Green College Thematic Lecture Series

The New Politics of Surveillance and Visibility

Edited by
Kevin D. Haggerty and Richard V. Ericson

UNIVERSITY OF TORONTO PRESS
Toronto Buffalo London
2006

15 Data Mining, Surveillance, and Discrimination in the Post-9/11 Environment

OSCAR GANDY, JR

Public awareness of the role that data mining and other forms of remote surveillance have come to play in the management of the concerns of daily life has increased rather dramatically in the United States following the immediate, and often quite hostile, response to the plans announced for the creation of a Total Information Awareness (TIA) program in the Department of Defense (Mittelstadt 2002). No small part of the concern being expressed was based on an assessment of the character of the program's head, the same retired Admiral John Poindexter who had been convicted of lying to Congress and other efforts towards derailing Congressional investigations of official misconduct in the Iran-Contra Affair (Moscoso and Edmonson 2003).

Although concerns about the use of data-mining techniques by corporations for the purpose of segmentation, targeting, and other forms of 'relationship management' have been expressed in the past (Danna and Gandy 2002), it was primarily in the wake of announcements about proposed TIA activities that similarities and differences between the uses, users, and risks associated with the technique became the focus of so much public attention (Leahy, Feingold, and Cantwell 2003; Markle Foundation 2002).

This chapter seeks to bring a critical light to bear on the ways in which the use of data mining and associated techniques of knowledge production seem likely to transform social relations within both the public and private spheres. After a brief introduction to the technology and its uses within business and government organizations we will turn our attention to the social concerns that are raised by such increased, widespread, and presumptively legitimate uses of computer-enhanced discriminatory techniques. The final section will focus on the

limitations to the development of responsive public policies that are inherent in the traditionally narrow framing of privacy legislation towards the protection of what has been termed 'individually identifiable information' (Hatch 2001; Lyon 2003; Pollack 2001; Skok 2000).

The Goals of Data Mining

Data mining is a process that has as its goal the transformation of raw data into information that can be utilized as strategic intelligence within the context of an organization's identifiable goals. At its core, data mining is concerned with prediction. Data-mining efforts are directed towards the identification of behaviour and status markers that serve as reliable indicators of a probable future. Because the development of data-mining applications is dominated by concerns about risk minimization or avoidance, the most relevant behaviours are those that can be assessed in terms of relative risk. The use of a data-mining resource to identify the best prospects – the 20 per cent of the customers who are likely to provide 80 per cent of the profits – can still be understood from a risk perspective by noting that selection of the best also means avoidance of persons and relations that are less desirable, or risky.

The identification of the best prospects or the highest risks necessarily involves some agreement about standards of value. In the business environment, an individual who is a 'good customer' under one system of valuation may be of considerably less value under another. Decisions within the realm of consumer credit provide some excellent examples (Leyshon and Thrift 1999). A customer who pays off her credit cards in full each month would surely be identified as a good 'credit' risk, in terms of the likelihood of default. That same customer would not be so highly valued in terms of her contribution to the bank's bottom line, as she would not incur any interest charges. A riskier customer, one who is late, and may even miss a payment on occasion, may actually be more valuable because of the interest and penalties he accrues (Adams, Hand, and Till 2001).

The traditional challenge for data miners is to determine which customers are more valuable, and therefore worth keeping. A somewhat different challenge is the determination of that point at which a previously 'good' customer becomes less desirable.

In one set of analyses based on some 90,000 customers, cluster analysis was used to divide the sample of customers into two classes. Pattern-searching techniques were then used to identify what the analysts characterize as 'interesting or unexpected observations' (Adams, Hand, and Till 2001: 1020). With this technique the analysts identified a class of customers who began as 'good,' and then rather suddenly switched to missing payments, which the analysts characterized as 'bad behaviour' under the first approach to valuation. The provision of 'warnings' about the likelihood of default by individuals is precisely the kind of predictive services that credit agencies have begun to offer their customers. Clearly, these behavioural analyses differ in important ways from the sorts of credit-scoring analyses that are relied upon in deciding whether to issue a card in the first place (Leyshon and Thrift 1999).

A great many terms that describe data-mining applications are broadly familiar, and while some may be unique to a particular organization or administrative specialty, fundamentally they refer to the same underlying task (Bowker and Star 1999). For example, although the identification of individual terrorists is of paramount importance to those who are involved in the pursuit of 'homeland security,' the assignment of individuals to classes, categories, or groups on the basis of an analysis of seemingly independent, and in some cases, anonymous, activities has become a routine administrative activity (Computer Science and Telecommunications Board 2000; Hobart and Schiffman 1998).

Data-mining applications in the area of security abound. Concern about the security of computers and other critical infrastructural resources has led to an accelerated push for the development of techniques to detect and prevent intrusion into sensitive systems (Computer Science and Telecommunication Board 2002). In the detection of attacks on networked computer systems, the challenge is one of identifying deviations from the normal flow of traffic to a server. Unlike the evaluation of traffic accidents or other discrete events that generate a reliable record, a great many attacks on computer systems go undetected, and as a result, analytical models that depend upon those distinctions are of limited use. It is here that the discovery of patterns within data that 'call attention to themselves' enables analysts eventually to arrive at distinctions between 'probes,' outright attacks, and legitimate variations in network traffic over time (Rothleder 2000).

Consumer profiling involves the attempt to assign individuals to one or more groups or segments on the basis of attributes that they share, or are assumed to share to some degree. The nature of consumer profiling has been explored in considerable detail by the Federal Trade Commission (2000) with regard to its use by online businesses. Online profiling refers primarily to the techniques used by companies with

an online presence to deliver banner advertisements and editorial copy to individuals.

