Moral Foundations of Data Mining

Kenneth W. Goodman
University of Miami, USA

INTRODUCTION

It has become a commonplace observation that scientific progress often, if not usually, outstrips or precedes the ethical analyses and tools that society increasingly relies on and even demands. In the case of data mining and knowledge discovery in databases, such an observation would be mistaken. There are, in fact, a number of useful ethical precedents, strategies, and principles available to guide those who request, pay for, design, maintain, use, share, and sell databases used for mining and knowledge discovery. These conceptual tools — and the need for them — will vary as one is using a database to, say, analyze cosmological data, identify potential customers, or run a hospital. But these differences should not be allowed to mask the ability of applied ethics to provide practical guidance to those who work in an exciting and rapidly growing new field.

BACKGROUND

Data mining is itself a hybrid discipline, embodying aspects of computer science, artificial intelligence, cryptography, statistics, and logic. In greater or lesser degree, each of these disciplines has noted and addressed the ethical issues that arise in their practice. In statistics, for instance, leading professional organizations have ratified codes of ethics that address issues ranging from safeguarding privileged information and avoiding conflicts of interest or sponsor bias (International Statistical Institute, 1985) to "the avoidance of any tendency to slant statistical work toward predetermined outcomes" (American Statistical Association, 1999).

It is in computer ethics, however, that one finds the earliest, sustained, and most thoughtful literature (Bynum, 1985; Johnson & Snapper, 1985; Ermann, Williams & Gutierrez, 1990; Forester & Morrison, 1994; Johnson, 1994) in addition to ethics codes by professional societies (Association for Computing Machinery, 1992; IEEE, 1990). Traditional issues in computer ethics include privacy and workplace monitoring, hacking, intellectual property, and appropriate uses and users. The intersection of computing and medicine has also begun to attract interest (Goodman, 1998a).

What is clear about this landscape is that terms we attach to issues — "privacy," for instance — can mask significant differences according as one uses a computer to keep track of warehouse stock or arrest records or real estate transactions or sexually transmitted diseases. Moreover, ethical issues take on somewhat different aspects depending on whether a computer and its data storage media are used by an individual, a business, a university, or a government. Atop this is the general purpose to which the machine is put: science, business, law enforcement, public health, or national security. This triad — content, user, and purpose — frames the space in which ethical issues arise.

MAIN THRUST

One can identify a suite of ethical issues that arise in data mining. All are tethered in one way or another to issues encountered in computing, statistics, and kindred fields. The question of whether data mining, or any discipline for that matter, presents unique or unprecedented issues is open to dispute. Issues between or among disciplines often vary by degree more than by kind. If it is learned or inferred from a database that Ms. Garcia prefers blue frocks, it might be the case that her privacy has been violated. But if it is learned or inferred that she has HIV, the stakes are altogether different. The ability to make ever-more-fine-grained inferences from very large databases increases the importance of ethics in data mining.

Privacy, Confidentiality, and Consent

It is common to distinguish between privacy and confidentiality by applying the former to humans' claim or desire to control access to themselves or information about them, and the latter, more narrowly, to specific units of that information. Privacy, if you will, is about people; confidentiality is about information. Privacy is broader, and it includes interest in information protection and control.

An important correlate of privacy is consent. One cannot control information without being asked for permission, or at least informed of information use. A business that is creating a customer database might collect data surreptitiously, arguably infringing on pri-

vacy. Or it might publicly — thought not necessarily individually — disclose the collection. Such disclosures are often key components of privacy policies. Privacy policies sometimes seek permission in advance to obtain and archive personal data or, more frequently, disclose that data are being collected and then provide a mechanism for individuals to opt out. The question whether such opt-out policies are adequate to give individuals opportunities to control use of their data is subject to widespread debate. In another context, a government will collect data for vital statistics or public health databases. Such uses, at least in democratic societies, may be justified on grounds of the implied consent of those to whom the information applies and who would benefit from its collection.

It is not clear how much or what kind of consent would be necessary to provide ethical warrant for data mining of personal information. The problem of adequate consent is complicated by what may be hypothesized to be widespread ignorance about data mining and its capabilities. As elsewhere, some solutions to this ethical problem might be identified or clarified by empirical research related to public understanding of data-mining technology, individuals' preference for (levels of) control over use of their information, and similar considerations. The U.S. Department of Health and Human Services has, for instance, supported research on the process of informed consent for biomedical research. Data mining warrants a kindred research program. Among the issues to be clarified by such research are the following:

- To what extent do individuals want to control access to their information?
- What are the differences between consent to acquire data for one purpose and consent for secondary uses of that data?
- How do individual preferences or inclinations to consent vary along with the data miner? (That is, it may be hypothesized that some or many individuals will be sanguine about data mining by trusted public health authorities, but opposed to data mining by [certain] business entities or governments.)

It should be noted that many information exchanges — especially including those generally involving the most sensitive or personal data — are at least partly governed by professionalism standards. Thus, doctor-patient and lawyer- or accountant-client relationships traditionally, if not legally, impose high standards for the protection of information acquired during the course of a professional relationship.

Appropriate Uses and Users of Data Mining Technology

It should be uncontroversial to point out that not all data mining or knowledge discovery is done by appropriate users, and not all uses enjoy equal moral warrant. A data-mining police state may not be said to operate with the same moral traction as a government public health service in a democracy. (We may one day need to inquire whether use of data-mining technology by a government is itself grounds for identifying it as repressive.) Similarly, given two businesses (insurance companies, say), it is straightforward to report that the one using data-mining technology to identify trends in accidents to better offer advice about preventing accidents is on firm moral footing, as opposed to one that identifies trends in accidents to discriminate against minorities.

One way to carve the world of data mining is at the public/private joint. Public uses are generally by governments or their proxies, which can include universities and corporate contractors, and can employ data from private sources (such as credit card information). Public data mining can, at least in principle, claim to be in the service of some collective good. The validity of such a claim must be assessed and then weighed against damage or threats to other public goods or values.

In the United States, the General Accounting Office, a research and investigative branch of Congress, identified 199 federal data-mining projects and found that of these, 54 mined private sector data, with 36 involving personal information. There were 77 projects using data from other federal agencies, and, of these, 46 involve personal information from the private sector. The personal information, apparently used in these projects without explicit consent, is said to include "student loan application data, bank account numbers, credit card information, and taxpayer identification numbers." The projects served a number of purposes, the top six of which are given as "improving service or performance" (65 projects), "detecting fraud, waste and abuse" (24), "analyzing scientific and research information" (23), "managing human resources" (17), "detecting criminal activities or patterns" (15) and "analyzing intelligence and detecting terrorist activities" (14) (General Accounting Office, 2004).

Is losing confidentiality in credit card transactions a fair exchange for improved government service? Research? National security? These questions are the focus of sustained debate.

In the private sphere, data miners enjoy fewer opportunities to claim that their work will result in collective benefit. The strongest warrant for private or for-profit