On the Efficacy of Data Mining for Security Applications

Ted E. Senator
SAIC[1]
3811 N. Fairfax Drive, Suite 850
Arlington, VA 22203, USA
1-703-469-3422
senatort@saic.com

ABSTRACT
Data mining applications for security have been proposed, developed, used, and criticized frequently in the recent past. This paper examines several of the more common criticisms and analyzes some factors that bear on whether the criticisms are valid and/or can be overcome by appropriate design and use of the data mining application.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – *data mining,* I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation, feature evaluation and selection, pattern analysis.* I.6.3 [**Simulation and Modeling**]: Applications. J.7.2 [**Computers in Other Systems**] K.4.1 [**Computers and Society**]: Public Policy Issues – human safety, privacy, use/abuse of power.

General Terms
Management, Measurement, Design, Economics, Security, Legal Aspects.

Keywords
Data mining, security, applications, pattern matching.

1. INTRODUCTION
Many data mining applications for solving security problems have been proposed and designed in the recent past, especially since September 11, 2001. [12] Some of these applications have been developed and deployed and are in use today. Others have been cancelled before they were deployed. [16] The reasons for cancellation have typically been because of concerns over effectiveness and/or concerns over societal impact, particularly with respect to privacy and civil liberties. Sometimes, the concerns have been expressed as a tradeoff between the benefits in terms of the amount of protection afforded and the costs in terms of the societal impact of using the system. Some critics of the use of data mining for security applications have simultaneously criticized systems both for being ineffective and for threatening civil liberties. Often – especially in the political and policy communities – the discussion of these issues is based on less than thorough analyses of the way these systems actually operate or the way they could operate if they were designed effectively.

This paper analyzes several of the criticisms directed at the effectiveness of various data mining applications for security by (1) clearly defining the distinct activities that are performed as part of data mining projects for security applications and (2) analyzing the criticisms in the context of these activities, while proposing alternative designs to those assumed by the critics. This paper does not address the vital issues of privacy and civil liberties that are raised by these systems, not because these issues are not important – they most certainly are fundamental considerations in the design and adoption of any system involving data mining for security – but simply because that is a separate topic that requires far more thorough analyses and discussion than has been addressed in the research and analysis reported here.

The main purpose of this paper is not to argue for any particular position with respect to the societal costs and benefits of using data mining for security applications; rather, it is to suggest ideas that would be part of a more thorough and principled framework within which to understand the inherent design issues, impacts, tradeoffs, and possibilities, in the hope that such a framework and understanding can be used to support rational and informed societal choices leading to effective security systems that respect privacy and civil liberties. This paper is offered in the spirit of [2] – to contribute to informed public debates and sound policy making that provide appropriate security and maintain civil liberties informed by careful analyses of alternatives and possibilities. It is hoped that the discussion and analysis in this paper will provide more of a mutual understanding between the technical and policy communities, at least in terms of the ability to communicate, discuss, and debate issues with a common understanding of what different terms mean and alternative solutions may imply.

The paper is based on actual and proposed data mining applications in the U.S. with which the author is familiar;

CSI-KDD'09, June 28, 2009, Paris, France.

[1] Affiliation is provided solely for identification purposes. The views and opinions expressed herein are not necessarily those of SAIC, any of is clients, or of any other organization with which the author has been or is affiliated. This work has been prepared independently of the author's duties and responsibilities as an employee of SAIC.

however, the general ideas discussed should apply equally well to non-U.S. applications. What may differ across countries is not the scientific and engineering principles on which such systems are based, but rather the values on which societal judgments are based and the corresponding legal, political, regulatory, and policy environments in which decisions are made regarding the benefits and costs of such systems.

The paper begins with a review of various definitions of data mining. It identifies several distinct but related activities that fall within these definitions. Next, it defines a model of data mining systems for security applications. This model forms the basis for the analysis that follows. The model is used to analyze several criticisms that have been levied at security applications of data mining. The paper then discusses metrics for the evaluation of security applications of data mining and finally concludes.

2. WHAT IS A SECURITY APPLICATION OF DATA MINING?

This section discusses what is meant by an application of data mining for security. As will be seen from the various examples that are cited, there is often much confusion about this very issue, and this confusion is a large contributor to misunderstandings about the effectiveness of data mining applications. The section begins with a comparison of definitions of data mining as used in the technical and the political/policy communities. It continues with a discussion of what data miners do and suggests a framework and terminology for distinct but related tasks. It then uses this framework to understand security applications of data mining and concludes with a discussion of the sources and role of patterns in security applications.

2.1 Definitions of Data Mining

There are a number of well-accepted definitions of data mining in the scientific community. Most of them center on the idea of pattern discovery. The most widely used definition, from [3] is that data mining is "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data." A newer definition from Jonas and Harper [5] defines data mining as "the process of searching data for previously unknown patterns and often using these patterns to predict future outcomes." Note how these definitions, cited in the same order as they were proposed, both focus on the discovery of patterns while the new definition adds the emphasis on the use of these discovered patterns, especially for prediction.[2] Note also how neither of these definitions mentions data collection, data aggregation or linking, or particular applications.

