Information
and Privacy
Commissioner/
Ontario

Data Mining:
Staking a Claim on Your Privacy

Ann Cavoukian, Ph.D.
Commissioner
January 1998

Information and Privacy Commissioner/Ontario

2 Bloor Street East
Suite 1400
Toronto, Ontario
M4W 1A8

416-326-3333
1-800-387-0073
Fax: 416-325-9195
TTY (Teletypewriter): 416-325-7539
Website: www.ipc.on.ca

This publication is also available on the IPC website.

Table of Contents

Preface

Globally, issues about informational privacy in the marketplace have emerged in tandem with the dramatic and escalating increase in information stored in electronic formats. Improvements and innovations in computer processing power, disk storage, and networks have been close to explosive. Very large databases with transactional information about every aspect of business are now measured in gigabytes and terabytes. Fuelled daily by massive amounts of data, the data volume is so huge that it has been estimated that businesses can only use seven per cent[1] of the data collected.

Much of this large mass of data is donated by the consumer in the course of conducting his or her daily personal business: withdrawing cash from ATMs; paying with debit or credit cards; using loyalty cards; borrowing money; writing cheques; renting a car or a video; making a telephone call or an insurance claim; and, increasingly, sending or receiving e-mail and surfing the Net. Since virtually all of these transactions or activities involve some form of electronic identification, each transaction captures some personal information about you and stores it in electronic form.

Speed, convenience, easy access, discounts, bonuses, awards, frequent flyer points — all of these have encouraged or eased the transition from social interaction to electronic interaction. Often, there is no longer a choice to be made, or if there is a choice, it will rarely match the speed, convenience, or, in a way — the *sense* of control, one gains through an electronic interaction.

The sharpening of the competitive edge to improve products and services now demands that businesses make sense of complex and voluminous data. This enables businesses to design effective sales campaigns, precision targeted marketing plans, and develop products to increase sales and profitability. In this context, a technology called "data mining" can be a valuable tool for business because it provides for the "efficient discovery of valuable, non-obvious information from a large collection of data."[2]

Although data mining can be extremely valuable for businesses, it can also, in the absence of adequate safeguards, jeopardize informational privacy. For this reason, the Office of the Information and Privacy Commissioner (IPC) has produced this report on data mining.

The report is aimed primarily at consumers and businesses. It covers the following: What is Data Mining; Examples of Data Mining; The Implications of Data Mining in the Context of Fair Information Practices; and Consumers and Businesses: Choices to Consider. (The "fair information practices" discussed refer to the basic principles of data protection first established in 1980 by the Organisation for Economic Co-operation and Development).

Generally speaking, we think that responsible data management in private sector businesses must be firmly based on fair information practices. The protection of personal information can be enhanced if: (1) consumers *choose* to voice and act on their expectations about the privacy of their personal information to the businesses with which they are transacting; and (2) businesses *choose* to adopt a culture of privacy through tangible everyday practices and through the use of privacy enhancing technologies. It comes down to a matter of choice for both consumers *and* for businesses.

Ultimately though, the IPC sees the need for government, businesses and consumers to share the responsibility in the management of the collection, use, retention and disclosure of personal information held by private sector businesses. We support a shared responsibility approach that is codified through government enactment of data protection legislation for private sector businesses, sustained by the business community adopting a culture of privacy; and strengthened by consumers taking greater control of their own personal information and voicing their privacy expectations to the business community.

Backdrop

The transformation of information storage from paper-based records to electronic formats has contributed to the mounting attention and concern about informational privacy in the marketplace. Two examples illustrate this point: 1) Canadian surveys show that increasingly, people are concerned about their privacy and the prospect of being watched;[3] and, 2) *Time* magazine pronounced the "The Death of Privacy" (cover story Canadian edition, August 25, 1997).

While we believe that *Time's* pronouncement to be somewhat exaggerated, it is true that the information age now upon us is bringing with it recurrent horror stories about loss of privacy and *dataveillance*, the exploits of hackers and breaches of security, and identity theft — the impersonation and fraudulent use of your identity by another. Certainly the Internet and electronic commerce have heightened concerns over privacy and security issues that were unthought of previously:

> As we move toward a more fully digital world, the cost of manipulating information approaches zero, and the hazards therein multiply. Even our privacy is in peril. The "clickstream" pouring into Web merchants — the information that you provide with clicks of your mouse … what music you listen to and where you like to eat — lets those merchants personalize their marketing, but it may be more information than you want to share widely. And some Web entrepreneurs collect this information and sell it. Supermarket scancards may be more convenient than coupons, but … they, too, "put a price on privacy." The activities in these examples are perfectly legal, of course, but they increase the potential for electronic malfeasance.[4]

Within this environment of a booming information economy, bringing with it a new set of challenges to data protection, two noteworthy responses have emerged in Canada:

- The Canadian Standards Association (CSA) released its *Model Code for the Protection of Personal Information* (the Code) in March 1996 — this Code, although voluntary, provides a national standard for the protection of personal information in non-government organizations. It is likely that the Code will form the foundation of any future data protection legislation in Canada covering the private sector. (Quebec is the only province in Canada that has legislation that sets out fair information practices for businesses operating in that province).

