



Networks Grand Challenge News and Science

Tracking Report

Update: August – October 2008 (Issue 2)

A selection of industry, funding, and research news relevant to the SNL Networks Grand Challenge Team

This issue of the Networks News and Science Tracking Report from Perspectives covers material primarily from August to October 2008, although some important material from other periods is included. For example, older material uncovered as part of the tracking research may be included if it has not been discussed in previous reports.

This document contains items abstracted (edited, direct text from or summaries of source material) from the news or other sources. Links are provided to the full text of source material. Occasionally, Perspectives' comments are included (indicated by italic type). Emphasis is ours and is indicated by bolding or underlining. Items of particular interest to us are indicated by a highlighted star.

The full text of any source item is available. Contact Ann Miksovic: ann@perspectivesweb.com or (505) 881-0370.

Table of Contents

I. Priority Applications	2
A. Terrorism / Intelligence Analysis.....	2
B. Cybersecurity.....	7
C. The Big Data Problem	10
D. Funding.....	10
II. Other Applications and Items of Interest	12
A. Epidemiology / Medical / Life Sciences / Pharma	12
B. Business Analytics.....	12
C. Other Items of Interest.....	13
III. Company News, in Brief	14
IV. Resources / Overviews	19
A. Reviews and Overview Articles	19
B. Resources on the Web.....	20
C. Conferences	20

I. Priority Applications

A. Terrorism / Intelligence Analysis

- **★ NRC Report: [All Counterterrorism Programs That Collect and Mine Data Should Be Evaluated for Effectiveness, Privacy Impacts; Congress Should Consider New Privacy Safeguards](#).** At the request of the DHS and the National Science Foundation, the report examines the **technical effectiveness and privacy impacts of data-mining and behavioral surveillance techniques**. All U.S. agencies with counterterrorism programs that collect or "mine" personal data ... should be required to systematically evaluate the programs' effectiveness, lawfulness, and impacts on privacy, says the report, "[Protecting Individual Privacy in the Struggle Against Terrorists: A Framework for Program Assessment](#)" ([executive summary](#), [report description](#)). Both classified and unclassified programs should be evaluated before they are set in motion and regularly thereafter for as long as they are in use, according to one recommendation in the report. It offers a framework agencies can use to assess programs, including existing ones.

The NRC report got wide coverage in the press, for example, an article in the *New York Times*, "[Study of Data Mining for Terrorists Is Urged](#)."

*Because this report has been reviewed and discussed at length by the NGC team (and a review memo has been prepared by Perspectives), the body of the NRC report is not described in detail here. Below, we list a set of **selected government activities** collected in an appendix to the NRC report which will no doubt be of interest to the team (full text available [here](#)):*

- **Total Information Awareness (TIA)**. Status – terminated. [An additional full chapter in the NRC report (full text [here](#)) is devoted to the TIA program.]
- Computer-Assisted Passenger Prescreening System II (**CAPPS II**) And Secure Flight. Status: CAPPS II abandoned; Secure Flight planned for deployment in 2008.
- Multistate Anti-Terrorism Information Exchange (**MATRIX**). Status: pilot program ended; no follow-on program started.
- **Able Danger**. Status: terminated in January 2001.
- Analysis, Dissemination, Visualization, Insight, And Semantic Enhancement (**ADVISE**) Status: Under development (some deployments decommissioned and aspects of the program were halted).
- Automated Targeting System (**ATS**). Status: In use.
- The **Electronic Surveillance Program**. Status: continuing subject to oversight by the Foreign Intelligence Surveillance Court.
- Novel Intelligence from Massive Data (**NIMD**) Program. Status: in progress.
- Enterprise Data Warehouse (**EDW**). Status: operational since 2000 and in use.
- Law Enforcement Analytic Data System (**NETLEADS**). Status: in use.
- Ice Pattern Analysis and Information Collection System (**ICEPIC**). Status: operating as pilot program as of July 2006; planned to enter full-scale operation in fiscal year 2008.
- Intelligence and Information Fusion (**I2F**). Status: in development.
- Fraud Detection and National Security Data System (**FDNS-DS**). Status: in use but without analytical tools to support data mining; support for data mining capabilities not expected for at least 2 years.

- National Immigration Information Sharing Office (**NIISO**). Status: in use without data mining tools.
 - Financial Crimes Enforcement Network (**FinCEN**) and **BSA Direct**. Status: FinCEN in use; BSA Direct withdrawn.
 - **Department of Justice Programs Involving Pattern-Based Data Mining**. Status: all programs under development or in use.
- [ACLU Demands NSA And DOJ Turn Over Spying Policy Records](#): Recent revelations suggest there are no adequate safeguards in place to protect innocent Americans from invasive surveillance. (10/15/2008) ... On October 9, *ABC News* reported that NSA officials have intercepted, listened to and passed around the phone calls of hundreds of innocent U.S. citizens working overseas, including soldiers, journalists and human rights workers from organizations like the International Red Cross and Doctors Without Borders ... The ACLU noted in a press release about its request for records: "The new information ... suggests there are no real safeguards in place to protect the privacy of Americans who are swept up in NSA surveillance, and that any safeguards that do exist are ineffective or largely ignored by NSA agents." (*More information at source.*)
 - ★ *There is a similar controversy in the UK: **Government faces fight from within for spy database***: A Home Office revolt is stalling a plan to store our e-mails and calls but a more sinister one may take its place. Jacqui Smith, the home secretary, faces a revolt from her senior officials over plans to build a central database holding information on every telephone call, e-mail and internet visit made in the UK. A "significant body of Home Office officials dealing with serious and organised crime" are privately lobbying against the plans, a leaked memo has revealed. They believe the proposals are "impractical, disproportionate, politically unattractive and possibly unlawful from a human rights perspective," the memo says ... ([full article](#) in the *London Times*).
 - ★ [Lockheed Martin Expands Registration for Intelligence Analysis Training](#). Beginning in October, registration for private intelligence analysis training courses will be significantly expanded. Intelligence analysis training is now being extended to the law enforcement community and U.S. citizens employed by federal or state governments, and U.S. corporations or organizations engaged in intelligence and security. The classes, conducted in Newington, VA at the Lockheed Martin Center for Security Analysis, are still available to the U.S. military, federal government, or any other organization within the intelligence community. "The decision to expand our registration is in response to the **growing demand for individually-focused expert training on analysis and data mining tools**," said Erik Kleinsmith, Lockheed Martin Program Manager for this initiative. "Organizations outside the intelligence community were unable to register for the classes in the past and similar training is often too expensive or unavailable for those wanting to participate as an individual."
 - [Contractors Augment Intelligence Agencies, Private Workers Cost More](#). (*Washington Post*) About a quarter of the nation's core intelligence workers are contractors, perhaps as many as 37,000 private employees who work side-by-side with civil servants as analysts, technology specialists and mission managers, according to a survey of government outsourcing by the **Office of the Director of National Intelligence**. ... The survey examined the national intelligence program's use of "core contract personnel" who were involved in intelligence activities during fiscal 2007. It does not include such workers as food-service employees or contract guards. It found that about 27 percent of the contract workers were involved in intelligence collection and operations. Just under a quarter were involved in information technology services. About one in five worked in analysis and production, and about the same proportion helped with administrative and support functions.

