January, 2006

American Statistical Association
Frequently Asked Questions Regarding the Privacy Implications of Data Mining

Media accounts of threats to individual privacy are now reported almost daily. Many of these allude to statistical work in the broad area of data mining, helping to shape images and public perceptions of statistical practitioners. These reports often make reference to the tension between the need to protect privacy and the desire to gather intelligence aimed at the marketing of products, winning elections, or thwarting terrorism. As members of the professional statistical community, we need to be concerned about the accuracy of these images and weigh in on these arguments. This commentary, in the form of answers to a set of Frequently Asked Questions (FAQs),[1] relies on the principles and uses of statistics broadly endorsed by the ASA.

Detecting terrorist activities is certainly a legitimate endeavor. There are hundreds of more mundane, but nonetheless important, applications of data mining techniques, however, that the public also needs to understand. The FAQs listed below, and our responses to them, are designed to help fill this vacuum of understanding by providing authoritative responses to questions that stand at the forefront of data mining.

Frequently Asked Questions

1. **What is data mining?**
2. **What types of data are being mined?**
3. **What are beneficial uses of data mining and what potential threats do these activities pose?**
4. **What are the controls (laws, policies, technologies) that help ensure the protection of people’s privacy?**
5. **How do statistical methods support data mining and what are their limitations?**
6. **What are the legal protections for federal data and how do they apply to data mining activities?**
7. **How does the ASA's ethical code (and those of other professional associations) treat data mining?**
8. **What are the threats posed by data mining to publicly disseminated statistics?**

[1] These FAQs were prepared by an ad hoc task force consisting of Jerry Gates (on behalf of the ASA Committee on Privacy and Confidentiality), Joe Salvo and John Gardenier (on behalf of the ASA Scientific and Public Affairs Advisory Committee), and Bill Seltzer (on behalf of the ASA Committee on Professional Ethics). This set of FAQs and responses greatly benefited from comments received from a number of individual members of ASA with expertise in the subject and from additional members of the three ASA Committees concerned.

1. What is Data Mining?

Within the discipline of statistics, data mining may be defined as the application of statistical methods to potentially quite diverse data sets, in order to clarify relationships (perhaps including some previously unknown), to estimate their strength, or to accumulate data about real or hypothetical entities (such as a person, a group of persons, a commercial enterprise, or other entities or events). The results may then be used to make statements about the real or estimated characteristics of these entities, or to test hypotheses related to one or more of the systems with which they interact.

Data mining relies heavily on statistical concepts and methods. Some are specifically relevant to data mining, such as regression trees. But a vast array of others are also used in data mining endeavors. Even when the data miners are not statisticians and think of their methods as artificial intelligence or computer science, statistical concepts are frequently embedded in their processes. A useful broad definition of data mining is, "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data."[1] It is important to note that data mining is contingent upon, and affected by, a range data management policy considerations including data collection, warehousing, sharing, and ownership. Each aspect affects personal and business privacy interests. To be used appropriately, all must be considered within an overall process of formulating sound privacy protection methods and policies. This paper does not deal with those wider policy considerations. It focuses on data mining itself.

In addition to relying on statistical concepts and methods, data mining relies on several assumptions. These include: (1) that one has access to a sufficient amount of data to be useful for one's purposes—often associated with business, government, or research interests, but sometimes just idle curiosity; (2) that there is reason to believe much of the data can be regarded analytically as "noise" but one or more "signals" of interest can be found by intelligent searching or "mining;" (3) that the use of various analytical tools, predominantly statistical tools, can extract and amplify these signal(s) and distinguish them in some reliable manner from the noise, and (4) that the uncertainties surrounding the conclusions drawn from any such analysis are examined and deemed acceptable. The signals, of course, are then applied to the problem at hand, such as identifying potentially profitable customers or business locations, risk factors for diseases, unauthorized use of credit cards, incidents of bioterrorism, or terrorist suspects.