- In September 1996, the Federal Government announced its commitment to the introduction of privacy legislation covering the private sector by the year 2000. Drafting of the bill is said to be underway.

"It has been estimated that the amount of information in the world doubles every 20 months, and the size and number of databases are increasing even faster."[5] Thus, it is with some sense of urgency that we have prepared this report on data mining with the view that, an entrenched culture of privacy in the business world will only come about if consumers speak up and convey their privacy expectations to businesses, and, if businesses truly believe that privacy protection makes good business sense.

What is Data Mining?

Data mining is a set of automated techniques used to extract buried or previously unknown pieces of information from large databases. Successful data mining makes it possible to unearth patterns and relationships, and then use this "new" information to make proactive knowledge-driven business decisions. Data mining then, "centres on the automated discovery of new facts and relationships in data. The raw material is the business data, and the data mining algorithm is the excavator, sifting through the vast quantities of raw data looking for the valuable nuggets of business information."[6]

Data mining is usually used for four main purposes: (1) to improve customer acquisition and retention; (2) to reduce fraud; (3) to identify internal inefficiencies and then revamp operations, and (4) to map the unexplored terrain of the Internet.[7] The primary types of tools used in data mining are: neural networks, decision trees, rule induction, and data visualization.

Although not an essential prerequisite, data mining potential can be enhanced if the appropriate data have been collected and stored in a *data warehouse* — a system for storing and delivering massive quantities of data. "Data warehousing is the process of extracting and transforming operational data into informational data and loading it into a central data store or warehouse."[8] The promise of data warehousing is that data from disparate databases can be consolidated and managed from one single database.

The link between data mining and data warehousing is explained as follows:

> Data Warehousing is the strategy of ensuring that the data used in an organization is available in a consistent and accurate form wherever it is needed. Often this involves the replication of the contents of departmental computers in a centralized site, where it can be ensured that common data definitions are in the departmental computers in a centralized site, where it can be ensured that the common data definitions are in use… The reason Data Warehousing is closely connected with Data Mining is that when data about the organization's processes becomes readily available, it becomes easy and therefore economic[al] to mine it for new and profitable relationships.[9]

Thus, data warehousing introduces greater efficiencies to the data mining exercise. "Without the pool of validated and scrubbed data that a data warehouse provides, the data mining process requires considerable additional effort to pre-process the data."[10] Notwithstanding, it is also possible for companies to obtain data from other sources via the Internet, mine the data, and then convey the findings and new relationships internally within the company via an Intranet.[11]

There are four stages in the data warehousing process:

> The first stage is the acquisition of data from multiple internal and external sources and platforms. The second stage is the management of the acquired data in a central, integrated repository. Stage three is the provision of flexible access, reporting and analysis tools to interpret selected data. Finally, stage four is the production of timely and accurate corporate reports to support managerial and decision-making processes.[12]

Though the term data mining is relatively new, the technology is not. Many of the techniques used in data mining originated in the artificial intelligence research of the 80s and 90s. It is only more recently that these tools have been applied to large databases. Why then are data mining and data warehousing mushrooming now? IBM has identified six factors that have brought data mining to the attention of the business world:

1. A general recognition that there is untapped value in large databases;
2. A consolidation of database records tending toward a single customer view;
3. A consolidation of databases, including the concept of an information warehouse;
4. A reduction in the cost of data storage and processing, providing for the ability to collect and accumulate data;
5. Intense competition for a customer's attention in an increasingly saturated marketplace;
6. The movement toward the de-massification of business practices.[13]

With reference to point six above, "de-massification" is a term originated by Alvin Toffler. It refers to the shift from mass manufacturing, mass advertising and mass marketing that began during the industrial revolution, to customized manufacturing, advertising and marketing targeted to small segments of the population.