In contrast to the definitions used in the scientific community, politicians have defined data mining both more broadly and more narrowly. These definitions are broader in so far as they include search and collection of data, and they are narrower in that they typically refer to security applications the purpose of which is to prevent terrorism. "Data Mining is a broad search of public and non-public databases in the absence of a particularized suspicion about a person, place or thing. Data mining looks for relations between things and people without any regard for particularized suspicion" according to U.S. Senator Russ Feingold on January 16, 2003. The U.S. Department of Defense Technology and Privacy Advisory Committee in March 2004 defined data mining as "searches of one or more electronic databases of information concerning U.S. person by or on behalf of an agency or employee of the government." Senator Feingold's proposed amendment to HR 5441 defined data mining as "a query or search or other analysis of 1 or more electronic databases, whereas – (A) at least 1 of the databases was obtained from or remains under the control of a non-Federal entity, or the information was acquired initially by another department or agency of the Federal Government for purposes other than intelligence or law enforcement; (B) a department or agency of the Federal Government or a non-Federal entity acting on behalf of the Federal Government is conducting the query or search or other analysis to find a predictive pattern indicating terrorist or criminal activity; and (C) the search does not use a specific individual's personal identifiers to acquire information concerning that individual." Senator Patrick Leahy, opening the Senate Judiciary Committee Hearing on the "Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs" on January 10, 2007, defined data mining as "the collection and monitoring of large volumes of sensitive personal data to identify patterns or relationships."

It is important to note how these definitions of data mining differ from the scientific definitions. First, they assume a particular purpose, namely, security applications. Second, they include the concepts of data collection, monitoring, and search. And third, they assume that pattern-based searches are being conducted to identify specific individuals who fit the patterns. The focus is not so much on pattern discovery, but rather on pattern matching; the end product is not the patterns themselves as in the scientific definitions, but rather the matches of the patterns to people (and perhaps also places, things, events, etc.) to predict something of interest having to do with security. Note that neither of these sets of definitions covers the primary activity of data mining researchers, i.e., developing new algorithms for pattern discovery.

In the U.S., the *Federal Agency Data Mining Reporting Act of 2007* ("Data Mining Reporting Act") requires the "head of each department or agency of the Federal Government" that is engaged in activities defined as "data mining" to report annually on such activities to Congress. The Data Mining Reporting Act defines data mining as "a program involving pattern-based queries, searches, or analyses of 1 or more electronic databases" in order to "discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity." According to [7] and [8], "the limitation to predictive 'pattern-based' data mining is significant because analysis performed … for counterterrorism and similar purposes is often performed using various types of link analysis tools. These tools start with a known or suspected terrorist or other subject of foreign intelligence interest and use various methods to uncover links between that known subject and potential associates or other persons with whom that subject is or has been in contact. The Data Mining Reporting Act does not include such analyses within its definition of 'data mining' because such analyses are not 'pattern-based.' Rather, these

[2] Prediction is actually used in two senses here. Prediction can mean "to infer some value about another entity in the database," or it can mean "to suggest something that might occur in the future based on an analysis of the past." As the physicist Niels Bohr said, "Prediction is very difficult, especially about the future."