2. What types of data are being mined?

Data that are mined pertain to individuals, businesses, or natural events or conditions (such as weather patterns or contamination). The types of personal data that are mined include age, race, sex, marital status, income, education, medical history, genetic information, employment, travel itinerary, and buying patterns. The data pertaining to individuals may be specific to an identified person; may be anonymized by removing direct identifiers such as name, address, or social security number; or may be aggregated over geographic, demographic, or other variables. These types of data come from sources such as internal government records supporting a program or activity, government records classified as public and open to review, and customer transaction records obtained by business.

The underlying data are often collected under a specific agreement or understanding as to how they will be used. Federal government records include tax returns, welfare applications, workman's compensation claims, social security earnings, and criminal investigations. Government records are legally protected from inappropriate access and use by the Privacy Act and other federal as well as state statutes (see further discussion under FAQ #6). Public records are open to the public and include voting registration records, driver's licenses, vehicle registrations, property taxes, and vital records such as births, deaths, and marriages. Most of these records have no legal constraints on their access and use, since they are determined to be of general importance to the public and their availability helps ensure government accountability. Business records include credit records, airline passenger records, insurance records, medical records, grocery store purchases, product warranties, and Internet site registration information. These records are often proprietary and limited to the business interests of the company or its subsidiaries. Businesses may also share their data with government agencies under contracts or subpoena. Data that are anonymized or aggregated provide privacy protection to the individual when used separately. When combined with individually identified data using suitable statistical techniques, these data may provide additional details (albeit approximations) about the individual that enhance their use.

In considering the types of data that are mined for particular types of information, it is important to recognize that they have often been collected for another purpose entirely. Thus, data that are sufficient for marketing products to customers may not be sufficient for approving insurance claims. Similarly, voting registration records may not be appropriate for investigating tax fraud. But medical records or genetic information may be quite useful in understanding drug effectiveness and potential adverse consequences from drug interactions. Even though the data may not be as precise as desired, statistical techniques involving regressions and predictive models can determine, and enhance, the degree of accuracy that can be obtained. In determining whether data are sufficient to be mined for a specific purpose, those planning or implementing any data mining activity must consider reliability of the data, estimates, and linkages involved in relation to the intended uses and the possible consequences to the individual should the results turn out to be wrong (i.e., yield a "false positive," as discussed further under FAQ #3).

3. What are beneficial uses of data mining and what potential threats do these activities pose?

Beneficial uses of data mining serve the public interest--for example, in the form of more efficient provision of goods and services by the government, by not-for-profit organizations, and by the private sector. In the federal government, the three most common applications of data mining are for improvements in service and performance; detecting fraud, waste, and abuse; and analyzing scientific and research information.[2] Understanding patterns in the failure rates of ship parts, for example, is key to creating and maintaining a supply line capable of ensuring that the fleet will not be disabled or impaired because parts are not available. Within the Veterans Administration, compensation and pension data are regularly mined to detect patterns that are indicative of abuse or fraud, making the allocation of benefits fairer for all.

At state and local levels, data mining is increasingly used to enhance public safety--an area in which program implementation is guided by insights gained from crime, health, and other public safety data, some of which are now being collected through enhanced 911 systems. In the private sector, data mining has been used for decades in market research, customer relations, supply chain analysis, financial analysis, and fraud detection. Corporations become more efficient when resources are more precisely allocated, and their products and services are more effectively targeted to customers. This includes fraud detection algorithms now commonly used by credit card and insurance companies, as well as analysis of failure rates for products like automobiles, where such detection leads to improvements in design and to the more efficient provision of spare parts.

When applied to human systems, information on the characteristics of persons becomes an important part of data mining efforts, because the characteristics of individuals--from age/sex to occupation and income--may be used to estimate behavioral propensities. In fact, evaluating these estimates of behavioral propensities at an aggregate level provides the very foundation for market research and management models that drive the provision of potentially valuable products and services. Since these are aggregates, a considerable degree of privacy protection is assumed especially for tabulations representing large numbers of people. With advances in database development and computing, however, certain organizations have moved in the direction of "micro-targeting," where the line between individual attributes and group behavior becomes blurred.