There are three basic steps in data mining:

> The first processing step is data preparation, often referred to as "scrubbing the data." Data is selected, cleansed, and preprocessed under the guidance and knowledge of a domain expert. Second, a data mining algorithm is used to process the prepared data, compressing and transforming it to make it easy to identify any latent valuable nuggets of information. The third phase is the data analysis phase where the data mining output is evaluated to see if additional domain knowledge was discovered and to determine the relative importance of the facts generated by the mining algorithms.[14]

Data mining differs from other analytical tools in the approach used in exploring the data relationships. Traditional database queries can answer questions like "what were my sales in Kenora in 1996?" Other analyses, often called multidimensional or online analytical processing, allow users to do more complex queries, such as comparing sales relative to plan by quarter and region for the prior two years.[15] In both cases, however, the results are simply figures extracted from the data or an aggregate of existing data. The relationship among these data is already known to the user, who, by framing the proper question, obtains the desired answer.

Data mining however, uses discovery-based approaches in which pattern-matching and other algorithms are used to discover key relationships in the data, previously unknown to the user.

> The discovery model is different because the system automatically discovers information hidden in the data — the data is sifted in search of frequently occurring patterns, trends, and generalisations about the data without intervention or guidance from the user… An example of such a model is a bank database which is mined to discover the many groups of customers to target for a mailing campaign. The data is searched with no hypothesis in mind other than for the system to group the customers according to the common characteristic found.[16]

Data mining usually yields five types of information — associations, sequences, classifications, clusters, and forecasting:

> Associations happen when occurrences are linked in a single event. For example, a study of supermarket baskets might reveal that when corn chips are purchased, 65% of the time cola is also purchased, unless there is a promotion, in which case cola is purchased 85% of the time.
>
> In sequences, events are linked over time. [For example][I]f a house is bought, then 45% of the time a new oven will be bought within one month and 60% of the time a new refrigerator will be bought within two weeks.
>
> Classification is probably the most common data mining activity today… Classification can help you discover the characteristics of customers who are likely to leave and provide[s] a model that can be used to predict who they are. It can also help you determine which kinds of promotions have been effective in keeping which types of customers, so that you spend only as much money as necessary to retain a customer.
>
> Using clustering, the data mining tool discovers different groupings with the data. This can be applied to problems as diverse as detecting defects in manufacturing or finding affinity groups for bank cards.

> All of these applications may involve predictions, such as whether a customer will renew a subscription … [f]orecasting, is a different form of prediction. It estimates the future value of continuous variables — like sales figures — based on patterns within the data.[17]

Generally then, applications of data mining can generate outputs such as:

- Buying patterns of customers; associations among customer demographic characteristics; predictions on which customers will respond to which mailings;
- Patterns of fraudulent credit card usage; identities of "loyal" customers; credit card spending by customer groups; predictions of customers who are likely to change their credit card affiliation;
- Predictions on which customers will buy new insurance policies; behaviour patterns of risky customers; expectations of fraudulent behaviour;
- Characterizations of patient behaviour to predict frequency of office visits.

As indicated above, data mining applications can be used in a variety of sectors: retail, finance, manufacturing, health, insurance, and utilities. Therefore across all sectors — if a business has data about its customers, suppliers, products, or sales, it can benefit from data mining. It is expected that data mining will be one of the greatest tools to be used by the business community in the next century as its ability to capitalize on the use of an already existing resource — information, becomes widely recognized, and the cost of data mining software goes down.

With regard to customers, the types of data that are needed to perform data mining applications are: 1) demographics, such as age, gender and marital status; 2) economic status, such as salary, profession and household income; and, 3) geographic details, such as city, street, province, rural/urban. All of these data types can be used to delineate particular sets or segments of customers that share similar interests and have common product requirements.

Examples of Data Mining

It has been estimated that the data mining market will reach more than $800 million by the year 2000.[18] The Gartner Group predicted that by the end of 1997, approximately 80 per cent of the Global 2000 (the world's largest 2,000 companies) will have or will be planning a data warehouse strategy that will likely incorporate data mining.[19]

By the year 2000, at least half of the Fortune 1000 companies worldwide will be using data mining.[20] Not surprising, when you think of the potential benefits to the businesses using various applications. Take, for example, the ability to scour data from multiple databases to predict future trends and behaviours: Blockbuster Entertainment uses it to recommend video rentals to individual customers;[21] American Express uses it to suggest products to its cardholders based on an analysis of their monthly spending patterns.[22]

MasterCard International uses it to extract statistics about its millions of daily cardholder transactions; furthermore, MasterCard plans to sell "a data warehouse of those transactions to its 20,000 business partners — banks and other companies, such as Shell Oil, that offer credit-card services."[23]

The Internet is also becoming an emerging frontier for data mining. Some technology companies provide "virtual" data mining services via the Internet. With access to an Internet server, it is possible to FTP (file transfer protocol) the data from the client's server and then conduct various data mining activities. (Alternately, if the client does not have access to an Internet server or if the data are too sensitive or voluminous, the data mining services can occur when the client provides a computer tape).[24]

