From Data to Knowledge to Action:
Enabling Advanced Intelligence and Decision-Making for America's Security

Randal E. Bryant
Carnegie Mellon University

Jaime G. Carbonell
Carnegie Mellon University

Tom Mitchell
Carnegie Mellon University

Computing Community Consortium
Version 6: July 28, 2010[1]

Large-scale machine learning can fundamentally transform the ability of intelligence analysts to efficiently extract important insights relevant to our nation's security from the vast amounts of intelligence data being generated and collected worldwide. Intelligence organizations can tap into rapid data analytics innovations that Internet industries and university research organizations are making through the use of unclassified research partnerships.

Our data-centric world

Our world is being flooded by ever-increasing volumes of data. With the Internet, videocameras, cell phones, satellites, and scientific instruments, it's estimated that close to one zettabyte (10^{21} bytes or one billion terabytes) of digital data are generated each year, and this number is rising steadily. Amidst all these data is information of great importance to understanding possible threats to our nation's security. Important data sources for intelligence gathering include:

- Satellite imagery
- Intercepted communications: civilian and military, including voice, email, documents, transaction logs, and other electronic data
- Radar tracking data
- Captured media (computer hard drives, videos, images)
- Public domain sources (web sites, blogs, tweets, and other Internet data, print media, television, and radio)
- Sensor data (meteorological, oceanographic, security camera feeds)
- Biometric data (facial images, DNA, iris, fingerprint, gait recordings)
- Photos, videos, and data collected by assets on the ground or in the air (e.g., unmanned aerial vehicles, or UAVs)
- Structured and semi-structured information supplied by companies and organizations: airline flight logs, credit card and bank transactions, phone call records, employee personnel records, electronic health records, police and investigative records, and much more

The challenge for our intelligence services is to find, combine, and detect patterns and trends in the traces of important information lurking among the vast quantities of available data in order to recognize threats and to assess the capabilities and vulnerabilities of those who wish to cause harm to our nation or disrupt our society. These challenges keep getting harder as the total

---

[1] Contact: Erwin Gianchandani, Director, Computing Community Consortium (202-266-2936; erwin@cra.org). For the most recent version of this essay, as well as related essays, visit http://www.cra.org/ccc/initiatives.

amount of data grows. The proverbial challenge of finding a "needle in a haystack" becomes more difficult as each haystack grows larger, and even more as the number of haystacks increases.

As a further challenge, the line between information that is relevant to intelligence and just ordinary data is becoming increasingly blurred. Terrorists use the same cell phone and email technology as billions of people worldwide. Our adversaries hide themselves and their operations among civilian populations both to escape detection and to shield themselves from reprisal. They also use the latest encryption and information hiding technology to disguise their tracks. As a result, some say that the problem of finding a needle in a haystack has been transformed into one of finding a "needle in a stack of needles." In fact, the problem is even more complex, since we are seeking to understand the combined implications of many data elements, rather than individual facts. The challenge becomes one of finding meaningful evolving patterns in a timely manner among diverse, potentially obfuscated information across multiple sources. These requirements greatly increase the need for very sophisticated methods to detect subtle patterns within data, without generating large numbers of *false positives* so that we do not find conspiracies where none exist.

As models of how to effectively exploit large-scale data sources, we can look to Internet companies, such as Google, Yahoo, and Facebook. They have created a multibillion-dollar industry centered on gathering vast quantities of data in many different forms (approximately 10–20 petabytes per day). The companies then index, analyze, and organize these data so that they can provide their customers with the information most relevant to their needs. These Internet pioneers are creating massive computer systems that dwarf the scale of the world's largest supercomputers, albeit with an architecture optimized more for data collection and analysis than for number crunching. Their investment in IT infrastructure is unprecedented; in 2007 alone, Google invested $2.4 billion building new data centers. They are developing technology for language translation, document summarization, social network analysis, mapping and geospatial analysis, parallel programming, and data management that can cope with and exploit the vast amounts of information being generated worldwide. They are hiring the best and brightest young people from our universities, lured by the opportunities to create new, highly visible capabilities and to have access to such rich information and computing resources.