With micro-targeting, the intersection of databases becomes focused at the individual level, enabling providers to tailor services and, in the case of recent elections, for example, target political appeals directly to individuals. Thus, administrative records collected and compiled for one purpose are being mined for other purposes, frequently without the direct knowledge of individuals, raising the specter of privacy breaches. The most egregious examples usually involve stories of adverse consequences from "false positives," where individuals may be incorrectly identified as being of high potential for some behavior, such as terrorist activity. This is not to say that use of data mining in counter-terrorism should be abandoned due to the potential for false positives. Rather, these activities must be done in a way that minimizes the potential for false positives and the effects on the victims when unavoidable false positives occur.

All data mining strategies that use information on human systems are potentially abusive, both by having individual information disclosed without consent and by linking records in databases that separately are not a threat to privacy but together give organizations the capacity to identify specific persons. It follows that the more complex the systems of linked databases the more serious are the threats to privacy and the more numerous the ethical dilemmas. It follows further that, as organizations pursue plans to become more efficient in their delivery of goods, services, and messages that support their cause, the inherent ethical dangers become increasingly inevitable. Most important, resolving the dilemmas involved is made even more difficult by the fact that laws regarding breaches of confidentiality have not done a good job of setting limits on the database development being used to create predictive models.

Some advocates of personal privacy point out that the legal system is not designed to protect individuals' privacy automatically. As a result, people who want to limit how much of their personal data end up in these massive data banks may adopt the strategy of not providing information to anyone except as required by law or as necessary to obtain some strongly desired service or benefit. Even in the latter case, individuals may be able to substitute a driver's license number for a social security number, for example, or deliberately omit personal data, such as "family income," from a form on which it is not legally required (such as a warranty form). Individuals may even choose to provide inaccurate information, such as an erroneous birth date, on the presumption that this will deter those seeking to target them. Carried to the extreme, however, the concern of such individuals with ensuring personal privacy may actually result in detrimental effects to the quality of survey data. Ironically, in an age of unprecedented capacity to compile, merge, analyze and disseminate data, such behavior may make data quality end up as a casualty of innovation.

4. What are the controls (laws, policies, technologies) that help ensure people's privacy is protected?

Many uses of data mining do not involve persons or organizations but rather address, say, properties of stars in the firmament or the genetic composition of insects. That said, there are still a lot of data in public and private hands that pose potential risks to the privacy of persons or organizations. These risks are not limited to actual identifications. Incorrect identification (the false positives noted earlier) can also threaten their lives, livelihoods, or reputations. Just as data mining technologies and methodologies are continuously evolving, the fairly new body of law, policy, and technology for privacy protection in data mining environments is also evolving.

Data mining activities are often limited by both mandatory and voluntary controls. Mandatory controls consist of legal restrictions on access and use of personally identified information and/or judicial means of redress for individuals who are falsely identified and harmed by the data mining activity. Voluntary controls consist of technical, methodological, and institutional (policy) approaches to limit the opportunity for inappropriate access and to ensure that the data mining methodology is sound and produces the highest likelihood of achieving the desired outcome. Three broad areas of technical/methodological controls have been used to protect the privacy of individuals when information about them is included in data bases subject to data mining: "(1) anonymization techniques that allow data to be usefully shared or searched without disclosing identity; (2) permissioning systems that build privacy rules and authorization standards into databases and search engines; and (3) immutable audit trails that will make it possible to identify misuse or inappropriate access to or disclosure of sensitive data."[3]

Each of these helps reduce the risk that a data mining activity designed for one use is used for another unrelated organizational purpose.

