support for the time being. It might be possible in future when we have partnership in US.
Training Options
We provide technical support only. Once users understand data mining and what they want to do, the rest is very simple and easy. Just reading quick-start help pages, users can start immediately. StarProbe has very intuitive graphical user interfaces. Users don't take much time in learning interfaces. Generally, playing with StarProbe following help pages for a few to several hours will make users quite proficient. Problems and questions can be asked as part of on-line technical support.

Data Input and Output Processes
Input from databases is automated interactively. However, input from data files requires user to specify input data format. Outputs are all in proprietary formats and can be re-opened from StarProbe. There are several export facilities. For example, charts and graphics can be made to image files (in GIF format). Textual outputs can be saved as files. In addition, predictive models and clustering outcomes can be applied directly to database tables.
Current Developmental Plans
Our current plan is to add association rules algorithms for the next version. (This is for telemarketing and may not have much value to your organization.) Beside this, our future development will be largely based on user feedbacks and requests. We will be adding and improving features what users ask, of course, provided that they are possible. Users will be able to use them immediately with the next releases. Currently we are also looking at ways to run StarProbe using web-browser's Java runtime. This will make it possible to deploy StarProbe across network without any client-side installation. Currently, StarProbe does run on web-browsers without any installations on client side. However, browser security manager blocks all I/O operations, and running StarProbe using browsers is useless at the moment.
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
There are two factors that StarProbe stands out from the rest: (1) Sophistication of clustering & related tools, and (2) Java-based cross-platform support. First, the real strength of StarProbe lies with the following core tools: Neural clustering, Hotspot analysis, Rule induction (or class/cluster profiling). StarProbe clustering is based on neural network known as Self Organizing Maps (SOM). SOM is how our brains learn and recognize patterns and features. It's very robust and powerful. (Otherwise, we won't be existing as intelligent beings!) StarProbe clustering does not require pre-processing of input data. Other clustering techniques are generally based on numerical distance calculation and require encoding of input data. This process makes clustering results more subjective that depends on quality of data encoding. SOM adjusts variations automatically as part of learning process. In addition, it is robust with noisy and fuzzy input data. Clustering results can be further analyzed using hotspot analysis and rule induction tools. These tools provide "IF-THEN" style production rules and hierarchical drill-down trees that might be useful for verbalizing cluster profiles. These, together with other many visualization tools, will significantly enhance analysis of clustering process. It is noted that hotspot analysis and rule induction tools can used with and without clustering. Other major commercial products such as SAS, SPSS, Angoss, etc., do not support these features. They provide clustering techniques based on numerical calculation which heavily depends on pre-processing of data. Though SAS EMiner provides SOM, it is not tailored for clustering. Second, StarProbe is written 100% Java 1.1.8. The same binary program code can be deployed on a wide range of workstations and servers, ranging from laptop to super computers that support Java, e.g., Windows, Mac OS X, Linux, Solaris, OS2, AIX, IRIX, etc. Data and output generated on a particular operating system can be copied to another operating system without conversions. Other major systems are developed to operate on a specific platform and limited to that operating system.
Toolset Strengths
The features described in the previous question are powerful and uniquely on StarProbe. This alone can justify StarProbe at least as complimentary system to other similar systems. However, the reverse may not be true: with StarProbe at hand, users may not need other similar systems. We strongly recommend to test above features along with other similar systems. StarProbe is a system to be considered at least as a complimentary tool to others.

Product	Vendor	URL
TextAnalyst	Megaputer	www.megaputer.com
Price	**Price Explanation**	
$1,290.00		

Description

A unique software tool for semantic analysis, navigation, and search of unstructured texts. Helps to quickly summarize, efficiently navigate, and cluster documents in your textbase. Breaking links representing weak relations in the original Semantic Network enables clustering of the textbase.

Web-Based System	Custom Output	Client Side Requirement
No	Yes	Yes
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	Yes	No
HP-Unix Server OS	**IBM AIX 5-2 or OS/400 Server OS**	
No	No	

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
VisuaLinks	Visual Analytics Inc.	www.visualanalytics.com
Price	**Price Explanation**	
$2,800.00		

Description

Used to discover patterns, trends, associations and hidden networks in any number and type of data sources. VisuaLinks presents data graphically uncovering underlying relationships and patterns. Its designed to expose clusters, or networks, of related information. Searches your data looking for particular types of data that are related by particular types of associations.

Web-Based System	Custom Output	Client Side Requirement
Yes	Yes	Yes
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	Yes	Yes
HP-Unix Server OS	**IBM AIX 5-2 or OS/400 Server OS**	
Yes	No	

Scalable

Yes, the tool is scalable. We are deployed in many locations one of which is 12 databases each with upwards of 30 million records with hundreds of user. It has been deployed it on Linux (Red Hat 9), Solaris, Windows 2000 Server. We have deployed where the client is running on windows and the server is running Solaris, in both applet configuration and application configuration. There are no limitations to the size of data. Basically, the data size is dependant on the underlying database that is used.

Data Access Speeds and Testing

We have people who have tested our product based on certain criteria from the Gartner group, the JIVA project which has been certified in intel networks, IRS and many more. These test have been marked as classified. A copy may be obtained by other classified individuals.

Development Code

VisuaLinks is developed completely in JAVA. DIG is developed in .NET

Configuration with Core ACE Tools

Windchill- no - Convera Retrieval Ware-yes - Oracle Database Systems- Yes - DOORS- Yes, through a jdbc driver. - ClearCase- no

Supported Data Types

Any relational data source, emails, document repositories, or websites.

Access Control Schemes

User name/password with permission attributes to read/write sources.

Help Systems Configuration

Our help is configured in a few different formats. We have Web-based, on our support site, which is accessible 24/7. The help consist of Knowledge-Based Articles, FAQ, and a "Help" guide which is also integrated into our software. We have People-based help, via phone and email, from the hours of 8:00 - 6:00 (EST) which often extends from our normal business hours. As well, depending on availability some support request can be made in advance for after hours.

Training Options

Our training is broken into 10 days, each class is strongly recommended. VisuaLinks 2 day Intro to VisuaLinks Training. 1 day Advanced VisuaLinks Training. 2 day Modeling VisuaLinks Training. DIG 1 day End User DIG Training 2 day Administrator DIG Training.

Data Input and Output Processes
VisuaLinks is able to perform queries to and from database's. Dig processes unstructured documents.
Current Developmental Plans
VisuaLinks is currently in two development phases - the first are patch releases for 3.0. The second is the next major release of 4.0 where we're adding many new features that clients have requested. DIG -is currently in version 2.0 with the next major release of 3.0 coming out in the later months of the fourth quarter.
Strategic Plan
A Non-Discloser Agreement.
Compare / Contrast with Similar Systems
Please see attached Comparison Chart.
Toolset Strengths
Please feel free to contact us for a live in person or WebEx demonstration of our VisuaLinks and Digital Information Gateway products.

Product	Vendor	URL
WordStat v4.0	Provalis Research	www.simstat.com
Price	**Price Explanation**	
$1,095.00		
Description		
A text analysis module specifically designed to study textual information. Includes numerous exploratory data analysis and graphical tools that may be used to explore the relationship between the content of the documents and information stored in categorical or numeric variables. Relationships among words or categories as well as document similarity may be identified using hierarchical clustering and multidimensional scaling analysis.		
Web-Based System	**Custom Output**	**Client Side Requirement**
No	Yes	Yes
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	No	No
HP-Unix Server OS	**IBM AIX 5-2 or OS/400 Server OS**	
No	No	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		

Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
Text Miner	SAS Institute Inc.	www.sas.com
Price	**Price Explanation**	
$0.00	Contact Vendor	
Description		
A suite of tools for discovering and extracting knowledge from text documents. Text documents can be clustered automatically into groups, classified into predefined categories and used in conjunction with structured data to build predictive models. Text clustering algorithms group documents into common themes and topics based on their content. Cluster summaries are easy to interpret in the context of the original text documents.		
Web-Based System	**Custom Output**	**Client Side Requirement**
Yes	No	Yes
Windows Client OS	**Windows Server OS**	**Solaris Server OS**
Yes	Yes	No
HP-Unix Server OS	**IBM AIX 5-2 or OS/400 Server OS**	
No	No	
Scalable		
Absolutely. SAS has no restriction on data size, depends on your hardware.		
Data Access Speeds and Testing		
Again, the speed of the solution depends on your hardware and network configurations. SAS has a department, referred to as the Enterprise Excellence Center, who focuses on hardware sizing and configuration if that, would be needed. Here is a brief description: The Sizing and Configuration Program is a program which will provide SAS Sales and their customers a recommendation of the hardware required for integrating a distinct SAS solution into the customer's infrastructure. The Enterprise Excellence Center will assist with assessing the client's current operations and in collaboration with our platform partners to determine a customer's computing needs. The resulting recommendations will become part of our technical marketing reference collateral. The Sizing and Configuration Program requires that SAS interacts directly with the customer and their IT organization to understand what is needed, the relative use of SAS products and their IT's hardware vendor platform and computing standards.		

Development Code
Enterprise Miner supplies complete scoring in SAS, C, Java and PMML. SAS is developed in C and C++ and the client piece of Enterprise Miner 5.1 is java. SAS itself is a 4th generation language (4GL) that is "flexible and extensible with an easy-to-learn syntax and hundreds of language elements and functions that support programming everything from data extraction, formatting and cleansing to data analysis, reporting and information delivery" (see http://www.sas.com/technologies/bi/appdev/base/). Enterprise Miner 5.1 can be extended easily with custom tools using just SAS code and XML.
Configuration with Core ACE Tools
SAS is ODBC and OLE DB compliant, making SAS flexible and allowing our solutions to access any ODBC or OLE DB compliant data source. SAS can connect to Oracle directly. http://www.sas.com/technologies/dw/etl/access/relational.html
Supported Data Types
supports any relational database; please refer to the following link: http://www.sas.com/technologies/dw/etl/access/index.html
Access Control Schemes
No Answer Provided
Help Systems Configuration
The SAS Tech Support is 24/7 and is web-based and people-based.
Training Options
Training is required, but recommended. - Predictive Modeling using SAS Enterprise Miner Software is the first course to offer. 3-days onsite is $2,850/day plus expenses for up to 20 students. Click on the title in the curriculum page for details. - Text Mining Using SAS Software has the Predictive Modeling course as a prerequisite. It is 1-day. An on-site course is $2,850/day plus expenses for up to 20 students. The same instructor can probably do both courses for a 4-day stretch of training. ** Our Training Dept. will need about 4 weeks to schedule and arrange these courses for your group **
Data Input and Output Processes
Processes can be automated via batch.
Current Developmental Plans
There is continuous development on this solution. SAS invests about 26% of total revenue into our R&D group, this is nearly twice the average investment of large software companies. This investment helps us be able to have continuous developments on all of our solutions. Customers are able to receive the latest versions as long as they are up to date on their maintenance.
Strategic Plan
The SAS Public Sector group was formed in order to focus on Federal, State, and Local government. We have had a lot of success in the government industry and several references. We work as a contractor to understand your business problems, determine a solution, and provide services via our Pilot Program. The pilot program is a way to determine the business requirements and implement/customize the solution to the organizational needs. Also, knowledge transfer via an on-site contractor and training is a very important part of the pilot program as well. We have had our clients' meet with established references to review over their solution and discuss how SAS has been beneficial. Also, we are definitely open to non-disclosure agreements if in the case that we would need to utilize your data for a demo or within the pilot program.

Compare / Contrast with Similar Systems
Please refer to the competitive overview document in my email, this provides a comparison between SAS Enterprise Miner & Text Miner to SPSS Clementine, Angoss, and IBM Intelligent Miner. I have included some general bullet points on the SAS solution below: - SAS provides flexible software that supports all steps necessary to address the business problems at hand in a single, integrated solution - Easy-to-use GUI interface helps both business analysts and statisticians - SAS provides all the components of data mining: data access, exploring, modifying, modeling and model assessment - Full control of model creation; no black box - Only solution that fully integrates text mining and data mining for more productivity - Our solution allows analysts to build more models faster which enables more collaboration between analysts - 40% of market share. . .have included an article that explains further
Toolset Strengths
Please refer to the attached documents.

10.4 Visualization Tools

Product	Vendor	URL
AnswerTree	SPSS Inc./2000	www.spss.com
Price	**Price Explanation**	
$1,495.00		
Description		
Contains four powerful algorithms, the widest choice of decision trees available. Displays models visually to allow you to easily see the groups that matter. The diagrams display a snapshot of the segments, patterns, and relationships in the data, that enable the user to make confident decisions. Built with scalability in mind, therefore the user can work with their data more efficiently.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	Yes
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
CoMotion	Maya Viz	www.mayaviz.com
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description

Provides sophisticated visualization components to create interactive, analytic, collaborative environments that bridge the gap between intelligence and knowledge management. Provides full access to data and a clear visual environment to explore it. Performs routine and exploratory analysis. Move data from one visualization to another.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
Honeycomb Analyzer	The Hive Group/1991	www.hivegroup.com
Price	**Price Explanation**	
$75,000.00	$75,000 per processor	
Description		
Transforms data from a database into an information map. All data are presented in a treemap. Filters allow end-users to eliminate irrelevant data-elements. Graphical icons highlight key attributes of the data		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	Yes
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
Insightful S-PLUS	Insightful Corporation/1995	www.insightful.com
Price	**Price Explanation**	
$2,400.00		
Description		
4,200 data analysis functions that include the most comprehensive set of robust and modern methods available anywhere. The user can import their data, select statistical functions and display results. Easy to examine and visually explore data, run functions one step at a time and visually compare models for fit.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	No	No
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	No
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
IN-SPIRE	Pacific Northwest National Laboratory	http://in-spire.pnl.gov/about.html
Price	**Price Explanation**	
$500.00	Government Application price only	
Description		
A discovery tool that integrates information visualization with interaction and query capabilities. Quickly and automatically conveys the gist of large sets of unformatted text documents such as technical reports, web data, newswire feeds, and message traffic. Allows the user to easily identify trends, anomolies, and relationships in huge volumes of text.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	No
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		
Training Options		
No Answer Provided		
Data Input and Output Processes		
No Answer Provided		
Current Developmental Plans		
No Answer Provided		
Strategic Plan		
No Answer Provided		
Compare / Contrast with Similar Systems		
No Answer Provided		
Toolset Strengths		
No Answer Provided		

Product	Vendor	URL
Miner 3D Enterprise	Dimension 5/1995	http://miner3d.com
Price	**Price Explanation**	
$1,195.00		

Description

Empowers the user to understand trends and relationships in their data. An all-visual intuitive tool, that helps the user to work effectively without extensive training. Integrated model builders automatically create charts on currently available data. Data points are visualized as graphic objects with properties capable of carrying information. Users have complete control over almost every part of the visualization space.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	Yes

Scalable

Miner3D is a data visualization software tool allowing creation of a broad range of different visualization scenes, from fine, complex but computer resources demanding scenes to simple and very fast data visualization scenes. Miner3D is perfectly scalable in the mean of amount of data as well as in quality of scenes. Miner3D is adaptable software, it continually monitors the computer's performance and when it detects a slower response it automatically downgrades the quality of visualization, so then the interactivity is always good. Our users come from a very broad range of industries and use it in many different tasks. We know about customers who work regularly with 500.000 rows and the computer handles such a huge data sets in real-time on single computer screen. Considering the continual growth in performance of personal computers, there is no worry about its capability of handling very large data sets in future.

Data Access Speeds and Testing

Unfortunately, I don't have access to any independent tests results. We can share with you an externally-researched study that was focused mostly on issues like ease-of-use, deployment and support.

Development Code

The source code is written in C/C++. We also use OpenGL libraries for graphics.

Configuration with Core ACE Tools

From the listed environments we directly support ORACLE. If your systems can be set up to deliver XML, CSV or TXT output, then even better - we can load data very easily. Also, if the other platforms are accessible via Microsoft ADO or ODBC, or via a data pipelining tools (SciTegic), then we also can access it.

Supported Data Types

Currently we support common data types - integers, real numbers, scientific format, currencies, date, time, text strings, links, all in a wide range of formats. From Version 5 (planned to be released in November 2004) we will support also images, pictures and textures. With a chemistry-specific software libraries we will support also chemical structures, 2D drawings of molecules. Similarly we can support also other specific areas… In future we plan to support also video, sounds and speech input.

Access Control Schemes
Miner3D is not a comprehensive application environment, but rather a user interface software technology. Applications built on/with Miner3D require integration with target operating systems, database systems that provide access control schemes. It is recommended to develop access rights schemes and permission policies with deployed operating system and through Miner3D interfaces provide visual access to control the schemes.
Help Systems Configuration
By now it is a combination of online help with web and email support. We plan however to open a US-office (we are Europe-based company) and within the next few months we should have a sales office providing also telephone support and consultancy services for North American customers.
Training Options
The training for developers is not necessarily to be too extensive, because we use standards and commonly accepted methods and technologies. Usually a couple of days of email support was sufficient for effective deployment. Training for end-users depends on your application and thus you should provide it internally with your own people.
Data Input and Output Processes
No Answer Provided
Current Developmental Plans
We now finalize Version 5 that unifies our two main product lines into a single software code delivering both advanced data analysis capability with component distributed and scriptable architecture required for Web-based applications.
Strategic Plan
We will enhance the software interface, improve support for data sources, default graph types, we will continue improving the advanced analytical tools.
Compare / Contrast with Similar Systems
A/ Web software like Adobe Atmosphere, Macromedia, VRML plugins: those kind of products emphasize on the "visual effects" part and underestimate the information value of data visualization. Miner3D provides information centric environment, fast and intuitive user interface where a user can focus fully on data. B/ Desktop software like Spotfire: Miner3D is not so comprehensive and contain less mathematical methods for data analysis, but is more compact, easier to use and has also very good price/value ratio. C/ Addins like MS-Chart in MS Excel/Access: Miner3D provides live visualizations instead of static dead-looking charts. User has always the option to do visual data queries with real-time response of the graphic.
Toolset Strengths
Miner3D is probably not the ultimate application development system that will solve you everything. We can however deliver and excellent user interface for your next information system with intuitive navigation, information access, searching, monitoring capability with full deployment of advanced graphics, sound and voice. We will be happy to contribute to your future application as a sub-contractor or provider of user interface technology.

Product	Vendor	URL
MineSet 3.1.1	Purple Insight Ltd./2003	www.purpleinsight.com
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description

Reveals the hidden value in your data warehouse with tools for both data mining and data visualization. Allows business users to enjoy visual interpretation of complex data mining algorithms. Scalability that handles massive amounts of data. Contains visualization tools unique to the industry.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	No	Yes

Scalable

MineSet™ is architected as a multi-threaded application, which enables unequalled analytical performance. There are now 20 patents that protect various aspects of the MineSet™ software suite. MineSet™ is an advanced scalable Client-Server tool set. The user interface (the Client) runs on the workstation or PC, and the data-mining engine (the Server) runs separately, either on the same machine or on a central more powerful data processing machine. More than one Client can connect to the same Server. MineSet™ is optimized for the ultimate scalable performance on multiple CPU machines. The Server can break intensive compute operations into parallel computational paths and so scale to take full advantage of available compute resource. The user has full control to configure the parallelization used by the Server. In addition to parallelization, MineSet™ offers very large memory support and data handling capability using a 64-bit implementation. The amount of data (size) handled by MineSet is function of the available compute resource. MineSet uses a highly parallel architecture and can handle large data sets. The size of data should not be a major factor--given enough hardware, MineSet can scale to data sets with hundreds of millions of rows and beyond.

Data Access Speeds and Testing

There is no formal industry standard for benchmarking Data-Mining and analysis tools. As the practical performance of an installation of MineSet is dependant on data type and available hardware resource. MineSet offers better performance and visualization techniques when managing and manipulating very large data sets such as the 800 million-record HCFA data set. MineSet combines the software's visual approach and high performance computing to speed navigation through more than 2.2 million gene sequences in Incyte's databases. Scientists quickly spot trends in gene expression data and easily locate novel targets for drug research. With MineSet, they have a tool that lets them focus on the science, not the computer.