Within the world of intelligence, we should consider computer technology as a way to augment the power of human analysts, rather than as a way to replace them. Computers can make analysts more efficient by reducing the volume of data that they must review through document filtering and summarization. Computers can eliminate language barriers through automatic translation and multilingual search. Computers can support collaboration between analysts with information sharing tools. They can also help analysts be more vigilant by automatically generating notifications when suspicious activities are detected, updating the activity detector based on analyst feedback. Such capabilities call for careful consideration of human factors in designing computer technology for intelligence applications.

Machine learning technology as a driver

As described in our overview whitepaper[2], machine learning is the core technology for extracting meaningful insights from large quantities of data. Rather than viewing the flood of data as a burden, machine learning views it as an unprecedented opportunity, gaining stronger results with increasing amounts of data. It is the only technology that can cope effectively with the huge volumes, the noisiness, and the changing nature of real-world data.

The core idea of machine learning is to first apply structured statistical analysis to a data set to generate a *predictive model* and then to apply this model to different data streams to support different forms of analysis. Some specific examples of how machine learning can be applied to intelligence and security applications include:

- **Language translation:** Translating documents from one human language to another. Modern translation systems create sophisticated statistical models of how the words, phrases, and syntactic structures of one language map to another. These models are generated via a training process operating on bilingual and monolingual corpora containing trillions of words of text. Given new documents, the model is then applied to generate translated versions. This approach has proved far more robust and reliable than more traditional rule-based translators, and it will automatically adapt to new terms and phrases, such as "improvised explosive devices." State-of-the-art systems are still not as good as expert human translators, but they are able to generate translations that capture the major points of a document, and therefore they can be used to filter a collection of documents down to those that should be translated by humans. For some languages, including Arabic, machine translation is moving beyond this "triage" capability to produce semantically reliable outputs. Current systems also work better for "major" languages, having large training corpora but are less effective for minor or tribal languages.
- **Knowledge extraction:** Creating a database of statistically validated facts from unstructured and potentially unreliable sources, such as Internet web pages. Learning algorithms construct and refine these databases by iteratively gathering facts with increasing certainty as more sources are combined. They rely on the property that many facts are stated in multiple locations and so they get statistically reinforced, while false information has a much lower rate of occurrence and is likely to be contradicted. These databases can provide the "real-world" knowledge required for a computer program to achieve true intelligence. Simple applications include allowing search engines to recognize synonyms and paraphrases of search terms, and to disambiguate context-dependent words (e.g., "Apple" can be either a fruit or a computer company).
- **Document summarization:** Extracting the sense of a document, or more interestingly a group of topically-related documents, and establishing the main points of consensus and divergence. This can greatly improve the productivity of humans trying to screen large document collections and enable the tracking of overall trends on what topics are of most importance to people and what their opinions are. Efficient summarization can greatly reduce the volume of information that an analyst must evaluate. Modern summarization methods focus on task-driven summaries, extracting the information of interest to the analyst for preferential inclusion.

[2] http://www.cra.org/ccc/docs/init/From_Data_to_Knowledge_to_Action.pdf.