A different form of voluntary control is institutional policy that promotes fair information practices. Widely accepted principles that promote fair information practices include notice and consent, and access and correction.[4] Notice and consent ensure that individuals are informed about the purposes of—and provide implicit or explicit consent to—the collection, use, retention, and disclosure of their personal information. Some of the data that are mined come from public records, such as voting or driver's licenses records, in which case notice and consent for secondary use is not required. Much of the data, however, are not public (e.g., government records) and implicit or explicit consent is often required--in the case of federal databases, by the Privacy Act. Fair Information Practices are advanced in some cases by federal, state, or local law; by policies of groups of organizations (including professional associations such as the Association for Computing Machinery); or by individual firms, laboratories, or persons. To support these fair information practices, responsible organizations implement employee awareness through regular training, encourage external oversight through committees or boards, include on their Web site information about data mining efforts as well as data sources, establish controls that limit the collection and retention of data to the purpose of the intended effort, and provide for remedies in the case of individuals who are treated unfairly.

How do organizations decide on their voluntary control strategy? The choice of controls is often influenced by their cost and likely benefit in ensuring the viability of the data mining effort. Different organizations and agencies will use different controls depending on their mission and resources. Decisions related to ensuring quality of the data being mined, for instance, are affected by the intended use when compared with potential liability. Decisions related to data retention are influenced by the value of the data and their potential for secondary uses. Decisions related to openness about the data mining activity and sharing of source data can be influenced by national security or, for private firms, by competitive advantage. Decisions about openness and sharing may be the most important ones. The National Association of State Chief Information Officers (NASCIO) notes that a lesson learned from Total Information Awareness (TIA), Computer Assisted Passenger Prescreening Program (CAPPS II, and Multistate Anti-Terrorism Information Exchange (MATRIX) is that transparency as to the project's purpose, the reasons why information is collected, how it is used, who will have access to it, how it will be secured, and whether individuals can access and correct their personal information are all key in formulating privacy policy and the detailed rules necessary both for the management of the data themselves and for setting the allowable scope and purpose(s) of any data mining.[5]

5. How do statistical methods support data mining and what are their limitations?

Statistical theory and methods, as noted, are at the core of virtually all data mining applications. Indeed, the proper understanding and use of statistical theory and methods is essential in ensuring that data mining applications yield sound and reliable results.

Statistical theory and methods, for example, are central to the classification, clustering, and modeling issues involved in most data mining applications—in testing the appropriateness of the models developed, providing guidance on the quality of the variables used and on how best to link different data sets, as well as developing and interpreting measures of statistical confidence associated with different conclusions based on data mining applications. Statistical theory and methods also play an important role in reducing the risk of confidentiality disclosures and privacy violations.

Data mining applications, which nearly always involve making use of information drawn from multiple data sets, are particularly subject to limitations of data and methods. The procedures used to combine the individual data sources may themselves introduce error and uncertainty. Moreover, the diverse characteristics of the data sets involved can give rise to multiple sources of error that may interact with one another in unknown ways.

The underlying sources of error may include, among others, coverage and content errors, the possibly different time references of individual data sets, and the additional uncertainty introduced when some of the data sets are based on samples. These sources of error and uncertainty emphasize the importance of ensuring that the necessary statistical expertise is involved in data mining applications. Such expertise will help users to be maximally confident in the results obtained and to avoid drawing ill-founded conclusions. These issues become even more acute when data mining serves as the basis for inferences about individuals or policy decisions about population subgroups defined along religious, racial, ethnic, ancestry, or linguistic lines.

6. What are the legal protections for federal data and how do they apply to data mining activities?

Several federal laws afford protection to personally identified data. First, the Privacy Act requires that federal agencies inform individuals on data forms and, more generally, through the Federal Register, about any databases they maintain from which individually identified records can be retrieved. Federal agencies must, in addition, explain the purpose of and uses for personal information at the time it is collected. The Privacy Act also permits individuals to access and correct their personal information so they know what information about them is being held by others and can be assured that the information that can be used to make decisions about them is accurate. Second, under the Computer Matching and Privacy Protection Act, agencies must ensure that a Data Integrity Board (DIB) approves matches of individually identified data across databases and report in the Federal Register the nature of all such matches. In recognition of the fact that personally identified information collected for statistical purposes cannot be used to disadvantage individuals, both of these laws provide important exceptions. The Privacy Act limits individuals' rights to access and correct their personal information, since there is (by definition) no possibility for their own identifiable information to be used to their own disadvantage. The Privacy Act also places unique restrictions on the collection and use of Social Security Numbers and further requires that agencies have the authority to collect the numbers and that they explain their particular uses. The Computer

Matching and Privacy Protection Act permits matches for purely statistical purposes without prior DIB review and approval.