Development Code

C++/Java

Configuration with Core ACE Tools
MineSet provides native support for Import/Export for over 27 different flat file and statistical formats via StatTransfer from Circle Systems, Inc. This wide range of import/export formats enables transfer of data between applications such as Windchill, RetrievalWare and DOORS. MineSet supports direct connection to Oracle, Sybase and Informix running on any major platform and connectivity to ODBC-compliant data sources including SQL Server and DB2. MineSet provides an API and plug-in interface for accessing external analytic algorithms and functions plus API's to other OLAP tool vendors.
Supported Data Types
1-2-3, Access, ASCII - Delimited, ASCII - Fixed Format, dBASE, Excel, Epi Info, FoxPro, Gauss, JMP, LIMDEP, Matlab, MiniTab, Mineset, OSIRIS, Paradox, Quattro Pro, SAS data file, SAS Transport, S-PLUS, SPSS Data, SPSS Portable, Stata, Statistica, SYSTAT MineSet supports direct connection to Oracle, Sybase and Informix running on any major platform and connectivity to ODBC-compliant data sources including SQL Server and DB2.
Access Control Schemes
MineSet client and server are licensed via a FLexLM license file. The license could be configured to be floating or node-locked and either open or, limited to a specific user.
Help Systems Configuration
MineSet ships with a comprehensive online tutorial, user guide, reference and interface manuals. Purple Insight provides a full range of online and/or telephone based support packages.
Training Options
The training options are: self teach, class room training or bespoke knowledge transfer: Self Teach - The step by step tutorial that ships with MineSet can give an overview of the main MineSet features in less than a day. The tutorial includes a 'further exploration' section that takes user through more advanced features, the time required depends on the users requirements. Class Room Training Course/Knowledge Transfer - Purple Insight offers consulting, knowledge transfer and training through Purple Insight consultants or partners. Training can be delivered as a standard 3 day training class or as knowledge transfer with the content tailored to the customer's requirement. Training can be delivered at the customer site or Purple Insight can provide facilities.
Data Input and Output Processes
Running MineSet in batch mode allows the software to perform operations without bringing up any visualizations. Batch mode can be particularly useful in projects requiring lengthy computations that need to be done frequently. For instance, the computations can be run at night so the data will be ready the next morning. Batch mode operation is controlled/configured by scripts.
Current Developmental Plans
Our next maintenance release will include a modular structure to the current suite of tools ranging from a lightweight client with only the visualisation tools up to the current Enterprise Edition. Our next major release will be v4.0, features being tested for inclusion in v4.0 include a control API for build/execution of mining histories, new visualisation tools for high dimensional data sets, new tools for clustering, interactive decision tree induction and anolomy detection and automatic build and compare of classifiers.

Strategic Plan
Purple Insight is a software and services company focused on delivering its premier Visual Data-mining application MineSet together with associated services including consultancy, project management, training and installation to customers, and ensuring they maximize returns from the fastest growing asset in their business -- data. From local and wide area network infrastructure, to global communication systems, from data warehouses to corporate databases to distributed enterprise-wide applications. Wherever the data is stored, our aim is to help businesses make better use of this vast amount of information to make better business decisions. In addition, we intend to expand our global partnerships to develop and deliver MineSet software based applications and associated services into new markets, industries and business sectors.
Compare / Contrast with Similar Systems
In comparison to other data-mining tools (such as Business Objects, SAS, Cognos, Intelligent Miner etc) MineSet combines the core of the most commonly used analysis techniques with its own unique strengths: award winning visualizations and scalability.
Toolset Strengths
See attachment

Product	Vendor	URL
PowerAnalyzer	Informatica Corporation/2000	www.informatica.com
Price	**Price Explanation**	
$50,000.00	$50,000 to indefinite	
Description		
A robust visualization solution designed for the rapid deployment of IT and corporate data in real time through dashboards and custom applications. Accelerates implementations. Customize to unique requirements. Ensures security, performance, and scalability. Transforms data into immediate, accurate, and understandable information.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	No	Yes
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		

Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
Statistica	StatSoft, Inc./1990	www.statsoftinc.com
Price	**Price Explanation**	
$795.00		
Description		
Provides the most comprehensive array of data analysis, data management, data visualization, and data mining procedures. Techniques include: predictive modeling, clustering, classification, and exploratory techniques. Offers the speed and capacity to handle datasets/designs of practically unlimited size and unusual comprehensiveness of its procedures.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	Yes
Scalable		
Yes The performance, flexible customizability, and the wide selection of options that can be tailored to your needs mentioned in the previous section would not be possible if STATISTICA did not feature the advanced technologies that drive all functions of the application. Multithreading and distributed processing architecture delivers unmatched performance (offered in the Client-Server version) including super-computer-like parallel processing technology that optionally scales to multiple server computers that can work in parallel to rapidly process computationally intensive projects. From low ended Windows Desktop systems to MultiCPU configurations. Limited by your Windows CPU file systems sizes.		
Data Access Speeds and Testing		
We have benchmarks of comparision to some of our competitors for some of our modules.		

Development Code
C++ for Com Objects with Java WebSTAT applets
Configuration with Core ACE Tools
We are used in a wide variety of applications by our customers who use Windchill, Oracle, and ClearCase. With our open systems you could easily use Covera Retrieval Ware or Doors from Hologram to access our technology from Com, ODBC, OLE, SOAP or other standards. Some of our customers include Catapiller, John Deer which are large PTC WindChill uses and we are Oracle Parters.
Supported Data Types
We support all Windows supported data types and we are fully OLE compliant.
Access Control Schemes
We use the Windows Access Control which is quite extensive but can be extended with other Access Control Methods.
Help Systems Configuration
We offer service for the North America 9 to 6 CST Monday thru Friday. We also have service available thru our foreign offices as well. This includes unlimited Email and Phone support. Our STATISTICA Help is also available online and is extremely extensive. We also offer a STATISTICS text book online http://www.statsoft.com/textbook/stathome.html StatSoft technicians can answer general questions regarding the usage of the software, limitations of specific procedures, software and hardware compatibility of StatSoft applications, as well as all instances when the user believes that a StatSoft application malfunctions. Free technical support does not include training in the use of the software, or help with writing macros and other programs in the STATISTICA environment, it does not cover statistical and other forms of consulting, questions related to the analyses of specific data sets, or theory of data analysis. The standard technical support for enterprise products does not include help with the integration of StatSoft products into existing computer infrastructures, or any issues related to custom installation and custom configuration. All these services are available, however, through StatSoft consulting, system integration and deployment services (please consult your regional StatSoft office regarding the custom consulting services).
Training Options
StatSoft offers both introductory and advanced training courses in major cities in the United States and overseas as well as on site. StatSoft's training classes offer: Practical hands-on experience with the program, An introduction to real-world example applications, Energetic, helpful, knowledgeable instructors, Comprehensive take-home course manual, Personal attention, small class size, Interactive, class-paced learning.
Data Input and Output Processes
Our tools offer a wide range of options for automation of data input and output from a wide range of data input and output formats.
Current Developmental Plans
We are currently shipping Version 7 and have development plans to enhance V7 and have specifications for V8 and wide vision for V9.
Strategic Plan
Yes we do and will share it with you after a nondisclosure.
Compare / Contrast with Similar Systems
We offer the widest array of Anlytics for Statistical Analysis, Data Mining, Quality Control, and Web Analytics under a common, open architecture that is easily customizable, easy to use, and high performance.

Toolset Strengths

STATISTICA. STATISTICA is not just "another advanced data analysis package." It offers not only the speed and capacity to handle datasets/designs of practically unlimited size and unusual comprehensiveness of its procedures - fully integrated with highest quality graphics (that won for STATISTICA the name of "the King of data visualization tools," see Reviews). STATISTICA offers much more. A unique record of recognition. This summary contains many positive words, but they are not just subjective views of a manufacturer. STATISTICA has something ultimately objective to back all of them up - something that no competing product can claim: The STATISTICA line of products has received the highest rating in EVERY published (independent) comparative review in which it was featured since its first release in 1993. In the entire history of the software industry, very few products - in any category - have ever achieved this status. COM and SOAP-based architecture, high-end technologies. STATISTICA is based on the COM and SOAP-based architecture and high-end technologies, that are usually not found in such "vertical market" applications as data analysis software. As a result, STATISTICA offers unique functionality and usability features that currently no other competing product can offer. Benefits of the STATISTICA technology: 1. For advanced users: Power, Scalability, Compatibility. The powerful and unique - in data analysis software - full implementation of the COM architecture, scalability to multiple CPU's, a complete Web-integration with support for distributed processing and multi-tier Client-Server architecture, support for XML, in-place processing of large databases on remote servers via the IDP technology, and the fully integrated Visual Basic make STATISTICA a perfect foundation (or a component) of global computing infrastructures (such as the Internet Information Delivery Systems or large, multi-user enterprise installations). STATISTICA offers one of the largest and richest development environments available in the entire software industry, with more than 11,000 data analysis and graphics functions directly exposed to end users and developers. It can be fully customized including even such low-level routines as the "event handling" - and most importantly, all these customizations can be done using the built in Visual Basic - the most widely known and used computer language in the world (support for other languages, such as C++ or Java, is also provided. These features also make STATISTICA a perfect tool to tackle the most demanding problems in data analysis, data mining, or QC/SPC applications. 2. For occasional users and novices: Simplicity, Customizability, Quickness, and Quality. The unique technologies of STATISTICA are also offering a lot to a desktop user who may not want to know what Visual Basic is or what the "enterprise business intelligence system" is. For example: - Because of these technologies, the STATISTICA desktop products are uniquely user friendly and flexible; they can also be extremely "simple" (e.g., run by anyone from an Internet browser). - Every action can be recorded (in the background) into a reusable, modifiable macro with a click of the mouse, and then instantly assigned to a toolbar button (the macro will have the industry standard Visual Basic format, but you do not even need to know that). - Every aspect of the STATISTICA user interface can be adjusted to your needs by dragging controls with the mouse. Countless program options can be as hidden or as exposed as you want. - You can make "your own STATISTICA" in minutes, but even if you do not want to change anything, you can still rapidly navigate through the simple "Quick tab" templates, designed to minimize the number of steps. 3. Quality and comprehensiveness. In addition to all these benefits - no other application can match the quality of implementation of every detail of STATISTICA - its graphics, responsiveness/speed, elegance, and built-in intelligence. Every aspect of STATISTICA is designed to offer the ultimate level of functionality and a large part of that functionality normally can be found only in designated, specialized applications, if they existed at all, and those that do - would not be integrated and are

usually not as comprehensive and well designed. In fact, you would need many of such specialized applications to get the functionality that is a part of every STATISTICA product, and offered even in the entry level STATISTICA Base. For example: - The graphics engine, a part of every STATISTICA product, offers more choices and options than any designated graphics package on the market. - The STATISTICA Query facility (allowing you to access external databases), just one of many parts of every STATISTICA product, includes features, power, and performance that are offered only in designated, expensive database querying products (such as BusinessObjects®). - The STATISTICA Visual Basic language (that is built into every product in the STATISTICA family), offers one of the largest and richest development environments available in the industry (with more than 11,000 data analysis and graphics functions, depending on the version, and professional programming tools).

Product	Vendor	URL
Thinkmap SDK	Thinkmap, Inc./1997	www.thinkmap.com
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description

Enables organizations to incorporate data-driven visualization technology into their enterprise Web application. Allows users to make sense of complex information in ways traditional interfaces are incapable of. Composed of a number of loosely coupled components that can be quickly reconfigured to fullfil many different visualization tasks. Template includes Spider, Hierarchy, Clustering and Chronology.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes
Access Multiple Data Sources	**Java Applets**	**Export to MS Office**
Yes	Yes	No

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

10.5 XML Conversion Tools

Product	Vendor	URL
ECS Engine	Exegenix Canada Inc.	www.exegenix.com
Price	**Price Explanation**	
$0.40	$0.20-$0.40 per Kilocharacter output	
Description		
An extensible, modular technology that can be adapted to meet the content conversion needs of any organization. All content goes through four processes in order to uncover the documents structure and generate valid XML that can then be transformed to meet specific customer needs. The XML file is designed for ease of use in XSL Transformation scripts. Can be integrated with both traditional and XML-enabled applications.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	No	No
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

Import MS Office	Automated Publishing Process
Yes	No
Batch Processing	**Customizable**
Yes	No

Scalable
No Answer Provided
Data Access Speeds and Testing
No Answer Provided
Development Code
No Answer Provided
Configuration with Core ACE Tools
No Answer Provided
Supported Data Types
No Answer Provided
Access Control Schemes
No Answer Provided
Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided

Data Input and Output Processes
No Answer Provided
Manual Data Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
W2XML	DocSoft, Inc.	www.docsoft.com
Price	**Price Explanation**	
$259.95		
Description		
A Word to XML conversion software. Converts DOC, RTF, HTM, and more to well-formed XML. Completely scalable and allows for custom XML exporting by allowing you to create and apply custom XSLTs to modify the standard output. The software comes with an XSLT that allows you to output Docbook-compliant XML.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No
Import MS Office	**Automated Publishing Process**	
Yes	Yes	
Batch Processing	**Customizable**	
Yes	Yes	
Scalable		
No Answer Provided		
Data Access Speeds and Testing		
No Answer Provided		
Development Code		
No Answer Provided		
Configuration with Core ACE Tools		
No Answer Provided		
Supported Data Types		
No Answer Provided		
Access Control Schemes		
No Answer Provided		
Help Systems Configuration		
No Answer Provided		

Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Manual Data Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
xDoc XML Converter	CambridgeDocs LLC	www.cambridgedocs.com
Price	**Price Explanation**	
$2,495.00		
Description		
Converts PDF files to XML. Converts WordPerfect into XML, PDF, or HTML. Has a point-and-click interface to make the process of transforming content from legacy formats into meaningful XML simpler. Able to manage your data coversion project without writing custom code or manually converting documents. Contains a powerful rules engine that is used to extract content from existing sources.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
Yes	Yes	Yes

Import MS Office	Automated Publishing Process
Yes	Yes
Batch Processing	**Customizable**
Yes	Yes

Scalable
No Answer Provided
Data Access Speeds and Testing
No Answer Provided
Development Code
No Answer Provided
Configuration with Core ACE Tools
No Answer Provided
Supported Data Types
No Answer Provided
Access Control Schemes
No Answer Provided

Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Manual Data Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
xmlspy 2004	Altova	www.xmlspy.com/download_spy_enterprise.html
Price	**Price Explanation**	
$1,248.74		
Description		
Contains robust, intelligent XML editing features of Text View, including code completion, syntax coloring and built-in XML validation and wellformedness checker. Has a built-in Authentic View to allow developers to create customized views and data input forms. A powerful XSLT Debugger allows a developer to troubleshoot problematic XSLT stylesheets node-by-node, viewing node sets, testing Xpath expressions, inspecting variables, and setting breakpoints. Encompasses the entire XML development life cycle, starting with application and data modeling in WSDL and XML Schema, all the way to XML transformation, storage, and syndication.		
Web-Based System	**Windows Client OS**	**Windows Server OS**
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

Import MS Office	Automated Publishing Process
Yes	Yes
Batch Processing	**Customizable**
Yes	Yes

Scalable
No Answer Provided
Data Access Speeds and Testing
No Answer Provided
Development Code
No Answer Provided

Configuration with Core ACE Tools
No Answer Provided
Supported Data Types
No Answer Provided
Access Control Schemes
No Answer Provided
Help Systems Configuration
No Answer Provided
Training Options
No Answer Provided
Data Input and Output Processes
No Answer Provided
Manual Data Processes
No Answer Provided
Current Developmental Plans
No Answer Provided
Strategic Plan
No Answer Provided
Compare / Contrast with Similar Systems
No Answer Provided
Toolset Strengths
No Answer Provided

Product	Vendor	URL
X-Style	CELI S.R.L.	http://www.celi.it/english/index.html
Price	**Price Explanation**	
$0.00	Contact Vendor	

Description
Automatically converts any document form any pre-existing format into XML. Ablility to realize completely automated conversions from textual formats, HTML and Microsoft Word documents into XML. An effective way to remove both obstacles from the path leading to XML integration. Word styles are mapped into XML elements and attributes according to a set of rules encoded in the customization phase.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

Import MS Office	Automated Publishing Process
Yes	Yes
Batch Processing	**Customizable**
Yes	Yes

Scalable
The tool is best suited to deal with large amount of data. As it presupposes manual configuration for targeting a specific DTD, it is worth only if the documents to be processed are in a relevant number. (In Italy it is adopted to convert past legislation to XML).
Data Access Speeds and Testing
We do not any independent testing. In any case performance depends on the amount of linguistic processing to be performed in order to match the target DTD.
Development Code
The system is 100% Java. CELI has full ownership on the source code.
Configuration with Core ACE Tools
Our tool does not access any third party software. In terms of integration, configuring it to work in the context of a web application is trivial, as it can be accessed either via java API or via HTTP (web service).
Supported Data Types
HTML/PDF/WORD/PURE TEXT
Access Control Schemes
No Answer Provided
Help Systems Configuration
People based. We follow our customers for all the duration of the software setup on a project base.
Training Options
In order to use the system, no training is required, as it performs automatic XML conversion. By contrary, configuring it presupposes know how about the following topics: XML, XSLT, Human Language Technologies (Conversion is content based, thus a computational linguist needs to train the Information Extraction System Sophia 2.1 (cf. info on http://www.celi.it/english/sophia.htm)).
Data Input and Output Processes
Via script, Via web input. Different forms of data input can be envisaged.
Manual Data Processes
The output needs to be revised by an XML competent persons in those cases in which the information contained in the input document were not necessary to complete XML conversion. Of course, if the target DTD is not too strict such a process is mostly optional.
Current Developmental Plans
To extend it beyond the legal and pharmaceutical domain.
Strategic Plan
Co-operation with governmental organization can go up to releasing the source code, assuming that all guarantees are provided.
Compare / Contrast with Similar Systems
Contrary to the totality of XML conversion systems, X-Style is based on content analysis in order to perform XML markup. The advantages of such an approach are robustness and independence on the adopted formatting system. The disadvantage is represented by the fact that the system needs manual configuration in order to work properly.
Toolset Strengths
Our tool is well suited for big organizations which have to translate big amounts of documents to specific DTDs. It is not suited for translation of small sets of documents as configuration costs would overcome the benefits. Also it is worth mentioning that XML conversion can be run as a service. Under this hypothesis configuration costs will be supported by CELI.

Product	Vendor	URL
ClearTags	ClearForest Corp.	www.clearforest.com
Price	**Price Explanation**	
$200,000.00		

Description

Powerful research tools that present a single-screen view of complex inter-relationships as well as Web-based monitoring and visualizations enabling users to gain new insights from news and research content. Bridges the gap between text and business intelligence. Tags and extracts information form inside text, that can then be incorporated into any business intelligence system or publishing database. Automatically categorizes documents and structures entities contained within the text. Generates metadata in a highly customizable XML format, that can populate datamarts, used for business intelligence in a third-party analytic tool or within ClearForest Analytics.

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes
Solaris Server OS	**HP-Unix Server OS**	**IBM AIX 5-2 or OS/400 Server OS**
No	No	No

Import MS Office	Automated Publishing Process
Yes	Yes
Batch Processing	**Customizable**
Yes	Yes

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Manual Data Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
RLO-Xtractor	Multimedia Design Corporation	www.mmdesigncorp.com

Price	Price Explanation
$0.00	Contact Vendor

Description

Automatically extracts multimedia elements from your PowerPoint presentation for easy re-use. Every content element (text box, graphic, video clip, etc.) becomes an abstracted object that can be referenced by XML, resulting in truly Re-usable Learning Objects (RLOs) that can be used independently in presentations, sales material or testing applications. Single step translation of PowerPoint presentations into XML-tagged objects and files. Open XML specifications for flexible import controls.

Web-Based System	Windows Client OS	Windows Server OS
No	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
No	No	No

Import MS Office	Automated Publishing Process
Yes	No

Batch Processing	Customizable
Yes	No

Scalable

No Answer Provided

Data Access Speeds and Testing

No Answer Provided

Development Code

No Answer Provided

Configuration with Core ACE Tools

No Answer Provided

Supported Data Types

No Answer Provided

Access Control Schemes

No Answer Provided

Help Systems Configuration

No Answer Provided

Training Options

No Answer Provided

Data Input and Output Processes

No Answer Provided

Manual Data Processes

No Answer Provided

Current Developmental Plans

No Answer Provided

Strategic Plan

No Answer Provided

Compare / Contrast with Similar Systems

No Answer Provided

Toolset Strengths

No Answer Provided

Product	Vendor	URL
KeyView Software Developer Kits	Verity	www.verity.com

Price	Price Explanation
$0.00	Contact Vendor

Description

Gives your applications the ability to interact with the widest range of intellectual capital and languages possible - up to 295 file formats in 70 languages. Converts the most file formats possible to valid XML or Web-ready HTML. KeyView Export dynamically converts documents to well-formed, valid XML using a predefined Verity Document Type Definition (DTD). The XML output can be displayed in standard browsers using cascading style sheets (CSS) or extensible stylesheet language (XSL).

Web-Based System	Windows Client OS	Windows Server OS
Yes	Yes	Yes

Solaris Server OS	HP-Unix Server OS	IBM AIX 5-2 or OS/400 Server OS
Yes	Yes	Yes

Import MS Office	Automated Publishing Process
Yes	Yes

Batch Processing	Customizable
Yes	Yes

Scalable

The KeyView Export SDK is a thread safe toolkit that supports scalable deployments. Multiple applications of Keyview can be implemented in parallel to support scalable deployments.

Data Access Speeds and Testing

We don’t have any third party tests.

Development Code

The KeyView Export SDK is written in C and provides APIs for development in C, Java, and COM.

Configuration with Core ACE Tools

No customization of the KeyView Export SDK APIs are required to operate in different application environments. Oracle is a user of the Keyview filters. That means they OEM it into their product. Verity has hundreds of vendors that OEM the Verity filters like Oracle, Lotus, Documentum, etc.