- [The U.S. spent \\$4 billion more on spying in the fiscal year ending Sept. 30 than during the previous year, the director of national intelligence said.](#) Spending on strategic intelligence by the CIA, the NSA and other U.S. intelligence organizations, reached \$47.5 billion in fiscal 2008, compared with the \$43.5 billion appropriated for fiscal 2007. The majority of the money supports electronic eavesdropping, wiretapping and the **vast, high-speed data-mining operations of the NSA**, which has headquarters at Fort Meade. The budget also funds satellite photo reconnaissance and collection and analysis by agents of the CIA, the Drug Enforcement Agency, the Treasury Department and others.
- ★ [Data mining: Know-alls – Electronic snooping by the state may safeguard liberty and also threaten it.](#) (*The Economist*) This article reviews the history of government programs to collect information on individuals for terrorist identification and tracking purposes, including Total Information Awareness (TIA) and others. *Because of The Economist's international perspective, the article is excerpted at length below (emphasis added):*

... The staggering, and fast-growing, information-crunching capabilities of data-mining technology broaden the definition of what is considered suspicious. In June America's Departments of Justice and Homeland Security and a grouping of American police chiefs released the "Suspicious Activity Report – Support and Implementation Project." Inspired in part by the approach of the Los Angeles Police Department, it urges police to question people who, among other things, use binoculars, count footsteps, take notes, draw diagrams, change appearance, speak with security staff, and photograph objects "with no apparent aesthetic value."

Companies, and especially credit-reporting firms, generally enjoy more latitude than government bodies do in making personal information available to third parties. They find intelligence agencies are eager clients. Chris Westphal, head of **Visual Analytics**, a firm in Poolesville, Maryland that operates data-mining software for security and intelligence agencies, says the data provided by such firms is "very significant". **Narayanan Kulathuramaiyer**, an expert in data mining at **UNIMAS**, a **Malaysian university**, says companies are selling database access to intelligence and law-enforcement agencies "at a level you would not even imagine."

Legal challenges to governments' use of personal information held by companies have reached high courts in many countries, including America's Supreme Court.

... profiling increasingly relies on "sentiment analysis." **Hsinchun Chen**, head of the **Artificial Intelligence Lab** at the **University of Arizona** says this technique, which he performs for American and international intelligence agencies, is an emerging and booming field. The goal is to identify changes in the behaviour and language of internet users that could indicate that angry young men are becoming potential suicide-bombers.

... Data-mining may be bad for national security as well as for civil liberties. The software is often modelled on the fraud-detection applications used by financial institutions. But terrorism is much rarer. So spotting conditions that may precede attacks is harder. **Mike German**, a former FBI agent who now advises the **ACLU**, says intelligence agencies too readily believe in the "snake oil" of total information awareness, which drains effort from more useful activities such as using informers and infiltrators.

... Last year two pattern-detection programmes, ADVISE and TALON, run respectively by America's Department of Homeland Security and the Pentagon, were shut down

following privacy concerns and irregularities. Privacy advocates, however, say that other programmes continue – and many are operated, with minimal oversight, by the National Security Agency.

In July, after fierce debate, Congress imposed new limitations on government wiretapping when it renewed the expiring Foreign Intelligence Surveillance Act (FISA) sought by President George Bush after September 11th. The main law governing data mining, this has provided the administration with broad and unprecedented electronic-spying powers. But civil-liberties lobbies such as Amnesty International and Human Rights Watch say the renewed, restricted law leaves largely untouched far-reaching secret “black” programmes, run by the NSA, which crunch data on great numbers of people, including millions of Americans. Much of that is personal financial information collected by the Treasury.

In 2001 American-led forces routed the Taliban in Afghanistan, destroying al-Qaeda training camps there. Berndt Thamm, who advises Germany’s armed forces on terrorism, says that in retreat the Islamists left valuable clues about their online communications and electronic plotting. It is in following up these leads that data mining and pattern analysis can, and should, be used. Such techniques, says Mr Thamm, are “the only answer” to jihadist extremists. That is the argument which the strenuous objections of civil libertarians need to overcome.

- **★ IEEE International Conference on Data Mining (ICDM 2008) Data Mining Contest:** The contest this year is on “**Radioxenon monitoring for verification of the Comprehensive Nuclear-Test-Ban Treaty.**” This contest concerns data mining techniques to verify worldwide compliance of the global ban on nuclear tests, which can be detected by measuring the amount of special xenon isotopes. These isotopes are also emitted during various legal activities. The problem of attributing a specific observation of airborne concentrations of radioxenon to an explosion is twofold. Firstly, in the first few weeks after an explosion the relative concentrations of the four isotopes are expected to be released in “fingerprint” relative concentrations quite distinct from other background sources. Since the CTBT stations are not located at the source of the explosion, the radioxenon is detected at a location which can be well over a thousand kilometres away. This atmospheric transport process can take weeks, which can increase the complexity of this signature. Secondly, one can never observe radioxenons emitted purely from an explosion source but admixtures of this gas with the radioxenons released from all background sources. These 2 points above constitute an interesting data mining problem for the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). (*More on the contest [here](#). Results of the contest are to be announced at the conference in December.*)

- ★ **VAST Competition:** The VAST competition (described [here](#); detailed [here](#)) has announced the winners of the competition, with three teams (denoted below with **highlighting**) selected to participate in the 2008 VisWeek (October 2008) to work on a new smaller problem with professional analysts:

TEAM	AWARD (Challenge)
Beijing University of Posts and Telecommunications, China Qi Ye, Tian Zhu, Deyong Hu, Bin Wu, Nan Du, Bai Wang, "Exploring Temporal Communication in Mobile Call Graphs"	Social Network Accuracy (Cell Phone mini challenge)
Fraunhofer Institute, Germany Natalia Andrienko, Gennady Andrienko, "Analysis of Movements with Geospatial Visual Analytics Toolkit"	Tool Integration (Evacuation Traces mini challenge)
Oculus Info Inc., Canada Lynn Chien, Annie Tat, Pascale Proulx, Adeel Khamisa, William Wright, "Visual Analysis with nSpace2" and "GeoTime"	Support for Diverse Analytic Techniques (Grand Challenge)
Palantir Technologies, USA Jason Payne, Jake Solomon, Ravi Sankar, Bob McGrew, "The Future of Analysis"	Interactive Visual Analytic Environment (Grand Challenge)
Drexel University and Pennsylvania State University, USA Donald Pellegrino, Chi-Chun Pan, Anthony Robinson, Michael Stryker, Junyan Luo, Chris Weaver, Prasenjit Mitra, Chaomei Chen, Ian Turton, Alan MacEachren, "Visualization and Collaboration in the VAST 2008 Challenge"	Data Integration (Grand Challenge)
Southern Illinois University Edwardsville, USA Dennis Bouvier, Britian Oates, "Staining for Information Discovery"	Innovative Trace Visualization (Evacuation Traces mini challenge)
SPADAC Inc., USA Benjamin Holland, Lisa Kuchy, Jason Dalton, "A Geo-Temporal Analysis of the Migrant Boat Dataset"	Analysis Summary (Migrant Boat mini challenge)
University of Bari, Italy Adalberto Simeone, Paolo Buono, "Discovering the Terrorist"	User Testing to Obtain Consensus (Evacuation Traces mini challenge)
University of California, Davis, USA Carlos D. Correa, Tarik Crnovrsanin, Christopher Muelder, Ryan Armstrong, James Shearer, Kwan-Liu Ma, "Visual Analytics of Cell Phone Data using MobiVis and OntoVis"	Intuitive Social Network Graphs (Cell Phone mini challenge)
University College Dublin, Ireland Michael Farrugia, Aaron Quigley, "Animating Multivariate Dynamic Social Networks"	Node-Link Animation (Cell Phone mini challenge)
University of Maryland, USA Adam Perer, "Using Social Action to Uncover Hidden Structure in Social Networks over Time"	Time Visualizations of Cell Phone Activity (Cell Phone mini challenge)
University of Zagreb, Croatia, VRVis Research Center and Virginia Tech, USA Ranko Miklin, Tomislav Lipic, Mario Beric, Zoltan Konyha, Wolfgang Freiler, Kresimir Matkovic, Denis Gracanin, "Geo-Temporal Analysis of Migrant Boats"	Simple and Effective Integrated Display (Migrant Boat mini challenge)
Vision Systems & Technology, Inc., USA Edward Swing, "Solving the Cell Phone Calls Challenge with the Prajna Project"	Effective Toolkit Integration (Cell Phone mini challenge)

Some additional detail on the work by the Penn State team and their winning entry is available [here](#). "Our specific work at Penn State emphasizes 'geovisual analytics,'" said Alan MacEachren, professor of geography, director of the GeoVISTA Center at Penn State and the principal investigator of NEVAC. ... "Most of the visualizations were built on Improve... The team built novel analytic components and coupled them with the Improve front-end ... approach helped us identify insights into the data that would not have been possible either by visualization or by analytic techniques in isolation."

- ★ **IACA** (the [International Association of Crime Analysts](#)) has an interesting [list of software for crime analysis](#) listed on their website. IACA also supports training for criminal intelligence analysis (e.g., see [here](#)). This professional association was formed in 1990; membership is around 1,500.
- ★ **Even Spies go to Trade Conferences**. (*Washington Post*) ... an **ODNI conference** ... to promote using **open sources** of information such as the Internet and television broadcasts as part of the intelligence process. ... The gathering reflected the intelligence community's evolution since the Sept. 11, 2001, terrorist attacks. ... intelligence analysts and government policymakers ... increasingly turning to the torrents of information available on the Internet and through other non-classified sources. **Booz Allen Hamilton** offered a service called **InTrack** to help collect, monitor and process data collected from the Internet and other sources. **LexisNexis** promoted a system for sending automated warnings of trouble abroad. There were companies selling translation systems, Web search tools and data-mining supercomputers. One of the more popular booths was **Google's**, though exactly what it wanted to sell the intelligence community was not clear.

... Booz Allen's InTrack service seems to derive almost oracular insights of the sort government leaders crave. The impact of a tsunami in Taiwan on global communications? The fallout from a spike in tuberculosis? Links between unexplained incidents and terrorism? "[InTrack's mission is to collect, monitor, process, and combine data with existing 'trusted' sources and baseline patterns to perform robust analyses,](#)" the company's flier says.

At the LexisNexis station, a slick brochure touted a "data analytics supercomputer" as an instant solution to intelligence dilemmas. The brochure claims the system can manage [hundreds of terabytes of data](#) – the equivalent of many times all the holdings in the Library of Congress. "Unparalleled linking technology and analysis uncovers key connections and relationships," it says.

The Google booth displayed a high-definition video of a virtual car driving through an exact digital representation of San Francisco – streets, buildings and all. But Google exhibitors said they were not allowed to tell a reporter why the company was there or what it did for intelligence or anything else.

B. Cybersecurity

- ★ **Social Network Analysis and Cyber Warfare: An Open Source Project** (posted on the *IntelFusion Blog* in August).

...Thanks to the ideas that Bob Gourley has recently expressed in his "Social Media and the National Security Professional," and to industry contacts made via this blog and through Twitter, ... [members of the IntelFusion blog] launched a **social network analysis of Russian cyber warfare activities**. Palantir Technologies has generously offered us the use of its very impressive analytic platform to conduct our research. We'll be looking not only at network data involved in past cyber warfare attacks (Chechnya, Estonia, and Georgia), but incorporating semantic analysis of Russian

hacker blogs in an effort to uncover connections that may not be readily apparent. If this model proves efficacious, we'll launch a second effort examining Chinese cyber warfare/espionage activities. This is a pure grass roots effort using only open source data pulled from the Web. All the participants are volunteers. Regular updates will be posted here, and our findings will be published in the appropriate venues. *(The source contains several updates. Material which we considered newsworthy from those updates is incorporated elsewhere in this report.)*