Statistical data (both for individuals and for businesses) are also protected by federal agency statute and by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002. Statistical data are defined as data produced from information collected for statistical purposes. The term "statistical purpose" means the description, estimation, or analysis of the characteristics of groups, without identifying the individuals or organizations that comprise such groups. "Statistical purposes" also describes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support these purposes.[6] Information collected by the U.S. Census Bureau, for example, is covered by Title 13, United States Code. Sections 9 and 214 of that law make all personally identified information confidential and provide strict penalties for improper disclosure. The law also ensures that this information is used only for statistical purposes and not to make determinations about any particular individual. Several other statistical agencies, including the Bureau of Transportation Statistics and the National Center for Health Statistics, have specific legislation that similarly protects information collected for statistical purposes. The CIPSEA extends similar legal protections afforded data collected by these agencies to all surveys undertaken by the federal government when those surveys are done under a pledge of confidentiality and the information provided may only be used for a statistical purpose.

Public use microdata files from federal survey (or census) statistical data are sometimes mined even though they do not directly or indirectly identify individual respondents. To meet legal requirements, data files that are produced for research use undergo a thorough review and are modified to ensure that unique attributes are disguised. These microdata files are broadly disseminated and often made available through the Internet. Those undertaking data mining activities will sometimes use them to study attributes of populations that can be associated to other data files they are mining. This is appropriate. They should not be used to try to estimate the attributes of specific individuals, however. This is because, since these micro data files have usually been explicitly altered to protect against such applications, such estimated attributes are likely to be unreliable for the intended purpose.

7. How does the ASA's ethical code (and those of other professional associations) treat data mining?

Like most statistical methodologies, data mining by itself is ethically neutral. This is particularly so because the term data mining is a generic one referring to a wide range of procedures, involving diverse data sets and carried out for numerous purposes. For these reasons there are no specific references to data mining in the ASA's "Ethical Guidelines for Statistical Practice," adopted by the ASA Board in 1999, which is available on line at the ASA's website (www.amstat.org) and in print from the ASA office. It also needs to be understood that, whether one is dealing with data mining or some other topic, ethical standards and legal requirements are not the same thing. While there is a very large overlap between the unlawful and the unethical, the two concepts are not equivalent.

From the perspective of statistical practice, data mining raises three quite different sorts of ethical issues: (a) the suitability and validity of the methods used in any given data mining application, (b) the degree to which confidentiality and privacy obligations are respected, and (c) the overall aims of a given data mining application. Each of these is addressed in the ASA's *Ethical Guideline for Statistical Practice.*

Suitability and Validity. Several provisions of the ASA's ethics guidelines address issues of the suitability and validity of methods used in any statistical application, including data mining. These include, in section II.A,

> "2. Guard against the possibility that a predisposition by investigators or data providers might predetermine the analytic result. Employ data selection or sampling methods and analytic approaches that are designed to assure valid analyses in either frequentist or Bayesian approaches.
>
> "4. Assure that adequate statistical and subject-matter expertise are both applied to any planned study. If this criterion is not met initially, it is important to add the missing expertise before completing the study design.
>
> "5. Use only statistical methodologies suitable to the data and to obtaining valid results. For example, address the multiple potentially confounding factors in observational studies, and use due caution in drawing causal inferences.
>
> "7. The fact that a procedure is automated does not ensure its correctness or appropriateness; it is also necessary to understand the theory, the data, and the methods used in each statistical study. This goal is served best when a competent statistical practitioner is included early in the research design, preferably in the planning stage."