Supported Data Types

CONTAINER FORMATS Microsoft Outlook (MSG) 97, 2000, 2002 (XP), 2003 DISPLAY FORMATS Adobe Portable Document Format (PDF) 1.1 (Acrobat 2.0) to 1.5 (Acrobat 6.0) GRAPHICS FORMATS AutoCAD Drawing format (DWG) 13, 14, and 2000 - extracts text only AutoCAD Drawing format (DXF) 13, 14, and 2000 - extracts text only Encapsulated PostScript (EPS) (raster only) TIFF header only Enhanced Metafile (EMF) no specific version Graphic Interchange Format (GIF) 87, 89 JPEG File Interchange Format no specific version Lotus AMIDraw Graphics (SDW) no specific version Lotus Pic (PIC) no specific version Macintosh Raster (PICT/PCT) 2 MacPaint (PNTG) no specific version Microsoft Windows Bitmap (BMP) no specific version PC PaintBrush (PCX) 3 Portable Network Graphics (PNG) no specific version SGI RGB Image (RGB) no specific version Sun Raster Image (RS) no specific version Tagged Image File (TIFF) 5 Truevision Targa (TGA) 2 Windows Animated Cursor (ANI) no specific version Windows Metafile (WMF) 3 WordPerfect Graphics 1 (WPG) 1 WordPerfect Graphics 2 (WPG) 2, 7 MULTIMEDIA

FORMATS MPEG-1 Audio layer 3 (MP3) ID3 versions 1 and 2 - metadata only PRESENTATION FORMATS Applix Presents (AG) 4.0, 4.2, 4.3, 4.4 Corel Presentations (SHW) 6, 7, 8, 10, 2000, 2002, 11 Lotus Freelance Graphics (PRE) 2, 96, 97, 98, Millennium Edition R9, 9.8 Lotus Freelance Graphics 2 (PRE) 2 Microsoft PowerPoint for Windows (PPT) 95 through 2003 Microsoft PowerPoint for PC (PPT) 4 Microsoft PowerPoint for Macintosh (PPT) 98 Microsoft Project (MPP) 98, 8, 2000, 2002 (XP) - metadata only Microsoft Visio (VSD) 5, 6 (2000), 2002 (XP), 2003 - metadata only Microsoft Visio XML format (VDX) 2003 - text only OpenOffice (SXI, SXP) 1, 1.1 - text only StarOffice (SXI, SXP) 6, 7 SPREADSHEET FORMATS Applix Spreadsheets (AS) 4.2, 4.3, 4.4 Comma Separated Values (CSV) no specific version Corel Quattro Pro (QPW, WB3) 6, 7, 8, 10, 2000, 2002, 11 Lotus 1-2-3 (123) 96, 97, Millennium Edition R9, 9.8 Lotus 1-2-3 (WK4) 2, 3, 4, 5 Lotus 1-2-3 Charts (123) 2, 3, 4, 5 Microsoft Excel for Windows (XLS) 2.2, through 2003 Microsoft Excel for Windows XML format 2003 - text only Microsoft Excel for Macintosh (XLS) 98 Microsoft Excel Charts (XLS) 2, 3, 4, 5, 6, 7 Microsoft Works Spreadsheet (S30,S40) 1, 2, 3, 4 OpenOffice (SXC) 1, 1.1 - text only StarOffice (SXC) 6, 7 - text only WORD PROCESSING & TEXT FORMATS Microsoft Word all versions ANSI (TXT) all versions ASCII (TXT) all versions HTML 2.0, 3.2, 4.0 IBM DCA/RFT (Revisable Form Text) (DC) SC23-0758-1 Rich Text Format (RTF) 1 through 1.7 Unicode Text 3, 4 XHTML 1.0 Generic XML 1.0 - text only

Access Control Schemes

Not applicable to this product. Access control is supported through the calling application.

Help Systems Configuration

N/A

Training Options

Full documentation and sample code is provided in the product. Minimum training required is experience with the C, Java, or COM API's and the KeyView Export SDK documentation.

Data Input and Output Processes

Data input and output automation is controlled through the calling application. What that means is that the calling application will pass the data to Keyview/export through our API set.

Manual Data Processes

Data input and output are controlled through a set of apis in the KeyView Export SDK.

Current Developmental Plans

We continue to enhance the KeyView suite of products. We are in the process of adding JAVA APIs as well as adding new data types.

Strategic Plan

Verity has a strategic plan and an NDA is required to review this plan in more detail.

Compare / Contrast with Similar Systems

The KeyView suite of SDKs provide support for filtering, conversion, and viewing of over 200 different file types without the native application. Verity has outstanding performance in the following areas: 1. Scalability 2. Performance 3. Coverage of different file formats 4. Documentation 5. Robustness 6. Technical support 7. Quality

Toolset Strengths

No Answer Provided

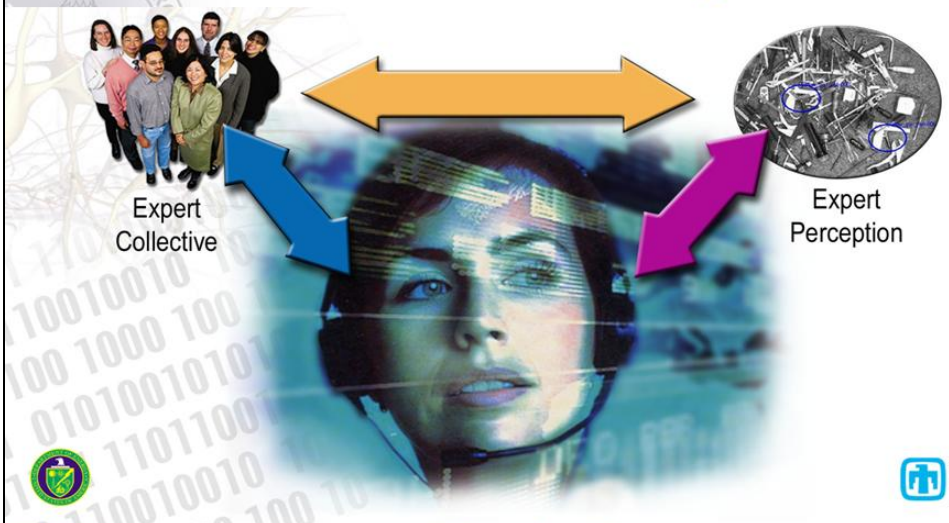
11.0 Appendix 4 - Agent – Based Experience Overview Briefing

Agent Based Experience at Sandia National Laboratories and Oak Ridge National Laboratory

Leon Chapman, PhD
505-284-9631
LDChapm@sandia.gov

The New Idea:

Enable a User to Create a Cognitive Collective of Diverse Individual Experts Having Extraordinary Perception to Help Him/Her Detect, Interpret, and Act on Meaningful but Hidden Patterns



Who uses our Agent Technology?

Sponsors Receiving Functional Agent-based Solutions

- U.S. Department of Energy
- U.S. Army (RDECOM)
- U.S. Navy Pacific Command (PACOM)
- U.S. Navy Sixth Fleet Command (C6F)
- U.S. Missile Defense Agency
- Lockheed Martin Tactical Air Systems
- Lockheed Martin Aircraft and Logistics Centers
- International Atomic Energy Agency
- U. S. Department of Homeland Security
- Tennessee Department of Homeland Security
- Battelle Memorial Institute
- U.S. Government Customer: "Railroad"
- U.S. Government Customer: "A"
- U.S. Government Customer: "C"
- Three contractor analyst organizations

Software Agent Technology
ORNL

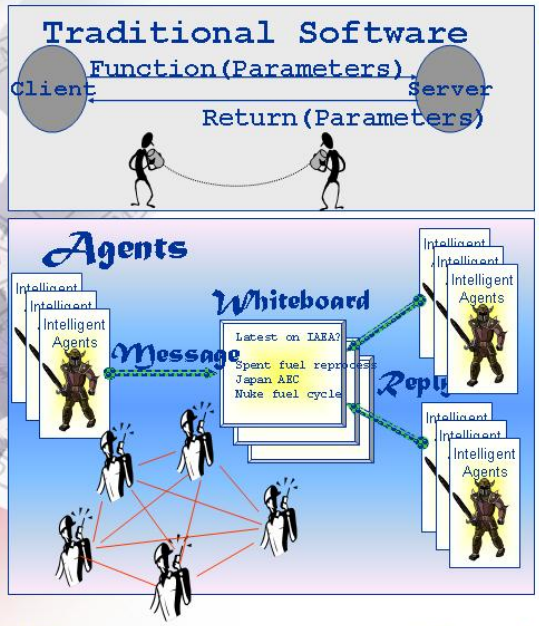
Mark Elmore
865-241-6372
ElmoreMT@ornl.gov

Leveraging the Capabilities of the
Future Force Integrated Support Team
for ACE

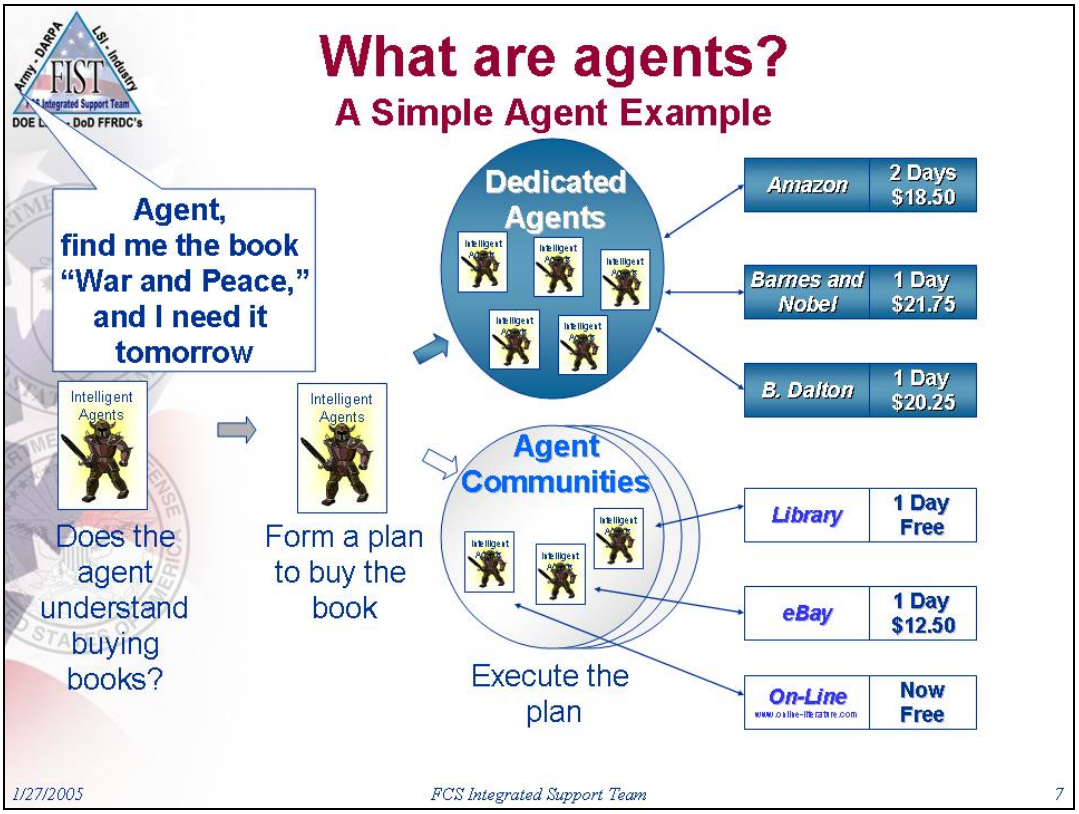
Why SW Agents as a Solution?

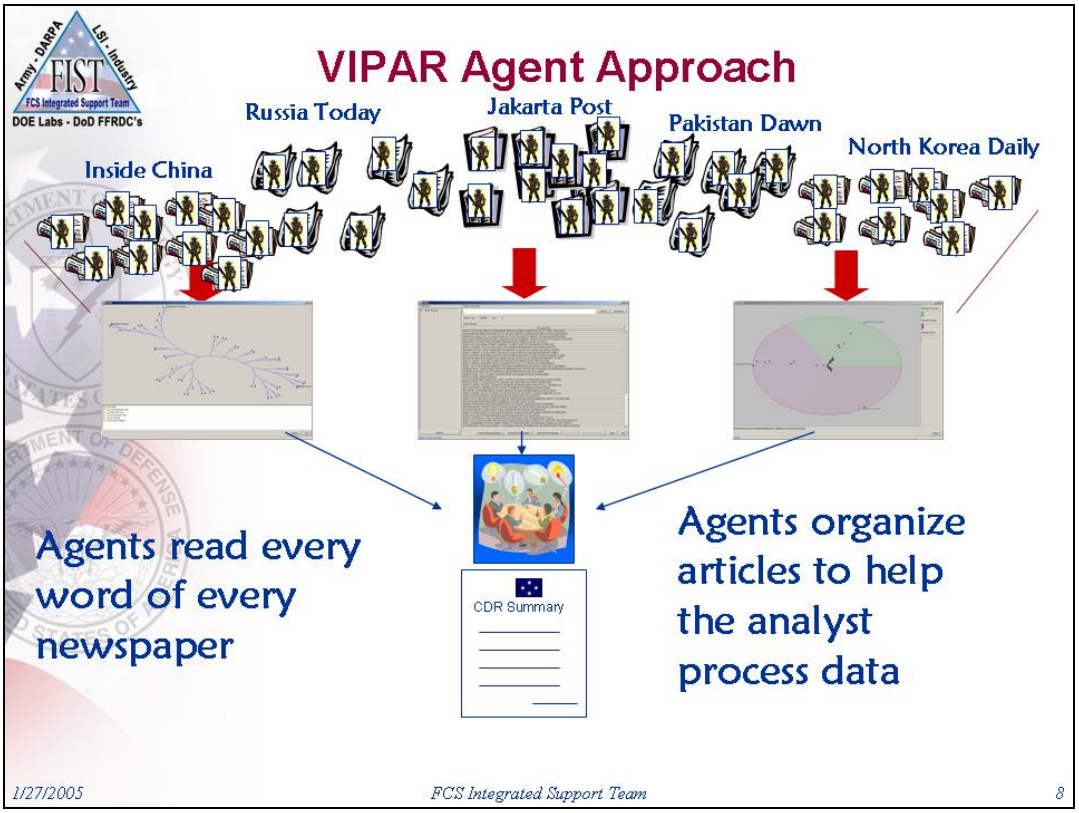
- For problems involving very large data sets
- For problems that may require real-time solution
- Adaptability to almost any application
- Robust and fault tolerant
- Extensible and Expandable
- Top-notch DOE scientists in the field
  - Delivering *effective* agent-based solutions

Why Agents?



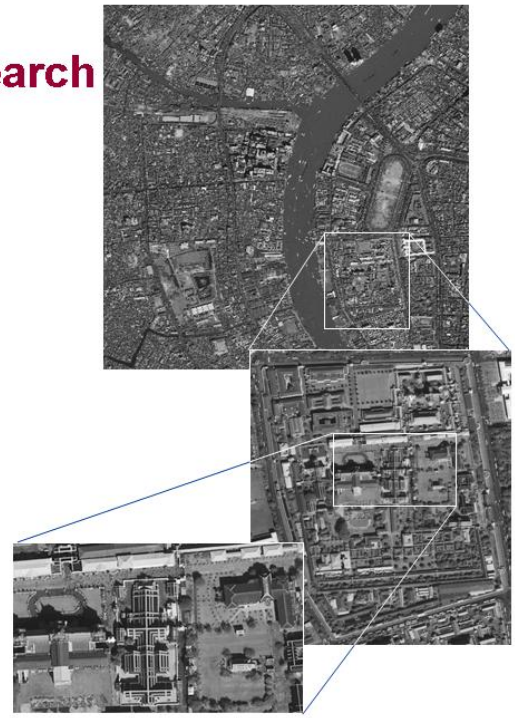
- **Traditional**
  - Client-server
  - Low-level messages
  - Synchronous
  - Move data to processing
  - Can not do the job!
- **Agent breakthroughs**
  - Peer-to-peer topology
  - Whiteboard coordination model
  - Encapsulated messaging
  - High-level message protocols
  - Asynchronous
  - Move processing to data
  - Distributed
    - Parallel processing
    - Fault Tolerant


What are agents?
A Simple Agent Example
Agent, find me the book "War and Peace," and I need it tomorrow
Intelligent Agents
Does the agent understand buying books?
Intelligent Agents
Form a plan to buy the book
Dedicated Agents
Amazon
2 Days $18.50
Barnes and Nobel
1 Day $21.75
B. Dalton
1 Day $20.25
Agent Communities
Library
1 Day Free
eBay
1 Day $12.50
On-Line
Now Free
Execute the plan
1/27/2005
FCS Integrated Support Team
7


VIPAR Agent Approach
Russia Today
Jakarta Post
Pakistan Dawn
North Korea Daily
Inside China
Agents read every word of every newspaper
CDR Summary
Agents organize articles to help the analyst process data
1/27/2005
FCS Integrated Support Team
8

Agent-based Image Search and Retrieval

- Where are the enrichment facilities?
  - More satellite image data than ever
  - Fewer trained analysts
  - Higher stakes
- I2IA is being developed so agents can find "features" in satellite images and remote sensing data



Name: Kentucky Highway
ETA: 6-9-04 / 4:35a
Previous Port: Japan
Next Port: Japan
ETD: 6-9-04 / 4:50p

Class: Car Carrier
Owner: United Car Transport Corp
Call Sign: ELJS7

Locations:
Jun-10, 32.3 / -118.8, 6am
Jun-10, 31.9 / -120.8, 12pm
Jun-10, 31.8 / -122.8, 6pm
Jun-11, 31.7 / -124.8, 12am

Sandia Agent Based Work

Shannon Spires
svspire@sandia.gov

Laurence Phillips
LRPhill@sandia.gov

AISL Research

Sandia's Advanced Information Systems Laboratory (AISL) conducts research in situated machine intelligence for applications in cybersecurity, C^2, robotics, distributed energy control, and other areas where large amounts of sensitive data must be processed quickly and reliably in hostile environments.

Our work is focused on how Intelligent Agents (IA) can be used to provide security, integrity and robustness to networks, access control, and policy enforcement in the context of decentralized, federated data sources, owners, and consumers.

AISL Technologies

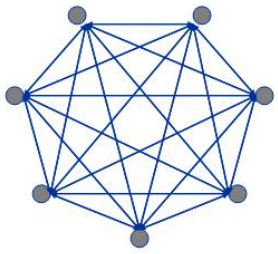
- Situated software/hardware agents
- Decentralized, distributed control for robustness
- Inherent (decentralized) information security
  - Authentication *Who wants to know and who gets to see/control what?*
  - Data Integrity *How much can I trust this data?*
  - Byzantine failure detection *Which nodes have failed or been compromised?*

Crypto Protocol Suite

Using our Standard Agent Architecture, we have implemented several tools that support secure dynamic collaboration in the face of adversaries

Reliance on n-person-rule prevents unilateral control

- **Group signature algorithms** allow agents to enforce n-agent-rule
- **Decentralized key generation protocols** enable agents to form groups able to run these algorithms
- **Key revocation and recovery** protocols allow groups to change over time



Multi-Party Signature Generation

Our Approach: I³ + Security

- **Introspection** *Provide each element with self-awareness and group awareness*
- **Integration** *Interconnect disparate infrastructure control elements*
- **Intercession** *Develop controls to regulate operation of the integrated system*
- **Security** *Detect data node failures, validate data quality, authenticate messages, withold data from those without need-to-know*

Security-related work

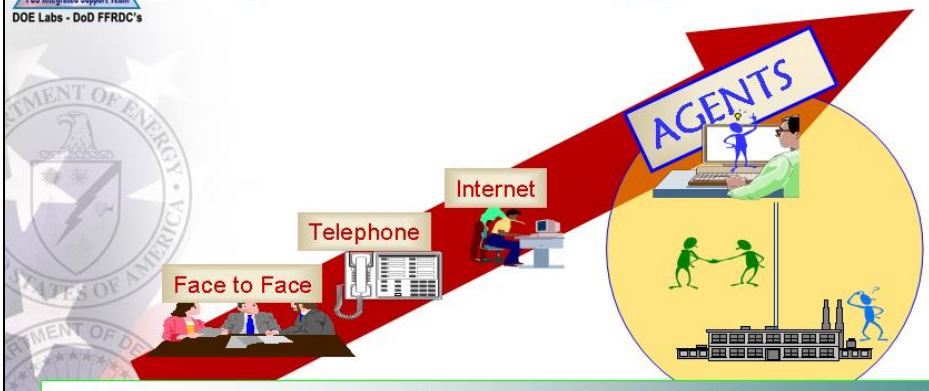
Application of agents to Secure C², Network IDS, Robotics

The threats of concern have been:
- **Malicious insiders** (throughout lifecycle)
- **Co-opted computers, software**

We want to ensure
- **Confidentiality**
- **Integrity** (and mission success)
- **Availability**
- **Access Controls** (denial of assets)

Some protections only make sense against certain adversaries

Agent Technology Trend



"The Semantic Web… will open up the knowledge and workings of humankind to meaningful analysis by ***software agents***, providing a new class of tools by which we can live, work and learn together"

Tim Berners-Lee, James Hendler, Ora Lassila

"The Semantic Web" Scientific American 5/2001

12.0 Appendix 5 - Government Development of Software Agents

Intelligent software agent technology has steadily found its way into mainstream military applications. The military, intelligence and law enforcement communities see software agents as a tool for dealing with the daunting task of having to retrieve and monitor huge amounts of data in ongoing investigations or to prevent potential problems (from terrorist activity to insider trading).