- **Social network analysis:** Constructing and analyzing graphs of communication and interaction patterns between individuals, possibly on a massive scale. From this work, an analyst can identify collaboration patterns and determine how groups are structured, who is in control, and in what ways they are vulnerable to disruption (motivated by both offensive and defensive concerns). Machine learning algorithms can help the analyst by discovering subgraph patterns that suggest potential new links that may not yet have been observed in the data. Temporal derivatives of social network structures can be analyzed to track evolving relations and power structures.
- **Image/video analysis:** Extracting features from image and video data, including where it originated based on its content, any identifiable text or symbols (signs, license plates, etc.), any faces and their identities, etc. A major research challenge is to provide image search capabilities comparable to what can be done with text.
- **Audio analysis:** Extracting features from audio data, including speech recognition, speaker identification, language identification, and mood identification (e.g., is the person in an Internet video speaking with a hostile voice?).
- **Trend identification:** Detecting, presenting, and validating or refuting patterns of information to determine evolving trends and their nature (e.g., unique, cyclic, etc.), as well as possible causal linkages among trends and supporting evidence.
- **Active learning:** Determining where information is lacking and which data would be most productive to acquire. Machine learning algorithms can identify conditions where the statistical models show a high degree of uncertainty and the decision outcomes have high sensitivity. This information can be used, for example, to determine where best to deploy further satellite surveillance, human assets, or signal intercepts. A recent variant called *proactive learning* weighs the information gathering costs and the expected reliabilities of different information gathering operations, attempting to optimize resource-bounded intelligence gathering.

Several important features of machine learning programs are worth noting. First, the quality of their results keep improving as more data are collected and the generated results are evaluated. For example, the Netflix video subscription service continually improves its ability to predict what movies a customer would like to see based on the ratings it collects from the customer and from millions of other subscribers. It refines its model by evaluating how well its past suggestions matched the subsequent viewer ratings. Second, machine learning algorithms have the unique ability to gain useful insights from millions of pieces of information, each of which has little significance and may in fact be incorrect or misleading. For example, the fact that an unidentified caller phoned from Karachi to London at 9:00 on December 7, 2010 is unlikely to be meaningful. But, if we can track millions of calls being made worldwide, the resulting social network graph can reveal important patterns that could indentify the operations of a terrorist operation.

Technical Challenges

Performing sophisticated data analysis at such a massive scale requires entirely new approaches to computer system design, programming languages, machine learning algorithms, and to the application technology itself. Fortunately, much of the technology base can be adapted from rapid advances being made in the commercial sector and in research laboratories.

- **DISC technology:** Coping with large data volumes and complex analysis tasks requires data management and computing capabilities at massive scales. Creating, programming, and operating such systems require approaches that differ greatly from traditional, high-performance systems. Our earlier paper on Big-Data Computing[3] describes the design of data-intensive scalable computing (DISC) systems targeting applications that are characterized by the large amount of data that must be collected, stored, and analyzed. These systems differ in their fundamental structure and operation from traditional high-performance computing systems. This technology has spread just in the last three years from a handful of Internet companies (notably Google and Yahoo) to a number of "Web 2.0" companies, as well as commercial, governmental, and university research laboratories. This spreading has been greatly enhanced by the availability of the *Apache Hadoop* open source software infrastructure, providing a combined file system and programming support for DISC systems.
- **Scalable machine-learning algorithms:** The entire field of machine learning is very young and evolving rapidly. Many of the classical algorithms require time proportional to the square of the number of data elements, which might be feasible for a data set with one million elements, requring trillions of computations, but not for data sets much beyond that level. New algorithms exploit regularities, sampling methods, sparsification (via principal components, kernels, etc.) and other techniques to obtain performance levels closer to linear complexity, but much remains to be done to cope with massive heterogeneous data sources with billions or trillions of elements. Principal among these is to make effective use of parallel computing resources, requiring new ways of formulating machine learning problems, algorithms, and complex indexing structures.
- **Trustworthiness:** Data sources vary with respect to their reliability. For instance, much hasty "news" is later retracted, modified or worse yet left uncorrected in light of later information. Dynamic aerial or satellite surveillance is subject to weather and adversarial obfuscation. Sensors may malfunction or de-calibrate. Translations may be erroneous. Social networks may lack crucial links. Misinformation may be inserted into a data stream. A new direction in machine learning research involves *learning under uncertainty*, combining potentially confirmatory or contradictory evidence from multiple sources, and updating estimates of source reliability, conditioned on exogenous factors (e.g., weather, sensor quality, etc.).
- **Scalable application technologies:** Similarly, many of the technologies needed for intelligence applications (e.g., language translation, document analysis, image processing, pattern detection, cross-source analysis, network analysis, etc.) are well studied and have undergone numerous refinements, but the notions of what is feasible and what are the best approaches are being radically revised due to the availability of new forms of data, new computational platforms, and new approaches based on statistical machine learning. For example, in 2005, a team from Google achieved major breakthroughs in machine translation when it used a 1000-processor cluster to train its statistical language models with over one trillion words of text, almost two orders of magnitude beyond what had been attempted in the past, and as a result achieved the top score in a NIST evaluation.
- **Cross correlation and information fusion:** Intelligence operations have traditionally partitioned their efforts according to the types of data sources being analyzed. Surveillance satellites are operated by one organization (NGA), while communications