Such provisions also include, in section II.C,

> "2. Report statistical and substantive assumptions made in the study.
>
> "5. Account for all data considered in a study and explain the sample(s) actually used.
>
> "6. Report the sources and assessed adequacy of the data.
>
> "7. Report the data cleaning and screening procedures used, including any imputation.

> “8. Clearly and fully report the steps taken to guard validity. Address the suitability of the analytic methods and their inherent assumptions relative to the circumstances of the specific study. Identify the computer routines used to implement the analytic methods.
>
> “9. Where appropriate, address potential confounding variables not included in the study.
>
> “12. Report the limits of statistical inference of the study and possible sources of error. For example, disclose any significant failure to follow through fully on an agreed sampling or analytic plan and explain any resulting adverse consequences.”

Privacy and Confidentiality. The ASA ethics guidelines address privacy and confidentiality obligations in section II.D, “Responsibilities to Research Subjects (including census or survey respondents and persons and organizations supplying data from administrative records, as well as subjects of physically or psychologically invasive research).” Among the pertinent provisions are

> “1. Know about and adhere to appropriate rules for the protection of human subjects, including particularly vulnerable or other special populations who may be subject to special risks or who may not be fully able to protect their own interests. Assure adequate planning to support the practical value of the research, the validity of expected results, the ability to provide the protection promised, and consideration of all other ethical issues involved. Some pertinent guidance is provided in key references 3 - 7 at the end of this document for U.S. law, the U.N. Statistical Commission, and the International Statistical Institute. Laws of other countries and their subdivisions and ethical principles of other professional organizations may provide other guidance.
>
> “3. Avoid excessive risk to research subjects and excessive imposition on their time and privacy.
>
> “4. Protect the privacy and confidentiality of research subjects and data concerning them, whether obtained directly from the subjects, from other persons, or from administrative records. Anticipate secondary and indirect uses of the data when obtaining approvals from research subjects; obtain approvals appropriate for peer review and for independent replication of analyses.”

Overall Aims of the Application. When considering the overall aims of any data mining application, two provisions of the Preamble to the ASA ethics guidelines are particularly pertinent. The first is from a section on “Statistics and Society”:

> "Statistical tools and methods, like many other technologies, can be employed either for social good or for evil. The professionalism encouraged by these guidelines is predicated on their use in socially responsible pursuits by morally responsible societies, governments, and employers. Where the end purpose of a statistical application is itself morally reprehensible, statistical professionalism ceases to have ethical worth.

The second is from a section "Shared Values":

> "All statistical practitioners are obliged to conduct their professional activities with responsible attention to: 1. The social value of their work and the consequences of how well or poorly it is performed. This includes respect for the life, liberty, dignity, and property of other people."

Other professional associations in statistics and allied fields also contain guidance applicable to data mining. Given the central role that data processing technology plays in data mining, the Association for Computing Machinery's 1992 "Code of Ethics and Professional Conduct" available at http://www.acm.org/constitution/code.html provides particularly relevant guidance. See, for example, two of its "moral imperatives":

> "1.1 Contribute to society and human well-being. -- This principle concerning the quality of life of all people affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures. An essential aim of computing professionals is to minimize negative consequences of computing systems, including threats to health and safety. When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and will avoid harmful effects to health and welfare."
>
> "1.2 Avoid harm to others. -- ... This principle prohibits use of computing technology in ways that result in harm to any of the following: users, the general public, employees, employers ... Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event the responsible person or persons are obligated to undo or mitigate the negative consequences as much as possible. One way to avoid unintentional harm is to carefully consider potential impacts on all those affected by decisions made during design and implementation."

Two additional sets of guideline relating to statistics exist at the international level--the International Statistical Institute's *1985 Declaration on Professional Ethics*, which is available at: http://www.cbs.nl/isi/ethics.htm, and the United Nations Statistical Commission's 1994 Fundamental Principles of Official Statistics, which is available at: http://unstats.un.org/unsd/goodprac/bpabout.asp. Although, like the other available guidelines, neither of these sources directly discusses data mining, the advice provided roughly parallels that contained in the ASA guidelines.