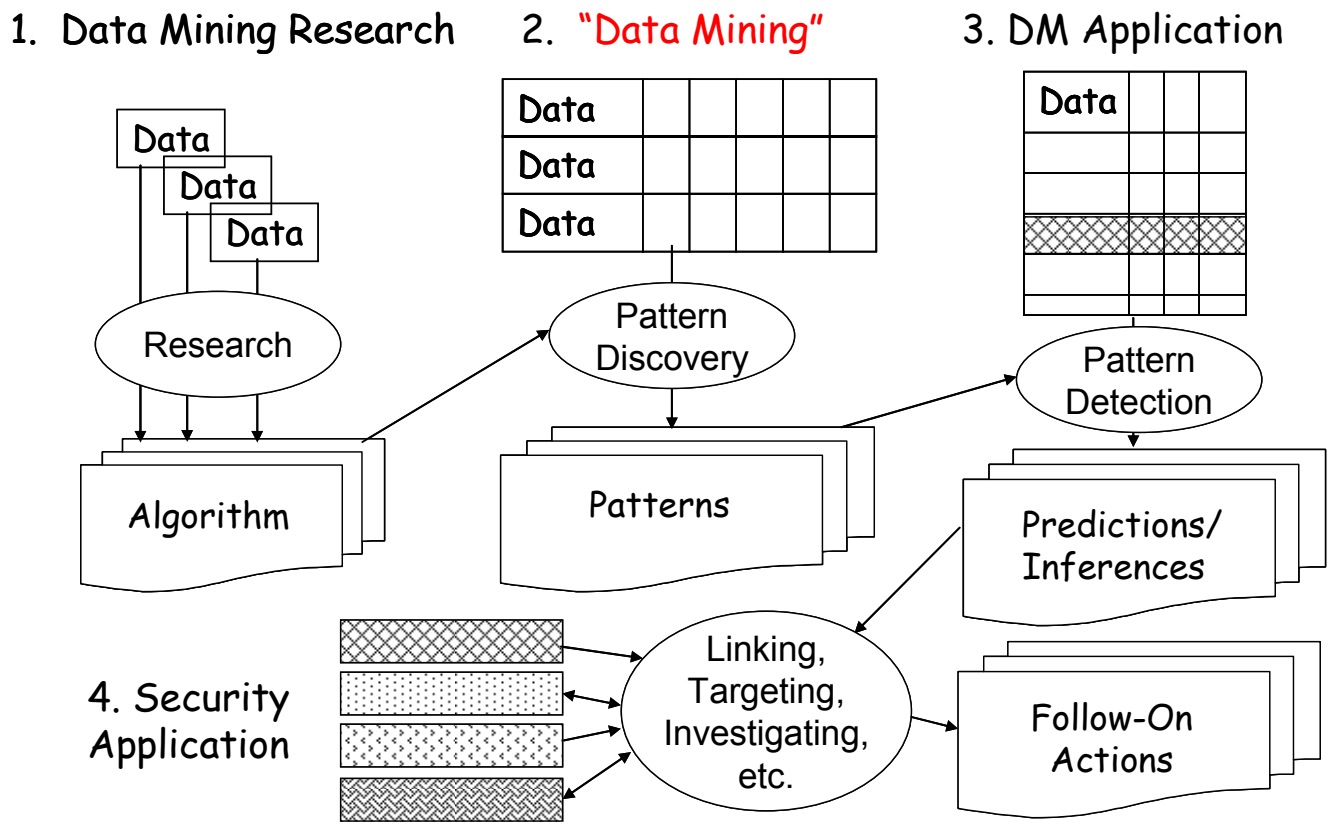

Figure 1. Data Mining Activities

analyses rely on inputting the 'personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals,' which is excluded from the definition of the act."

2.2 What Data Miners Do

Based on the above definitions, it appears that data miners engage in three distinct but related activities: (1) data mining research, the primary focus of which is algorithm development, (2) data mining itself, whose primary focus is pattern discovery, and (3) data mining applications, whose primary focus is predicting or inferring the value of a feature for some purpose. (In the case of security applications of data mining, the feature is typically a likelihood that a particular person is high-risk.) Finally, the data mining application developer may also engage in a fourth activity: (4) the design and development of other aspects of an end-to-end system that makes use of the predicted feature for some particular purpose. This end-to-end system can involve a variety of data sources and analytical and investigatory techniques; may result in many alternative downstream analyses, decisions and actions; and typically involves human analysts and other actors. Figure 2 of [15] provides an example of how this end-to-end process may occur in the context of law enforcement investigations. (It is important to note that the result of a positive match to a pattern that may be indicative of increased risk is usually and appropriately a more thorough analysis by a human analyst; rarely if ever is a pattern match relied on for any consequential action, nor should it be.)

Each of the activities performed by data miners has distinct data requirements and distinct products, as depicted in figure 1. The figure is intended to depict not only the distinct activities, but also the different data needs and uses for each activity. Data mining researchers typically identify a set of databases that have a characteristic that has not been previously exploited for effective pattern discovery. They acquire several – or as many are as readily available – databases that share this characteristic and develop an algorithm that takes advantage of this characteristic and results in the discovery of more effective patterns. (Note that the effectiveness of the patterns is defined with respect to some particular application task.) The databases the data mining researchers use need not have information that identifies the entities in the databases, although they often do need to maintain unique identifiers for some classes of patterns. The databases need not be complete or even close to complete with respect to the actual populations, although some degree of representativeness is highly desirable. Multiple databases that are about a diverse set of domains are strongly preferred in order to demonstrate the widespread applicability and utility of the newly developed algorithm.

The second activity, the actual mining of data, uses various algorithms to develop models, or, synonymously, to discover patterns. This step is sometimes called knowledge discovery, and the resulting patterns or models are referred to as knowledge. This activity typically is based on a single database, or at least a single

"virtual database" in so far as the analysis is concerned.[3] All fields of a database are relevant here, as the purpose of pattern discovery is to determine which of the fields are relevant and which are not for the detection of the phenomena being modeled; however, it does not necessarily require records referring to all elements of the population, just a large enough sample with enough examples of the phenomena of interest. This activity of data mining may more generally be termed "data analysis" – it could use, for example, statistical or other techniques. The result of this mining of data is a set of patterns that have predictive value. This is the activity that conforms to the widely accepted definition of data mining in the technical community.

The third activity in which data miners engage is the actual prediction itself. Predictions are made using the patterns discovered by the second activity on new data elements that had not been used in the pattern discovery. This activity would typically be widely applied to all members of the population of interest, but would require only those fields or attributes that have been determined to be relevant during the actual data mining analysis. Each record in the database is matched against the pattern and an inference or prediction is made. These inferences or predictions may then be used, typically in conjunction with additional information that has been collected based on knowing the identity of the individual whose particular records matched the pattern, to make a further determination of interest and take appropriate actions, outside the scope of but resulting from the data mining application. An example of such inference might be the assignment of a credit score to a new applicant, based on a match to patterns that were discovered during the second activity and were determined to be useful in predicting credit risk.

The fourth activity type is the actual end-to-end security application. This activity is not performed by scientists or engineers, but by some organization with an operational mission. The organization may be responsible for screening applicants for some purpose, or in a non-security context it might be responsible for marketing a particular product or service. Such an organization does not care about the source of the knowledge used in its systems; it cares simply about their effectiveness. This knowledge may be the result of patterns discovered by data mining, or it may arise from other sources, such as a deep understanding of the domain, a formal set of regulations, etc. Often, these domain-specific applications do not incorporate any data mining algorithms or any patterns that were discovered using such algorithms; why they do not is an issue both for application developers, who could potentially build more effective systems by taking advantage of data mining techniques and results, and for data mining researchers, who could potentially provide more useful technology for real applications.

The next section of this paper explores end-to-end security applications in more depth.

[3] Note that for purposes of simplicity, we ignore the field of distributed data mining. It is recognized that distributed data mining techniques can be used to discover local patterns that can later be merged; the relevant question in this field is not how databases can be split vertically for pattern discovery, but rather how they can be split horizontally and the patterns combined.