[3] http://www.cra.org/ccc/docs/init/Big_Data.pdf.

surveillance is performed by another (NSA). Much can be gained by combining many forms of information to get a coherent assessment of a situation. For example, assisting soldiers in countering a group of insurgents could involve simultaneously analyzing images being captured by helmet- and vehicle-mounted cameras with the patterns of communication observed by intercepting cellphone communications, after applying automated translation and summarization to these communications. Detecting a pending terrorist attack might involve simultaneously monitoring text and video postings on websites, newsfeeds from Al Jazeera, and both intercepted email and phone calls, as well as correlating this information with known or inferred terrorist social networks. Cross-source, cross-media and cross-agency detection of emerging patterns, possibly indicative of incipient threats requires information fusion technology, whose enablers are common meta-data creation, machine learning in the large, and trend detection technologies.

- **Information and identity hiding:** The ability of data mining to extract subtle features from data can reveal information that was intended to be hidden, especially by cross-correlating multiple data sets. For example, researchers at Carnegie Mellon were able to determine the social security numbers for a number of individuals using combinations of death records, voter registration information, and other publicly available information. Researchers at the University of Texas were able to identify individual Netflix customers from a data set that Netflix had attempted to anonymize and release as part of a research challenge, by correlating ratings with ones posted in the Internet Movie Database. These technologies can be invaluable when they are used to uncover a criminal or terrorist, but they can also be used by our adversaries to identify covert sources and methods, or by identity thieves to extract information about ordinary citizens from supposedly anonymized, publicly available data sets. Better methods for anonymizing data in the face of cross correlation, and better methods for assessing the vulnerability of information to discovery is of crucial importance as more data become available online and our methods of analyzing data become more sophisticated.
- **Maximizing human effectiveness:** Much more work needs to be done on how to present information to human analysts in ways that allow them to understand subtle patterns and grasp the degrees of certainty to which the data support different possible conclusions. Furthermore, there is a great opportunity to develop *mixed-initiative* learning approaches, in which the computer and human analyst explore the data interactively, combining the unique abilities of each (e.g., the ability of the computer to discover statistical regularities over vast data sets, and the ability of the human to identify which of the many data regularities are truly important, and to suggest follow up hypotheses). Many aspects of social networking technology can be applied to supporting group collaboration among intelligence analysts. One unique aspect would be the need to maintain different forms of trust and confidentiality across organizations and possibly with intelligence services in other countries.

The talent challenge

Intelligence agencies and the companies that produce the analytic tools they use face a major challenge in recruiting people with the necessary training and talent in large-scale machine learning. Being productive in this area requires a strong background in algorithms, statistics, computer systems, databases, and parallel computing. Most people with traditional computer

science backgrounds lack the grounding in mathematics and statistics needed to understand the capabilities, limitations, and best ways to use different machine learning methods. On the other hand, those with backgrounds in statistics and mathematics are unfamiliar with the programming and data management techniques required to work with massive data sets. Many of the core algorithms have only been invented in the last decade, and others are being invented and refined every week. Commercial data mining tools can be characterized as well-designed tools encapsulating yesterday's algorithms—they augment but do not substitute for the high levels of expertise required to make use of scalable machine learning.