Internet websites can be a further source of data for companies who want to know more about visitors to their own websites. For example:

> The Chicago Tribune Co. publishes a variety of services on the Web and on America Online Inc, … many of which are focused on classified marketing. The Chicago Tribune uses data mining to analyze customer behaviour as they move through its various sites.[25]

WalMart is often described as a pioneering leader in data mining and data management:

> WalMart captures point-of-sale transactions from over 2,900 stores in six countries and continuously transmits this data to its massive 7.5 terabyte data warehouse. WalMart allows more than 3500 suppliers to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities.[26]

Other companies supplement their customers' transactional information with external data such as postal codes to do a market basket analysis:

> Practically every retailer now records all the details of each POS (Point of Sale) transaction for stock keeping purposes. Sometimes these are supplemented by customer information. *Home Depot*, for example supplements the data with ZIP or postal code of the purchaser. Sometimes the cashier may also enter the sex and appropriate age of the customer into the cash register. Affinity cards and credit card numbers can be used to track repeat customers. *Market Basket Analysis* is the analysis of the data that this generates with a view to improving the performance of the retail outlet.[27]

Another example of what data mining can do involves the directed targeting of customers for new products, at a fraction of the cost:

> A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small test mailing, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.[28]

In the health care field, data mining applications are growing quickly. Applications can be used to directly assist practitioners in improving the care of patients by determining optimal treatments for a range of health conditions. Data mining is used to assist caregivers to distinguish patients who are statistically at risk for certain health problems so that those patients can be treated before their conditions worsen. Data mining can also be used to detect possible fraudulent behaviours of health providers as well as health service claimants. For example, patterns of care indicating that a particular practitioner is ordering too many diagnostic tests or conducting tests that are inappropriate may be identified through data mining; similarly, patterns, associations and overpayments for claims made by patients can be discovered through this process.

As much of the literature suggests, however, data mining is not a magic bullet nor a simple process. It also presents challenges that go well beyond the technical:

> Many data management challenges remain, both technical and societal. Large online databases raise serious societal issues. To cite a few of the societal issues: Electronic data interchange and data mining software make it relatively easy for a large organization to track all of your financial transactions. By doing that, someone can build a very detailed profile of your interests, travel, and finances. Is this an invasion of your privacy? Indeed, it is possible to do this for almost everyone in the developed world. What are the implications of that?[29]

In the next section we will explore the implications of data mining in the context of a set of principles designed to protect and guide the uses of personal information, commonly referred to as "fair information practices."

The Implications of Data Mining in the Context of Fair Information Practices

Around the world, virtually all privacy legislation, and the policies, guidelines, or codes of conduct used by non-government organizations, have been derived from the set of principles established in 1980 by the Organisation for Economic Co-operation and Development (OECD). These principles are often referred to as "fair information practices," and cover eight specific areas of data protection (or informational privacy). These are: (1) Collection Limitation; (2) Data Quality; (3) Purpose Specification; (4) Use Limitation; (5) Security Safeguards; (6) Openness; (7) Individual Participation; and (8) Accountability.

Essentially, these eight principles of data protection or fair information practices codify how personal data should be protected. At the core of these principles is the concept of personal control — the ability of an individual to maintain some degree of control over the use and dissemination of his or her personal information.

Concerns about informational privacy generally relate to the manner in which personal information is collected, used and disclosed. When a business collects information without the knowledge or consent of the individual to whom the information relates, or uses that information in ways that are not known to the individual, or discloses the information without the consent of the individual, informational privacy may be violated.

Data mining is a growing business activity, but from the perspective of fair information practices, is privacy in jeopardy? To determine this, we reviewed data mining from a fair information practices perspective. As discussed below, we have identified issues with five of these principles.

Data Quality Principle

Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete, and up-to-date.

Any form of data analysis is only as good as the data itself. Data mining operations involve the use of massive amounts of data from a variety of sources: these data could have originated from old, current, accurate or inaccurate, internal or external sources. Not only should the data be accurate, but the accuracy of the data is also dependent on the input accuracy (data entry), and the steps taken (if in fact taken), to ensure that the data being analyzed are indeed "clean."

This requires a data mining operation to use a good data cleansing process to clean or scrub the data before mining explorations are executed. Otherwise, information will be inaccurate, incomplete or missing. If data are not properly cleansed, errors, inaccuracies and omissions will continue to intensify with subsequent applications. Above all else, consumers will not be in a position to request access to the data or make corrections, erasures or deletions, if, in the first instance, the data mining activities are not known to them.