Before continuing the discussion, however, two other points are worth mentioning. First, while this discussion of data mining activities is depicted in terms of propositional data, the basic ideas apply to relational data as well. Second, we note that these four distinct activities are often conflated not only by policymakers, end users, and other stakeholders, but also by data miners themselves. In particular, data mining researchers tend to view every activity other than data mining research as "applications," while those responsible for end-user applications tend to group activities one and two as "research." [4]

2.3 Security Applications of Data Mining

With these four activities in mind, we see that what is typically referred to as a security application of data mining may combine aspects of several activities but usually emphasizes some combination of the third and fourth – the detection of entities in the database that match particular patterns of interest and their use in an end-to-end application process for evaluating risk, initiating and/or conducting investigations, and taking appropriate actions. Such an application could involve no automated pattern matching at all, or it could be totally dependent on such automated pattern matching. Many real applications, such as the one described in [14], combine these aspects. While the use of the application for its intended purpose may not include any algorithm development or pattern discovery, its development process may benefit from these activities. The application may also include such activities to enable continuous updating of the patterns to account for changes in behavior. It may also occasionally take advantage of the first activity, if improved algorithms can result in the discovery of more effective patterns.

Key issues in the design of such applications are (1) what is the purpose of the application?, (2) what data sources are available, appropriate, and useful for the intended purpose?, (3) what techniques will best accomplish these purposes with the available data and patterns?, (4) what additional justifications are required to acquire additional data?, (5) what records are kept after an analysis is performed?, and (6) what follow-on actions are allowed as a result of the application? The purpose of the application is determined by some need, having to do with an organization's mission and independent of the consideration of any use of data mining. The application may use data from which useful patterns have been mined, or it may not, depending on whether such sources and patterns exist, whether it is appropriate to use such sources for the intended purpose, and whether other sources are more useful for the intended purpose. In the context of security applications, data that have been collected for security purposes (e.g., existing law enforcement and intelligence databases) are often useful and appropriate sources; other data sources such as commercial transactions are often neither useful nor appropriate. The selection bias that results in inclusion in such a security database in the first place may, in fact, be viewed as a prima facie indicator of a high-risk entity; this is why techniques that start from known risky individuals and "connect the dots" are often most effective for security applications. Other pre-screening techniques, such as observations of suspicious

[4] In fact, this confusion often manifests itself not only in policy debates, but also in acceptance criteria for data mining conferences and journals.

behavior or setting off alarms for carrying potentially dangerous material provide a similar selection bias that may be at least as effective as more abstract pattern matching for purposes of a particular security application. Techniques that are useful for security applications may include pattern matching, link analysis, anomaly detection, and others; often, some combination of these techniques is most effective. Often the security application includes additional data collection about entities for whom more information is justified based on initial indicia of risk or suspicion; this additional data collection may involve additional entities who are somehow "connected" to known entities or it may involve collection of additional types of information through a subject-based query on a known entity to enable an accurate determination of that entity's status. (Such additional collection based on subject-based queries is depicted by the bi-directional arrows in activity 4 in figure 1.) The application may store or discard various intermediate results about a particular entity; these data retention issues are crucial because of the potential long-term effect on an individual. If an individual is determined to be low-risk only after an extensive analysis of additional data, pertinent data about that individual could be discarded at a cost of having to repeat the analysis in the future; however, if such data are retained, then it would be essential to prevent the use of such additional data for any purpose other than avoiding a more detailed analysis of such an individual in the future. Finally, the application exists in the context of some business process and some set of authorizations and authorities, both of which determine what follow-on actions may result from use of the application. It is typically this last issue that is of most societal concern, for this is when consequences can occur. It is important to note that these consequences are not the direct result of the use of data mining to discover patterns; rather, they are the result of policies and procedures that are adopted by a user organization with regard to the results of a security application.

A key feature of all security applications is that they are multi-stage processes. Each stage passes along the riskiest entities to the subsequent stage for more detailed analysis and discards the low-risk entities. The more detailed analysis may incorporate additional data sources – data sources that are more expensive to obtain or data sources the use of which is restricted until some additional justification exists. The more detailed analysis may, and often does, result in a conclusion that the entity under consideration is, in fact, not as risky as determined by the previous stage and, therefore, cause the entity to be removed from the risky category.

A crucial issue in the design of any security application is the source and role of patterns. Pattern detection may be an effective technique, especially when applied to existing law enforcement and intelligence data and used to detect low-level activities and combine such low-level activities into higher-level plans or organizations. It may be less useful when applied to screening of individuals. An important point is that the utility of any pattern must be established and verified empirically before such a pattern is used as a component of a security application. Patterns may come from data mining, but also from other sources. For example, patterns used in [13] resulted from an analysis of market regulations and hypotheses about possible schemes to engage in improper market behavior, and these patterns were deployed only after a rigorous and iterative process of modification and validation. Patterns may also come from external sources; for example, knowledge of a newly confirmed or suspected adversarial technique could result in the development and use of a pattern for its detection. Patterns may also arise from anomaly detection techniques; in this case, normal patterns of activity are removed from the data and what remains is considered unusual and potentially suspicious.