A variety of techniques, many of which are hidden from, or not disclosed, to website visitors are used to gather transaction-generated details about them (Pollack 2001). Among the most frequently used techniques are 'cookies' or relatively durable markers that are placed on the visitor's hard drive. These cookies make it possible for firms, or their representatives, to 'track' individuals as they move to different sites around the web (Helling 1998). As testimony revealed to the FTC, data gathered from this 'clickstream' is often combined with additional information acquired from third-party sources. 'This enhanced data allows the advertising networks to draw a variety of inferences about individual customer's interests and preferences. The result is a detailed profile that attempts to predict the individual consumer's tastes, needs, and purchasing habits and enables the advertising companies' computers to make split-second decisions about how to deliver ads directly targeted to the consumer's specific interests' (Federal Trade Commission 2000: 5).

Some users of customer-profiling systems report dramatic improvements in the productivity of their websites. In some cases, the source of the improvement was a fairly rough index of consumer interests based on little more than a zip code (Stevens 2001). In other cases, profiling systems are continually updated through the incorporation of multiple streams of information.

The provision of targeted advertising is not the only use of data-mining techniques within the networked environment. The profiling of visitors to corporate websites can also influence business decisions about the cost and quality of the services to be offered to that visitor (Danna and Gandy 2002). The use of data-mining techniques to facilitate this kind of discrimination in e-commerce transactions is referred to as 'weblining' (Stepanek 2000) because of its similarity to familiar but illegal forms of racial discrimination known as 'redlining.' Its name is derived from a common practice in financial service markets of drawing red lines around neighbourhoods where loans would not to be made or property insurance would not be underwritten (Nier 1999).

Firms that use data-mining in support of 'customer relationship management' (CRM) are interested in differentiating among existing and potential customers in terms of their estimated total value. Reportedly, there are very few segments of the market that mine customer data more actively than firms in the gaming industry. According to one

source, 'Casino operators gather and store data from loyalty programs, direct mail databases, socioeconomic data, and credit card issuers, among other sources.' Harrah's Entertainment is an exemplary user of data-mining for CRM. Harrah's gathers data from its nearly five million 'loyalty' program members and uses transaction-generated information from its loyalty cards as well as slot machine data and information derived from food and beverage sales (CFO Research Services 2002: 11).

Data-mining Techniques

Data mining, or knowledge discovery in databases, as the insiders would have it known, is a small but rapidly expanding specialty within the field of applied mathematics that seeks to derive meaningful intelligence from the analysis of patterns within sets of data, including the virtual mountains of data that are generated by a host of routine interactions enabled by networked computers (Fulda 2000).

Data mining is both descriptive and predictive, and although evaluation is not always explicit, the process of identification nearly always includes some form of evaluative assessment (Gandy 2000). Descriptive analysis may involve the assignment of 'objects' or entities into pre-existing categories in accordance with well-developed rules for classification (Bowker and Star 1999). Objects of interest may also be classified into categories that are the products of the analysis reflecting the application of statistical or other criteria for the determination of boundaries (Norusis 1994).

At the descriptive level, analyses common to data-mining can provide condensed summaries of massive amounts of information. We readily understand common statistical measures, such as means or averages, as summaries. We also understand correlation coefficients as summary statements about relationships between objects, or the attributes of objects. At a multivariate level, we might understand factor scores as a summary statement about relations between a large number of attributes and a large number of objects that might be persons or groups of persons. We are perhaps less familiar with summaries at the level of discourse.

Data-mining systems applied to the analysis of text can produce summaries of varying length (Mani, Guilder, Clifton, and Concepcion 2000). Not long ago, the MITRE Corporation announced the development of a data-mining tool for text processing that is capable of processing a large number of news stories, identifying the key 'players' in the

news, and generating a concise summary. In one of the examples provided in a promotional article, 60,000 news stories were condensed into an eighty-six-word summary (Mani et al. 2000: 3).

At the descriptive level, the process of discovery is often aided through the use of techniques that facilitate the visualization of patterns and relationships in data. These patterns may allow distinctions to be made between persons, behaviours, and outcomes on the basis of relations between the attributes of each. Increasingly these relationships may be discerned in spatial terms with the aid of geographic information systems (Baker and Baker 1999; Phillips and Curry 2003). Online analytical processing tools (OLAP) make it easier for subject area specialists who are not necessarily familiar with advanced statistical techniques to pursue an analysis of complex multivariate relationships merely by manipulating objects on screen, or responding to queries about alternative models or conditions (Roberts-Witt 2001).

Clustering techniques are frequently used to assign objects of analysis to groups or classes on the basis of criteria that can be adjusted on the basis of theory and experience. While classification into categories or groups may be the ultimate goal of these analytical efforts, a preliminary stage in the process may emphasize the discovery of patterns of association, or covariation. Unlike traditional statistical methods where specific tests are used to evaluate a hypothesis, statistics are used within data-mining efforts to generate the hypotheses that might be explored through other means, or through subsequent analyses of enhanced data bases (Carbone 2000).

It is also likely that the associative links which emerge through analysis will make even more sense when they are examined over time. Thus, insurers may decide to classify automobile accident claims as problematic on the basis of a determination of the temporal relationship between the calls the insured made to his doctor and those he made to his attorney.