- [Cybercrime toll mounts for businesses](#). 1,387 IT professionals [were] surveyed by security firm Finjan. 91% of respondents call cybercrime a "major business risk," and 73% say they are more concerned about data theft than about downtime and loss of productivity from malware. In addition, 25% of respondents admit to data breaches in their organizations.
- [Cyber security threats grow in sophistication, subtlety and power](#) – [The latest version of an] annual report from Georgia Tech Information Security Center identifies five evolving cyber security threats, and the news is not good. GTISC interviewed a range of industry security experts to explore the threats and the available countermeasures. The five are malware, botnets, cyber warfare, threats to VoIP and mobile devices, and the "evolving cyber crime economy." In all five areas, attackers are becoming increasingly sophisticated ... ["Emerging Cyber Threats Report for 2009: Data, Mobility and Questions of Responsibility will Drive Cyber Threats in 2009 and Beyond,"](#) is available [online](#).
- [Social network providers should enhance security controls available to users and businesses, says analyst firm Gartner](#). Emerging threats from virtual environments include **social network analysis tools** that allow easy integration of data from a variety of sources, and potential flaws in user interfaces and media formats such as QuickTime, AVI and MP4. "These threats are exacerbated by the speed at which new features are developed and implemented by the providers of virtual environments, without a long-term testing process to identify security flaws," said [Gartner analyst Andrew] Walls. ... Gartner said virtual worlds, social networks and mapping environments will merge into highly integrated online environments over the next 10 years. "Organisations cannot block social networks and virtual worlds because they will become the base infrastructure for business and personal interaction in the future," said Walls. "Now is the time to build security tools and infrastructure that enable the organisation to benefit from them."
- **DOD's DNI** (Director of National Intelligence) has released "[VISION 2015: A Globally Networked and Integrated Intelligence Enterprise](#)" (available [here](#)).
- The Air Force's "**Cyber Command**" may move forward, but in a very diminished capacity. The Air Force suspended the Cyber Command in August; as of mid-October ([source](#)):

The Air Force says it will back off its ambitious plan to set up a separate command for cyber space ... Placing cyber operations under Space Command will "mean a diminution in its marquee value from Broadway to Off-Broadway," the source said. "But the command will still have warfighting capabilities." ... The Cyber Command was a victim of the Air Force's neglect of its nuclear mission, Defense sources said.

- [National Cyber Security Initiative will have a dozen parts](#). (*GovExec.com*) President Bush's largely classified government wide cybersecurity initiative will have a dozen components designed to better protect computer networks and systems, and to improve information technology processes and policies, a Homeland Security Department official said. ... Some analysts estimate that it will cost as much as **\$40 billion** to implement the cybersecurity plan. ... The Trusted Internet Connections program is the most established

piece of the initiative. ... The other 11 components ... are aimed at ... the following areas: Intrusion detection; Intrusion prevention; Research and development; Situational awareness, specifically through the **National Cyber Security Center**, ...; Cyber counter intelligence; Classified network security; Cyber education and training; Implementation of information security technologies; Deterrence strategies; Global supply chain security; and Public / private collaboration. DHS has started to improve collaboration with the private sector, dubbed "Project 12." Department officials have met with executives from the banking, telecommunications and energy industries, among others. The 12 projects will allow the federal government to take a broad view of cybersecurity.

The public face of the program continues to be a mix of fairly mundane blocking and tackling – e.g., the TIC effort and a focus on “traffic sensors ... [which] will eventually be implemented in all agencies to detect malicious software and alert DHS to security breaches in real time.”

- **★ U.S. urged to develop offensive cyberwar capabilities.** The United States needs to do more to develop an offensive cyberwar capability, rather than just focus on defending its networks from attack, says the chairman of the House Cybersecurity Subcommittee. "The best defense is a good offense, and an offensive (cyberwar) capability is essential to our national defense," Rep. James Langevin, D-R.I., told *United Press International*, calling it "a necessary deterrent."

... Langevin, the chairman of the House Homeland Security Subcommittee on Emerging Threats, Cybersecurity and Science and Technology, and a member of the House Intelligence Committee, also **called on the White House to declassify much more of its Comprehensive National Cybersecurity Initiative, or CNCI**, and said the **Department of Homeland Security should be stripped of its lead role in defending the nation's computer networks.** ...

Former intelligence official Suzanne E. Spaulding told the hearing that focusing on offensive capabilities and giving a lead role to the military might make it harder for the United States to work with other countries on cyber issues, where the lines separating crime, terrorism and warfare are often hard to draw. "My concern is that [the Defense Department] has been so vocal ... that it will be very difficult for that department to develop and sustain the trust necessary ...".

- **FYI: MIT Lincoln Laboratory software aims to thwart cyber hackers.** In response to the chronic cyber threat of hackers, MIT Lincoln Laboratory researchers are developing a software tool to identify the most vulnerable points in a computer network. The tool aims to make it possible for system administrators to focus on parts of a network that are most prone to attack, instead of securing all parts of the network. **NetSPA** (for Network Security Planning Architecture) uses information about networks and the individual machines and programs running on them to create a graph that shows how hackers could infiltrate them. System administrators can examine visualizations of the graph themselves to decide what action to take, but NetSPA also analyzes the graph and offers recommendations about how to quickly fix the most important weaknesses.

C. The Big Data Problem

- [The big data dump](#) – A deluge of electronic information may overwhelm American civil justice. (*The Economist*) ... The process of e-discovery starts when the adversaries in a lawsuit demand to see all sorts of information in their search for relevant nuggets. Each side then has to identify all the laptops, smart-phones, memory sticks, network servers ... [and probably] request logs from online-service providers ... The results then have to be indexed and reviewed by humans. This usually falls to the junior staff at law firms, some of whom are so fed up with the drudgery that they have quit the profession altogether.
- [Start-Ups Bring Google's Parallel Processing To Data Warehousing](#). **Aster** and **Greenplum** have made Google's MapReduce compatible with SQL for use in the parallel data warehousing systems based on open source PostgreSQL. MapReduce is a software framework that allows users to complete many instances of tasks simultaneously across many computers, often commodity servers. ... "We took the model that Google used, farms of commodity servers, and we offer a database that can be deployed on massive farms of clusters and can transform hundreds of individual servers into a single database..." Both Aster and Greenplum have made MapReduce compatible with SQL for use in the parallel data warehousing systems based on open source PostgreSQL ... many of the uses of Greenplum will likely be for data sets in the tens of terabytes.
- [New Visualization Laboratory debuts on UT Austin's main campus](#). Texas Advanced Computing Center (TACC) unveiled a modern, reconfigured Visualization Laboratory (VisLab) capable of reproducing terascale (one trillion bytes) data sets with exceptional clarity and resolution. ... a 20.5' x 7.5' display environment ... offering 184 million pixels, powered by a cluster of 23 high-end Dell XPS systems, ... Stallion enables datasets to process on a massive scale, and allows the interactive visualization of substantial geometries.



D. Funding

- **★ Air Force:** Request for Information (RFI) – **Information Fusion and Understanding Applied Research, Experimentation and Demonstration**. Responses are due **November 30, 2008**. Full Synopsis [here](#).