Universities are only now starting to generate a sizable number of students who are well trained to work in this area. Very few of these, however, end up with jobs that support intelligence activities. Consider the case of computer science Ph.D.s. Of the 1,877 degree recipients in 2008, 978 of them – over 50 percent – were foreign nationals, generally excluding them from classified work. Graduates with training in machine learning were heavily recruited by Internet companies, by Wall Street hedge funds, and by Web 2.0 startups. (After all, machine learning is one of the most important recent technological innovations in computer science, and so the demand for talent far outstrips the supply.) Whereas the talent pipeline from U.S. universities to defense-related industries and government agencies was very strong during the Cold War, it is much reduced today. Due to the shift of funding for university research away from defense agencies, such as the Defense Advanced Research Projects Agency (DARPA) (whose funding has significantly shifted from academia to the private sector), graduate students are increasingly funded by the National Science Foundation (NSF) and other sources and therefore are much less likely to be exposed to the needs and career opportunities in intelligence- or defense-related industries and agencies.

Suggestions for Federal investment

The intelligence community stands to gain much by adopting large-scale machine learning to process and analyze the wealth of available information sources. By fusing many forms of information, and by processing data at a massive scale, insights can be gained that would be missed by more traditional approaches. Automated methods can vastly improve the productivity and vigilance of human analysts in coping with the ever-increasing amount and complexity of information that must be evaluated in a timely manner.

There is a strong overlap in the core technology used by both the private sector and the intelligence community in this area. On one hand, that puts intelligence organizations at a disadvantage in recruiting talent, but it also means that they can leverage the research and technology development taking place in U.S. universities and industry. One important property of the technology described here is that it can generally be developed and tested on unclassified data sets and then adapted to meet the needs of the intelligence community. Data sources such as commercial satellite images, images downloaded from photo sharing websites, blog posts, news feeds, social networks, and documents retrieved from the Internet can serve as suitable proxies for classified information. The intelligence community should also recognize that funding university research has the beneficial side effect of providing an opportunity to develop relationships with faculty and students across many institutions, some of whom would then be more inclined to pursue careers in support of intelligence organizations.

Specific suggestions include the following:

- Invest in unclassified research on intelligence-relevant applications of machine learning (language translation, speech recognition, image analysis, pattern detection from heterogeneous sources, active and proactive learning, learning under uncertainty, etc.)
- Invest in unclassified research on the supporting technology: DISC systems, data-oriented programming languages, large-scale databases, scalable machine learning algorithms, machine translation, text and image mining, reliable anonymization, and relevant aspects of human-computer interaction.
- Provide surrogate data for large-scale learning and analytics from multiple sources. This work may entail pre-negotiating with providers to make the data available to researchers, much as the Linguistic Data Consortium (LDC) currently does for language resources.
- View these investments not just as ways to acquire technology, but also to develop relationships with student and faculty as a way to improve the pipeline of talent into companies and agencies supporting intelligence operations.

The intelligence community has already developed mechanisms for engaging with university researchers, including funding through Intelligence Advanced Research Projects Activity (IARPA), by channeling funding through agencies such as NSF and DARPA, and by creating research centers in partnership with universities. Much of this work can be done in unclassified environments and with unclassified data, but with agencies providing guidance based on experimentation with actual data, and by having some of the key researchers participate in classified briefings. This approach to research is perhaps less comfortable than simply working within a classified environment, but it will lead to more creative and up-to-date approaches, tapping into the rapid innovations taking place in both university and industrial research organizations.

Summary

Large-scale machine learning has the potential to fundamentally accelerate our ability to extract important insights relevant to our nation's security from the vast amounts of information being generated and collected worldwide. Rather than being overwhelmed by these data volumes, machine learning has the advantage that "more is better." This principle has been well established by Internet companies, and it is certainly applicable to intelligence analysis in everything from text mining and translation to obfuscated pattern detection and network analysis. The intelligence community can only realize these possibilities, however, by taking maximum advantage of unclassified work taking place in universities and industry, and by using its resources to develop and tap the necessary pool of talent. Doing so is critical to safeguarding the interests of the U.S. in the twenty-first century.