As an area of applied statistics, data-mining involves the application of increasingly sophisticated techniques to larger and larger pools of data by more and more decision makers within organizations. This diffusion of analytical capacity has been enabled in part by increases in the diffusion of relatively inexpensive computing power to desktop and even to laptop computers (Darling 1997; Lawson 2002). Commonplace statistical resources used for prediction (or explanation in very polite company) include multiple regression and discriminant analysis.

More sophisticated techniques involve the use of so-called neural networks that can process data autonomously to generate clusters, and with sequential adjustment, develop fine-tuned predictive algorithms. Neural networks are said to provide advantages over traditional statistical techniques like multiple regression because assumptions about relations between data are not as restrictive. Unlike multiple regressions, neural network techniques are able to include a large number of variables that comes closer to reflecting both the complexity and the uncertainty that characterize routine business decisions (Garver 2002).

Neural networks or a related approach known as decision tree analysis may be used to identify the variables, or attributes of objects that are most important or influential in the prediction of some outcome. A relatively mundane example should suffice to illustrate the use of these techniques in the business environment.

Success in the competitive fast food environment is based in part upon superior knowledge about what governs consumers' decisions, and about what aspects of the experience produces a level of satisfaction high enough to increase the probability of repeat sales. Determining the relative importance of different attributes of pizza delivered to the home was the task at hand in one case study (Garver 2002). Semantic differential scaling techniques were used to capture customer ratings of aspects of the experience, including price, temperature of the pizza, and the 'friendliness' of the person delivering the product. While the neural network analysis and regression techniques both identified 'taste' as the most important factor, 'amount' only emerged as relatively important in the neural network analysis (Garver 2002).

Of course, only an experiment in which the model's parameters can be adjusted and tested against actual consumer behaviour would determine whether the hypotheses generated by the analyses win empirical support.

Data and the Real World

It is important to keep in mind that data-mining produces statements about the relationships between data as well as statements about the relationships between phenomena in the real world. Some analytical approaches common to data-mining efforts use insights derived from observed relationships between data to make statements about phenomena that have not been, and perhaps cannot be, observed directly (Gigerenzer et al. 1989; Smith 1998). Data mining is said to differ from ordinary information retrieval in that the information which is sought

does not exist explicitly within the database, but must be 'discovered' (Fulda 2000: 106). In an important sense, the relationships between measured or 'observed' events come to stand as a 'proxy' for some unmeasured relationship or influence (Baard 2002).

For example, factor analysis programs generate estimates of the correlation, or association, between measured variables and unmeasured factors or influences. In the area of media studies, numerous examples can be found of factor analytic models that draw inferences about unmeasured tastes and preferences within population segments on the basis of the structural relationships that can be observed between the programs that hundreds of people report viewing with varying degrees of regularity (Frank and Greenberg 1980; Webster, Phalen, and Lichty 2000).

A fairly sophisticated application of this technique has been described by Predictive Networks in relation to their 'Digital Silhouettes' product (Predictive Networks 2001). In addition to traditional demographic indicators that include gender, occupation, and race, the 'anonymous predictive models of individuals' includes ninety categories of inferred 'content affinities.' The 'Digital Silhouette' technology can be seen as an example of a product or service that continually updates the profiles of individual web users. The technology is promoted as a resource for advertisers interested in reaching a particular type of consumer. This ideal type is assumed to share common affinities and demographics, and on the basis of the client's specification, according to the promoters, 'Predictive Networks selects the Digital Silhouettes most likely to respond to the ad' (Predictive Networks 2001: 3). As in many such applications, the accuracy or precision with which individuals are described by these profiles is said to improve over time. The brochure claims that as 'the user continues to surf the web, his or her Digital Silhouette becomes a refined composite of the demographic and affinity distributions of all the web sites he or she has visited' (Predictive Networks 2001: 3).

It is important to note here that each individual's profile is shaped not only by her own behavioural attributes, but also by the program's characterization of the profiles of all the other individuals within the database who visit particular websites. As a result, these profiles are fundamentally relational, or comparative, rather than *individual* identities (Gandy 2000).

The application of data-mining techniques for predictive purposes uses data gathered in the past in order to generate descriptions about

events that may occur in the future. In the context of ordinary business applications, such predictions may be concerned about determining which individuals will respond affirmatively to a particular offer in the context of a particular marketing campaign (Adams, Hand, and Till 2001).

Applications in the realm of national security are a bit more specialized, but the underlying assumptions, resources, and techniques do not vary substantially, and both may be concerned with the development of behavioural profiles.

The Sources of Data

At the heart of public concerns about data-mining is the extent to which the technology invites the integration of data from previously independent data record systems. Concerns about the development of centralized government data banks expressed during early periods of concern about an emerging 'surveillance society' (Flaherty 1989) have been replaced by concerns about the sorts of 'virtual databases' enabled by high speed networking and distributed processing capacity (Solove 2001). Although different approximations of 'fair information practices' have sought to restrict the secondary use of information to the purposes for which it was initially gathered, there is an extremely limited record of success in this regard (Sovern 1999).

Heightened concern has been expressed about the enhanced capacity of government agencies to create virtual databases because of their recently increased access to public and private record systems. Both government agencies and private businesses have continued to utilize massive amounts of data from a host of sources, limited primarily by well-placed concerns about data quality and standardization (Gandy 2001).

The expansion in the pool of available data is being affected rather dramatically by changes in both technology and public policy. The growth in the number of databases under the administrative control of government agencies reflects more than the growth of government oversight and responsibility. It also reflects the impact of changes in the cost structure of data capture, storage, and transmission.