We consider two examples of security applications to illustrate these ideas. First, imagine an application for screening passengers at airports. In such an application, there is no a priori reason to suspect that people who choose to fly on airplanes are more likely than the general population to be dangerous. Rather, the concern is with the possibility of any particular, dangerous person being aboard an aircraft. In such an application, initial screening might be, and often is, based on a physical inspection, close behavioral observations, detailed questioning (in the case of El Al Israel Airlines), or some combination, rather than on pattern matching. Additional data might be considered for people who somehow appear suspicious on one of these tests. This situation contrasts with an application that might be used to determine where to focus investigatory resources based on people who appear in lawfully collected intelligence databases. In such an intelligence application, the initial indicia of risk come directly from the fact that a person is included in the database itself; hence, pattern-based analysis is likely to be a useful tool.

Finally, it must be noted that a security application will almost always and should always include strict audit functions, controls on use, and review mechanisms to ensure that the application is being used solely for its intended purpose and is not being abused in any way. In fact, data mining techniques independently applied to the audit logs are themselves one method to detect, deter, and guard against abuses of security applications themselves.

3. CRITICISMS OF SECURITY APPLICATIONS OF DATA MINING

Security applications of data mining that have received the most criticism include Total Information Awareness (TIA), Computer-Assisted Passenger Prescreening System (CAPPS II), Multistate Anti-Terrorism Information Exchange (MATRIX). [12] These systems/projects were all cancelled after expenditures of millions of U.S. dollars., because of concerns both about privacy and civil liberties and about their effectiveness. [16] Secure Flight, a follow-on to CAPPS-II, and the Department of Homeland Security's Analysis, Dissemination Visualization Insight and Semantic Enhancement (ADVISE) System were also cancelled due to security vulnerabilities and privacy concerns, respectively. [16] Even research programs that incorporated a full measure of privacy protection and had the sole purpose of determining whether a particular algorithm, technique or approach could develop patterns that indicate terrorist activity were reported, although they did not meet the requirements of the Data Mining Reporting Act [7], and were later cancelled according to [8]. These research programs were an example of the first activity depicted in figure 1.

Criticisms of the effectiveness of data mining security applications appear in [1], [5], [9], [10], and [11]. Analyses of some of these criticisms are contained in [4], [6], [14], and [15]. This section of this paper summarizes the criticisms and analyzes how they may be addressed in security applications of data

mining, using the model presented in section 2 as the basis for distinguishing separate, and different activities.

3.1 Too Many False Positives

The simplest criticism of security applications of data mining is frequently expressed as "too many false positives." In particular, while it is accurately noted that for events that occur far less frequently than the accuracy of a classifier, most instances of positive results will be false positives. This criticism is addressed in detail in [14]; a multi-stage classification architecture preceded by a high-risk population selection and followed by link analysis is shown to be one method of mitigating this problem of too many false positives. The example of a 99.9 percent accurate classifier applied to a population of 300 million entities containing only 3,000 true positives, i.e., 0.001 percent, would yield over 100 times more false positives by itself. However, with multi-stage classification techniques consisting of two independent stages at 99 percent and 99.9 percent accuracy and assuming 5 percent of the population in a high-risk group that was 10 times more likely to be positive, almost all groups of reasonable size would be detected. The "false-positive" criticism is also addressed in [4] in the context of relational data, ranking classifiers, and multi-pass inference.

The flaw in the false-positive criticism is that it assumes a single-stage classifier. As should be clear from the discussion in section 2.3 of this paper, no serious security application would be this simplistic, if only because any credible application designer would be aware that such an approach would not work. An initial classifier *might* be used in some applications as a first level screener to manage a large workload; such a classifier would have to be tuned to minimize false negatives. The application would rely on subsequent stages to rule out false positives. These subsequent stages would likely employ a combination of techniques.

3.2 Nobody Does That Anymore

This criticism suggests that matching known patterns is not useful. The flaw in this criticism is that even past threats are still dangerous if they can be executed again. It is important to prevent instances of known attack patterns, or they can be reused by the attackers. There is no a priori reason to assume that a known attack pattern will not be reused; in fact, if something is successful, human nature suggests trying it again. Active detection of indicators of past patterns – and publicizing the ability to do so, although not the details of how it is done – will not only detect such patterns, but also deter individuals from trying them again. This is why we still have to take off our shoes at airports even though there have not been any publicized accounts of attempted use of shoe bombs in quite a while; if we did not have to have our shoes inspected, then shoe bombing might return as it is a proven and low-cost attack method. Further, there may be many potential adversaries who are capable of executing only a single type of attack; preventing them from using that method removes them from the potential population of adversaries. And finally, detecting known attack patterns prevents potential increases in the population capable of using and motivated to use that attack type by avoiding the possibility of copycat attacks. In the context of section 2.3, this criticism relates to the choice of patterns to use in the security application. These patterns are easy to specify precisely because they are known, and therefore, pattern detection is a useful technique for this security measure.

3.3 It Won't Be Perfect

Some systems are criticized because they will not be perfect – there is no way at an acceptable cost to prevent *all* potential attacks. This criticism is often explained in terms of the cost of a false negative – if even one terrorist attack occurs because it is not detected, the cost to society would be astronomical. (This situation is frequently contrasted with the cost of a false negative in a marketing or fraud detection application, in which case, the right thing to do is minimize the cost across a large number of cases, in contrast to security applications where the goal is to prevent all false negatives.) What this criticism ignores is the fact that *no* system is or can ever be perfect; rather, the goal is to maximize effectiveness at a fixed or minimal cost (in terms of effort to develop and use the system, in terms of disruptions to normal functions, and in terms of the impact on privacy and civil liberties). Comparing alternative resource allocations to maximize effectiveness is the subject of [6]. The right question to ask is not "Is this system perfect?" but rather "How does this system increase our overall security in the context of all our other systems?" An effective security application will be part of a layered defense that uses a multitude of techniques with uncorrelated errors; such a design will be most effective at providing maximum security for a fixed resource allocation.