The military has been in the front of this agent stuff," confirms Dr. Noel Greis, director of the Center for Logistics and Digital Strategy at the University of North Carolina's Kenan-Flagler Business School. "In the commercial world, the interest in agents [really took off] during the e-commerce era, and most of the applications focused on the interface between the customer and the process." (Amazon.com, for example, uses intelligent agents to automatically recommend new titles based on the customer's previous purchases.) [Source: "They're in the army now…" *Technology Review*, July 2004]

The military immediately envisioned how agents can help with internal processes, especially logistics—the deployment of people, ammunition, fuel, water. In the military, agents are part o this whole digitization of the battlefield and systemization of operations. They need information captured in real time so they can see how the environment is changing dynamically and make sense of the information coming in.

2.1 DARPA

Defense Advanced Research Projects Agency (DARPA) has long been a primary sponsor of research on AI.

12.1 UltraLog

Perhaps the most promising intelligent software system being developed in the United States is **UltraLog** http://www.ultralog.net/ . Project UltraLog was established by DARPA in 2001 to take over from the Advanced Logistics Project that ran there from 1996 to 2001.

The UltraLog project's stated goal is "to create a comprehensive capability [that] will enable a massive scale, trusted, distributed agent infrastructure for operational logistics to be survivable under the most extreme circumstances." In other words, researchers are seeking the best way to use smart technology for maximum operational efficiency in combat situations—with the ability to continue operating at 80-percent capacity with up to 45-percent information infrastructure loss, for example.

UltraLog's not the only project under way, however. Others include Control of Agent-Based Systems (CoABS, also under DARPA), and Log Net, a joint-venture between the University of North Carolina, software vendor Saffron Technology of Morrisville, N.C., and Boeing Inc., headquartered in Chicago, which is developing technology for the U.S. Marine Corps. The Office of Naval Research announced in August 2003 a $5.74 million contract with the University of Southern California and Vanderbilt University to use agent software developed by computer

scientists at the two schools to handle Navy and Marine Corps pilots' schedules. There is also a more ambitious and long-term project called Future Combat Systems, fostered by the U.S. Army. [Source: "They're in the army now…" *Technology Review*, July 2004]

12.2 Control of Agent-Based Systems (CoABS)

The goal of DARPA's CoABS program is to build taskable software robots that cut the amount of time warfighters spend manipulating information systems by a factor often, rather than focusing on the mission. In addition, information that otherwise might be overlooked has greater likelihood of being incorporated into the overall picture. Lt. Cmdr. Dylan Schmorrow of DARPA's Information Technology Office is program manager for Control of Agent-Based Systems (CoABS).

DARPA's Control of Agent Based Systems program has put in place a JINI-based software infrastructure to support agent control and collaboration. JINI is not an acronym, but the name of a Java network technology of Sun Micro Systems.

2.2 Sandia National Laboratory

12.3 Sandia Intelligent Agent Program

Sandia National Laboratories has developed "Intelligent Agent" software that scans for "unusual" computer activity. Once detected, this surveillance program can automatically shutdown machines that engaged in such behavior.

Sandia National Laboratories has focused on developing agents that reason about information security of transactions between agents. In addition, the agents can relieve the operators from many routine information management tasks and detect errors in the data or lack of timely arrival of required information.

12.4 Intelligent agents and power grid coordination

Intelligent agents and multi-agent systems promise to take information management for real-time control of the power grid to a new level. Sandia has presented a concept for intelligent agents to mediate and coordinate communications between Control Areas and Security Coordinators for real-time control of the power grid.

This work was funded by the Assistant Secretary of Energy Efficiency and Renewable Energy, Office of Power Technologies of the U. S. Department of Energy.

[Source: "Agent Concept for Intelligent Distributed Coordination in the Electric Power Grid," Douglas C. Smathers and Steven Y. Goldsmith, SAND2000-1005, March 2001]

12.5 Sandia's Cyberagent

Sandia National Laboratories has developed Cyberagent, intelligent agent software that challenges computer intruders. According to Sandia, what distinguishes Cyberagent from other programs is that it integrates security functions with normal services such as FTP, WWW, and browsers. They're all in each agent. Cyberagent is used generically to mean an intelligent network software agent—that is, a software agent operating on the Internet.

Cyberagent offers a plethora of defensive cybertools. Using a sophisticated pattern- recognition system, Cyberagent can pick up and store the memory of very faint probes (almost indistinguishable from system noise) that attackers use to try to take over computers in a group. It can remove from the network a computer taken over by a hostile insider or close a system's "gates" to prevent it from being flooded with repetitive requests. Cyberagent also has prohibitions on live programs such as the I-Love-You virus entering an e-mail system. Also, decentralized control of the algorithm operating each Cyberagent makes each one autonomous yet cooperative. So, no single point of attack can bring down the collective.

Steve Goldsmith, the project's lead scientist estimates that a consumer release of the program—which is still in the laboratory stage—is three years away. "The basic agent program will be ready for specific applications in security-critical businesses and government next year, but the agent must be trained to protect a wider variety of services before it can be used by the average household."

Other members of the Sandia team include Shannon Spires, Hamilton Link, Brian Murphy-Dye, Brad Nation, Pat Gilfeather, and Gabi Istrail.

Work on the program was initially funded by Sandia's Laboratory-Directed Research and Development program in its grand challenge called Engineered Collectives. Current funding is from DOE Defense Programs.

2.3 Lawrence Livermore National Laboratory

12.6 "A System for Building Intelligent Agents that Learn to Retrieve and Extract Information"

Authors: Tina Eliassi-Rad Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Box 808, L-560, Livermore, CA 94551 USA. E-mail: eliassi@llnl.gov; Jude Shavlik, Computer Sciences Department, University of Wisconsin-Madison, 1210 West Dayton Street, Madison, WI 53706 USA. E-mail: shavlik@cs.wisc.edu

[Source: User Modeling and User-Adapted Interaction archive, Volume 13, Issue 1-2, February -May 2003, Pages: 35 – 88]

ABSTRACT: We present a system for rapidly and easily building instructable and self-adaptive software agents that retrieve and extract information. Our Wisconsin Adaptive Web Assistant (WAWA) constructs intelligent agents by accepting user preferences in the form of instructions.

These user-provided instructions are compiled into neural networks that are responsible for the adaptive capabilities of an intelligent agent. The agent's neural networks are modified via user-provided and system-constructed training examples. Users can create training examples by rating Web pages (or documents), but more importantly WAWA's agents uses techniques from reinforcement learning to internally create their own examples. Users can also provide additional instruction throughout the life of an agent. Our experimental evaluations on a 'home-page finder' agent and a 'seminar-announcement extractor' agent illustrate the value of using instructable and adaptive agents for retrieving and extracting information.

12.7 Sapphire Project: Large Scale Data Mining and Pattern Recognition

The Center for Applied Scientific Computing at the Lawrence Livermore National Laboratory is developing scalable algorithms for the interactive exploration of large, complex, multi-dimensional scientific data. By applying and extending ideas from data mining and pattern recognition, we are developing a new generation of computational tools and techniques that will be used to improve the way in which scientists extract useful information from data. [Contact: kamath2@llnl.gov -- Chandrika Kamath, (925) 423-3768]

2.4 Oak Ridge National Laboratory Applied Software Engineering Research Group

"VIPAR: Advanced Information Agents discovering knowledge in an open and changing environment"

Thomas E. Potok, Mark Elmore, Joel Reed, and Frederick T. Sheldon, Proceedings *of the IIIS Agent Based Computing*, Orlando, July 27-30, 2003.

ABSTRACT: Given the rapid evolution of information technology, most people on a daily basis are confronted by more information than they can reasonably process. The challenge to organize/classify and comprehend immense amounts of information is vitally important to the scientific, business, and defense/security communities (particularly when projecting the future evolution of information technology. For example, the defense/security community is faced with the daunting challenge of gathering and summarizing information so that military/political leaders can make informed decisions and recommendations. One such group, the Virtual Information Center (VIC) at US Pacific Command, gathers, analyzes, and summarizes information from Internet-based newspapers on a daily basis (largely a manual, time and resource intensive process).

This paper discusses the VIPAR project3, which has addressed this need. Intelligent agent technology was chosen 1) to utilize the ability for broadcast as well as peer-to-peer communication among agents, 2) to follow rules outlined in an ontology, and 3) because of the ability for agents to suspend processing on one machine, move to another, and resume processing (persistence). These strengths are well suited to addressing the challenges of automatically gathering Internet-based information.

The VIPAR system is a multi-agent system that demonstrates the ability to self-organize newspaper articles in a manner comparable to humans. The VIPAR system demonstrates the important ability where agents use a flexible RDF ontology to monitor/manage Internet-based newspaper information. Moreover, VIPAR extends this capability by dynamically adding/clustering new information entering the system. The VIPAR system includes thirteen information agents that manage thirteen different newspaper sites. Results from the project show that VIPAR can organize information in a way comparable to human organized information and validates the agent approach taken.

I2IA – Image To Intelligence Archive

I2IA is an agent-based system that autonomously manages a massive but dynamic image data archive and transforms that to an intelligence archive to aid national security needs. I2IA is being built to use agents to find "features" in satellite images and remote sensing data, using
Agent Technology
Geoconformance Technology
Image Analysis Technology
[Contact: potokte@ornl.gov – Thomas E. Potok]

Virtual Information Process Research Agent (VIPAR)

To develop intelligent agents which comb the internet intelligently, develop appropriate XML formats for insertion in the captured data, organize the data into information clusters, and lastly organize the clusters into "knowledge bins" for use by the analysts in decision making. [Contact: potokte@ornl.gov – Thomas E. Potok]

13.0 Appendix 6 - University Research in Software Agents

13.1 Carnegie Mellon University

Intelligent Software Agents

http://www.cs.cmu.edu/~softagents/
The Robotics Institute at Carnegie Mellon University has projects which address the infrastructure issues of multi-agent systems, including agent communication, interoperability, and security.

Text Miner: Text Classification for Intelligent Agent Portfolio Management

http://www-2.cs.cmu.edu/~softagents/text_miner.html

13.2 Massachusetts Institute of Technology

Media Lab: Software Agents

http://agents.www.media.mit.edu/groups/agents/
Software agents research group at MIT. Projects cover many issues concerning software agents, including collaborative filtering, user modeling, learning, multi-agent systems, privacy and security. The group is led by Prof. Pattie Maes.

MIT Media Lab, Software Agents Group

Current efforts to build the Semantic Web have been based on creating machine readable metadata, using XML tags and RDF triples to formally represent information. It is generally assumed that the only way to create an intelligent Web agent is to build a new Web, a Web specifically for machines that uses a unified logical language. MIT's group has approached solving this disparity between humans and machines from the opposite direction, by enabling machines to understand and reason on natural language statements, and giving them knowledge of the world we live in.

13.3 University of Melbourne

Intelligent Agentlab –

http://www.cs.mu.oz.au/agentlab/
The Intelligent Agent Laboratory is in the Department of Computer Science and Software Engineering, The University of Melbourne. Current major research activities are in the areas of multi-agent systems, agent frameworks, and information agents.

13.4 University of Maryland

IMPACT: Interactive Maryland Platform

http://www.cs.umd.edu/projects/impact/
A Multi-agent platform developed at University of Maryland. Each agent has several declaratively specified components such as a decision Component, a security Component, and a metaknowledge component

13.5 University of Minnesota

The Department of Computer Science and Engineering at the University of Minnesota has been working on intelligent logistics software.

13.6 University of North Carolina-Chapel Hill

The Center for Logistics and Digital Strategy at The Kenan Center, UNC, has a pilot project on intelligent logistics Software. The director of the Center is Dr. Noel Greis.

13.7 Iowa State University

An Agent-Based Environment for Integrating and Analyzing Plant Genomic Databases

http://www.cs.iastate.edu/~honavar/ailab/projects/biodkn.html
Agents for ontology assisted data integration and data mining from distributed, heterogeneous plant genomic databases.

MySpiders

While search engines have become the major decision support tools for the Internet, there is a growing disparity between the image of the World Wide Web stored in search engine repositories and the actual dynamic, distributed nature of Web data. We propose to attack this problem using an adaptive population of intelligent agents mining the Web online at query time. We discuss the benefits and shortcomings of using dynamic search strategies versus the traditional static methods in which search and retrieval are disjoint. This paper presents a public Web intelligence tool called MySpiders, a threaded multiagent system designed for information discovery. The performance of the system is evaluated by comparing its effectiveness in locating recent, relevant documents with that of search engines. We present results suggesting that augmenting search engines with adaptive populations of intelligent search agents can lead to a significant competitive advantage. We also discuss some of the challenges of evaluating such a system on current Web data, introduce three novel metrics for this purpose, and outline some of the lessons learned in the process.

[Source: Complementing search engines with online web mining agents. By: Menczer, Filippo. Decision Support Systems, May2003, Vol. 35 Issue 2, p195, 18p;]

13.8 Pennsylvania State University

Pennsylvania State University is working on the vision of employing intelligent agents to assist analysts at the U.S. Department of Homeland Security, using a technique called Collaborative Agents for Simulating Teamwork. A distinguished feature of this technique is that it enables an agent to dynamically infer information requirements of its teammates from a shared mental model about the structure and the process of the team, which is then expressed in a knowledge representation language.

[Source: PROACTIVE INFORMATION GATHERING FOR HOMELAND SECURITY TEAMS. By: Kogut, Paul; John Yen, Paul; Yui Leung, Paul; Shuang Sun, Paul; Rui Wang; Mielczarek, Ted; Hellar, Ben. Communications of the ACM, Mar2004, Vol. 47 Issue 3, p48, 3p]

13.9 University of Connecticut

The Mobile Software Agent Group at the University of Connecticut

http://www.engr.uconn.edu/dcg/

14.0 Appendix 7 - Industry Development of Software Agents

Intelligent agent research has matured to the point that commercial products are available. Most of the commercial products deal with information search and management over the Internet. Over thirty companies sell products and services which utilize agent technologies. A few of the companies are Kinetoscope, Mergent.com, Nearlife, Net Perceptions, NetBotz, Neuromedia, Inc., Nimble.com, Reticular Systems, Inc., Soar Technology, Thiink Ltd, Tryllian Inc., and Zoesis, Inc.

14.1 Artificial Life, Inc.

www.artificial-life.com

Founded in 1994, Artificial Life, Inc. (OTC Bulletin Board: ALIF - news) develops, markets, and supports intelligent, award winning software robots and intelligent agents technology and provides knowledge mining technology for life science applications and bio-computing solutions. The company offers a variety of products for applications such as self-help, self-service, consultative selling, e-CRM, e-Finance, Portfolio Management and Mobile Computing. Major customers of Artificial Life include, among others, Advance Bank, Credit Suisse First Boston, Eagle Star, Liechtenstein Global Trust, MobilCom, PricewaterhouseCoopers, UBS and ZDF. Artificial Life, Inc. is headquartered in Hong Kong.

14.2 Avalon

Cable television network Court TV has installed an archive and automation system based on intelligent data management on Sun Microsystems Inc.'s server. The archive is based on an Avalonidm (Intelligent data management).

[Source: Verdict Is In on Avalon. Authors: Kerschabaumer, Ken Source: Broadcasting & Cable; 8/4/2003, Vol. 133 Issue 31, p24, 1/4p]

Blackboard Technology

Overview: A blackboard architecture has three major components:

- a hierarchically organized global memory or database called a blackboard which saves the solutions generated by the knowledge sources;
- a collection of knowledge sources that generate independent solutions on the blackboard using expert systems, neural networks, and numerical analysis;
- a separate control module or scheduler which reviews the knowledge sources and selects the most appropriate one.

The advantages of a blackboard include separation of knowledge into independent modules with each module being free to use the appropriate technology to arrive at the best solution with the most efficiency. An additional advantage of the independent modules is the potential for using separate computing units for the independent knowledge sources, thus allowing distributed computing. This approach allows for rapid prototyping of complex problems and simplifies long-term system maintenance

14.3 Comet Way

http://www.cometway.com/
Contact: Comet Way, Inc.
4551 Forbes Ave
Pittsburgh, PA 1521
Comet Way is a private company that builds autonomous agent-based software that makes everyday things smarter.

14.4 Cycorp

http://www.cyc.com/
Suite 100
3721 Executive Center Drive
Austin, TX 78731
telephone: +1 (512) 342-4000
fax: +1 (512) 342-4040
Cycorp was founded in 1994 to research, develop, and commercialize Artificial Intelligence.

Product: The Cyc Knowledge Server is a very large, multi-contextual knowledge base and inference engine developed by Cycorp. Cycorp's goal is to break the "software brittleness bottleneck" once and for all by constructing a foundation of basic "common sense" knowledge--a semantic substratum of terms, rules, and relations--that will enable a variety of knowledge-intensive products and services. Cyc is intended to provide a "deep" layer of understanding that can be used by other programs to make them more flexible.

14.5 Engenia Software, Inc

http://www.engenia.com/
1800 Alexander Bell Drive Suite 100
Reston, Virginia 20191
Phone: 703-234-1400
Fax: 703-234-1430
Provides innovative enterprise solutions that use intelligent "software agents" to integrate the planning, execution and reporting functions of complex organizational processes. With Engenia's non-invasive, end-to-end process management solution, projects are streamlined and productivity is optimized, resulting in increased innovation, customer satisfaction, and profitability.

14.6 Agent Orientated Software

www.agent-software.com/
Key personnel:
Dr Andrew Lucas, managing director
Andrew Hodgson, V.P. Technical
Paul Maisano, JACK Product Manager

Intelligent agent system: JACK

Applications: Used on UAVs (unmanned aerial vehicles) A personal digital agent (PDA) is linked to its flight control system. The PDA carries intelligent agent software, decision-making software which makes the route decision. The use of the software on UAVs was tested by Australia's Defence Science and Technology Organisation. [Source: Flight International, August 3, 2004 p22]

14.7 21st Century Systems, Inc.

http://www.21csi.com/enter.htm

Key Personnel:
Dr. Alexander D. Stoyen is Founder and CEO, Alexander.Stoyen@21csi.com

Funding: 21st Century Systems Inc. has been funded by Navy and DOD SBIR Program

Synopsis: Founded in April 1996, 21st Century Systems, Inc. ™ (21CSI™) is a privately held, growing company. 21st Century Systems is a pioneer in intelligent agent based decision support software technology. An eight year old DOD contractor, 21st Century Systems, Inc. is applying its revolutionary technology to a wide spectrum of mission-critical systems, from dismounted soldier to Navy ships to Command and Control centers.

Recent News: Dr. Stoyen testified before the House Armed Services Committee on July 21, 2004 on the subject of Small Business Technology and Innovation.

Contracts: The Army has awarded a Phase III SBIR contract to 21CSI® to further develop a software product for a tactical decision support system.

Commercial spin-off: Recently, 21CSI has spawned off a separate entity, AgentKind™, Inc., to better support the AEDGE™ product family.

Products:
21st Century Systems, Inc.'s products have application across the full spectrum of mission areas. Our decision support software can be used when one has too much information, not enough information, uncertainty of information, in real-time situations, post-op, tactical or security, and so on. Some of the representative SBIR Phase III applications of our adaptive technology are as follows.

- The Consolidated Undersea Situational Awareness System focuses on decision support under uncertainty, for the Submarine Fleet.
- The Advanced Battlestation with a Decision Support System (ABS/DSS) provides tactical decision support for command and control personnel aboard Aircraft Carriers.
- A CUSAS variant, Shipboard Automated Reconstruction Capability (SHARC) is going to provide Navy platforms (beginning with the Virginia class submarine) with mission reconstruction and training tools.

- SituSpace provides a Single Integrated Space Picture for the Army and (under consideration) Joint space operations.
- HiRSA, or High Resolution Situational Awareness is being fielded in August with the 1 Marine Expeditionary Force in Iraq beginning in August, and provides a high-resolution 3D command and control platform for urban tactical operations.
- ExLoG21 provides the intelligent decision support capability for autonomic logistics systems, supporting expeditionary logistics operations.
- SentinelNet creates a fused, intelligent network of sentries, sensors, computers and command and control center personnel, and will provide Anti-Terrorism/Force Protection command and control for Navy ships and shore installations.
- SASA provides weapon-target pairing and automated ingress/egress routing for tank crews.
- Webster is an intelligence analysis tool that continuously monitors, interprets and correlates web-based open source information, providing analysts with correlated hypotheses on terrorist activities occurring worldwide.

http://researchweb.watson.ibm.com/able/

14.8 IBM

Agents are a chief focus of a wide-ranging project at IBM's Research Div., called the Agent Building & Learning Environment (ABLE) at the IBM T.J. Watson Research Center. Under ABLE, IBM's strategy is to build stronger, engineering-caliber versions of AI software "to make AI more ruggedized, more successful," says Nagui Halim, head of distributed computing research at IBM. These will be used to create an integrated toolbox of AI methods, plus an expert-system manager for picking the right tool or combination of tools for each specific job. These will be used to create an integrated toolbox of AI methods, plus an expert-system manager for picking the right tool or combination of tools for each specific job. "We believe ABLE could have significant commercial impact in the near term as the nervous system helping to control the bones and muscles of conventional systems," Halim adds.