Purpose Specification Principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

Use Limitation Principle

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with the Purpose Specification Principle except: a) with the consent of the data subject, or b) by the authority of law.

Purpose Specification means that the type of personal data an organization is permitted to collect is limited by the purpose of the collection. The basic rule is that data collected should be relevant and sufficient, but not excessive for the stated purpose. In other words, restraint should be exercised when personal data are collected. *Use Limitation* means that the purpose specified to the data subject (in this case, the consumer) at the time of the collection restricts the use of the information collected. Hence, the information collected may only be used for the specified purpose unless the data subject has provided consent for additional uses.

Data mining techniques allow information collected for one purpose to be used for other, secondary purposes. For example, if the primary purpose of the collection of transactional information is to permit a payment to be made for credit card purposes, then using the information for other purposes, such as data mining, without having identified this purpose before or at the time of the collection, is in violation of both of the above principles. The primary purpose of the collection must be clearly understood by the consumer and identified at the time of the collection. Data mining, however, is a secondary, future use. As such, it requires the explicit consent of the data subject or consumer.

The *Use Limitation Principle* is perhaps the most difficult to address in the context of data mining or, indeed, a host of other applications that benefit from the subsequent use of data in ways never contemplated or anticipated at the time of the initial collection. Restricting the secondary uses of information will probably become the thorniest of the fair information practices to administer, for essentially one reason: at the time these principles were first developed (in the late 70s), the means by which to capitalize on the benefits and efficiencies of multiple uses of data were neither widely available nor inexpensive, thus facilitating the old "silo" approach to the storage and segregated use of information.

With the advent of high speed computers, local area networks, powerful software techniques, massive information storage and analysis capabilities, neural networks, parallel processing, and the explosive use of the Internet, a new world is emerging. Change is now the norm, not the exception, and in the quickly evolving field of information technology, information practices must also keep pace, or run the risk of facing extinction. Take, for example, the new directions being taken intending to replace the information "silos" of old, with new concepts such as "data integration" and "data clustering." If privacy advocates do not keep pace with these new developments, it will become increasingly difficult to advance options and solutions that can effectively balance privacy interests *and* new technology applications. Keeping pace will enable us to continue as players in this important arena, allowing us to engage in a meaningful dialogue on privacy and future information practices.

The challenge facing privacy advocates is to address these changes directly while preserving some semblance of *meaningful* data protection. For example, in the context of data mining, businesses could easily address this issue by adding the words "data mining" as a primary purpose at the time of data collection — but would this truly constitute "meaningful" data protection? Take another example: when applying for a new credit card, data mining could be added to the purposes for which the personal information collected on the application form would be used. But again, would this type of general, catch-all purpose be better than having no purpose at all? Possibly, but only marginally so.

The quandary we face with data mining is what suggestions to offer businesses that could truly serve as a meaningful primary purpose. The reason for this lies in the very fact that, at its essence, a "good" data mining program cannot, in advance, delineate what the primary purpose will be — its job is to sift through all the information available to unearth the unknown. Data mining is predicated on finding the unknown. The discovery model upon which it builds has no hypothesis — this is precisely what differentiates it from traditional forms of analysis. And with the falling cost of memory, the rising practice of data warehousing, and greatly enhanced processing speeds, the trend toward data mining will only increase.

The data miner does not know, cannot know, at the outset, what personal data will be of value or what relationships will emerge. Therefore, identifying a primary purpose at the beginning of the process, and then restricting one's use of the data to that purpose are the antithesis of a data mining exercise.

This presents a serious dilemma for privacy advocates, consumers, and businesses grappling with the privacy concerns embodied in an activity such as data mining. To summarize, the challenge lies in attempting to identify as a primary purpose, an as yet, unknown, secondary use. We offer some suggestions on how to address this issue in the next section.

Openness Principle

There should be a general policy of openness about developments, practices, and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

The principle of openness or transparency refers to the concept that people have the right to know what data about them have been collected, who has access to that data, and how the data are being used. Simply put, it means that people must be made aware of the conditions under which their information is being kept and used.

Data mining is not an open and transparent activity. It is invisible. Data mining technology makes it possible to analyze huge amounts of information about individuals — their buying habits, preferences, and whereabouts, at any point in time, without their knowledge or consent. Even consumers with a heightened sense of privacy about the use and circulation of their personal information would have no idea that the information they provided for the rental of a movie or a credit card transaction could be mined and a detailed profile of their preferences developed.