This RFI announcement is an exploratory request ... This announcement is not a request for proposals.

The Information and Fusion Core Technical Competency is defined as the continuous process that provides world-wide situational awareness in order to enable decision superiority. Activities include research, development, integration and testing of innovative technologies that enable continuous assessment of global conditions and events; that establish and maintain battlespace situational awareness

Some of the subtopics: **Data Mining**, Temporal Entity Databases, Automatically Harvesting Information about Networks, Dynamic Networks Over Time, **Abnormality Detection in Counter Terrorism Evidence**, and **Advanced Analysis**.

- The U.S. **Army Research Laboratory (ARL)** announced that it is developing preliminary plans for a potential Network Science Research and Technology Center of Excellence consisting of four components: a Social Cognitive Networks Research Center, an Information Networks Research Center, a Communication Networks Research Center, and an Integration Center for evaluation and analysis of research outputs and overall management of the Center of Excellence. (More details [here.](#))
- **WebTAS:** Capable sources are sought to perform software research and development for the RIE Division ... AFRL Rome has a requirement to conduct software development and system support for the **Web-Enabled Temporal Analysis System Tool Kit (WebTAS TK)**. WebTAS supports Air Force efforts in providing Department of Defense commands, services and various Federal Agencies with a software toolkit that provides a general data visualization and analysis infrastructure for the analysis of temporal, spatial, entity and association information. The flexible WebTAS architecture provides access to multiple data sources, data mining, and collaboration tools that facilitate trend, link, pattern, and distributed analysis to support activity prediction and customer unique capabilities like Strategic Worldwide Integration Capability (SWIC), Predictive Analysis and Network-centric Analysis for Collaborative Intelligence Assessment (**PANACIA**) and Master Air Attach Plan (MAAP) Toolkit. (More details [here.](#))
- *(National Labs Not Eligible)* **NIH (R01): Exceptional, Unconventional Research Enabling Knowledge Acceleration (EUREKA).** ... National Library of Medicine seeks innovative applications in the following two 'discovery science' areas, including "**Integrated discovery mining for biology and medicine,**" which requires employing two or more sources to generate a new and meaningful hypothesis in biomedical science, capable of being tested by bench or clinical research. One source must be full-text published biomedical literature; the other source should be either (1) a database storing primary data from basic biomedical research or (2) data drawn from the electronic health records used for routine clinical care or from the data accumulated for a clinical research project. ...The rationale for EUREKA is that for science to move forward in leaps rather than in incremental steps, investigators must have opportunities to test unconventional, potentially paradigm-shifting hypotheses ... (More details [here.](#))
- **★ FYI: Counter-Improved Explosive Devices Blogging** (HSHQDC-09-R-00004; Response date is November 28, 2008). Sources Sought [Notice](#). DHS. ... As the use of the internet by terrorists has increased, blogging and message boards have played a substantial role in allowing communication among those who would do the United States harm. In order to better counter the use of IEDs, it is necessary to identify speech acts in near to real-time which [precede] the decision by terrorists to use an IED. *[Capabilities of interest include]* ... 2) developing objective, systematic data collection and retrieval techniques to gather data on a near real-time basis from blogs and message boards. ... metrics for determining the impact factor and usage patterns of the blogs and message boards. 3) identifying blogs and message boards utilized or favored by groups that engage in violent or terrorist activity to include in the study. Blogs and message boards must be representative of various characteristics of the larger populations of interest. and 4) collecting quantitative and qualitative data from the bloggers to evaluate such issues relating to knowledge of the preparation and execution of violent activities, including IED attacks.

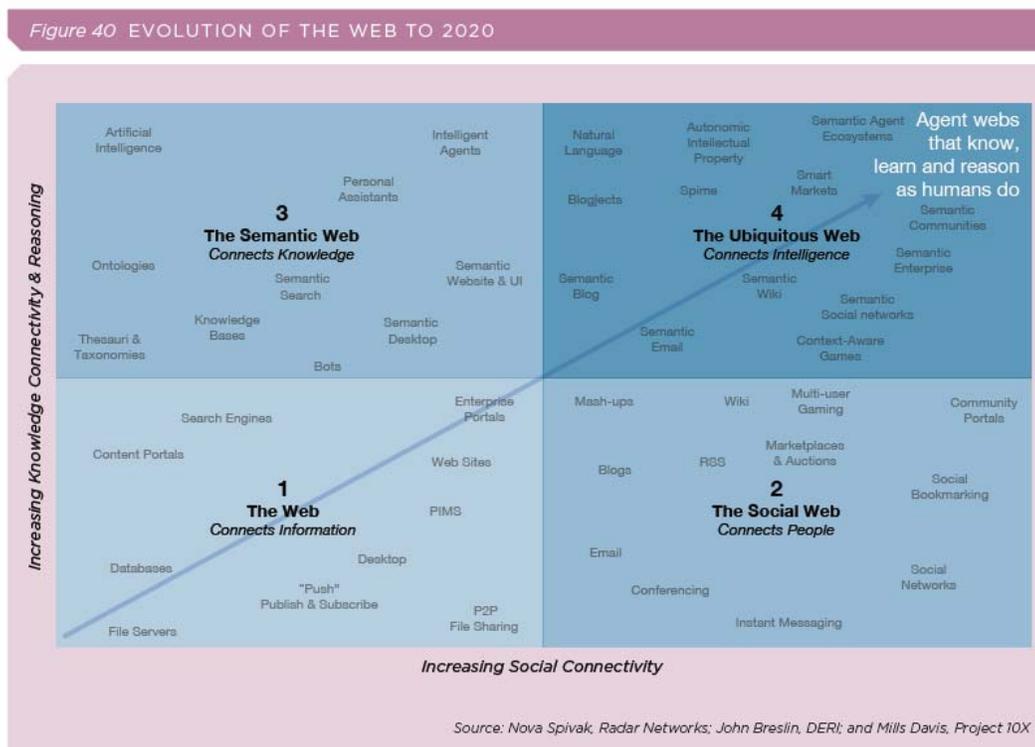
II. Other Applications and Items of Interest

A. Epidemiology / Medical / Life Sciences / Pharma

- [Eidetics Launches Groundbreaking Data Visualization Service ... Better Drug Development Decisions](#). Eidetics, a global healthcare market research firm ... launched a new interactive tool provides pharmaceutical brand teams and clinical researchers with a panoramic view of complex data, utilizing advanced data mining techniques and Bayesian networks. ...organize and analyze hundreds of variables at once, and instantly visualize statistically significant relationships. ... access to valuable public-domain patient-level data collected and maintained by the US Centers for Disease Control and Prevention (CDC), such as the National Ambulatory Medical Care Survey (NAMCS) and National Health and Nutrition Survey (NHANES).