Joe Bigus is the project leader.
[Source: "Smart Tools," ***The America's Intelligence Wire***, March 25, 2003]

14.9 Livewire Logic, Inc.

Contact:
2700 Gateway Centre Blvd., Ste. 900
Morrisville, NC 27560
Phone: 919-234-2144
Fax: 919-234-2145

http://www.realdialog.com

Key Personnel:
Dr. James Lester, Founder
CEO, Michael Lough
VP, Technology, Bradford Mott

Company type: Private

Size: small (30 employees)

Company Synopsis: Founded in 2000, LiveWire Logic develops web-based dialog software such as RealDialog (web-based two-way dialog interface) and RealDialog Agent (automated intelligence response agent). RealDialog helps reduce redundant calls and emails arriving into support centers and improves online selling opportunities. The company's founding team specializes in intelligent virtual agents and computational linguistics, and is recognized as leaders in advanced artificial intelligence research.

Product: RealDialog

Recent News: Water Pik Technologies, Inc. to Present Use of Automated Customer Self-Service Agents from Livewire Logic (Business Wire, 06/04/2004)

14.10 Lockheed Martin Advanced Technology Laboratories

http://www.atl.external.lmco.com/
Contact:
Lockheed Martin Advanced Technology Laboratories
3 Executive Campus
6th floor South
Cherry Hill, NJ 08002

They have been involved in over a dozen intelligent, mobile agent programs supporting all branches of the military. Programs for which the Advanced Technology Laboratories has developed intelligent agent technologies include:

- DARPA's Control of Agent-Based Systems (Multi-Agent Common Operating Environment)
- DARPA's Control of Agent-Based Systems (Cooperating Agents for Specific Tasks)
- DARPA's Human Computer Interactions (Domain Adaptive Information System)
- DARPA's Small Unit Operations (CyberAngels)
- DARPA's Joint Logistics Advanced Concept Technology Demonstration (CyberExpress)
- DARPA funding for the Domain Query Ontology
- U.S. Communication-Electronics Command's Logistics Command and Control Advanced Technology Demonstration
- DARPA's Listen Compute Show-Marine
- DARPA Communicator
- U.S. Air Force Rome Laboratory's Agent-based Decision Aids for Mobility Operation

14.11 NeurOK, LLC

www.**neurok**.com/

Contact:
Executive Plaza Center
2010 Corporate Ridge
Suite 700, McLean, VA 22102

NeurOK has developed its proprietary agentware technology allowing software agents to learn the meaning of words and understand the context of documents. Intelligence for software agents means first of all the capability of understanding to some extent the meaning of information. This implies understanding the context in which the words are used, since the same word has different meanings in different contexts. The context in its turn is defined by the same words. Thus one faces a recursive problem: the context is defined by the words, and the meaning of the words depends on context. NeurOK's core semantic technology resolves the above recursive problem. The solution is based on proprietary patent pending learning algorithm, which automatically builds up the self-consistent set of semantic categories, allowing for comparison of the meaning of pieces of information: queries, phrases, texts. This set of categories depends strongly on the learning corpus. Collections with different topics will result in different set of categories, each optimized to the given subject. It makes it possible to educate the agents having their expertise in various areas. Note, that this learning technique is completely language independent. The meanings of the words are extracted from the way they are used, independent on the nature of the language. Moreover, using multilingual corpus one can teach agents to understand relations between the words in different languages, paving the way for cross-language applications. Information sorting and filtering is one of the basic operations in Knowledge Management. NeurOK provides the complete solution for these applications. Provided the existence of directory, NeurOK's Semantic Server can automatically rubricate the incoming information stream, redirecting information according its content. This feature may be useful in many eBusiness applications.

14.12 Nu Tech Solutions Inc.

www.**nutechsolutions**.com/
Corporate Headquarters:
Charlotte, NC
NuTech Solutions, Inc.
8401 University Executive Park Drive
Suite 102
Charlotte, NC 28262

Phone: (704) 549-4480
Toll-Free: (800) 526-6784

Products: Ascape is NuTech's framework for creating generative, or "Agent Based Models."
NuTech Solutions' Merix software is a fraud detection system that uses intelligent agents. Merix' knowledge discovery system automates the process of predicting new fraud schemes as they begin. Once Merix is confident that it has detected a new pattern of fraudulent activity, it automatically 'learns' by integrating that new knowledge into the system.

Mergers/Acquisitions: By early 2003, Santa Fe's BiosGroup was acquired by NuTech Solutions Inc. in Charlotte, N.C.

Key People:
Dr. Stuart A. Kauffman, co-founder
Tom H. Wilson, Jr., new CEO effective April 1, 2004.
Kauffman tackled complexity using an army of ``ants" -- small software entities called agents that are modeled after the dynamics of ant colonies. His explorations with agents helped Southwest Airlines Co. improve its cargo operations and streamlined the supply chain at Procter & Gamble Co. Southwest more than $10 million in the first year of operation using the BiosGroup agents.

14.13 Roke Manor Research

www.**roke**.co.uk/

A Siemens subsidiary based on the south coast of Britain, Roke Manor Research has come up with an automated way of keeping an eye out for change. The research lab's "video motion anomaly detection system" analyses live CCTV images for out-of-the-ordinary events-bringing them to the operator's attention by highlighting them on screen.

According to Richard Evans, a senior researcher on the project, the system works in two parts. In the first, a video motion processor tries to find localized features in the live image that are distinct enough to be tracked from frame to frame. These features may be something that a human could recognize-for instance, the edge of a car bumper (fender). The reference points may be meaningless to humans; what matters is that they can be tracked.

The second part of the system builds up a statistical history of how such features normally move through the image, tracking their speed and direction. When the CCTV image changes, the system can check against what it has "learnt" to decide whether the new event is so unusual that it should be brought to an operator's attention.

[Source: Title: Who watches the watchers? , Economist, 00130613, 9/6/2003, Vol. 368, Issue 8340]

14.14 Saffron Technology

http://www.saffrontech.com/

Product: SaffronOne is a proprietary digital associative memory engine that is able to capture instantaneous associations among very large quantities of data. With incremental learning, the agent learns in real time as information is encountered.

Applications: Boeing's LogNet Project uses SaffronOne intelligent agents to assume roles in logistics management. LogNet is an intelligent information system that provides an integrated view of the military logistics environment. The Boeing LogNet system will enable logisticians to gain "situational awareness" of forces around the world, and to make better decisions about how to support them.

Saffron Technology's co-founder, chairman and chief scientist, Manny Aparicio, says one obstacle to adoption in the commercial sector is the issue of trust, especially with learning agents—ones that gather information about habitual human preferences and responses over time. People are nervous about delegating responsibility to a piece of software, Aparicio says. But, once they learn to trust it, things change. "The user can have it set so it just makes recommendations, but then they might get bored seeing it's right all the time and let it build a whole re-supply strategy," Aparicio says. There's always a safety net, he adds. "Whenever an agent sees a situation it doesn't know, it can still send an exception back."

14.15 Semaview

http://www.semaview.com/home/home.html

Semaview is a Toronto based company developing products by and for the semantic web.

14.16 SonicBoomerang Inc.

www.sonicboomerang.com.

Located in Toronto, SonicBoomerang Inc. is Canada's leading provider of intelligent agent software and services.

Company Type: Private

Product: ClassPro; document classification software

Recent News:

CanWest Interactive announced today that infomart.ca has gone live with ClassPro, SonicBoomerang Inc.'s state-of-the-art document classification software. With over 1,700 topics or geographies classified, this represents one of the largest topic classification systems implementations in the world to date.

In March 2003, the Information Highways and the e-Content Institute named SonicBoomerang's ResearchAssistant the Most Innovative Application of 2002.

[Source: TORONTO--(BUSINESS WIRE)--Oct. 23, 2003]

15.0 Appendix 8 - Commercial Agents Tabular View

Product Name	Product URL Page	Agent Description Instance on Home Page
Artificial Life, Inc.	http://www.artificial-life.com/	Referential
Avalon	http://avalon.apache.org/	Referential
Cycorp	http://www.cyc.com/	Referential
Livewire Logic, Inc.	http://www.realdialog.com	Web Page Content
NeurOK, LLC.	http://www.neurok.com/	HTML Source Code
Nu Tech Solutions, Inc.	http://www.nutechsolutions.com/	Referential
Roke Manor Research	http://www.roke.co.uk/	Referential
Saffron Technology	http://www.saffrontech.com/	Referential
Semaview	http://www.semaview.com/home/home.html	Referential
NetAngels	http://www.netangels.com/	Referential
Agent Oriented Software	http://www.agent-software.com/shared/home/	Web Page Content
Topics and Search '97	http://www.verity.com/	Referential
News Page, First! And Hoover	http://www.individual.com/login.php	Referential
San Jose Mercury News	http://www.mercurynews.com/mld/mercurynews/	Referential
21st Century Systems Inc.	http://www.21csi.com/enter.htm	Web Page Content
Comshare	http://www.performance.geac.com/	Referential
Agent Knowledgebase Associates	http://akainc.com/	Referential
Movie Critic	http://www.moviecritic.com/	Referential
Broad Vision	http://www.broadvision.com/bvsn/bvcom/demand/home.do	HTML Source Code
Web Browser Intelligence, IBM	http://www.ibm.com/us/	Web Page Content
Fido, the shopping dog	http://www.shopfido.com/	Referential
Amazon Books	http://www.amazon.com/	Referential
Edify	http://www.edify.com/	Referential
CometWay	http://www.cometway.com/	Web Page Content
British Telecom Laboratories	http://www.labs.bt.com/	Web Page Content
AgentWare Systems, Inc.	http://www.agentwaresystems.com/	Web Page Content
Plan-b-media.de	http://www.planb-media.de/english/	Referential

Product Name	Product URL Page	Agent Description Instance on Home Page
Engenia	http://www.engenia.com/	Web Page Content
Living Systems	http://www.living-systems.com/pages/index.html	Referential
DataBots	http://www.imagination-engines.com/databots.htm	Web Page Content
Topia Ventures	http://www.topiaventures.com/	Web Page Content
WebWrappers Corporation	http://www.webwrappers.com/	HTML Source Code
NovoMind	http://www.novomind.com/index_ht_en.html	HTML Source Code
Sonic Boomerang, Inc.	http://www.sonicboomerang.com/	Web Page Content

Key:

Referential: When there is a reference to a web page of a Company that uses software agents, but there is no mention about it on the home page or in the HTML source code.

HTML Source Code: When there is a reference to a web page of a company that uses software agents, but there is no mention about it on the home page. However there is mention about software agents in the HTML Source Code.

Web Page Content: When there is a reference to a web page of a company that uses software agents, and it is mentioned in the content of the company home page.

16.0 Appendix 9 - Government Activity Related to KDDM

16.1 Lawrence Livermore National Laboratory

Data Foundry

Introduction

DataFoundry is an ongoing research effort to improve scientists' interactions with large data sets. Because this is a broad goal, the project focus evolves with the needs of the scientists using DataFoundry. Efforts within the current scope include:

- improving access to distributed, heterogeneous data, for example through multi-database or data warehousing techniques
- reducing the size of the data sets being analyzed, for example by filtering data \
- providing novel ways of interacting with the data, for example allowing a broad range of user defined queries
- determining appropriate ways to store data for efficient retrieval.

Source: http://www.llnl.gov/CASC/datafoundry/

The DataFoundry **Ad-Hoc Query Project** is an ongoing effort to help scientists explore terabytes of scientific simulation data by permitting ad-hoc queries over the data while reducing data storage requirements and access times. To accomplish this, we are combining database and data analysis techniques to develop an infrastructure that evaluates queries against mathematical and statistical models of the data instead of against the full data set. By utilizing multiple models of varying complexity, we expect to support several types of queries including both simple SQL-like range queries and complex similarity matching ones.
Source: http://www.llnl.gov/CASC/datafoundry/AdHocQuery.html

Large Scale Distributed Data Access Project The Internet is becoming the preferred method for disseminating scientific data from a variety of disciplines. This has resulted in information overload on the part of the scientists, who are unable to query all of the relevant sources, even if they knew where to find them, what they contained, how to interact with them, and how to interpret the results. Thus instead of benefiting from this information rich environment, scientists become experts on a small number of sources and use those sources almost exclusively. Enabling information based scientific advances, in domains such as functional genomics, requires fully utilizing all available information. As part of a collaborative effort with Georgia Tech, SDSC, and NCSU we are developing an end-to-end solution using leading-edge automatic wrapper generation, mediated query, and agent technology that will allow scientists to interact with more information sources than currently possible.
Source:

Project Leader

Terence Critchlow
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
critchlow@llnl.gov
Phone: (925) 423-5682

Sapphire: Data Mining and Pattern Recognition for Large and Complex Science Data

The Sapphire project is developing scalable algorithms for the interactive exploration of large, complex, multi-dimensional scientific data. We are applying and extending ideas from data mining and pattern recognition in order to improve the way in which scientists extract useful information from data. To address the challenges that arise when data mining techniques are applied to massive and complex data sets, we are focusing on the following research areas:

- Image processing techniques for de-noising, object identification, and feature extraction
- Dimension reduction techniques to handle multi-dimensional data
- Scalable algorithms for classification and clustering
- Parallel implementations for interactive exploration of data
- Applied statistics to ensure that the conclusions drawn from the data are statistically sound

Source: http://www.llnl.gov/CASC/sapphire/pubs/UCRL-JC-132151.abs.html

Project Leader

Chandrika Kamath
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
kamath2@llnl.gov
Phone: (925) 423-3768

Potentially Relevant LLNL LDRD Projects

Concealed Threat Detection at Multiple Frames-per-Second (02-ERD-061)

Automated Imagery Data Exploitation (AIDE) (02-ERD-034)

Techniques for Judging Intent Behind Cyber Attacks (03-ERD-012)

Surveillance, Prediction and Insight for Decision making for Earliest Response (SPIDER) (03-ERD-046)

Detection and Tracking in Video (03-ERD-031)

ViSUS: Visualization Streams for Ultimate Scalability (02-ERI-003)

Enabling Large-Scale Data Access (02-ERI-007)

Image Content Engine (03-SI-003)

Program of Simulations and Experiments for Assessment of Rapid Multipurpose Cargo-Scanning Technologies (02-ERD-064)

16.2 Sandia National Laboratories

VxInsight

V_xInsight(TM) is a tool for discovering relationships within large databases. While most data retrieval tools and most data mining tools are able to find information in a database, they only tell you about the data elements. **V_xInsight**(TM) reveals the implicit structure of the data. **V_xInsight**(TM) can help analysts uncover strategically important connections and patterns making it an important knowledge management tool.

Source: http://www.cs.sandia.gov/VIS/

Project Leader

George Davidson
Data Analysis and Visualization
gsdavid@sandia.gov
(505)844-7902

Potentially Relevant SNL LDRD Projects

Adaptive Awareness for Personal and Small Group Decision Making
Principal Investigator: PANCERELLA,CARMEN M.
Project Manager: WAGNER,JOHN S.

An Advanced Learning Model for Agent Behavior
Principal Investigator: PRYOR,RICHARD J.
Project Manager: DAVIDSON,GEORGE S.

Research and Development of Mathematical Algorithms to Define and Quantify Critical Infrastructure Interdependencies
Principal Investigator: STAMBER,KEVIN L.
Project Manager: SNYDER,LILLIAN A.

War on Terrorism
Principal Investigator: MENDENHALL,FREDERICK T.
Project Manager: WOODARD,JAMES B

Cognition Driven Augmented Analyst
Principal Investigator: WAGNER,JOHN S.
Project Manager: WOODALL,TOMMY D.

Detecting and Tracking the Active Insider Using 3D Detection Technology

Principal Investigator: NELSON,CYNTHIA L.
Project Manager: ORTIZ,STEPHEN

Extensibility of Knowledge-Based Human Agent Simulation
Principal Investigator: FORSYTHE,JAMES C.
Project Manager: JORDAN,SABINA ERTEZA

Cognitive Models Applied to Human Effectiveness in National Security Environments
Principal Investigator: DOSER,ADELE BEATRICE
Project Manager: WAGNER,JOHN S.

Mathematical Analysis of Deception
Principal Investigator: COHEN,FREDERICK BERNARD
Project Manager: HESS,BARRY V.

Development of Computational Algorithms and Inversion Capabilities for Transport/reaction Simulations of Chemical/biological/radiological Terrorist Attack Scenarios in Support of Homeland Security
Principal Investigator: VAN BLOEMEN WAANDERS,BART G.
Project Manager: FINLEY,RAY E.

Graduated Embodiment for Sophisticated Agent Evolution and Optimization
Principal Investigator: BOSLOUGH,MARK B. E.
Project Manager: DAVIDSON,GEORGE S.

Adaptive Awareness for Personal and Small Group Decision Making
Principal Investigator: PANCERELLA,CARMEN M.
Project Manager: WAGNER,JOHN S.

Improving Human/System Interactions in Systems-of-Systems
Principal Investigator: TUCKER,STEPHENSON
Project Manager: WAGNER,JOHN S.

System of Systems Modeling and Analysis
Principal Investigator: ANDERSON,DENNIS J.
Project Manager: CRANWELL,ROBERT M.

Understanding Communication in Counterterrorism Crisis Management
Principal Investigator: JOHNSON,MICHAEL M.
Project Manager: HIRANO,HOWARD H.
The Endowment of Simulator Agents with Human-Like Episodic Memory
Principal Investigator: FORSYTHE,JAMES C.
Project Manager: HIRANO,HOWARD H.

A Model of Infrastructure Interdependency Using Communication Agents
Principal Investigator: BARTON,DIANNE C.
Project Manager: NELSON,JENNIFER

Research and Development of Mathematical Algorithms to Define and Quantify Critical Infrastructure Interdependencies
Principal Investigator: BEYELER,WALTER E.
Project Manager: STAMBER,KEVIN L.

Augmented Cognition: Next-Generation Intelligent Systems
Principal Investigator: ELLIS,LARRY J.
Project Manager: SKOCYPEC,RUSSELL D.

Winning the War: A Systems Approach to Defending Our Borders
Principal Investigator: WOODALL,TOMMY D.
Project Manager: HORSCHEL,DANIEL S.

Self Organizing Software Research and Development
Principal Investigator: OSBOURN,GORDON C.
Project Manager: HAYS,GERALD N.

16.3 Oak Ridge National Laboratories

Collaborative Management Environment (CME)

CME is a research project funded by the Department of Energy (DOE) to investigate advanced information technologies for improved management of research information across the DOE complex of national laboratories. The CME system is capable of storing research proposal information in such a way that it can be queried based on general keywords, and by field-specific keywords. The research proposal information is then viewable over the Internet. Since each National Laboratory has a different presentation form for the research proposal information, the system presents this information in its original format.
Source:

Frank V Damiano
DAMIANOFV@ornl.gov

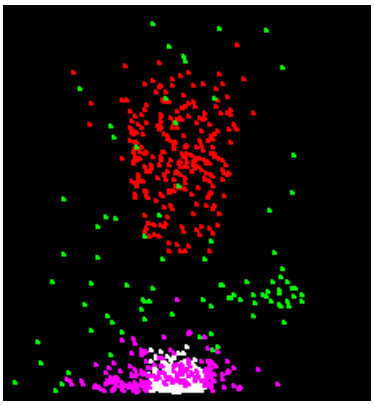
Data Mining at ORNL

Research Emphasis

- Dimension reduction and feature visualization in large data sets
- model search algorithms
- Advanced Statistical techniques
- Information Visualization techniques
- Algorithms for information discovery
- Change and Outlier Detection for Networks of Heterogeneous Sensors
- Parallel and distributed processing

Source: http://www.csm.ornl.gov/dm/

Analysis of Huge Datasets



Methodology for **analysis of huge datasets** by extracting and classifying features worthy of further investigation: paper.pdf (270 Kbytes), paper1 (2,642 Kbytes), paper2 (2,771 Kbytes), paper3 (184 Kbytes), prototype software. Several methods are developed. Among them, local modeling with global analysis can be viewed as a generalization of chaotic time series analysis. The "fireworks" on the right are the result of cluster analysis applied to parameters of local models on an atmospheric radiation measurement time series. Such analysis can identify and classify unusual segments of large data series. In most cases, each cluster has an application-specific interpretation because the local models are application-specific.

Source: http://www.csm.ornl.gov/~ost/compstat.html

Image to Intelligence Archive (I2IA)

The volume of digital image data is exploding Increasing number of satellites and airborne sensors with very high spatial, spectral, and temporal resolutions. We propose to develop an agent-based system that autonomously manages a massive but dynamic image data archive and transforms that to an intelligence archive to aid national security needs.