3.4 It Will Just Make Them Try Something Else

Many bad guys are intelligent adversaries. They can be very creative in creating attacks of different types. This criticism typically suggests that there is no point in preventing one type of attack because another equally costly attack can easily be devised and substituted. However, this criticism ignores the fact that not every bad guy is capable of creating a new attack method. Preventing known attacks forces adversaries to spend time developing new attack methods, acquiring new capabilities and resources, and training new attackers. This prevention of known attack types, therefore, has a real cost to adversaries. And not doing so would have a huge cost in morale to those being attacked repeatedly by the same methods with no effective response. One technique for increasing security in the face of potential new attack types includes red-teaming potential new attack types and incorporating such patterns in the security application. A second, related technique is to use patterns corresponding to variants of known attack types, based on the assumption that variants of previous attacks are likely to be tried by an intelligent adversary because they involve minimal change. A third technique to detect new attack types is to decompose the known attacks into required constituent activities, and then create new patterns based on novel recombinations of these lower-level activities.

3.5 It Will Make Them Try Something More Complicated and Serious

This criticism suggests that prevention of low-consequence attacks will result in more devastating attacks as adversaries creatively invent new methods. This is a variant of the previously discussed criticism – not only will prevention of some attack types cause other attack types to be used, but the new attack types

will be more serious than those that have been prevented. What this criticism ignores is that more serious attacks are typically more complicated; they require far more planning, capabilities, training, and resources that less serious simpler attacks. This additional complexity typically involves a longer time to plan and prepare for the attack, the involvement of more people in the plan, and, perhaps most important, more interactions with non-conspirators. All of these factors make it easier to detect the more complicated and serious attack before it is executed – only one starting point is needed and many more are available. Not only is there a cost of something new, but there is an additional cost of something more complicated.

3.6 It Will Make Them Try Something New That You Haven't Thought Of

A further criticism is that effective detection of known attack patterns ignores detection of new attack patterns that have not yet been conceptualized by those responsible for security applications. This criticism is countered by several observations: (1) that even novel attack patterns involve low level activities that arouse suspicion (think of the flight training prior to 9/11), (2) that starting from known subjects can lead to other bad guys (this is the essence of link analysis), and (3) that novel attacks are difficult and expensive to devise. It is this last observation that is key – by detecting previously used attack patterns, those responsible for security are forcing the bad guys to adapt constantly. Every attack they try is new, and is being tried for the first time. This greatly increases the probability that an attack will not be successful – who gets everything right on the first try? Forcing their adversaries to invent more complicated and novel attacks makes their tasks as difficult as possible. It also forces adversaries to test components of a new attack, which converts these component activities from novel actions to repeated ones and makes them amenable to detection techniques that rely on the use of automated pattern discovery to detect repeated sequences of related activities.

3.7 There's Not Enough Training Data

Often, data mining applications for security are compared to applications in credit card fraud detection. The discussion typically has an advocate for data mining applications who cites the high effectiveness in real time of scoring credit card transactions and a critic who points out that that there are a multitude of examples from which a system can learn the indicators or patterns of fraud in credit cards compared to few examples of terrorist attacks.[5] Jonas and Harper [5] make this argument quite effectively, pointing out that there are "a relatively small number of attempts every year and only one or two major terrorist incidents every few years – each one distinct in terms of planning and execution – that there are no meaningful patterns that show what behavior indicates planning or preparation for terrorism."

There are certainly and fortunately a small number of examples of successful terrorist attacks and known disrupted attacks and, presumably, a larger but not extremely large number of unknown disrupted attacks, but nowhere near the amount needed for statistically valid pattern discovery. However, as in many types of fraud detection applications, the components of such attacks are similar. They all involve financing, acquisition of material, recruitment of participants, communication between the participants, etc. While these activities occur frequently and predominantly for legitimate reasons, when combined in particular contexts, they can potentially provide enough cause for further information collection and analysis, enabling the type of link analysis that Jonas and Harper advocate. Improvements in data mining algorithms that would enable the learning of usefully discriminative patterns from minimal training data is a challenge for the research community; while pattern-based data mining may be inadequate at present and even for the foreseeable future, new techniques may prove to be useful at some point in the future. Before they would be deployed or even considered for inclusion in a security application, such techniques would have to be subject to a rigorous cost-benefit analysis, including considerations of data use and privacy implications. In all likelihood, such techniques would be useful only in combination with link-analysis techniques, referred to as "subject-based data analysis" and contrasted with "pattern-based data analysis" by Jonas and Harper. As the first step in a mass screening system, such predictive data mining is unlikely to be useful for the reasons pointed out by Jonas and Harper. However, despite their use of the term "predictive data mining" to describe what would be ineffective, they really are arguing against only a particular design choice rather than against the entire set of data mining techniques described in section 2 of this paper.

3.8 They Can Reverse Engineer the System

The Carnival Booth algorithm has been proposed as a way that bad guys can reverse engineer a security system. [1] This is a serious criticism, and it deserves a serious and thorough analysis.