Record systems expand because the systems that enable the capture, storage, and retrieval of information have become more efficient, less expensive, and less demanding of specialized training. The fact that the data stored within these systems has also been determined to have

economic value beyond its primary function plays a less important role in determining whether and how the information will be stored by government agencies.

However, a somewhat different logic may explain the rapid growth in the size and scope of private databases. Indeed, it may be the case that the need to meet government demands for information related to taxation and regulatory oversight explains much of the initial impetus for the development of computerized record-keeping systems within the private sector. The discovery that there was additional economic and strategic value in transaction-generated information (TGI) is of fairly recent vintage. Hundreds of articles are being published each year designed to remind business managers, especially those involved in electronic commerce, of the value to be derived from mining the 'gold' within their own record systems, and dozens more compare various techniques and analytical products in terms of their efficiency and ease of use (Greening 2000).

The direct marketing industry, a primary consumer of TGI about individuals, is said to have evolved into its present form in response to the enhanced potential for target marketing that was made possible by the establishment of postal zip codes in the 1960s (Harper 1986). This form of marketing has since come to support a segment of the business services industry that includes list vendors, consolidators, and other data managers that bill customers and clients in excess of $3 billion each year (Solove 2001: 1408).

Many database vendors also provide clients with access to lists derived from government databases, including those derived from voter registration files. One vendor, Aristotle International, offered political campaign managers lists of voters enhanced with additional information about ethnicity, voting frequency, motor vehicle ownership, and estimates of household income in addition to data about whether the registrants owned, or rented their homes (Aristotle 2002; Pressman 2000).

A major competitor to Aristotle offered an online service that allowed clients to specify the sorts of voters they wanted to reach. The vendors of a product called VISION (Voter Information System Integrated Online Network), made what has emerged as an obligatory nod in the direction of concern about privacy before going on to describe the nature and benefits of the services they provide. Privacy concerns are supposed to be reduced because VISION clients are invited to specify the criterion they want used in generating a list of prospects, rather than making requests for information about individually identified persons.

In one scenario provided in their online brochure (I-Centrix 2002), a client might request a list of all males over forty-five who are registered Democrats, who have an income above $70,000, who have voted in past Presidential elections, and who also happen to live in the 3rd Electoral District of Troy, New York. If the customer is willing to pay the established price for the 125 voters who would be produced by this multivariate sort, pre-printed mailing labels as well as phone numbers might be supplied. The VISION service claimed to have on offer an additional three hundred variables that might be used to define the preferred class of voters to be targeted by a specialized appeal (I-Centrix 2002: 2).

Of course, there are no guarantees that the information acquired from database vendors or consolidators will be accurate. On occasion, the consequences of almost certain errors can be quite substantial. In the case of ChoicePoint, a database company that claims to be primarily in the 'credential verification' business, we find an organization that combines identification with classifications along a number of critical dimensions. Unfortunately for the nation, and perhaps for the world, a ChoicePoint subsidiary provided credential verification services for the county election boards in Florida. Many of their characterizations of American citizens were grossly incorrect. As a result, an estimated eight thousand potential voters, many of them African Americans, were listed as convicted felons, and thereby denied the right to vote in a hotly contested national election. While not all of those who were excluded on the basis of this classification error were likely to have voted for Al Gore, they are far less likely to have supported George W. Bush. In any event, their numbers far exceeded the margin of victory that history has been forced to record (Coppola 2002).

ChoicePoint is now emerging as one of the largest private forensic laboratories in the country. Derek Smith, ChoicePoint's chief executive, seems driven to add DNA and other biometric identifiers to the list of commercially available data about individuals they will use in their identification and certification services (Coppola 2002).

Although we might not ordinarily consider the providers of search engines as being very similar to the list vendors that charge fees for providing access to those lists, it is clear that search engines also provide access to the information in databases. The developers of search engines have invested in the development of quite sophisticated algorithms for searching and indexing the rapidly expanding caches of information stored on servers around the globe. Google is readily iden-

tified as a leader within this very dynamic industry. Its search engines support in excess of 150 million searches per day (Fraim 2002).

Following its acquisition of Deja.com, a major innovator in the exploitation of Usenet newsgroup archives, hundreds of millions of personally identified comments, critiques, and representations became available for search, analysis, and commercial exploitation by Google (Google 2001). Fraim (2002) suggests that Google's demonstrated ability to segment searches into national and linguistic groups can easily be expanded into more meaningful 'communities' that might include corporations, universities, and neighbourhoods defined by zip codes. The fact that Google uses 'cookie' technology in ways that link searches to searchers has raised serious privacy concerns, even though Google claims that they currently limit the identification of searchers to broad aggregates (Olsen and Kane 2003). Nevertheless, analysis of the word and subject rankings that could be based on the millions of searches initiated within and between these 'aggregates' has strategic potential that is not lost on data miners within business and government.

Businesses are particularly concerned about the ways in which their firms and products are 'positioned' in relation to other firms at any moment in time. They are concerned about their status and visibility not only among the population at large, but among particular segments of the population. The analysis of searches and commentary can provide them with continual assessments of their relative position.

Policy makers are always interested in knowing about the salience of particular concerns among the electorate. Like corporate strategists, they are concerned about the salience of issues and policy options among elite segments of the population that are more likely to respond to what they see as threats or opportunities (Ferguson 2000). These assessments are increasingly easy to produce with the aid of text-mining software. Indeed, an analysis of elite or expert discussions and debates ongoing in Usenet newsgroups is precisely the sort of application that text-mining products like MITRE's 'Top Cat' have been designed to support (Mani, Guilder, Clifton, and Concepcion 2000).