Source: http://www.csm.ornl.gov/~v8q/Homepage/Projects/I2IA.htm

Project Leader

Dr. Thomas E. Potok
Applied Software Engineering Research Group Leader
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
865-574-0834
potokte@ornl.gov

Virtual Information Process Research Agent (VIPAR)

To develop intelligent agents which comb the internet intelligently, develop appropriate XML formats for insertion in the captured data, organize the data into information clusters, and lastly organize the clusters into “knowledge bins” for use by the analysts in decision making.
Source: http://www.csm.ornl.gov/~v8q/Homepage/Projects/vipar.htm

Project Leader

Dr. Thomas E. Potok
Applied Software Engineering Research Group Leader
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
865-574-0834
potokte@ornl.gov

16.4 Pacific Northwest National Laboratory

Starlight Information Visualization Technologies

The Pacific Northwest National Laboratory's *Starlight Information Visualization System* (Starlight) is a forerunner of an emerging new class of information system, one that couples advanced information modeling and management functionality with a visualization-oriented user interface. This approach makes relationships that exist among the items in the system *visible*, enabling exciting and powerful new forms of information access, exploitation, and control. The product of over six years of information visualization research, Starlight is simultaneously a powerful information analysis tool and a platform for conducting advanced visualization research.
Source: http://starlight.pnl.gov/

A Method for Generating Analyses of Categorical Data

Methodologies for browsing, retrieving and viewing categorical data objects are disclosed. The data are treated in a probabilistic domain by presuming that there is an underlying probabilistic structure that naturally distinguishes the object on which the categorical data measurements were derived. High dimensional representations for the measurements are constructed. These representations can be processed in a computer system to browse, retrieve and view the objects. Two methodologies are also presented for obtaining a holistic view of a collection of objects based on multiple, similar categorical measurements of those objects. One of these methodologies is similar in spirit to, but significantly different than, a principal-components view of a collection of data objects. The other is based on a general strategy for constructing the representation from categorical data, is also described.
Source: http://availabletechnologies.pnl.gov/infotechenergy/amet.stm

Human Interface Workspace (HI-SPACE)

The key to developing the next-generation human-to-information interface is to move beyond the limitations of computer monitors as our only view of the electronic information space and keyboards and mice as our only interaction devices. Our physical information space, which includes desks, tables, and other surfaces, should be our view into the electronic information space. People perform physical interactions with information every day by picking up a book, building a model, or writing notes on a page. Similar interactions need to be developed for electronic information. The Human Interface Workspace (also known as HI-SPACE) is being developed to support leading-edge human computer interaction features, such as:

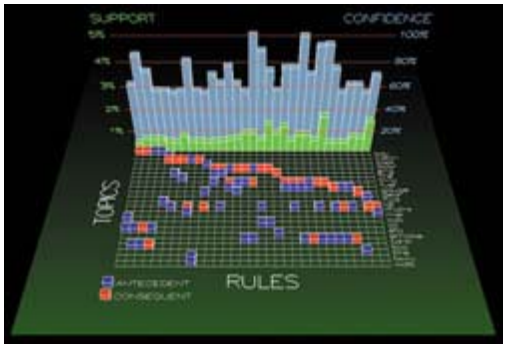
Pacific Northwest's researchers envision HI-SPACE as a new physical workspace for presenting and analyzing information, one in which all elements integrate to enhance the decision-making process.

- taking advantage of the redundancy of multimodal input (gesture and speech recognition)
- allowing more natural direct interactions with the information
- supporting group interactions with the same data at the same time and in the same space
- enabling users in different locations to interact with each other and with the same data set

- supporting fluid transfer of data and interaction between physical and electronic spaces
- maintaining an unencumbered work environment.

Source: http://www.pnl.gov/cse/computersci/hispace.htm

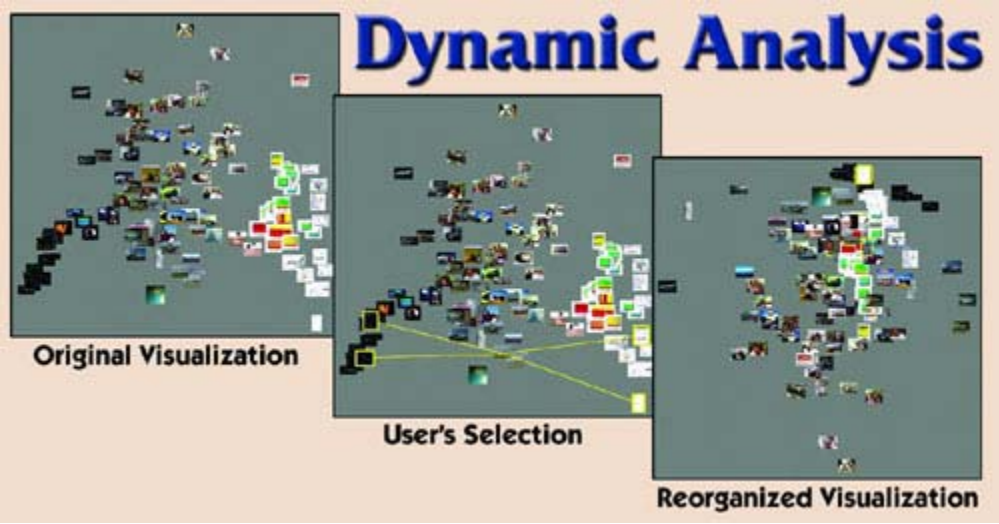
Data Mining



Pacific Northwest National Laboratory's researchers have developed tools that combine data mining with visualization to allow users to simultaneously study patterns, associations, and other complex relationships in large collections of data. These tools not only uncover hidden and unpredicted relationships among documents, people or concepts, they also identify the absence of relationships where they were expected. The visualization shows statistical patterns identified by the data mining engine, making associations easier to grasp and explore.
Source:

Dynamic Analysis

Dynamic analysis gives users the opportunity to define which images in a collection should be considered similar. The tool reorganizes and displays the images based on the user's guidance. This approach could be used to analyze image databases including photos, images captured by research or medical instruments, or satellite imagery, as well as applied to text or other kinds of data.

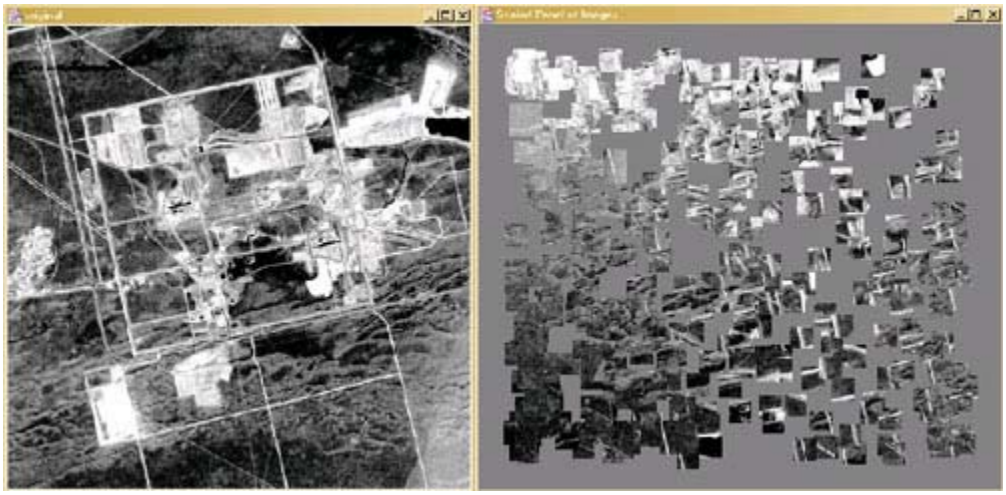


Source: http://www.pnl.gov/cse/computersci/dynamic.htm

Image Classification

At home, sorting a shoebox of photos can be a daunting task. Pacific Northwest National Laboratory has developed a solution for sorting, classifying and organizing image and video data in similar situations. This technique automatically analyzes images and calculates a summary of the collection. It creates a summary of snapshots, commercial photo collections, electronic image

libraries or even surveillance videos. We also have approaches for automatically classifying objects within images that could be built directly into instruments such as microscopes or satellites.



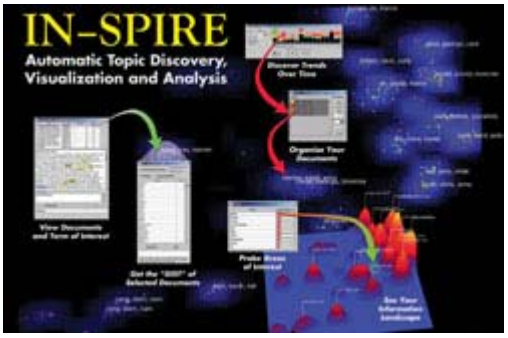
Source: http://www.pnl.gov/cse/computersci/imgclassification.htm

Imaging Architecture

Existing image processing tools are predominantly workstation based, and require imaging algorithms to be coded in a specific manner (for example, language) for every different tool. This greatly inhibits computational performance and makes it prohibitively costly to rapidly deploy new imaging algorithms across multiple image processing environments. In Pacific Northwest National Laboratory's (PNNL) Imaging Science and Technology (ISAT) program, a solution to these problems is being developed. This approach creates a software architecture that integrates existing image processing technology with extended imaging services and data management capabilities provided by the emerging scientific Grid environment. This component-based architecture extends the capabilities of existing tools to leverage high performance computing resources, and only requires a single, sharable implementation of a new algorithm to be constructed. In this manner, scientists can continue to work with their preferred image processing technologies, while seamlessly exploiting the computational and data resources on the Grid.
Source: http://www.pnl.gov/cse/computersci/imaging.htm

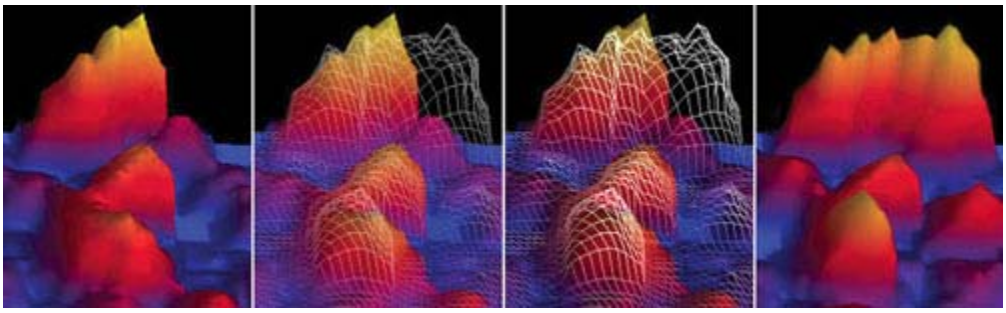
IN-SPIRE (Spatial Paradigm for Information Retrieval and Exploration)

IN-SPIRE quickly and automatically conveys the gist of large sets of unformatted text documents such as technical and patent literature, marketing and business documents, web data, accident and safety reports, newswire feeds and message traffic. By clustering similar documents together, this Windows-based software unveils common themes and reveals hidden relationships within the collection. Building upon nearly a decade of Pacific Northwest National Laboratory's research for the U.S. Government, IN-SPIRE allows analysts to spend more time exploring the information they find most relevant and less time sifting through the masses of irrelevant documents. IN-SPIRE automatically analyzes a multitude of text files, determines the key topics or themes in each, and creates a signature for each document in the collection. IN-SPIRE's two visualizations display representations of the documents so those with similar or related topics appear closer together. The Galaxy visualization uses the metaphor of stars in the night sky, with each star representing individual documents. The ThemeView™ visualization uses a three-dimensional terrain map to provide a high-level overview of the data.



IN-SPIRE allows users to quickly and easily discover trends, key issues, and hidden information relationships in large volumes of text. Full Image (jpg 109KB)

Users explore and interact with these visualizations to gain insight. While standard query and trend analysis tools identify documents relevant to a query, IN-SPIRE is fundamentally different. Its powerful combination of conventional query tools and visualization techniques not only finds documents of known interest, it also uncovers their relationship to other documents in the collection. A document viewer allows users to go between the high-level abstraction and the original document text for advanced exploration of the details. A trend analysis tool lets users see new concepts grow in importance over time.

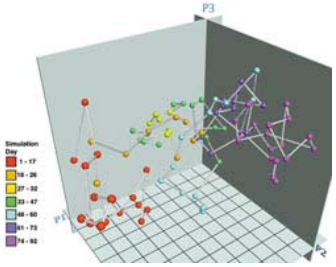


IN-SPIRE displays change over time.

Source: http://www.pnl.gov/cse/computersci/inspire.htm

Simplifying Scientific Datasets

Scientific datasets are extremely large, which makes analyzing, computing, and learning from the data extremely time-consuming and resource-intensive. Pacific Northwest National Laboratory has created techniques that reduce the burden of working with generated data, such as climate models, and observed data, such as that gathered by scientific instruments. The essence of the large data set is captured in a representation that is a fraction of the original size. These smaller, visual representations of the data can be analyzed and compared with other data sets quickly and easily.
Source: http://www.pnl.gov/cse/computersci/datasets.htm

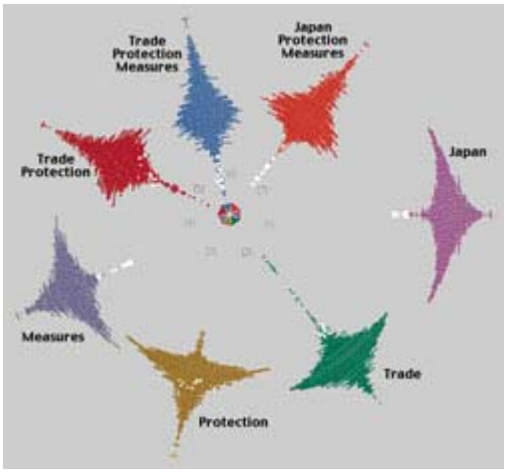


Simplifying very large datasets for scientific computation.

SURMISE™

Rather than automatically revealing similarities among the data and presenting images for investigation, some of the tools developed by Pacific Northwest National Laboratory (PNNL) allow users to define relationships and theories they would like to explore or test. With these technologies, people can manipulate the analysis without having to be familiar with the details behind the scenes.

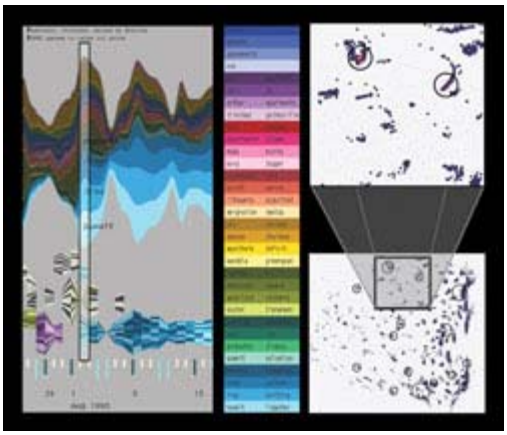
SURMISE™ is PNNL's first visualization technique to explore multiple queries or hypotheses within the same collection of documents. Like a detective, SURMISE™ analyzes a dataset in the context of multiple hypotheses specified by the user. The tool then rapidly evaluates how the data relates to those hypotheses, grouping relevant information and identifying information that is not relevant to support the theory.
Source:
http://www.pnl.gov/cse/computersci/surmise.htm



SURMISE™ enables users to explore multiple queries with the same collection of documents.

ThemeRiver™

In some situations, such as market research or national security assessments, discerning how topics and themes change over time can bring a new level of understanding. The patented ThemeRiver™ technology identifies sequential patterns, trends and temporal relationships within large collections of documents and analyzes them over time. A collection of documents or other data could be displayed as a "river" or ribbon of different colors that flows across a period of time. Within the river, color-coded "currents" identify themes and widen or narrow depending on their relative strength. Events that cause major changes in the themes are automatically extracted and displayed as "Event Rocks™™" along the river.



ThemeRiver™ displays a collection of data as a "river" that flows across a period of time.

Source:
http://www.pnl.gov/cse/computersci/themeriver.htm

Toolkit for Collaboratory Development

The Toolkit for Collaboratory Development is a cross-platform software suite that can be used to rapidly develop Internet-based scientific collaboratories. The Toolkit can be readily integrated with instruments and analysis tools, allowing for the development of environments tailored to your needs. The Toolkit's Collaborative Research Environment (CORE2000) provides an extensible suite of tools (including audio, video, screen sharing, and whiteboard) for group discussions, brainstorming, data analysis, and training. In addition, the Toolkit offers secure remote control capabilities and remote laboratory cameras. The Toolkit's secure Web-based Electronic Laboratory Notebook ensures that all of the data, notes, sketches, molecular structures under consideration, etc., are always available from any desktop. The Toolkit allows you to conduct experiments at facilities across the country without leaving your office. In addition, the Toolkit can be used to support remote lectures, demonstrations, and laboratory experiences for students and faculty.

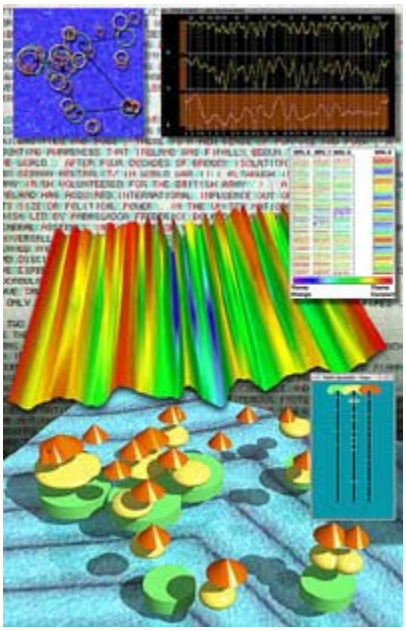
Source: http://www.pnl.gov/cse/computersci/toolkit.htm

TOPIC ISLANDS™

While many of our visual analytic tools were created to look at collections of multiple documents, TOPIC ISLANDS™ was designed specifically to help users understand the main points of individual large documents without having to read every page. This interactive tool automatically creates excerpted summaries, similar to an outline. TOPIC ISLANDS™ permits users to query large documents for concepts—without having to use exact terms or keywords—and quickly finds the sections where those concepts appear.

Pacific Northwest National Laboratory's wavelet technology, TOPIC-O-GRAPHY, serves as the basis for the TOPIC ISLANDS™ software. The underlying TOPIC-O-GRAPHY technology can "read" a large document, analyze its contents and provide useful summaries and outlines. This technology applies wavelet transforms to a custom digital signal constructed from words within a document. This multi-resolution approach extracts the thematic flow from text into a fuzzy kind of architecture that is similar to the outline format used often in writing.

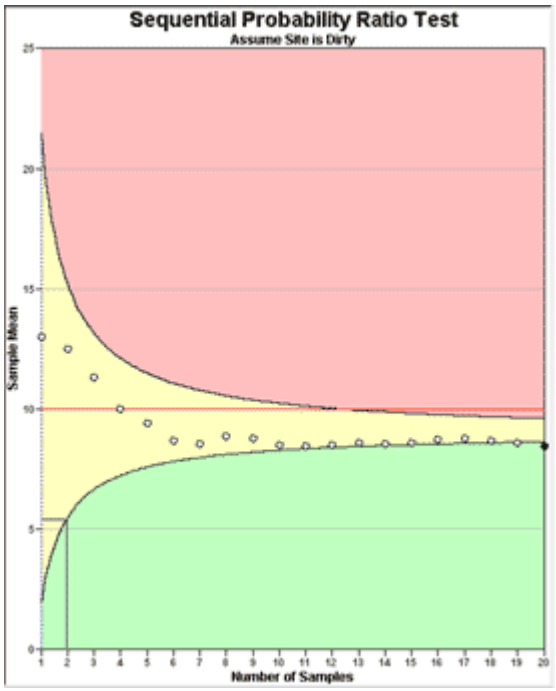
Source: http://www.pnl.gov/cse/computersci/topicislands.htm



TOPIC ISLANDS™ helps you understand the contents of a document without having to read it word for word.

Visual Sample Plan

The Visual Sample Plan (VSP), developed at Pacific Northwest National Laboratories (PNNL), uses simple tools to help define sample schemes for characterizing data. It is applicable for any two-dimensional sampling, specifically surface soil, building surfaces, water bodies, and other similar applications.



VSP Output.

Specifically designed for the collection and analysis of environmental data, this technology provides statistical solutions to sampling design, mathematical and statistical algorithms, and a user-friendly visual interface. VSP helps researchers to reach conclusions that are statistically defensible—while minimizing data analysis costs, and without requiring mastery of complex new software tools.

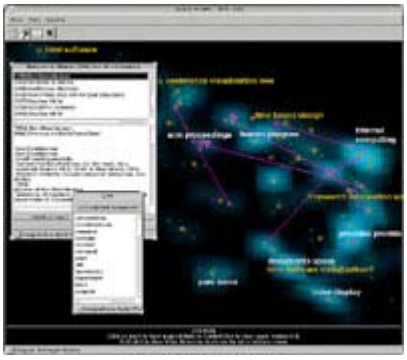
VSP is tailored to the environmental professional who values cost-effectiveness, simplicity, accuracy, and defensible methods. It offers numerous benefits to the user:

- Uses world-class statistical and mathematical algorithms applicable to environmental statistics.
- Interacts with the user through familiar visual interfaces such as site maps and building plans.
- Provides immediate feedback of the projected results of selected statistical sampling plans by overlaying random sampling locations or grids directly onto the site map or building plan.
- Provides projected number of samples, total sampling costs, and sampling locations in appropriate coordinates.
- Provides graphic decision tools such as graphs of probability of hot spot detection vs. total sampling costs.
- Creates simple, clear, and visually appeasing output.
- Allows nonparametric and parametric sampling designs.
- Generates MARSSIM supported sampling designs for soils and building surfaces.