In order for the process to become open and transparent, consumers need to know that their personal information is being used in data mining activities. It is not reasonable to expect that the average consumer would be aware of data mining technologies. If consumers *were* made aware of data mining applications, then they could inquire about information assembled or compiled about them from the business with which they were transacting — "information" meaning inferences, profiles and conclusions drawn or extracted from data mining practices.

Ultimately, openness and transparency engender an environment for consumers to act on their own behalf (should they so choose). Consumers could then make known to the businesses they were transacting with, their expectations about the collection, re-use, sale and resale of their personal information.

Individual Participation Principle

An individual should have the right: a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him; b) to have communicated to him, data relating to him i) within a reasonable time, ii) at a charge if any that is not excessive, iii) in a reasonable manner and iv) in a form that is readily intelligible to him; c) to be given reasons if a request made under subparagraph (a) and (b) is denied, and to be able to challenge such denial; and d) to challenge data relating to him and, if the challenge is successful, to have the data erased, rectified, completed or amended.

Data mining operations are extremely far removed from the point of transaction or the point of the collection of the personal information. As data mining is not openly apparent to the consumer, then the consumer is not aware of the existence of information gained through a data mining application. This prevents any opportunity to: 1) request access to the information, or 2) challenge the data and request that corrections, additions, or deletions be made.

Consumers and Businesses — Choices to Consider

Consumers

In the United States, media coverage of public concerns about informational privacy matters began around the start of this decade with the uproar that erupted over *Lotus Marketplace: Households*.[30] This was an early and perhaps defining demonstration of the public's sensitivity about informational privacy. In 1996, the Lexis-Nexis incident drew massive attention to how people feel about their personal information: Lexis-Nexis, an online information service in Dayton, Ohio was accused of making social security numbers and other personal information widely available in its P-TRAK locator service. Then in 1997, after an electronic firestorm, America Online backed off of its plan to rent out its subscribers' telephone numbers.[31] In each of these cases, businesses quickly responded to a public outcry from their customers and either withdrew their products or changed their policies.

However, in order for consumers to react (and businesses to respond), consumers must have knowledge and awareness that something they could potentially choose to object to is actually occurring. The invisible nature of data mining (to the consumer) eliminates this possibility. In order for data mining to fall into line with fair information practices, the first step for consumers must be an awareness that any large business they are transacting with could be carrying out data mining activities. For some consumers, this knowledge will make no difference; for others, it will matter a great deal.

Once consumers are equipped with knowledge, it is up to each individual to decide for him or herself what matters, and based on that, what choices they want to make about assuming control over the uses of their personal information.

Concerned consumers *can* choose to take responsibility by informing businesses of their requirements and expectations regarding privacy. To assist in framing privacy-related questions relating to data mining, consumers may wish to consider the questions below. Then it is up to the consumer to decide what course of action, if any, to take.

As a consumer:

- Do you expect to be informed of any additional purposes that your personal information may be used, beyond the primary purpose of the transaction?
- Do you expect the option to say "no" to secondary or additional uses of your personal information, usually provided in the form of opting-out of permitting the use of your personal information for additional, secondary uses? Or, do you expect an opportunity to "opt-in" to secondary uses?

- Do you expect a process to be in place that gives you the right to access any information a business has about you, at any point in time?
- Do you expect a process that permits you to challenge, and if successful, correct or amend any information held by a businesses about you, at any point in time?
- Do you expect an option to have your personal information anonymized for data mining purposes and/or, an option to conduct your transactions anonymously?

For those consumers who wish to have greater control over the use and circulation of their personal information, we suggest the following initiatives:

- Ask to see a business's privacy or confidentiality policy. Assess it against your expectations of how you want your personal information handled. If the policy does not meet your expectations, contact the business and inform it of your expectations. If no policy exists, inform the business that you expect respectful and fair handling of your personal information.
- Give only the minimum amount of personal information needed to complete a transaction. If you are in doubt about the relevance of any information that is requested, ask questions about why it is needed, and ask that *all* of the uses of the requested information be identified.

Businesses

Businesses need a corporate will to adopt a culture of privacy — piece-meal or theoretical approaches will not be effective in responding to consumers' concerns. Ultimately, the impact of various technologies on privacy, including data mining, can only be averted by instilling a culture of privacy within the organization.

"Instilling a culture of privacy" means that businesses will have to tackle the conflict between the "use limitation" principle and the secondary uses of personal information arising out of data mining. It may be advisable for businesses to provide a multiple choice opt-out selection whereby consumers are given three choices: the choice of not having their data mined at all; only having their data mined in-house; or having their data mined externally as well. (Studies have shown that less concern is expressed over the *internal* secondary uses of one's data by the company collecting the data, but far greater resistance to having data disclosed externally for use by unknown parties).