The U.S. *Patriot Act* (2001) provided the U.S. government with substantially improved access to data within record systems in government agencies that are not generally concerned with law enforcement or national security. The guidelines issued less than a year later for the FBI reflected the increased ease with which government agencies can expect to gather data from public and private sources (Center for Democracy and Technology 2002). Where previous guidelines restricted

agency investigations to cases where there is at a minimum some suspicion of criminal activity, the new guidelines reflect a greater appreciation of the ways in which data-mining techniques can be used 'as the basis for generating the suspicion of criminal conduct in the first place' (Center 2002: 1). Under the new guidelines, the FBI appears to be authorized to conduct 'fishing expeditions' on their own or with the support of commercial firms that are better equipped, and have far more experience in data-mining.

Although there are signs that some private corporations as well as librarians (Hull 2002) are resisting government attempts to gain access to the TGI in their files, there is also evidence that many corporations are voluntarily providing government agencies with access to information out of some heightened sense of patriotism. Of course, on occasion, this patriotic urge seems to coincide with corporate self-interest.

Reputedly acknowledged as the world's largest retailer, eBay.com has also been identified as an organization that is quite willing to share information about the transactions of its more than 60 million registered users if security personnel at the company have their suspicions aroused by the behaviour of buyers or sellers, or by the inquiries of government officials. Such arousal is likely because security personnel at eBay routinely use their own data-mining resources to continually search for patterns of suspicious behaviour that might indicate a risk of fraud.

Although eBay's 'director of law enforcement and compliance' reported that the company receives approximately two hundred requests for information about users each month, he noted that the overwhelming majority of these requests were informal, rather than official subpoenas or court orders (Dror 2003). While eBay doesn't provide unhindered access to its entire corporate database, which reportedly contains all the TGI the company has produced since 1995, a request from a law enforcement officer is likely to be rewarded with a complete history of a specific target's interactions with the eBay servers. This report may include the contributions that registrants may have made to the company's many discussion groups.

In those cases where there were explicit legal barriers to the sharing of financial information (Janger and Schwarts 2002), such as information that eBay gained access to through its acquisition of the PayPal transaction clearance service, security officials at eBay were reportedly quite willing to provide law enforcement with sufficient 'hints' about how to produce the appropriate court order (Dror 2003).

TIA and Data-mining Unleashed

Of course, eBay is not alone in its willingness to provide government investigators with information about customers that they believe might be useful in the identification of terrorists or other miscreants (Baard 2002; Wylie 2003). As one astute observer noted, 'It didn't take long for data-management companies to realize that if their software could find links in customer buying patterns and improve retailers' inventory decisions, perhaps it could find links among the government's vast terrorism-related intelligence warehouses and enhance the government's ability to prevent the next attack' (Franklin 2002).

Many entrepreneurs rushed to offer their services to government, and hundreds more responded to requests for proposals designed to meet the information needs of the Defense Advanced Research Projects Agency (DARPA) Total Information Awareness (TIA) program (Electronic Privacy Information Center 2003). According to the initial solicitation, 'DARPA is soliciting innovative research proposals in the area of information technologies that will aid in the detection, classification, identification, and tracking of potential foreign terrorists, wherever they may be, to understand their intentions, and to develop options to prevent their terrorist acts' (Defense Advanced Research Projects Agency [DARPA] 2002).

At the core of the research initiative was the development of 'ultra large all-source information repositories' that were later defined as 'large scale databases covering comprehensive information about all potential terrorist threats.' What DARPA was in search of was 'innovative technologies needed to architect, populate, and exploit such a database for combating terrorism.' After noting that DARPA already had 'on-going research programs aimed at language translation, information extraction from text' and other techniques, the agency made it clear that it was inviting 'new ideas for novel information sources and methods that amplify terrorist signatures and enable appropriate response.'

Because the successful proposals were classified, it was not possible to determine precisely what the twenty-six early 'winners' had promised to deliver. It is easy, however, to imagine the sorts of data-mining resources that will emerge from the Palo Alto Research Center (PARC), given its title: 'Knowledge-based tracking of content change in growing collections of text documents' (EPIC 2003).

Although the U.S. Senate managed to call a moratorium on the use of

data-mining by the departments of Defense and Homeland Security (Gross 2003) because of the unprecedented level of intrusion into the lives of average citizens that the program seemed to invite (Mack 2002), it is also clear that the underlying rationale for expanding the use of the technology has many supporters. Indeed, while a recent report on national security from the Markle Foundation emphasized the need to balance the value of strategic intelligence against the risks to civil liberties, there was little room for doubt that the members of the Task Force endorsed the widespread use of data-mining in support of national security (Markle Foundation 2002).

The Markle group (Markle Foundation 2002: 45) agreed that there was a need for 'an information collection and analysis strategy that will allow us to detect and prevent dangers from these unanticipated sources.' Specifically, the working group identified the sorts of tools that might be developed and deployed in an effort to anticipate, and thereby prevent, some 'terrorist' action in the future. In their view, theoretically driven scenarios could be assessed in relation to empirically generated probabilities that have been derived from data-mining applications. Indeed, from their perspective one 'advantage of computer-generated scenarios is that very large numbers could be created without the unintended constraints of human definitions of "likely" or "plausible," thus decreasing the chances of surprise through novelty' (Markle Foundation 2002: 52).