The Carnival Booth algorithm is developed and analyzed in the context of the CAPPS system. The conclusion is that selecting individuals for increased scrutiny can actually decrease security because the individuals can probe the selection algorithm to determine who is likely to be selected and who is not. The analysis assumes a fixed percentage of people who can be subject to secondary screening at airports due to a fixed amount of screening resources and the need to keep passengers flowing through the system at a reasonable rate. Using the Carnival Booth algorithm, a terrorist group can determine who is not likely to be selected for increased scrutiny and then use that person to execute an attack. It suggests that this algorithm would be most effective when there is a diverse population of potential attackers and presents anecdotal information that this is indeed the case.

Essentially, the Carnival Booth algorithm works if a terrorist can determine that his chance of being selected for secondary screening is less than average; for example (using the numbers from [1]) if 8 percent of passengers are subject to secondary screening and 2 percent are selected randomly, then a terrorist has to reduce his chance of being selected for enhanced screening to

[5] This is similar to a scene from "Fiddler on the Roof." In the scene, Tevye hears an argument between his neighbors Perchik and Mordcha, and after hearing each of their positions, says "you are right" and "you are also right." Another character, Avram, says, "He's right and he's right? They can't both be right." Tevye replies, "You know, you are also right." As in this scene, who is really right in this situation?

less than 6 percent. While an individual potential attacker can not change his chance of being selected under this model, a terrorist group leader could use a population of potential attackers and select those who do not get selected on a large number of probing flights. A potential attacker is not reducing his actual chance of being selected; rather, he is decreasing the uncertainty in his estimate of his chance of being selected by repeatedly probing the system. Once some potential attacker is determined to have a lower than average chance of being selected for increased scrutiny, he is given the mission to execute an attack.

What are the flaws in this strategy? For one, it requires multiple recruits rather than a single recruit for each position on the attack team. While there may be *one* recruit whose profile causes him to be less likely than average to be selected for increased scrutiny, it is unlikely that, on average, the recruits will be less likely than average to meet the selection criteria. In fact, one can make an argument that people who are subject to terrorist recruitment are actually, on the average, *more* likely to be subject to increased scrutiny, especially if those designing the selection criteria have insights into what makes someone susceptible to terrorist recruitment. The Carnival Booth algorithm also assumes that even if a recruit is selected for increased scrutiny, he will not be arrested when he is not actually on an attack mission, because he will not be in possession of any suspicious materiel. This assumption ignores the fact that the recruit knows he is on a probing mission for the terrorist group and may behave in a way that arouses increased suspicion. Even if he is allowed to fly, his behavior may result in his being the subject of additional information collection – i.e., the starting point for a link analysis. The Carnival Booth algorithm, therefore, shares some characteristics with the classic gambler's strategy of doubling every losing bet – without an infinite amount of resources, it will eventually fail. The fact that there must be a reasonably large number of recruits for the Carnival Booth algorithm to result in one recruit with a lower than average selection probability creates additional risk of mistakes or exposure to the terrorist group. And the probing activity of the recruits could result in adaptation of the profiles used for selection, which would defeat the supposed advantage of the Carnival Booth algorithm, especially if this adaptation occurred as frequently as the probes.

3.9 You Can't Catch the Lone Wolf

This criticism is the most serious of all that have been proposed. A capable individual acting alone, who devises and executes a serious new attack scheme, will likely be able to evade detection. Reducing the possibility that this scenario can occur would seem to be a critical aspect of providing increased security. Tighter controls on dangerous materials, separating components needed to create weapons, and reducing the motivations of people to engage in terrorist activities would seem to be the most effective strategies for this difficult problem. In a sense, this criticism says that a security application won't be able to detect someone who manages to avoid all of its data sources and analytical techniques. Such an interpretation is obviously true but not particularly insightful.

4. METRICS

Any paper with the subject of the efficacy of data mining applications for any purpose whatsoever is incomplete without at least a brief discussion of metrics. This paper is no different. We briefly present a number of metrics that are either implicit or explicit in the evaluation of data mining applications for security. Ultimately, it is not analyses of the type discussed herein, but rather rigorous metrics-based experiments that will establish the efficacy of alternative designs and techniques for security applications.

Because of the high cost of a terrorist attack, the typical metric for a security application is the number (or probability) of a false negative; i.e., failure to prevent an attack. This is typically traded off against the probability of a false positive. Because of the vastly unequal costs of false negatives (extremely high) compared to false positives and the vastly different numbers of true positives (extremely low) compared to true negatives in the population, the likelihood of misclassifications must be weighted by the costs and frequencies to determine the overall costs of a security application. The benefits of the security application are expressed in terms of threats averted.

Because the benefits of a security application depend on the assumed distribution of threats in a population, it is desirable to have a metric that illustrates its effectiveness independent of this distribution. A metric with this property is discussed in [13].

Other metrics that may be used to evaluate alternative system designs include the distribution of costs (e.g., is it better or worse to inconvenience one person a lot or many people a little?), minimizing the worst-case outcome (i..e, maximizing the likelihood of preventing the most serious threats even if this means increasing the chances that more instances of less serious threats will occur).

Another class of metrics relates to the various criticisms. How quickly can new patterns be discovered, validated, and deployed? What is the value in preventing previous attack patterns compared to detecting new ones? How can security forces cause maximum disruption to attackers while minimizing costs to those whom they are protecting?