B. Business Analytics

- [The Challenges of Business Intelligence](#). As BI has evolved, the greatest challenge has been how to integrate data on different systems accumulated from different vendors over many years. This challenge is being addressed in the following key ways: 1) "Service-oriented architecture" (SOA)3) The big vendors – IBM , Microsoft , Oracle and SAP – which already offered some BI solutions, have moved into performance management by acquisition. ... As Gartner puts it: "The data warehouse has evolved from an information store supporting the running of reports and ad-hoc queries into an analytics infrastructure repository." ... The article continues with an extensive discussion of the activities of the larger vendors.
- ★ ["Digital Disruptions – Technology Innovations Powering 21st Century Business"](#) (full text) prognosticates on a variety of issues, including the increasing important of networks (pp. 69-79). The articles are from a "Leading Edge Forum" run by consulting firm CSC.



C. Other Items of Interest

- **[Collaboration: Group theory. What makes a successful team?](#)** (*Nature*, October 8, 2008 – full text freely available [here](#).) A look at research that uses massive online databases and network analysis to come up with some rules of thumb for productive collaborations. ... Sociologist **Brian Uzzi** of **Northwestern University** in Evanston, Illinois, and his colleagues analyzed more than 2 million patents, along with nearly 20 million papers published since 1955. They found that in the early 1950s, the most cited paper in any year was more likely to have been written by a single author than a team, but this pattern reversed decades ago. ... **[Debbie] Duran** of NIH does expect network studies to be an important part of what she calls "the **emerging science of science management**." [The NIH already uses data-mining tools devised by the company Discovery Logic, based in Rockville, Maryland, to see how grants connect to papers, citations, patents and products.](#) Duran suggests that in the future, network analysis could be used to track the spread of new ideas, work out the best ways to disseminate information or to target particularly well-connected individuals to work on emerging issues. "I think, hope and believe that this will become useful," she says. **Network visualization of possible interest:** By mapping which authors collaborated with whom and when, **Katy Börner**, who studies networks and information visualization at **Indiana University** in Bloomington, show how networks extend with time and how the impact of both an author and a partnership can grow. See [animation](#).
- "The Structure of Information Pathways in a Social Communication Network" is a paper presented at the 2008 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, by Gueorgi Kossinets and Jon Kleinberg of **Cornell** and Duncan Watts of **Yahoo! Research** ([full text paper](#)). A write up of the presentation, as well as a video of the talk may be found [here](#).

Abstract: Social networks are of interest to researchers in part because they are thought to mediate the flow of information in communities and organizations. Here we study the temporal dynamics of communication using on-line data, including e-mail communication among the faculty and staff of a large university over a two-year period. We formulate a temporal notion of "distance" in the underlying social network by measuring the minimum time required for information to spread from one node to another – a concept that draws on the notion of vector-clocks from the study of distributed computing systems. We find that such temporal measures provide structural insights that are not apparent from analyses of the pure social network topology. In particular, we define the network backbone to be the subgraph consisting of edges on which information has the potential to flow the quickest. We find that the backbone is a sparse graph with a concentration of both highly embedded edges and long-range bridges – a finding that sheds new light on the relationship between tie strength and connectivity in social networks.
- **[New Algorithm Significantly Boosts Routing Efficiency of Networks](#)**. The XL algorithm developed by computer scientists at **UC San Diego** significantly outperforms standard link-state and distance-vector algorithms, speeding routing in computer and communications networks. It promises to significantly boost the efficiency of network routing. The algorithm increases network routing efficiency by suppressing updates from parts of the system – updates which force connected networks to continuously re-calculate the paths they use in the great matrix of the Internet.

III. Company News, in Brief

- **21st Century Systems:**

21CSI received a **\$7.43 million deal** to expand its Webster-Agent Case Expert technology across the Defense Department's networks from its current isolated use at the Strategic Command's Joint Information Operations Warfare Command. Webster-ACE is a set of versatile software applications which allows military users to discover, connect, and visually present seemingly unrelated information in support of intelligence and information operations. ([Source](#))

21CSI Spins-off **Eelios, Inc.** Eelios is... taking to the commercial market ... technology that 21CSI has developed and fielded for use in a variety of government programs ... an **initial focus [is] on intelligent personal weather monitoring** and alerting. (More detail [here.](#))

- **Aperio CI:**

Aperio CI plans to release a new product later this year aimed at tapping social networks. **SNAP (Social Networks Analysis & Propensities)** ... allows operators to examine their subscriber base and identify social networks. It builds the social networks by analyzing each subscriber's on-net call and text traffic and generating links between different subscribers... by building social networks it can identify the highly influential subscribers. (More detail [here.](#))

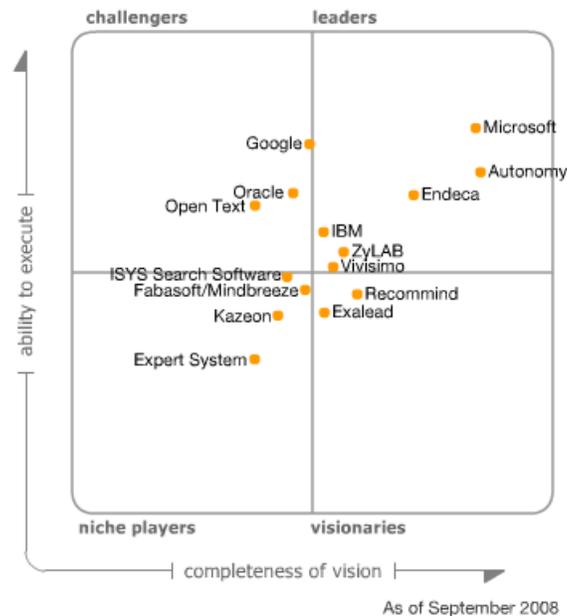
- **Autonomy:**

According to a recent report by IDC, Autonomy is a major player in the worldwide "data search and discovery" market. IDC says Autonomy has a 15.5% market; Google has 8.5%, Fast Search & Transfer (now owned by Microsoft) holds 5.7%, and Microsoft is at 3.6%. IDC said the search and discovery software market grew to £1.1bn in 2007, a growth rate of 28%. (More detail [here.](#))

Autonomy/Virage unveiled a new **Situational Awareness Web Portal** to control and monitor their security and surveillance activities through a rich user interface accessible over a standard Internet connection. This Portal not only provides security officers with a secure, comprehensive view and analysis of security scenarios, but it also allows them access to all relevant intelligence information. (More on the overall platform [here.](#))

- **★ Blue Spider Analytics:** The company ([website](#)) has recently released a new beta version of its product to registered users. The company has also added some online webcasts demonstrating their product – see [Workspace Overview](#), [Extract Stepped Network](#), and [Extract Edge List](#). In addition, the company conducted live demonstrations of their new network analysis software at the 2008 IACA & FCIAA Joint Training Conference in October.
- **Endeca:** Gartner has released a magic quadrant report on "**Information Access Technology**," which they define as "technologies [that] access applications such as document management, Web content management and relational database management systems to provide users with insight into their contents." "Leaders" include **Microsoft**, **Autonomy**, and **Endeca** (the In-Q-Tel funded company). The full report is available [here](#), including an enumeration of the strengths and weaknesses of the companies shown below.