We are of course familiar with the strategic advantages that can be derived from the relaxation of traditional boundaries around the possible or imaginable that are based on naive concepts of reasonableness. On the other hand, it is important for us to evaluate the risks that flow from the design of intervention strategies that might be pursued if government and corporate planners are no longer constrained by traditional conceptions of 'fairness,' 'justice,' 'equality,' or 'legitimacy.' I have discussed these outcomes elsewhere with regard to myriad ills that accompany the use of TGI to facilitate discrimination (Danna and Gandy 2002; Gandy 2000, 2001). Among the most serious of these concerns are those related to the ways in which discrimination in information markets reinforces disparities in the level and impact of participation in the public sphere by segments of the population. These disparities are likely to become sharper still because of the ways in which the strategic resources provided by data-mining may also be used to shape the 'preferences' for information that individuals come to recognize and express as their own (Zarsky 2002/3).

The Limits of Privacy Policy

I have argued in the past that the traditional policy framework generally relied upon to protect individual and collective interests in privacy are hampered by an emphasis on the use of 'individually identified information' (Gandy 1993). Individuals are placed at risk of discrimination by virtue of their membership in groups, rather than specifically on the basis of their individual identities (Gandy 1996, 2000). This 'categorical vulnerability' (Gandy 2001) is especially problematic in the context of a degraded public sphere because individuals may not be aware of the groups to which they have been assigned by various data mining analyses (Bruening and Schwartz 2002). Indeed, it is the tendency among those who engage in data mining and related approaches to the development of behavioural profiles to keep those profiles secret that limits the utility of the traditional privacy torts. Because of the assumption that information about individuals is made public on a widespread and routine basis, the interests of individuals in such information is minimized at the same time that corporate and government interest in the strategic intelligence generated through analysis of such data are protected by security classification and intellectual property regimes (Belgum 1999; Hatch 2001).

In the case of corporate actors, short-term strategic advantage is realized in part by the ability of competitors to treat such classificatory schemes as trade secrets (Montana 2001). For somewhat different reasons, government agencies that develop profiles that they believe will be useful in identifying smugglers, terrorists, or other subjects of interest tend to resist attempts by the public or civil libertarians to gain access to the details about the composition of these profiles.

Critics have suggested that the use of profiles as alternatives to traditional markers of stigmatized identity (Loury 2002) makes it possible for an increase in discriminatory acts that would be declared illegal if they relied solely or primarily on the use of 'suspect categories' like race, gender, or national origin (Belgum 1999). Because these meaningful categories are submerged or replaced by coefficients assigned to variables or explanatory factors, the names that have been assigned to these factors are less likely to attract the heightened scrutiny of the courts.

It should be clear, however, that these profiles, or 'signatures' of group membership (McCarthy 2000), have much in common with traditional stereotypes or schema that have been developed over time for

subordinate groups. However, in the case of racial stereotypes, we have evidence that the targets of stereotypic thinking have a high degree of awareness of the ways in which they have been characterized and evaluated by those higher in the social hierarchy (Sigelman and Tuch 1997). It seems unlikely, however, that very many of the targets of data mining know much about the ways in which they are being classified. They are largely unaware of the ways in which they are the victims or the beneficiaries of such remote sensing.

In the absence of knowledge about the groups to which they have been assigned, individuals are less likely to engage in political organization or in the development of social movement organizations (Eskridge 2001). This is the case despite the suggestion that computer networking would make such organizational efforts less difficult to develop and sustain (Hula 1999; Kreimer 2001).

I suspect that it will be difficult to mount a sustained challenge to the use of data-mining and other discriminatory techniques that are at the heart of what I refer to as 'the panoptic sort' (Gandy 1993). The best we can hope for is transparency – that citizens and consumers are routinely and effectively informed about the nature and extent to which their options and opportunities are being shaped by the application of statistical techniques.

It seems unreasonable to expect that those who use these techniques will be the best sources of public awareness of the consequences of their use. Therefore it will be particularly important for journalists, scholars, and advocates of informed choice to be able to convey a sense of the array of individual and collective risks that flow from the use of data-mining and other discriminatory techniques.

REFERENCES

Adams, N.M., D.J. Hand, and R.J. Till. 2001. 'Mining for Classes and Patterns in Behavioural Data.' *Journal of the Operational Research Society* 52:1017–24.

Aristotle International, Inc. 2002. US voter lists. Aristotle International, Inc. http://www.aristotleonline.com/voterlists.html.

Baard, E. 2002. 'Your Grocery List Could Spark a Terror Probe: Buying Trouble' *Village Voice*. 30 July. http://www.oas.villagevoice.com/issues/0230/baard.html.

Baker, K., and S. Baker. 1999. 'Divide and Conquer.' *Journal of Business Strategy* 20:16–18.