8. What are the threats to publicly disseminated statistics posed by data mining?

Participation in a democratic society is based upon open provision of information. Publicly disseminated statistics from government agencies and other organizations, for example, are key to a well-informed electorate. Data on crime, levels of disease, home ownership, tax assessments, military service, business productivity, and a whole host of other demographic, social, and economic items have all been disseminated through a variety of media. Similarly, efforts to make government and businesses more efficient frequently use data on creditworthiness and consumer behavior. Such data are used by groups within our society to evaluate need, conduct oversight, and provide critical analysis that permits citizens to draw independent conclusions.

While the era of electronic dissemination has made access to these data easier for organizations and individuals, this very access now poses threats to data dissemination. Electronic data files have combined with data linking and data mining technologies in ways that raise the specter of privacy breaches. Once the exclusive purview of large organizations, powerful desktop computers have enabled individuals to unlock the potential of data and actively use data mining technologies. Further, businesses and other organizations will frequently "push the envelope," in order to attain a competitive advantage. This has created a major dilemma for those who electronically disseminate data and, for legal and/or practical reasons, feel an obligation to minimize the potential for confidentiality breaches. Those in the business of providing data, inside and outside government, must balance the fact that data mining technologies have the capacity not only to lead to great insights that are of benefit to mankind but also, in and of themselves, to pose threats to personal privacy. Particularly in those cases where the objective of a data mining exercise is to make inferences about the behavior of real individuals, there needs to be a careful balancing of the risks and benefits involved.

Because of the inherent dangers to personal privacy, data disseminators should have a common interest in developing standards that prevent individual disclosure but permit data to be mined to realize its benefits. The basic dilemma faced by those who disseminate data is that data mining methods used to "clean" and "detect" patterns useful for applications can also be used to break disclosure measures that protect individuals. It is ironic that the very access afforded by the Internet and by database technologies generally, creates the very conditions for suppressing information, thus potentially limiting dissemination of data. This is because data are so accessible in digital form, that they create unprecedented opportunities for conditions that increase the potential for confidentiality breaches.

Governments and other organizations seeking to protect themselves from legal threats can opt to limit public release of data, especially public use files consisting of anonymized records for persons or households (i.e., microdata files). Such a step to limit the risk of disclosure would mean the loss of valuable information to policy makers, academic researchers, businesses, and others, To compensate for this loss, federal agencies are offering controlled access and use arrangements, such as through research data centers where data access and use are strictly controlled. However, these too have their limitations since they are create obstacles such as the need to be located near the data center and to use computers and software provided by others.

Drawing distinctions between “safe” and “dangerous” data release procedures based on current real threats is well documented [1].[7] What is less understood is the impact that the perception of new threats plays in decisions to release data publicly. To a greater degree than ever, this has caused arguments about “proper and responsible” dissemination to move well beyond the statistical realm; however, statisticians remain at the table as active participants in this discussion.

[1] W. Frawley and G. Piatetsky-Shapiro and C. Matheus, Knowledge Discovery in Databases: An Overview. AI Magazine, Fall 1992, pgs 213-228

[2] U.S. Government Accountability Office, "Data Mining: Federal Efforts Cover a Wide Range of Uses" GAO report number GAO-04-548, 2004.

[3] Dempsey, James X. And Rosenweig, Paul, “Technologies that Can Protect Privacy as Information is Shared to Combat Terrorism.” Washington, DC: Center for Democracy and Technology, May 26, 2004

[4] See AICPA/CICA Privacy Framework American Institute of Certified Public Accountants and Canadian Institute of Chartered Accountants, November 2003 (rev. March 2004).

[5] “Think Before You Dig: Privacy Implications of Data Mining and Aggregation.” NASCIO, September 2004

[6] Confidential Information Protection and Statistical Efficiency Act. Section 502(9)

[7] Federal Committee on Statistical Methodology, Statistical Policy Working Paper 22, Report on Statistical Disclosure Limitation Methodology, 1994