Figure 1. Magic Quadrant for Information Access Technology



Source: Gartner (September 2008)

- **Detica / BAE:**

[Fraudsters hiding under an avalanche of data.](#) **Tom Black**, the chief executive of **Detica Group**, spoke to July's Homeland and Border Security conference and provided some info on company focus. "Every day three billion emails, eight billion text messages; every month six billion Internet searches. In 2007: 281 exabytes of data. By 2011: 1.8 zettabytes. What's an anti-terrorism squad to do?" ... Detica began in 1971 as Smith Associates, and spent the 1970s and 1980s working on various Cold War projects. "The core of our service to clients then – governments and national security – was cracking tough problems, usually involving large amounts of data," says David Porter, head of Detica's Security and Risk. Now, the company's clientele is half government, half commercial, but the line of work is generally the same. In addition, says Porter, Detica is retained by banks and other organisations to spot good customers and detect credit card fraud, money laundering, transport ticket fraud, and so on. "Anywhere that you're looking through a stream of transaction data to spot something 'interesting', where 'interesting' depends on your point of view."

★ [Spy chiefs plot £12bn IT spree for comms überdatabase.](#) Billions of pounds of public money will soon be up for grabs for private IT contractors ready to serve the **Interception Modernisation Programme** [described elsewhere as the *UK's version of "Total Information Awareness"*] - UK spy chiefs' plan to store details of every call, email, text and web browsing session. (Detica is considered a shoe-in contractor for this program, and was described in this article as "The Most Important IT Company Most People Have Never Heard Of.") According to the article's sources (who have knowledge of systems that have long allowed GCHQ to eavesdrop on phone calls), Detica owns and operates the current "black box" infrastructure under contracts funded by the secret intelligence budget.) (See also [article](#) in *Computer Weekly*.)

[Her Majesty's Revenue & Customs purchases Detica NetReveal® licence.](#)

[Detica reports contract with global insurer, RBS Insurance.](#)

- **i2 Ltd.:**

[Silver Lake Sumeru completes acquisition of i2.](#) Silver Lake Sumeru (SLS) announced that it has completed the acquisition of i2 in a transaction valued at approximately \$185 million.

[Environmental Crime:](#) A report from the UK's Environmental Investigation Agency (EIA), "[A Threat To Our Future](#)," says the authorities are not taking environmental crime seriously enough. The agency has installed i2's software to help it infiltrate global networks of these criminals.

- **InferX:**

[InferX Corporation Approved as Subcontractor on U.S. Navy SeaPort-e Contract.](#) InferX will serve as a sub-contractor to the Iirus Group, Inc. (www.iirusgroup.com) ... [to] "provide proven predictive analytic tools and technologies to enable the Navy to meet critical operational needs."

- **IBM:**

[New IBM center to focus on social software.](#) Building on the momentum in social networking for business, IBM Corp. is opening the IBM Center for Social Software in Cambridge, Mass., to research and quantify the effects of social software on workplace productivity. ... partnerships with **Harvard University** and the **Massachusetts Institute of Technology** Media Lab.

- **ISS (Intelligent Software Solutions):** *(Most news items below could not be directly hyperlinked; access them from the "news" tab at the top of their homepage. For more information on any of the products mentioned below, use the "solutions" tab accessible from the homepage.)*

AFRL Awards ISS a \$20.5 Million Task for Operations Support and Research and Development Efforts. ...ISS will undertake efforts related to research, net-centric software development, and support to ongoing combat operations around the world.

ISS awarded \$49.9 million dollar increase in Air Force contract for Secure Information Access Analysis and Dissemination contract vehicle. Through the SIAAD vehicle, ISS provides enterprise solutions and services to a wide range of government customers within the Defense Department, Department of Homeland Security and intelligence agencies.

- **Lockheed Martin:**

LM has developed **Contrail**, a context-aware computing framework that gives the intelligence community support to capture, retrieve and share contextually relevant information at reduced time and cost. Integrated into an intelligence agency's computing infrastructure, Contrail's software builds an explicit, machine-understandable representation of analysts' contexts by monitoring how they handle information. The technology then builds a personalized software model that automatically tags newly found data, enabling analysts to later re-find that needed intelligence using metadata, content, or context at time of storage. ... developed as ... part of the Collaboration and Analyst/System Effectiveness program sponsored by the Intelligence Advanced Research Projects Activity. ([Press release](#))

- **Tibco Inc. / Spotfire:**

[Tibco Unveils 'Network Analytics' For Spotfire](#): The add-on... presents a visual representation of interactions among people, business processes or other entities within Spotfire, the company's in-memory database/analytics software for business activity monitoring. In general, [Spotfire](#) lets users take a set of instance data and interactively scatter-plot any metrics against each other, zoom in on a region of interest, filter instances with a slider in real-time and list the instances in the set. ... The software **provides help in any type of relationship analysis**, from identifying e-mail or cell phone traffic patterns to performing **social network analysis** to analyzing counter-party or supply-chain relationships. Spotfire 2.2. starts at about \$1,000 per seat. ([Information page](#) on network analytics.) (Additional sources [here](#) and [here](#).)

- **Blog posts** of interest:

- [20 Useful Visualization Libraries](#)
- [Visualization Companies Leading the Way](#): Tableau Software, Stamen Design, Spotfire/Tibco, and Kitware are covered.
- [DataDepot](#): From Microsoft Research, a “set of tools for collaboratively uploading, sharing, and analyzing data. You can use DataDepot to track personal data, to explore public data, and to engage with scientific data.”
- [“Epidemiology” of the Credit Crisis.](#)” With a short empirical investigation, Reginald Smith (MIT – Sloan School of Management) have come to some interesting complex networks (nodes in here are financial stocks) over time, since the beginning of the financial crisis in August 10, 2007, till today. His rather simple econophysics study (draft [PDF link](#)) demonstrates that the losses in certain markets, in this case the US equity markets, follow a cascade or “epidemic” flow like model along the correlations of various stocks.