5. CONCLUSIONS

What can we conclude? Is data mining useful for security or not? What aspects of data mining are likely to be useful and what aspects are likely to be ineffective? What criticisms are valid because of the requirements of security applications, and what criticisms really just point out ineffective designs? Where might additional research yield useful new techniques, and where is it unlikely to do so?

In some areas, the jury is still out. Data mining algorithms have not yet resulted in the ability to discover patterns that can predict terrorism or other security threats that manifest themselves rarely and as a complex set of related events. They have been effective at discovering patterns that can detect common events that occur more frequently, such as cellular telephone or credit card fraud. A challenge for the research community is to design algorithms that can extend the range of feasible applications. Even as this range is extended, it is extremely unlikely that completely automated pattern discovery will be useful by itself for the detection of terrorist events. However, automated pattern discovery tools may be able to aid in the discovery of patterns of activity that are components of such threats and that can be incorporated into security applications. These security applications would have to include other techniques as well in

order to be useful for their specific purposes. So while data mining will not be an entire solution, it can be a useful component of such a solution.

The hardest threat to detect is the threat of a capable, intelligent adaptive adversary acting alone. Therefore, the most effective strategy is one that makes this threat increasingly unlikely. The other threats, of less capable adversaries, non-adaptive adversaries, and less-intelligent adversaries, can be effectively countered by appropriately designed and deployed data mining applications as a key part of a multi-layered prevention and detection system. Data mining can be one technique for pattern discovery, but it is only a part of the design and deployment of an effective security application. And other techniques such as starting from known subjects and performing link analyses as well as detection of dangerous materials and discouraging terrorist recruitment are at least as important. While patterns may be useful to guide a search, following connections from known risky subjects matters more.

Finally, it is important once again to note that effective security applications are complex systems that must have a clearly defined purpose, clearly specified authorities and authorizations, appropriate, available and useful data, and clear and manageable business procedures in addition to effective technologies if they are to succeed. And they must respect all aspects of privacy, civil liberties, and other considerations regarding the use and retention of data for specific purposes. Even with all these constraints, it is possible to design security applications that can be useful and to continue research into how to do so.

6. ACKNOWLEDGMENTS

I thank the many colleagues with whom I have had the opportunity to discuss and refine many of the ideas in this paper over the years. In particular, I thank Henry Goldberg for helping to develop many of the ideas discussed in the paper and David Jensen for helping to develop the model used in figure 2 as well as for much useful discussion and feedback. Responsibility for the ideas in the paper is, of course, solely that of the author.

7. REFERENCES

[1] Chakrabarti, S. and Strauss, A. "Carnival Booth: An Algorithm for Defeating the Computer-Aided Passenger Screening System," *First Monday* Vol 7., No. 10, 7 October 2002. http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/992/913

[2] Executive Committee on ACM Special Interest Group on Knowledge Discovery and Data Mining. "Data Mining" Is NOT Against Civil Liberties. June 30, 2003 (revised July 28, 2003). http://www.sigkdd.org/civil-liberties.pdf

[3] Fayyad, U.M., Piatesky-Shapiro, G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, 1-30. Menlo Park, CA: AAAI Press 1996.

[4] Jensen, D., Rattigan, M., and Blau, H. Information Awareness: A Prospective Technical Assessment. In *Proceedings of the Ninth ACM SIGKDD International Conference o Knowledge Discovery and Data Mining (KDD-2003)*. (Washington, DC, USA, August 24-27, 2003). ACM Press, New York, NY, 2003, 378-387.

[5] Jonas, J. and Harper, J. Effective Counterterrorism and the Limited Role of Predictive Data Mining, Policy Analysis No. 584. Cato Institute (December 11, 2006). http://www.cato.org/pubs/pas/pa584.pdf

[6] McLay, L.A., Jacobson, S.H, and Kobza, J.E. Making Skies Safer: Applying Operations Research to Aviation Passenger Prescreening Systems. OR/MS Today. October 2005.

[7] Office of the Director of National Intelligence. Data Mining Report. 15 February 2008. http://www.fbiic.gov/public/2008/feb/ODNI_Data_Mining_Report.pdf

[8] Office of the Director of National Intelligence. Data Mining Report. January 31, 2009. http://www.dni.gov/electronic_reading_room/ODNI_Data_Mining_Report_09.pdf

[9] Paulos, J., Do the Math: Rooting Out Terrorists is Tricky Business. *Los Angeles Times*, January 23, 2003.

[10] Schneier, B. and Hawley, K. Interview with Kip Hawley (July 30, 2007). http://www.schneier.com/interview-hawley.html

[11] Scientific American (editorial). Total Information Overload. *Scientific American*, March 2003, 12.

[12] Seifert, J.W., Data Mining and Homeland Security: An Overview. Congressional Research Service (Order Code RL31798) Updated January 27, 2006. http://www.au.af.mil/au/awc/awcgate/crs/rl31798.pdf

[13] Senator, T.E. Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation's Advanced Detection System (ADS). In *KDD-2000*: 44-53.

[14] Senator, T.E. Multi-Stage Classification. In Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05). (Houston, TX, November 27-30, 2005).

[15] Taipale, K. Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data, *Columbia Science and Technology Law Review*. Vol. 5. No. 2 (Dec 2003). Available at SSRN: http://ssrn.com/abstract=546782

[16] Vijayan, J. House Committee Chair Wants Info on Cancelled DHS Data-Mining Programs: Millions Have Been Spent on Work That Was Eventually Abandoned. ComputerWorld, September 18, 2007. http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9037319