Belgum, K.D. 1999. 'Who Leads at Half-time?: Three Conflicting Visions of Internet Privacy Policy.' *Richmond Journal of Law and Technology* 6:1.
Bowker, G.C., and S. Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.
Bruening, P., and A. Schwartz. 2002. 'Privacy and Online Politics: Is Online Profiling Doing More Harm Than Good for Citizens in Our Political System?' Washington, DC: George Washington University Graduate School of Political Management. http://www.democracyonline.org.
Carbone, P. 2000. 'What Is the Origin of Data Mining?' *The Edge* 4:2–12.
Center for Democracy and Technology. 2002. *CDT's Analysis of New FBI guidelines: Center for Democracy and Technology*. http://www.cdt.org/wiretap/020530guidelines.shmtl.
CFO Research Services. 2002. *Mining the Value in CRM Data: A CFO Perspective*. Boston, MA: CFO Publishing Corporation. http://www.cfo.com.
Computer Science and Telecommunications Board. 2000. *Summary of a Workshop on Information Technology Research for Federal Statistics*. Washington, DC: National Academy Press.
– 2002. *Cyber-security Today and Tomorrow: Pay Now or Pay Later*. Washington, DC: National Academy Press.
Coppola, V. 2002. 'Derek Smith's Brave New World.' In *Georgia Trend* ChoicePoint, Inc. http://www.choicepoint.net/news/gatrend.html.
Danna, A., and O.H. Gandy Jr. 2002. 'All That Glitters Is Not Gold: Digging Beneath the Surface of Data mining.' *Journal of Business Ethics* 40:373–86.
Darling, C.B. 1997. 'Datamining for the Masses.' *Datamation*. (Feb.): 52–4.
Defense Advanced Research Projects Agency (DARPA). 2002. *BAA 02–08 Information Awareness Proposer Information Pamphlet*. Washington, DC: U.S. Department of Defense.
Dror, Y. 2003. 'Big Brother Is Watching You – And Documenting: EBay, Ever Anxious to Up Profits, Bends over Backwards to Provide Data to Law Enforcement Officials.' Haaretz.com. http://www.haaretz.com.
Electronic Privacy Information Center (EPIC). 2003. *EPIC Analysis of Total Information Awareness Contractor Documents*. Washington, DC: Electronic Privacy Information Center. http://www.epic.org/privacy/profiling/tia/doc_analysis.html.
Eskridge, W.N. Jr. 2001. 'Channelling: Identity-based Social Movements and Public Law.' *University of Pennsylvania Law Review* 150:419–525.
Federal Trade Commission, U.S. 2000. *Online Profiling: A Report to Congress*. Washington, DC: U.S. Federal Trade Commission, Bureau of Consumer Protection.

Ferguson, S.D. 2000. *Researching the Public Opinion Environment: Theories and Methods*. Thousand Oaks, CA: Sage.
Flaherty, D.H. 1989. *Protecting Privacy in Surveillance Societies: The Federal Republic of Germany, Sweden, France, Canada, and the United States*. Chapel Hill: University of North Carolina Press.
Fraim, J. 2002. 'Electronic Symbols: Internet words and Culture.' June 2002. http://www.firstmonday.org/issues/issue7_6/fraim/index.html.
Frank, R.E., and M.G. Greenberg. 1980. *The Public's Use of Television*. Beverley Hills, CA: Sage.
Franklin, D. 2002. 'Data Miners: New Software Instantly Connects Key Bits of Data That Once Eluded Teams of Researchers.' *Time.com*, 15 Dec. http://www.time.com/globalbusiness/printout/0,8816,400017,00.html.
Fulda, J. 2000. 'Data mining and Privacy.' *Albany Law Journal of Science and Technology* 11:105–13.
Gandy, O.H. Jr. 1993. *The Panoptic Sort: A Political Economy of Personal Information*. Boulder, CO: Westview Press.
– 1996. 'Legitimate Business Interest: No End in Sight? An Inquiry into the Status of Privacy in Cyberspace.' *University of Chicago Legal Forum 1996*: 77–137.
– 2000. 'Exploring Identity and Identification in Cyberspace.' *Notre Dame Journal of Law, Ethics and Public Policy* 14:1085–1111.
– 2001. 'Dividing Practices: Segmentation and Targeting in the Emerging Public Sphere.' In W.L. Bennet and R.M. Entman, eds., *Mediated Politics: Communication in the Future of Democracy*, 141–59. New York: Cambridge University Press.
Garver, M.S. 2002. 'Try New Data-mining Techniques.' *Marketing News* (16 Sept.): 31–3
Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Kruger. 1989. *The Empire of Chance: How Probability Changed Science and Everyday Life*. New York: Cambridge University Press.
Google, Inc. 2001. 'Google Acquires Usenet Discussion Service and Significant Assets from Deja.com.' http://www.google.com/press/pressrel/pressrelease48.html
Greening, D.R. 2000. 'Data mining on the Web. There's Gold in That Mountain of Data.' *New Architect*. http://www.webtechniques.com/archives/2000/01/greening.
Gross, G. 2003. 'US Senate Blocks Government Data-mining Plan.' *InfoWorld Daily News*. 24 Jan. Washington, DC. http://www.infoworld.com/article/03/01/24/030124hntia_1.html.