Is your business willing to:

- Have a privacy strategy that is,
  - based on fair information practices and entrenched through tangible actions;
  - resourced throughout all facets of the organization; and
  - evaluated and assessed so that ongoing adjustments and improvements can be made?
- Have an open and transparent relationship with its customers?
  - Do you inform your customers upfront as to how all information collected about them will be used and disclosed, and by whom?
  - Do you have a process that makes it easy for customers to find out what personal information you have about them and a process to challenge any information that may be incorrect, incomplete, inaccurate or out-of-date?
- Accept that some consumers do not want their personal information to be mined, and nuggets about their buying patterns extracted?
  - Do you advise consumers of all uses of their personal information and give them a range of opt-out choices about data mining such as: 1) no data mining; 2) data mining internally; 3) data mining internally and externally. Or, for maximum choice and control, do you provide consumers with positive consent — an opportunity to "opt-in" for specified secondary uses of their personal information?
- Use privacy-enhancing technologies that can anonymize information and securely protect privacy?

Although there is no data protection legislation governing the collection, use and disclosure of personal information in the private sector (with the exception of Quebec), there *are* resources that can provide practical ways for businesses to address the protection of personal information. Some notable sources are the following:

- The Canadian Standards Association's *Model Code for the Protection of Personal Information (CAN/CSA-Q830-96)* and its companion publication, *Making the CSA Privacy Code Work for You — A workbook on applying the CSA Model Code for the Protection of Personal Information (CAN/CSA-Q830) to your organization*.
- The use of privacy-enhancing technologies such as *blind signatures* (which build on public key encryption) and biometric encryption. Each of these technologies relies on the "blinding" of identity through advance forms of encryption. Similarly, through the use of

an anonymous database,[32] personal and identifying information may be encrypted and stored separately in different locations — enabling businesses to consolidate their databases and keep them secure, while protecting privacy at the same time.

- Previous publications of the IPC have also outlined a range of ways in which businesses can carry out responsible data management practices with respect to the impact of technology on consumers' personal information: *Privacy Protection Makes Good Business Sense* (October 1994); *Privacy and Electronic Identification in the Information Age* (November 1994); *Privacy-Enhancing Technologies: The Path to Anonymity*, 2 Volumes (August 1995); *Identity Theft* (June 1997), and *Smart, Optical and Other Advanced Cards: How to do a Privacy Assessment* (September 1997).

- Professor Roger Clarke's strategic approach to privacy and dataveillance as set out in *Privacy and Dataveillance, and Organisational Strategy*[33] — provides a framework that can be used by businesses to develop a culture of privacy.

A Final Word

The need for protecting and managing personal information has been likened to the management of natural resources:

> Personal information is a resource, exploited commercially but valued as an element of human dignity and enjoyment of one's private life. It is therefore to be protected and managed, not unlike the protection and management of other resources. As with early efforts to protect the environment in the absence of legislation, privacy protection currently relies on ancient common law principles that continue to adapt to new technological challenges to personal integrity, happiness and freedom. These principles have now found legislative expression in various statutes relating to environmental protection. Information, however, has some unique qualities in need of special regulatory and judicial attention.[34]

Looking ahead, consumers will not only want goods and services, but will increasingly want assurances that the information they provide to a business is, from a privacy perspective, protected. To deal with this need, a shared responsibility for the management of personal information will be essential, involving government, the business community and consumers. Only through shared responsibility, sustained by the business community through a culture of privacy, and strengthened by the voice of consumers, can personal information become a protected, managed and valued resource. We hope that this report will give all three parties — consumers, businesses, and government — incentives for action towards protecting personal information in the marketplace.

Finally, we believe that the tension between technology and privacy can be minimized if privacy safeguards are made a key consideration upfront, rather than as an afterthought. Although current data mining practices are somewhat beyond the "upfront" stage, there is still time to ease this "tension" before applications become widely commonplace. One short term approach, as suggested earlier, may be for businesses to provide consumers with choices in the form of multiple selection opt-outs. To explore further solutions on how to address the "primary purpose" dilemma that data mining presents, we are committed to an open exchange. We invite those of you with any ideas as to how to resolve this issue to contact us — we would welcome your comments and encourage an open dialogue.

End Notes

1. IBM, *Data Mining: Make your data work for you like never before*, at direct.boulder.ibm.com/bi/tech/mining/index.html. Viewed on July 7, 1997.

2. Joseph P. Bigus, *Data Mining with Neural Networks* (United States: McGraw-Hill, 1996), p. 9.

3. For more information, see the following surveys: The Equifax Canada Report on *Consumers and Privacy in the Information Age* (1995); The Ekos Research Associates Inc. survey, *Privacy Revealed: The Canadian Privacy Survey* (1993); *The Information Highway: What Canadians Think about the Information Highway* (1994); and, *Surveying Boundaries: Canadians and Their Personal Information* (1996).