Source: http://www.pnl.gov/cse/appliedmath/vsp.htm

WebTheme™

The information superhighway is littered with bits and pieces of information—some of which might be useful only when its relationship to other information on the Web is clear. The WebTheme™ tool rapidly identifies themes and relationships among thousands of pages of web-based text. WebTheme™ could be used for market research, harvesting related data using search terms, or following links derived from web addresses selected by the user.



WebTheme™ is an interactive tool that provides a visual display of the common themes in collections of web-based documents.

Source: http://www.pnl.gov/cse/computersci/webtheme.htm

Potentially Relevant PNNL LDRD Projects

Knowledge Discovery & Data Mining (LDRD FY 2002 research)

16.5 Lawrence Berkley National Laboratory

Scientific Data Management Integrated Software Infrastructure Center

Introduction

The Scientific Data Management Center focuses on the application of known and emerging data management technologies to scientific applications. The Center's goals are to apply and deploy software-based solutions to the efficient and effective management of large volumes of data generated by simulation and analysis of scientific applications. Our purpose is not only to achieve efficient storage and access to the data using specialized indexing, compression, and parallel technology, but also to enhance the effective use of the scientist's time by eliminating unproductive simulations, by providing specialized data mining techniques, and by automating time consuming tasks. Our approach is to work closely with application scientists in various domains on specific problems that will enhance their ability to achieve new scientific insights.

Source: http://sdmcenter.lbl.gov

Project Leader

Arie Shoshani
Computational Research Div [CRHPD]
Lawrence Berkley National Laboratory
(510)486-5171
AShoshani@lbl.gov

16.6 MITRE

Social Information Retrieval

Problem
Our research centers on developing new technology for tracking Internet-based networked organizations, and using those results to identify potential vulnerabilities and threats. Current information retrieval technology does not directly address the problem of detecting activist networks, assessing behavior, and tracking their evolution; new technology is needed to detect networks based on their structure and context.

Objectives
The main objective is to develop technology for a worldwide monitoring system used to detect the emergence of new groups (e.g., activists) and track the evolution of existing organizations based on their online activity. The focus will be on assessing an organization's behavior and its vulnerabilities.

Activities
We are exploring the confluence of information retrieval for collecting distributed information, social network analysis for determining network structure and characteristics, and dynamical systems modeling for determining network function or behavior. Work includes the development of advanced smart crawler collection tools that will use adaptive and cooperative searching techniques to provide efficient and high-coverage collection from the Web or other network search environments.

Impacts
This research will provide new tools for detecting emergent networked organizations in the open Web and enterprise environments, and will provide a basis for modeling their behavior and identifying critical nodes for assessing vulnerabilities and network robustness. Our initial work has already had impact on several sponsor mission areas.
Source: http://www.mitre.org/news/events/tech03/

Raymond D'Amore, Principal Investigator
703-883-1764
rdamore@mitre.org

Web Ontology Language (OWL)

Sponsor

DARPA Office: IXO
DARPA PM: Mr. Murray A. Burke

Problem
Interoperability is difficult when we use different terms; for example, “guarantee” versus “warranty.” Will a machine be able to dynamically realize that these terms mean the same thing? How do we define the semantics of a vocabulary in a way that will enable machines to

dynamically realize that two terms are the same or are related (and how they are related)? Solving this problem will go a long way toward achieving the vision of the Semantic Web.

Objectives
For several years DARPA has been working on an XML-based language to enhance interoperability by defining the semantics of vocabulary. An outcome of their work is the Web Ontology Language (OWL). MITRE has been asked to develop a PowerPoint-based tutorial that explains in a clear fashion the OWL and OWL-Light ontology languages.

Activities
A PowerPoint tutorial will be developed, including complete, validated examples. The culmination of the work will be a tutorial that is presented to DARPA.

Impacts
The resulting tutorial will be made available Internet-wide. It will enable the Internet community to quickly attain expertise in this technology. The consequence will be increased skill levels on ontologies, and (hopefully) increased interoperability. From a corporate perspective this will be great public relations for both MITRE and DARPA.

Source: http://www.mitre.org/news/events/tech03/

Roger L. Costello, Principal Investigator
781-271-6439
costello@mitre.org

Audio Hot Spotting

Problem
Large volumes of recordings require rapid retrieval of segments potentially relevant to a given query (audio hot spotting). Because of high automatic speech recognition (ASR) word error rates and the loss of important audio information in speech transcription, spoken document retrieval systems that simply combine ASR with information retrieval (IR) do not meet this need in real applications.

Objectives
We propose to research and develop audio-specific retrieval algorithms in critical domains by (1) exploiting multiple types of acoustic information from the audio signals; (2) exploring several adaptive techniques to improve existing ASR performance; and (3) fusing component technologies such as ASR, language/speaker identification, audio feature extraction, and information retrieval.

Activities
We will research algorithms and techniques to extend and improve ASR and audio feature extraction and to develop audio-based query algorithms making use of the multiple types of audio information. We will research and develop fusion algorithms to build an audio hot spotting system based on the extended ASR, audio feature extraction, language/speaker identification, and the new audio query language.

Impacts
Our research in audio hot spotting algorithms and prototype development will address the needs of MITRE sponsors with warehouses of recordings waiting for efficient retrieval. It will extend MITRE's information retrieval capability from text to include audio. The expertise gained through the research will equip MITRE to better advise industry developers and our sponsors on audio information retrieval topics and evaluation standards.
Source: http://www.mitre.org/news/events/tech03/

Qian Hu, Principal Investigator
781-271-2959
qian@mitre.org

Automated Information Discovery and Retrieval from Asian Language Sources

Problem
While several commercial capabilities exist to address particular facets of machine translation (MT) needs, emphasis has been placed on European-based languages. Furthermore, none of the existing COTS products are particularly well suited to the military environment. English translation of Asian languages is much more difficult than translation of European languages and has presented the MT community with significant challenges.

Objectives
This project will develop a capability to perform Chinese and Korean cross-language information retrieval, information discovery (ID), data mining (DM), and knowledge management (KM) in support of open source intelligence analysis. The project will develop a prototype capability that can support in-field experimentation with a broad spectrum of users.

Activities
We will provide an automated capability to translate electronic textual information between Chinese and English, and between Korean and English. We will characterize and subsequently retrieve information, based on user specified profiles, from Chinese- and Korean-language sources by means of a prototype analytic tool. A dictionary management capability will allow users to build, import/export, and aggregate custom dictionaries.

Impacts
This project has the potential for improving the efficiency and effectiveness of intelligence organizations currently impacted by foreign language translation issues. It is expected to provide the beneficiaries with needed interim capabilities and validation of the most fertile areas for the future application of government funds.
Source: http://www.mitre.org/news/events/tech03/

Ray LeBlanc, Principal Investigator
808-473-6439
rl@mitre.org

Foundations for Next Generation Information Access

Problem
Computerized support for information gathering is fragmented across multiple research communities, and integration is difficult due to the lack of an underlying formalism that cuts across the different technologies. Statistical techniques for individual components have been developed in isolation and without a common theoretical foundation. As a result we are left with a number of reasonably effective, semi-principled, incompatible techniques.

Objectives
The principal objective is the development of statistical foundations for information access. A successful foundation will comprise rigorous characterizations of the issues of modeling and estimation, together with principled methodologies for adapting to new languages, genres, information domains, auxiliary knowledge sources, and tasks.

Activities
We will develop simulations that model the stochastic generation of latent document features, observable document features, the determination of document relevance, and the distribution of query characteristics. We will perform exploratory data analysis on available research corpora to verify our models. A central focus will be on research into the importance of variance reduction and the potential benefits of various bias-variance strategies.

Impacts
This research is directly relevant to existing MITRE projects. The results will allow MITRE to develop information access systems incorporating new sources of evidence and to tailor information systems to meet specific military and intelligence needs. MITRE will then be strategically positioned to set the direction of research into, and development of, next-generation information access technology.
Source: http://www.mitre.org/news/events/tech03/

Warren Greiff, Principal Investigator
781-271-7421
greiff@mitre.org

An Emulation Facility for Networking and Distributed Application Development

Problem
Today's network simulation tools are inadequate to handle the volume of traffic necessary for application-level testing. Similarly, commercially available network emulation boxes are not sufficiently flexible to capture the nuances of military networks, particularly the mobile ad hoc nature of wireless networks such as the Multi-Sensor Command and Control Constellation (MC2C).

Objectives
This project will develop a real-time network emulation facility capable of accurately representing a heterogeneous military communications environment including delays, bandwidth constraints, and performance challenges. Application developers will be able to connect to the

network emulation facility to evaluate their system's performance in a variety of realistic settings. Network developers will be able to evaluate protocol improvements on an end-to-end basis.

Activities
The network emulator will be constructed in a spiral with an initial emphasis on developing a "thread" through a simple end-to-end effects-based emulation. This thread will require development of foundational elements such as the basic routing cluster, packet manipulation code, and control elements in addition to the infrastructure components of the high-level architecture, such as the path loss server and online instances of theater-level models.

Impacts
Although the development of major programs in terrestrial (e.g., Future Combat System), airborne (e.g., MC2C) and space-based (e.g., Transformational Communications System) communications will facilitate true network centric warfare, their simultaneous execution poses substantial challenges to the development of applications and network protocols. This project will give network architects and application developers a tool to perform engineering and design tradeoff studies for these future networks.
Source: http://www.mitre.org/news/events/tech03/

Ambrose M. Kam, Principal Investigator
781-271-5513
akam@mitre.org

Advanced Problem Analysis, Resolution, and Ranking

Problem
Advanced problem detection and resolution are part of the integrated sector suite concept for future air traffic management decision support systems (ATM DSS). SWRL (Severe Weather Resolution) is an integral component in the evolution of en route DSS capabilities. Research is needed to validate operational concepts, utility and acceptability, and system benefits.

Objectives
User confidence depends on validation of the "goodness" of the weather products. Research is needed to develop and integrate resolution algorithms (utilizing the weather products) with existing DSS functionality and to determine the operational benefits (efficiency, safety, capacity) of the weather products and resolution algorithms.

Activities
We will hold structured discussions with flight service personnel, aviation weather groups, pilots, and others. We will perform functional analysis of the weather products and continue development of the resolution algorithms. We will also carry out human-in-the-loop experimentation with operational personnel to help validate the resolution logic. Concurrently, we plan to develop methodologies and metrics for measuring performance and benefits.

Impacts
This work will contribute to the state of the art in integrating weather products into DSS automation at the sector level. It will provide for continuity in the evolution of en route DSS capabilities. It has potential for significant improvements in en route efficiency, safety, and capacity. The validation approach promotes cost-efficient utilization of existing assets and supports future technology transfer activities.
Source: http://www.mitre.org/news/events/tech03/

Win Heagy, Principal Investigator
703-883-6825
wheagy@mitre.org

Counter-Deception Decision Support

Problem
Denial and deception aim to disrupt an adversary's ability to "observe, orient, and decide," and to induce inaccurate impressions about friendly capabilities or intentions, causing the adversary to apply intelligence collection assets inappropriately, or fail to employ capabilities to best advantage. While the need for counter-deception (CD) is recognized, proposed solutions make little or no use of the psychology of deception and decision-making.

Objectives
We will develop a decision framework based on existing research on the psychology of deception, and integrate the framework with belief modeling tools to create a CD decision support system for intelligence analysts. Our hypothesis is that the psychology of decision-making and deception can be combined with existing belief management and planning technology to produce a CD decision support system.

Activities
In the Modeling phase we will construct a psychological framework of deception based on a deception taxonomy and deception cognitive model. In the Development phase we will develop tools for generating deception hypotheses and assessing the evidence of deceptions. The result will be a computational system that helps analysts to recognize potential deception moves, evaluate evidence, identify probable deceptions, and de-bias estimates. The Assessment phase will test the hypothesis through experiments with intelligence analysts.

Impacts
Research in CD will position MITRE to assist in several intelligence community initiatives. The research will position MITRE to develop systems to address several of our sponsor's identified "hard problems." It will also augment the Information Operations Planning Tool ACTD with deception planning aids.
Source: http://www.mitre.org/news/events/tech03/

Frank Stech, Co-Principal Investigator
703-883-5920
stech@mitre.org

Christopher Elsaesser, Co-Principal Investigator
703-883-6563
Chris@mitre.org

Indications and Warning for Countering Terrorism

Problem
The I&W community today relies on judgment and experience to identify threats. Watch centers are always "ON," with little time to reflect, test alternate hypotheses, or review processes, procedures, and methods. Once threats are identified, prioritization is a judgment call. Vestiges of Cold War I&W remain. Open source information is undervalued and not well understood, preventing analysts from taking advantage of completely new forms of analyses.

Objectives
Our objective is to have significant impact on directly funded projects at DIA, the combatant commands, OSD C3I, and other elements of the intelligence community (IC). We would also like to contribute to efforts in the law enforcement arena and provide threat identification for MITRE initiatives on homeland security.

Activities
We will assess operational I&W processes and methods currently employed in the IC; propose a universal model for terrorist assault, ambush, raid, and precision destruction operations; and validate the model through retrospective and predictive analysis. We will also examine and implement best analytical methods for threat assessment and risk measurement and develop a prototype that supports experimentation.

Impacts
As a result of current deficiencies, the public views the IC as either failing to provide warning (9/11) or overreacting to all types of threats without prioritizing them (e.g., crop dusting planes, banks in the northeast, malls, California bridges, theme parks, etc.). We seek to improve the I&W processes used by the IC and potentially those used by law enforcement.
Source: http://www.mitre.org/news/events/tech03/

Lisa Costa, Principal Investigator
813-831-5535 ext. 215
lahc@mitre.org

Mental Models in Naturalistic Decision Making

Problem
Computational models are needed to formalize the strengths and bounds of human thinking so that effective systems can be designed to advise people and automate functions. Computational models are also needed to simulate human behavior so that existing systems can be demonstrated and evaluated with fewer people in the loop.

Objectives
Our method is to measure and model human performance in a synthetic environment that poses prototypical tasks of command and control. These tasks include probabilistic risk assessment, dynamic resource management, and competitive/collaborative engagement.

Activities
Our tool is a micro world game called TRACS that offers the dual advantages of practical relevance and empirical rigor. Relevance is achieved via psychological correspondence between game tasks and real tasks. Rigor is achieved via mathematical comparisons between normative (optimal) performance and cognitive (human) performance.

Impacts
Our results are computational models of how, and how well, people make command and control decisions. These models provide a cognitive-scientific basis for designing systems that improve human-computer performance, and for designing agents that behave like people in large-scale simulations.
Source: http://www.mitre.org/news/events/tech03/

Kevin J. Burns, Principal Investigator
781-271-8762
kburns@mitre.org

Retrospective Source Analysis

Problem
Intelligence analysts rely substantially on source "reliability" ratings to determine which classified sources to use as primary information sources. Unfortunately, the overall accuracy and usefulness of sources of different reliability levels have never been measured; nor have the accuracy and usefulness of open sources and classified sources ever been compared.

Objective
The objective is to develop both a method for evaluating source usefulness and an initial assessment of the usefulness of some source types (e.g., local newspapers, international newspapers, etc.). The source types will be evaluated by accuracy (proportion of claims that are accurate), specificity (proportion of claims useful), and timeliness (sources that first report important events).

Activities
The major activity will be a structured retrospective analysis of archived open source (and possibly classified) reports. For specified topics and time periods, historical reports will be rated for accuracy, specificity and timeliness.

Impact
We anticipate that certain open sources will be as accurate and useful as highly rated classified sources. Such a result would provide new and very useful guidance to analysts on how to find and weigh sources in generating an assessment/conclusion.

Source: http://www.mitre.org/news/events/tech03/

Paul Lehner, Principal Investigator
703-883-7968
plehner@mitre.org

Reading Comprehension: Reading, Learning, Teaching

Problem
This project is addressing a three-stage grand challenge application for human language technology: building a system that can "learn to read," then "read to learn," and finally "teach to learn." It deals with issues of machine learning, knowledge acquisition, and instructional technology.

Objectives
First we will build a computer-based system capable of passing a third grade reading-comprehension test. Second we will build a system that will "read to learn," passing a test on that subject matter after having read the text. Finally we will build a system that can learn through interacting with a person, and, at the same time, help to teach the person.

Activities
We have applied prototype systems on reading comprehension tests designed for fourth to eighth graders with 30%–40% accuracy. We are improving the system to include more components. We will implement a reciprocal teaching demonstration, where the system plays the role of teacher (grading student answers) or the role of peer learner (answering questions posed by a real student).

Impacts
This research will open new areas of research, addressing issues of machine learning, breaking the knowledge acquisition bottleneck, developing new evaluation measures for understanding and learning, and creating new instructional technologies via learning companions and interactive teaching environments.
Source: http://www.mitre.org/news/events/tech03/

Lynette Hirschman, Principal Investigator
781-271-7789
lynette@mitre.org

TIDES (Translingual Information Detection Extraction Summarization)

Sponsor
DARPA Office: IAO
DARPA PM: Mr. Charles Wayne

Problem
Over the years, expanded trade and travel have increased the potential economic and political impacts of major disease outbreaks. Recently, biological terrorism has become a very real threat. Appropriate response to disease outbreaks and emerging threats depends on obtaining reliable and up-to-date information, which often means monitoring many news sources, particularly local news sources, in many languages worldwide.

Objectives
TIDES (Translingual Information Detection Extraction Summarization) aims to revolutionize the way that information is obtained from human language by enabling people to find and interpret needed information quickly and effectively, regardless of language or medium.

Activities
MITRE's role in the TIDES Integrated Feasibility Experiment-Translingual will cover data collection/processing/distribution, machine translation, and data exchange standards. We will also support the Total Information Awareness program through integration and training, geospatial and temporal normalization, topic tracking, and rapid domain portability. We will provide a test bed for research and continue to maintain and improve the operational MITRE Text and Audio Processing (MiTAP) systems.

Impacts
MiTAP has been deployed to four sites. It is being used to track global threats, including disease outbreaks and terrorist activity. MiTAP focuses on providing timely multilingual, global information access to over 450 analysts, medical experts, government users, and humanitarian organizations. A MiTAP product, the World Press Update, is distributed to hundreds of readers and decision-makers worldwide.
Source: http://www.mitre.org/news/events/tech03/

Laurie Damianos, Principal Investigator
781-271-7938
laurie@mitre.org

Building the Semantic Web

Problem
As the amount of information on the World Wide Web continues to grow, the value of automated tools capable of finding, filtering, and combining information in response to specific user requirements greatly increases. The largest barrier preventing more automated use of Web resources is that the semantics (meaning) of these resources is generally unavailable to automated agents.

Objectives
The objective of this project is to develop technical foundations for a “Semantic Web,” in which programs such as agents, search engines, or service brokers can identify and use World Wide Web resources (including both information and services) based on machine-readable representations of their semantics (meaning).

Activities
We are investigating language concepts for representing and processing semantic information that scale to the Web environment, and application areas that include eBusiness and disaster relief. We are participating in the World Wide Web Consortium's Semantic Web Activity, engaging in joint research with MIT's Context Interchange (COIN) project, and cooperating with researchers in DARPA's DAML (DARPA Agent Markup Language) program.

Impacts
This research addresses a key area of current Web technology development, impacting numerous MITRE programs dependent on Web technologies such as XML, as well as wider eBusiness and other communities addressing issues of large-scale interoperability. The research also provides technology transfer opportunities with a wide range of academic and industry R&D activities and standards groups.
Source: http://www.mitre.org/news/events/tech03/

Frank Manola, Principal Investigator
781-271-8147
fmanola@mitre.org

Data Integration as an Industrial Process

Problem
Data integration requires too much human time and skill. We need to industrialize, to create narrow-skill steps, each of which produces reusable knowledge rather than opaque code. To move from (easily evaded) mandates to natural incentives, we will explore "describe and generate" tools to make even the first connection easier. The approach should be incremental, driven by real interoperability needs, not special initiatives.

Objectives
Our goals are to refine the industrial approach and to move industry, the research community, and sponsors toward that vision. Specifically, we will extend (very scalable) profile-driven integration techniques to be compatible with commercial multi-database query tools, develop metrics to help project planners compare data integration techniques and judge tools' utility, and evaluate emerging describe-and-generate data integration research prototypes in real projects.

Activities
We will conduct experiments using research prototypes (e.g., IBM Research's Clio) with aviation, brain mapping, and tax administration data; conduct a survey of data integration practitioners to determine where the costs are the greatest; and adapt metrics to improve project planning. Subsequently, we will refine the metrics and perform further experimentation. Throughout, we will publish results and transition them to MITRE and sponsor projects.

Impacts
We will reframe a critical technology to reflect rarely addressed organizational realities. We will influence emerging industrial tools and researchers' agendas, and provide metrics where none previously existed. MITRE's sponsors will be aided in moving from doomed giant data integration initiatives to incremental progress.
Source: http://www.mitre.org/news/events/tech03/

Len Seligman, Principal Investigator
703-883-5975
seligman@mitre.org

Arnon Rosenthal, Co-Principal Investigator
781-271-7577
arnie@mitre.org

Neuroinformatics

Problem

The neuroscience community is accumulating a vast amount of human brain mapping data that does not reach its full scientific potential because it is generally confined to the originating lab. While data may exist that a researcher could use to explore a hypothesis, the investigator may be unaware of it or lack access to it.