Harper, R. 1986. *Mailing List Strategies*. New York: McGraw-Hill.
Hatch, M. 2001. 'Electronic Commerce in the 21st Century: The Privatization of Big Brother: Protecting Sensitive Personal Information from Commercial Interests in the 21st Century.' *William Mitchell Law Review* 27:1457–1502.
Helling, B. 1998. 'Web-site Sensitivity to Privacy Concerns: Collecting Personally Identifiable Information and Passing Persistent Cookies.' *First Monday*. http://www.firstmonday.dk/issues/issue3_2/helling/index.html.
Hobart, M.E., and Z.S. Schiffman. 1998. *Information Ages: Literacy, Numeracy, and the Computer Revolution*. Baltimore: Johns Hopkins University Press.
Hula, K.W. 1999. *Lobbying Together: Interest Group Coalitions in Legislative Politics*. Washington, DC: Georgetown University Press.
Hull, D. 2002. 'Libraries Face Privacy Test: Anti-terrorism Law Used to Demand Records.' *Mercury News*, 18 Oct. http://www.bayarea.com/mid/mercurynews/news/local/4312790/html.
I-Centrix. 2002. 'Voter Information System Integrated Online Network.' I-Centrix http://www.i-centrix.com/vision.html.
Janger, E.J., and P.M. Schwartz. 2002. 'The Gramm-Leach-Bliley Act, Information Privacy, and the Limits of Default Rules.' *Minnesota Law Review* 86:1219–61.
Kreimer, S.F. 2001. 'Technologies of Protest: Insurgent Social Movements and the First Amendment in the Era of the Internet.' *University of Pennsylvania Law Review* 150:119–71.
Lawson, N.J. 2002. 'The Emperor's New Clothes: Data mining, the Real Key to the Kingdom.' *Assessment Journal* (May/June): 49.
Leahy, P., R. Feingold, and M. Cantwell. 2003. Letter to Attorney General John Ashcroft. 10 Jan. U.S. Senate, Committee on the Judiciary.
Leyshon, A., and N. Thrift. 1999. 'Lists Come Alive: Electronic Systems of Knowledge and the Rise of Credit Scoring in Retail Banking.' *Economy and Society* 28: 434–66.
Loury, G.C. 2002. *The Anatomy of Racial Inequality*. Cambridge, MA: Harvard University Press.
Lyon, D. 2003. 'Surveillance as Social Sorting: Computer Codes and Mobile Bodies.' In D. Lyon, ed., *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, 13–30. New York: Routledge.
Mack, G. 2002. *Total Information Awareness Program (TIA) System Description Document (SDD)*, 1-150. Washington, DC: Defense Advanced Projects Research Agency
Mani, I., L. Van Guilder, C. Clifton, and K. Concepcion. 2000. 'Text Mining by Filter Composition.' *The Edge* 4: 1–4.
Markle Foundation, Task Force on National Security in the Information Age. 2002. *Protecting America's Freedom in the Information Age*. New York: Markle Foundation.
McCarthy, J. 2000. 'Phenomenal Data mining.' *Communications of the ACM* 43: 75
Mittelstadt, M. 2002. '"Orwellian" Data-mining Program Draws Fire from All Sides.' *Dallas Morning News*, 16 Dec. (accessed online, Lexis/Nexis, 13 Feb. 2003).
Montana, J. 2001. 'Data mining: A Slippery Slope.' *Information Management Journal* 35:50.
Moscoso, E., and G. Edmonson. 2003. 'Pentagon "Data mining" Program Fans PrivacyFears.' *Cox News Service*. 18 Jan. (accessed online, Lexis/Nexis, 13 Feb. 2003).
Nier, C.L. 1999. 'Perpetuation of Segregation: Toward a New Historical and Legal Interpretation of Redlining under the Fair Housing Act.' *John Marshall Law Review* 32:617–65.
Norusis, M.J. 1994. *SPSS Advanced Statistics 6.1*. Chicago, IL: *SPSS*.
Olsen, S., and M. Kane. 2003. 'Google Triggers Privacy Concerns.' *CNET News.com*. 3 Mar. http://znet.com/2100-1104-990724.html.
Phillips, D., and M. Curry. 2003. 'Privacy and the Phenetic Urge: Geodemographics and the Changing Spatiality of Local Practice.' In D. Lyon, ed., *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, 137–52. New York: Routledge
Pollack, M. 2001. 'Opt-in Government: Using the Internet to Empower Choice.' *Catholic University Law Review* 50:653–702.
Predictive Networks, Inc. 2001. 'Digital Silhouettes.' http://www.predictivenetworks.com
Pressman, A. 2000. 'Voters for Sale.' *The Industry Standard*. http://www.thestandard.com/article/0,1902,19864,00.html.
Roberts-Witt, S.L. 2001. 'Gold Diggers.' *PC Magazine*. http://www.pcmag.com/print_article/0,3048a%253D5356,00.asp
Rothleder, N. 2000. 'Data mining for IntrusionDetection.' *The Edge* 4:10–11.
Sigelman, L., and S.A. Tuch. 1997. 'Metastereotypes: Blacks' Perceptions of Whites' Stereotypes of Blacks.' *Public Opinion Quarterly* 61:87–101.
Skok, G. 2000. 'Establishing a Legitimate Expectation of Privacy in Clickstream Data.' *Michigan Telecommunication and Technology Law Review* 6:61–88.
Smith, B.C. 1998. *On the Origin of Objects*. Cambridge, MA: MIT Press.
Solove, D.J. 2001. 'Privacy and Power: Computer Databases and Metaphors for Information Privacy.' *Stanford Law Review* 43:1393–1462.
Sovern, J. 1999. 'Opting In, Opting Out, or No Options at All: The Fight for Control of Personal Information.' *Washington Law Review* 74:1033–1118.

Stepanek, M. 2000. 'Weblining.' *Business Week E.Biz.* 3 Apr.: EB26–33.
Stevens, L. 2001. 'It Sharpens Data mining's Focus – Instead of Building Data mining Applications with No Clear Goal, Companies Are Setting Priorities Up Front to Maximize ROI.' *Internet Week* (6 Aug.): 29.
U.S. *Patriot Act*. 2001. Public Law No. 107-56, 115 Stat. 272.
Webster, J.G., P.F. Phalen, and L.W. Lichty. 2000. *Ratings Analysis: The Theory and Practice of Audience Research*. Mahwah, NJ: Lawrence Erlbaum Associates.
Wylie, M. 2003. 'Homeland Security; With Marketers, Not Much Remains Private; Companies Own Data That Government Wants to Identify Terrorists.' *San Diego Union-Tribune* (5 Jan.): A1.
Zarsky, T.Z. 2002/3. '"Mine Your Own Business!": Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion.' *Yale Journal of Law & Technology* 5:1–55.