IV. Resources / Overviews

A. Reviews and Overview Articles

- ★ “Visual Data Mining – Theory, Techniques and Tools for Visual Analytics” in the book series *Lecture Notes in Computer Science*, v. 4404/2008 (Simeon J. Simoff, Michael H. Böhlen and Arturas Mazeika, Editors). The table of contents of this volume is shown below. ([Introduction](#) to volume; online access [here](#).)

Visual Data Mining: An Introduction and Overview 1 <i>Simeon J. Simoff, Michael H. Böhlen, and Arturas Mazeika</i>	Mining Patterns for Visual Interpretation in a Multiple-Views Environment 196 <i>José F. Rodrigues Jr., Agma J.M. Traina, and Caetano Traina Jr.</i>
Part 1 – Theory and Methodologies	
The 3DVDM Approach: A Case Study with Clickstream Data 13 <i>Michael H. Böhlen, Linas Bukauskas, Arturas Mazeika, and Peer Mylov</i>	Using 2D Hierarchical Heavy Hitters to Investigate Binary Relationships 215 <i>Daniel Trivellato, Arturas Mazeika, and Michael H. Böhlen</i>
Form-Semantics-Function – A Framework for Designing Visual Data Representations for Visual Data Mining 30 <i>Simeon J. Simoff</i>	Complementing Visual Data Mining with the Sound Dimension: Sonification of Time Dependent Data 236 <i>Monique Noirhomme-Fraiture, Olivier Schöller, Christophe Demoulin, and Simeon J. Simoff</i>
A Methodology for Exploring Association Models 46 <i>Alípio Jorge, João Poças, and Paulo J. Azevedo</i>	Context Visualization for Visual Data Mining 248 <i>Mao Lin Huang and Quang Vinh Nguyen</i>
Visual Exploration of Frequent Itemsets and Association Rules 60 <i>Li Yang</i>	Assisting Human Cognition in Visual Data Mining 264 <i>Simeon J. Simoff, Michael H. Böhlen, and Arturas Mazeika</i>
Visual Analytics: Scope and Challenges 76 <i>Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler</i>	
Part 2 – Techniques	
Using Nested Surfaces for Visual Detection of Structures in Databases 91 <i>Arturas Mazeika, Michael H. Böhlen, and Peer Mylov</i>	
Visual Mining of Association Rules 103 <i>Dario Bruzese and Cristina Davino</i>	
Interactive Decision Tree Construction for Interval and Taxonomical Data 123 <i>François Poulet and Thanh-Nghi Do</i>	
Visual Methods for Examining SVM Classifiers 136 <i>Doina Caragea, Dianne Cook, Hadley Wickham, and Vasant Honavar</i>	
Text Visualization for Visual Text Analytics 154 <i>John Risch, Anne Kao, Stephen R. Poteet, and Y.-J. Jason Wu</i>	
Visual Discovery of Network Patterns of Interaction between Attributes 172 <i>Simeon J. Simoff and John Galloway</i>	
Part 3 – Tools and Applications	
Immersive Visual Data Mining: The 3DVDM Approach 281 <i>Henrik R. Nagel, Erik Granum, Soren Bobbjerg, and Michael Vittrup</i>	
DataJewel: Integrating Visualization with Temporal Data Mining 312 <i>Mihael Ankerst, Anne Kao, Rodney Tjoelker, and Changzhou Wang</i>	
A Visual Data Mining Environment 331 <i>Stephen Kimani, Tiziana Catarci, and Giuseppe Santucci</i>	
Integrative Visual Data Mining of Biomedical Data: Investigating Cases in Chronic Fatigue Syndrome and Acute Lymphoblastic Leukaemia. 367 <i>Paul J. Kennedy, Simeon J. Simoff, Daniel R. Catchpole, David B. Skillicorn, Franco Ubaudi, and Ahmad Al-Oqaily</i>	
Towards Effective Visual Data Mining with Cooperative Approaches ... 389 <i>François Poulet</i>	

- **Make Molehills from Unstructured Data Mountains:** An article on the use of visual methods for dealing with large volumes of unstructured data. (Overview [article](#) at *Sci-Tech Today.com*)

B. Resources on the Web

The ACM [Transactions on Knowledge Discovery from Data](#) (TKDD) – which began publication in 2007 – welcomes papers on a full range of research in the knowledge discovery and analysis of diverse forms of data.

C. Conferences

- ★ The **DNI Open Source Conference 2008** was held in September. This two-day event explored a range of open source issues and best practices for the Intelligence Community and its partners. The [agenda](#) contains some links to presentations during this conference. While most of these were rather uninformative (and perhaps were drastically sanitized in their web-accessible versions), the [presentation](#) by **Gary LaFree** of the **START Center** at the **University of Maryland** was particularly interesting. He posed and addressed a number of questions with data from the Global Terrorism Database – for example, “Do U.S. attacks by terrorist groups fit into clear trajectories over time?” This final slide presented implications of the data:



The conference also included an "**Open Source Challenge**" in a session where representatives from academia; think tanks; industry; the media; government; and other diverse sectors were invited to submit ideas and solutions to current intelligence community challenges directly to the ODNI. Participants presented their solutions at the final plenary session of the conference.

- **INSA's** (Intelligence and National Security Alliance) **Analytical Transformation 2008** was held in September. The [agenda](#) covered such issues as “Development of an Integrated Nation-wide Analytic Framework,” ... and “Technological Transformation of the Analytic Environment.” Remarks by **Thomas Fingar**, the Deputy Director of National Intelligence for Analysis & Chairman, National Intelligence Council, are available [here](#). A *CNN* transcript of **Mike Wertheimer**, Assistant Deputy Director of National Intelligence for Analytic Transformation & Technology, discussing Analytic Transformation on *CNN* & *Federal News Radio*, is available [here](#). Wertheimer also discussed **A-Space**.
- **Upcoming Conferences:**
 - **IEEE International Conference on Data Mining series (ICDM)**, Pisa, Italy, December 15 - 19, 2008 ([call for papers](#); see also discussion of the data mining challenge described [earlier](#) in this report).
 - Inaugural **Predictive Analytics World** conference ([website](#)), February 18 - 19, 2009 in San Francisco, CA. “... sessions across verticals such as banking, financial services, e-commerce, entertainment, healthcare, high technology, insurance, non-profits, publishing, and retail. ... covers the gamut of commercial applications of predictive analytics, including response modeling, customer retention with churn modeling, product recommendations, online marketing optimization, behavior-based advertising, email targeting, insurance pricing and credit scoring.”