Objectives

The overall goals of this research, conducted in conjunction with an external NIH grant, are to design, prototype, and evaluate an information infrastructure to help realize the full potential of a growing store of human brain mapping data. In this initial undertaking, we focus on a system that enables the analysis, exploration, and dissemination of structural magnetic resonance imaging (MRI) data.

Activities

We have made significant progress toward development of a digital library, including schema integration, a data sharing policy space, and Web-based tools for exploring the data. The project focus is on medical image exploitation: designing query-by-example functionality, enhancing querying using data mining, and developing data quality metrics intrinsic to the neuroimagery. We are also continuing to acquire MRI data from our collaborators.

Impacts

This project provides an important public service to the neuroscience research and clinical communities. But the problems facing these communities are not unique; they are isomorphic to those facing many of MITRE's traditional sponsors who must manage and exploit large quantities of imagery. We expect our research to readily transition to our Treasury Department, DoD, and USGC sponsors.
Source: http://www.mitre.org/news/events/tech03/

Monica Carley, Principal Investigator
703-883-7045
mcarley@mitre.org

Automated Discovery of Innovative Tactics and Behaviors

Problem

Modeling and simulation plays a key role in the design, analysis, and implementation of new military concepts and systems. Effective modeling in this context requires the capability to quickly generate innovative twists on operational concepts, tactics, and possible threat responses. Currently, the only possibilities examined are those few that happen to come to mind for the human designers and analysts.

Objectives
Any technique that enables humans to systematically examine a broader range of options, or suggests alternatives they may not have considered, would greatly increase the effectiveness of these simulation-based activities. We will develop new machine learning techniques to address this need. Our hypothesis is that innovative tactics and behaviors can be learned automatically from experience in a simulation.

Activities
The research has already developed techniques that can learn rule-based reactive behaviors given feedback about outcomes. That capability is being extended to learn more structured, distributed behaviors (e.g., those requiring teamwork). The final improvement will address the knowledge representations needed to learn coordinated tactics in challenging simulated environments (e.g., RoboCup soccer and micro-air vehicle swarms).

Impacts
This research will develop new capabilities that will enhance the effectiveness of simulation technology in critical applications such as simulation-based acquisition and joint experimentation. If successful, these developments will also advance the state of the art in machine learning and produce several refereed publications.
Source: http://www.mitre.org/news/events/tech03/

Lashon B. Booker, Principal Investigator
703-883-7609
booker@mitre.org

16.7 National Institute of Standards and Technology

Statistical Analysis of IT Performance

Project Goal and Summary

Commerce at the beginning of the twenty-first century is seeing an information technology revolution driven by software innovation, by network systems, and by hardware advances. Assessing quality in each of these areas draws on statistical models that are developed expressly for this purpose. At NIST, statistical modeling and research address these areas with major collaborative efforts in software testing, in information retrieval, in network performance assessment, and in ongoing work on standards for electronics. Testing software often requires development of stochastic models for failure, where failure points are determined by the specs for the software, whether this be computer graphics meta file, for Java, for computer-aided design software, or for statistical computation software. Then statistical methods for inference must be derived for tests of conformance, measures of reliability, estimates of failure rates and of variability according to failure mode. In the case of testing algorithms for feature recognition, such as the identification of an individual person based on a poor or partial image, the first statistical issue is the design of the testing procedure using an accumulated data set of images. The goal of assessing the performance of an algorithm requires development of an overall performance measure and also construction of indicators of specific algorithm weaknesses together with statistically sound measures of uncertainty associated with each measure. To understand performance measurements on information retrieval systems, several statistical methods must be applied to discover whether differences among the retrieval systems are real, and then to characterize these differences with respect to such factors as topics searched or collections of documents. Similarly, network performance representation requires statistical designs and visualization tools for the utilization of an accumulating database, although the database characteristics differ sharply.
Source: http://www.itl.nist.gov/div898/itperf/homepage.htm

Principal Contact Information

Liu, Hung-Kung
hung-kung.liu@nist.gov

Bayesian Metrology Project

The NIST Statistical Engineering Division is developing the use of Bayesian statistical methods to solve NIST metrology problems as a NIST Competence Project.
Future activities in Bayesian metrology include

- Bayesian consensus mean
- Equivalence
- Performance comparison of solutions to combining results from multiple methods
- Combined uncertainty
- Bayesian analysis and consensus guidance for measurement curves and images from interlaboratory studies
- Addition of MCMC to StRD (Statistical Reference Datasets)

- Magneto-Optically trapped atoms
- Statistical analysis of the LADAR measurements
- Data assimilation and Bayesian design for physical/complex/high-dimensional process
- Statistical design and standardization of microarray experiments and data analysis
- High-dimensional modeling

Source: http://www.itl.nist.gov/div898/bayesian/homepage.htm

Nell Sedransk
Statistical Engineering Division
Information Technology Laboratory
301-975-2839
nell.sedransk@nist.gov

Research on Statistical Methods Project

Since the formation of the Statistical Engineering Division in 1947, division staff, through their interdisciplinary research with NIST scientists and engineers, occasionally encounter problems that cannot be addressed using existing, or textbook statistical methods. On such occasions, appropriate division staff conducts original research in mathematical and/or computational statistics, leading to new and more broadly applicable statistical methods. The division's unique contributions to the general methods of statistics tend to concentrate in areas where the measurement science activities at NIST present new challenges in planning and analyzing high precision data on high-accuracy measurement systems. So, many of the divisions original contributions fall into the following areas:

- Bayesian methods for metrology,
- statistical calibration and measurement assurance,
- experiment designs,
- components-of-variance estimation,
- methods for the design and analysis of interlaboratory comparisons, and
- measurement process control.

  The division also conducts limited research in some areas enabled by modern computing systems:

  1. computer intensive methods (bootstrap, permutation procedures, general distribution-free methods) and
  2. image analysis methods.

Source: http://www.itl.nist.gov/div898/projects/statmeth.htm

Dom Vecchia
Statistical Engineering Division
Information Technology Laboratory
301-975-2846
dominic.vecchia@nist.gov

Human Language Technology

Project Goal and Summary

To develop and apply metrics and testing to advance the state of the art of human language processing, including speech and speaker recognition, spoken language understanding, information search, retrieval, and filtering, and other advanced text processing techniques such as summarization and extraction. To increase communication and technology transfer between industry and academia in the field of human language technology. To meet these goals by (1) developing measurement methods and evaluation infrastructure, (2) providing reference materials, including test data, (3) coordinating community-wide benchmark tests within the research and development community, and (4) building prototype systems.
Source: http://www.itl.nist.gov/iad/programs.html#Human

Principal Contact Information
Harman, Donna K.
donna.harman@nist.gov

Text REtrieval Conference (TREC)

The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA), was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has the following goals:

- to encourage research in information retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

Source: http://trec.nist.gov/

Ellen Voorhees
TREC Project Manager
Retrieval Group
Information Accesss Division
Information Technology Laboratory
National Institute of Standards & Technology
U.S. Department of Commerce
(301) 975-3761
ellen.voorhees@nist.gov

The NIST Speaker Recognition Evaluation

The NIST Speaker Recognition evaluation is part of an ongoing series of yearly evaluations conducted by NIST. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. To this end the evaluation was designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible.
Source: http://www.nist.gov/speech/tests/spk/index.htm

Pallett, David S. Dr. -
(301) 975-2935
Information Access Division (894)
david.pallett@nist.gov

Topic Detection and Tracking

Topic Detection and Tracking research is being pursued under the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program.

TDT research develops algorithms for discovering and threading together topically related material in streams of data such as newswire and broadcast news in both English and Mandarin Chinese. The overview paper "Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation," (Wayne LREC2000) describes in more detail the TDT program, the TDT corpora (collections of broadcast news recordings and transcripts), and the TDT technology evaluation paradigm.

TDT research started with a pilot study in 1997 and has continued with four open evaluations in 1998, 1999, TDT 2000, TDT 2001, and TDT 2002. The TDT 1999 Workshop and TDT2000 Workshop web pages contains detailed information about the most recent evaluations plus copies of virtually all the presentations and papers from the workshops.

The TDT research applications keep track of topics, (events of interest), in a constantly expanding collection of multimedia stories.

TDT applications either organize vast amounts of data or facilitate large scale collections of non-text media. There are 5 research applications defined in the TDT Program.

1. Story Segmentation - Detect changes between topically cohesive sections
2. Topic Tracking - Keep track of stories similar to a set of example stories
3. Topic Detection - Build clusters of stories that discuss the same topic
4. First Story Detection - Detect if a story is the first story of a new, unknown topic
5. Link Detection - Detect whether or not two stories are topically linked

Shared resources, such as TDT corpora, language resources and evaluation software, provide the necessary tools to build a TDT application. Arguably, the most valuable resource made available to the community is the TDT corpora. The TDT corpora consist of broadcast news and newswire

texts sampled daily during most of 1998. The LDC exhaustively annotated the corpora by identifying which stories discuss a predefined set of topics.
Source: http://www.nist.gov/speech/tests/tdt/index.htm
Fiscus, Jonathan G.
Information Access Division (894)
(301) 975-3182
jonathan.fiscus@nist.gov

Automatic Content Extraction

The objective of the ACE program is to develop automatic content extraction technology to support automatic processing of human language in text form. The program is devoted to three sources types. These are namely newswire, broadcast news (with text derived from ASR), and newspaper (with text derived from OCR). ACE technology R&D is aimed at supporting various classification, filtering, and selection applications by extracting and representing language content (i.e., the meaning conveyed by the data). Thus the ACE program requires the development of technologies that automatically detect and characterize this meaning.
Source: http://www.itl.nist.gov/iad/894.01/tests/ace/index.htm
Pallett, David S. Dr. -
(301) 975-2935
Information Access Division (894)
david.pallett@nist.gov

A Distributed Agent-Based Lookahead Strategy for Intelligent Real-Time Decision-Making in Manufacturing

Project Goal and Summary

This extramural research project is led by Lookahead Decisions, Inc. with co-funding from the NIST Advanced Technology Program (ATP), which accelerates the development of innovative technologies for broad national benefit through partnerships with the private sector. Develop a software technology based on lookahead strategies -- strategies that follow a tree of possible resultant events -- to permit real-time decision-making in automated manufacturing systems based on data from the shop floor.
Source: http://jazz.nist.gov/atpcf/prjbriefs/prjbrief.cfm?ProjectNumber=00-00-4443

Principal Contact Information
Boudreaux, Jack C.
jack.boudreaux@nist.gov

A Phrase-Based Statistical Approach to Understanding and Translating Natural Language

Project Goal and Summary

This extramural research project is led by Sehda, Inc. with co-funding from the NIST Advanced Technology Program (ATP), which accelerates the development of innovative technologies for broad national benefit through partnerships with the private sector. Develop and demonstrate

technologies that will enable accurate machine understanding of human languages by isolating statistically significant phrases and mapping equivalencies in their usage.
Source: http://jazz.nist.gov/atpcf/prjbriefs/prjbrief.cfm?ProjectNumber=00-00-4826

Principal Contact Information
Omidvar, Omid Massoud
omid.omidvar@nist.gov

Adaptive Web Learning Guides

Project Goal and Summary
This extramural research project is led by Extempo Systems, Inc. with co-funding from the NIST Advanced Technology Program (ATP), which accelerates the development of innovative technologies for broad national benefit through partnerships with the private sector. Develop and test animated computer 'guides' and affordable tools for customizing these guides to enhance access to World Wide Web-based educational material in many fields and industries as well as K-12 public education.

Source: http://jazz.nist.gov/atpcf/prjbriefs/prjbrief.cfm?ProjectNumber=99-01-3055

Principal Contact Information
Liebergot, Harris L.
harris.liebergot@nist.gov

Interactive Software For Cognitive Skill Development

This extramural research project is led by Lexia Learning Systems, Inc. with co-funding from the NIST Advanced Technology Program (ATP), which accelerates the development of innovative technologies for broad national benefit through partnerships with the private sector. Develop new computer user interfaces and instructional methods grounded in concepts from cognitive-neurosience to enable a radically new class of educational software designed to develop fundamental cognitive skills.

Source: http://jazz.nist.gov/atpcf/prjbriefs/prjbrief.cfm?ProjectNumber=00-00-4002

Principal Contact Information
Liebergot, Harris L.
harris.liebergot@nist.gov

Spoken Language User Interface (SLUI) Toolkit

This extramural research project is led by BCL Technologies (Formerly BCL Computers) with co-funding from the NIST Advanced Technology Program (ATP), which accelerates the development of innovative technologies for broad national benefit through partnerships with the private sector. Develop a code-generating software 'toolkit' that will allow programmers to rapidly develop spoken-language input interfaces for new and existing applications without a detailed knowledge of linguistic theory.
Source:
http://patapsco.nist.gov/ext_npris/DetailsShort.cfm?OU_id=47&Proj_id=3710&criteria=SLUI

Principal Contact Information

Currens, Christopher
christopher.currens@nist.gov

17.0 References

1. Fayyad Usama, Piatetsky-Shapiro Gregory, and Smyth Padhraic, From Data Mining to Knowledge Discovery in Databases, AI Magazine, Fall 1996, pp. 37-54.

2. Definition provided by http://encyclopedia.thefreedictionary.com/

3. Lashkari, Y., Metral, M. and Maes, P., "Collaborative Interface Agents", *Proceedings of AAAI '94 Conference*, Seattle, Washington, August 1994

4. Lieberman, H., "Letizia: An Agent That Assists Web Browsing", *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, Montreal, Canada, August 1995

5. Rhodes, B., "Remembrance Agent: A continuously running automated information retrieval system", *Proceedings of The First International Conference on The Practical Application Of Intelligent Agents and Multi Agent Technology*, London, UK, April 1996, pp. 487-495

6. Gray, R., "Agent Tcl: A transportable agent system", In *Proceedings of the CIKM Workshop on Intelligent Information Agents, Fourth International Conference on Information and Knowledge Management (CIKM 95)*, Baltimore, Maryland, December, 1995

7. Foner, L., "Clustering and Information Sharing in an Ecology of Cooperating Agents", *Workshop Notes of the AAAI '95 Spring Symposium on Information Gathering from Distributed, Heterogeneous Environments*, Stanford University, California, March 1995

8. Corradi, A., Stefanelli, C., Tarantino, F., "How to Employ Mobile Agents in Systems Management", *Proceedings of the Third International Conference on The Practical Applications of Intelligent Agents and Multi-Agent Technology (**PAAM'98**)*, London, UK, March 23-25, 1998 Pages 17-26

9. Minar, N., "Designing an Ecology of Distributed Agents", *Masters Thesis at Massachusetts Institute of Technology, 1998*

10. Lesser, V., Horling, B., Klassner, F., Raja, A., Wagner, T., and Zhang, S. X., "A Next Generation Information Gathering Agent", In *Proceedings of the 4th International Conference on Information Systems, Analysis, and Synthesis; in conjunction with the World Multiconference on Systemics, Cybernetics, and Informatics (SCI'98); also available as UMASS Tech Report 98-72*. July, 1998

11. Chia, M.H., Neiman, D.E. and Lesser, V.R. "Poaching and Distraction in Asynchronous Agent Activities". In *Proceedings of the Third International Conference on Multi-Agent Systems*, pp. 88-95. 1998

12. Corkill, D., Lander, S., "Diversity in Agent Organizations", *In Object Magazine*, Volume 8, Number 4, pp. 41-47. May, 1998

13. Lieberman, H., Dyke, N., Vivacqua, A., "Let's Browse: A Collaborative Browsing Agent", *Knowledge-Based Systems*, Vol. 12, Dec. 1999, pp. 427-431

14. Minar, N., Kramer, K., Maes, P., "Cooperating Mobile Agents for Dynamic Network Routing", *Software Agents for Future Communications Systems*, Springer-Verlag, 1999, ISBN 3-540-65578-6

15. Yu, B., Singh, M., "A Multi Agent Referral System for Expertise Location", *In Working Notes of the AAAI Workshop in Intelligent Information Systems*, Pages 66-69, 1999

16. Lange, D., and Oshima, M., The Aglet book "Programming and Deploying Java Mobile Agents with Aglets", Addison-Wesley, http://cseng.awl.com/bookdetail.qry?ISBN=0-201-32582-9&ptype=0

17. Prasad, N., and Lesser, V., "Learning Situation Specific Coordination in Cooperative Multi-Agent Systems", In *Autonomous Agents and Multi-Agent Systems*, Volume 2, pp. 173-207. 1999

18. Yu, B., and Singh, M., "A social mechanism of reputation management in electronic communities", *In Cooperative Information Agents, CIA-2000*, Boston, MA, USA, pages 154---165, 2000

19. Horling, B., Lesser, V., Vincent, R., "Multi-Agent System Simulation Framework", In *16th IMACS World Congress 2000 on Scientific Computation, Applied Mathematics and Simulation.* August, 2000

20. Youll, J., Morris, J., Krikorian, R., and Maes, P., "Impulse: Location-based Agent Assistance", *Software Demos, Proceedings of the Fourth International Conference on Autonomous Agents (Agents 2000)*, Barcelona, Catalonia, Spain, June 3 - June 7, 2000

21. Morris, J., Ree, P., and Maes, P., "Sardine: Dynamic Seller Strategies in an Auction Marketplace", *Proceedings of the Conference on Electronic Commerce (EC '00)*, Minneapolis, MN, October 17-20, 2000

22. Woolf, B., Lesser, V., Eliot , C., and Klein, M., "A Digital Market Place for Education", In *Electronics Business and Education: Recent Advances in Internet Infrastructures, 2001*

23. Vincent, R., Horling, B., and Lesser, V., "An Agent Infrastructure to Build and Evaluate Multi-Agent Systems: The Java Agent Framework and Multi-Agent System Simulator", In *Lecture Notes in Artificial Intelligence: Infrastructure for Agents, Multi-Agent Systems, and Scalable Multi-Agent Systems*., Volume 1887. January, 2001

24. Singh, M., Yu, B., and Venkatraman, M., "Community-Based Service Location", *Communications of the ACM*, volume 44, number 4, April 2001, pages 49-54

25. Xing, J., Wan, F., Rustogi, S., and Singh, M., "A Commitment-Based Approach for Business Process Interoperation", *IEICE Transactions on Information and Systems*, volume E84-D, number 10, October 2001, pages 1324-1332

26. Cheng, Z., Singh, M., and Vouk, M., "Verifying Constraints on Web Service Compositions", In *Vipul Kashyap and Leonard Shklar (eds), Real World Semantic Web Applications, 2002*

27. Potok, T., Elmore, M., Reed, J., and Samatova, N., "An Ontology-based HTML to XML Conversion Using Intelligent Agents", *Proceedings of the Hawaii International Conference On System Sciences*, 1/2002

28. Potok, T., Elmore, M., Reed, J., and Sheldon, F., "VIPAR: Advanced Information Agents discovering knowledge in an open and changing environment", *Proc. 7th World Multiconference on Systemics, Cybernetics and Informatics, Special Session on Agent-Based Computing*, Orlando FL, July 27-30, 2003

29. Horling, B, Mailler, R., Sims, M., and Lesser, V., "Using and Maintaining Organization in a Large-Scale Distributed Sensor Network", In *Proceedings of the Workshop on Autonomy, Delegation, and Control (AAMAS03)*, July, 2003

30. The EEG is a tool that measures electrical behaviors in the brain at fidelity of tens of milliseconds. Research has demonstrated clear correlative relationships between specific cognitive processes and EEG signatures. In addition, early research indicates this relationship may be causal – where changes in the EEG signature of a particular portion of the brain may actually cause changes in the resulting cognitive behavior or performance (Klimesch, personal communication July, 2004).

31. We follow common usage by writing "COTS" when we mean "COTS technology".

Distribution List:

Internal:

5	1176	Chapman, Leon D., 15242
5	1176	Cranwell, Robert M., 15242
2	1176	Jordan, Danyelle, 15242
2	1176	Homan, Rossitza, 15242
2	1170	Skocypec, Russell D., 15240
1	1005	Vandevender, Alice A., 15240
1	1005	Nanco, Alan S., 15240
1	1188	Bauer, Travis, 15241
1	0455	Spires, Shannon V., 05517
1	0455	Phillips, Laurence R., 05517
1	9018	Central Technical Files, 8945-1
2	0899	Technical Library, 9616

External:

LTC Steve Bristow
PM FCS-ACE
9350 Hall Road, Attn: FCS-ACE
Ft. Belvoir, VA 22060-5526

Edwin P. Chamberlin, Jr.
9720 Euclid Ave. NE
Albuquerque, NM 87112

David M. Hetrick
Oak Ridge National Laboratory
P.O. 2008, MS 6315
Oak Ridge, TN 37831-6414

Mark T. Elmore
Oak Ridge National Laboratory
P.O. 2008, MS 6364
Oak Ridge, TN 37831-6364

Terry Gimbre
PM UA-ACE/PTC
9350 Hall Road, Attn: FCS-ACE
Ft. Belvoir, VA 22060-5526

Mark McCoy
PM UA-ACE
6501 E. 11 Mile Road
Warren, MI 48397-5000

Jim McNicol
PM UA-ACE/PTC
100 W. Big Beaver Suite 400
Troy, MI 48084

Jim N. Treadwell
Oak Ridge National Laboratory
P.O. Box 2008, MS6085
Oak Ridge TN 37831-6085