4. Marshall Jon Fisher, "moldovascam.com," *The Atlantic Monthly*, September 1997, p. 22.

5. Queen's University of Belfast, *What is Data Mining?*, at www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_2.html#HEADING2. Viewed on July 4, 1997.

6. Bigus, *Data Mining with Neural Networks*, p. 9.

7. Nick Wreden, *Communications Week Interactive*, February 17, 1997, at cmp-pub1.web.cerf.net/cw/cwi/pages/021797/650close.htm. Viewed on August 11, 1997.

8. Queen's University of Belfast, *What is Data Mining?*

9. Michael Bell, *A Data Mining FAQ*, at www.qwhy.com/dmfaq.htm. Viewed on August 26, 1997.

10. SAS Institute, *What is Data Mining?*, at www.sas.com/feature/4qdm/whatisdm.html. Viewed on July 25, 1997.

11. Ibid.

12. Conspectus, *The Data Warehousing Boom*, at www.pmp.co.uk/feb2.htm. Viewed on July 24, 1997.

13. IBM, *Data Mining — An IBM Overview*, at direct.boulder.ibm.com/bi/info/overview.htm. Viewed on July 7, 1997.

14. Bigus, *Data Mining with Neural Networks*, pp. 10–11.

15. Herb Edelstein, "Mining Data Warehouses," *Information Week*, January 8, 1996, p. 48.

16. Queen's University of Belfast, *What is Data Mining?*

17. Herb Edelstein, *Technology How To: Mining Data Warehouses*, at techweb.cmp.com/iw/.561/61oldat.htm. Viewed on July 24, 1997.

18. META Group, *Press Release on META Group Announces, "Data Mining Opportunities: 1996–1998*," at www.metagroup.com/newweb.nsf/Web+Pages/OldPR. Viewed on August 7, 1997.

19. Wreden, "The Mother Lode," *Communications Week Interactive*.

20. Kurt Thearling — The Data Intelligence Group, *From Data Mining to Database Marketing*, at www.santafe.edu/~kurt/wp9502.shtml. Viewed on August 7, 1997.

21. Author unknown, *Data Mining: What is Data Mining?*, at www.anderson.ucla. Viewed on July 25, 1997.

22. Ibid.

23. Barbara DePompa, "There's Gold in the Databases," *Information Week*, January 8, 1996, p. 54.

24. One reference is www.ultragem.com/ultrafaq.htm#howquestion. Viewed on July 4, 1997.

25. Wreden, "The Mother Lode," *Communications Week Interactive*.

26. Author unknown, *Data Mining: What is Data Mining?*

27. Bell, *A Data Mining FAQ*.

28. Pilot Software, *Data Mining White Paper — Profitable Applications* found at, www.pilotsw.com/dmpaper/dmindex.htm#dmapp. Viewed on July 4, 1997.

29. Jim Gray, *Data Management: Past, Present, and Future*, at www.research.microsoft.com/%7Egray/DB_History.htm. Viewed on August 12, 1997.

30. *Lotus Marketplace: Households* was a series of disks produced by Equifax and Lotus Development Corporation in 1990. On these disks (available to anyone for a price), were the names, addresses, buying habits and income information of roughly 120 million American consumers. Over 30,000 consumer enquiries and complaints lodged shortly after its release effectively cancelled the sale of the disks.

31. Robert Ellis Smith, "Rapid-Response Time," *Privacy Journal*, August 1997, p. 1.

32. Mytec Technologies Inc., a Toronto-based company, has developed a system which permits the consolidation of various databases but which keeps personal information under specific controls accessible only to those with a need to know. Mytec calls this system the

"Anonymous Database." In the Anonymous Database, an individual's private and identifying information are encrypted and stored separately in different locations. Mytec uses the information contained in the pattern of a person's fingerprint to code a number called a "Bioscrypt" which operates as the link between an individual's private information and identifying information. For further information, see www.mytec.com/applic/#database.

33. See Roger Clarke's paper *Privacy and Dataveillance, and Organizational Strategy*, at www.anu.edu.au/people/Roger.Clarke/DV/PStrat.html. Viewed on October 9, 1997. This paper presents a framework to guide businesses and governments towards adopting a strategic approach to privacy.

34. Ian Lawson, *Privacy and Free Enterprise: The Legal Protection of Personal Information In the Private Sector* (Ottawa: Public Interest Advocacy Centre, 1992), p. 442.