DATA MINING AND PRIVACY

Joseph S. Fulda

I. INTRODUCTION

Data mining is the intelligent search for new knowledge in existing masses of data. This article reviews what data mining technology can do and asks and answers the question, ‘Is data mining a violation of privacy that ought to be limited by law?’—tort law in the case of an individual actor and Fourth Amendment jurisprudence in the case of a state actor. Since data mining is a new technology and the existing data that are mined are presumed to be properly acquired, the answer would appear to be ‘No.’ However, we educe from three hypothetical cases of what might be called pre-technological data mining that our intuitions tell us that data mining is indeed a violation of privacy. We then review the case law—which does not agree with our intuitions—and briefly discuss alternative, technological means of protecting privacy in the face of advanced data mining techniques.

II. KNOWLEDGE DISCOVERY AND DATA MINING

Traditional information retrieval from databases returns “tuples”\(^1\) derived from fields of records or entire database records in response to a database query. The results of a database query are hence explicit in the database. Knowledge discovery using data mining techniques differs from ordinary information retrieval because what is sought and extracted (i.e., “mined”) from the data is often not explicit in the database.\(^2\) Rather, objects that “typically will not exist a priori”\(^3\) are discovered. The process of discovering such

---

\(^1\) See Jacob T. Schwartz, et al., Programming with Sets: An Introduction to SETL 50 (defining tuples as “finite ordered sequences of arbitrary elements”). See also id. §3.2 on the basics of tuples, §§3.6-3.7 on tuple operations, formers, and iterators, and §4.3.2 on tuple iterators. Tuples are also called vectors, arrays, lists, or \(n\)-tuples. Here we refer to just those data drawn from the fields of a database record requested by a database query. Thus although database records may have many fields, sub-tuples of these records will be returned in response to a query (search) of the database which will contain just those data that are in the relevant fields of the retrieved records.

\(^2\) Tomasz Imieliński and Heikki Mannila, A Database Perspective on Knowledge Discovery, 39 COMM. OF THE ACM 58, November 1996, at 60.

\(^3\) Id.
patterns, which is termed “data mining” when considered apart from the necessary concomitant parts of the knowledge discovery process, may:

(a) **classify** data into preëxisting categories;

(b) **cluster** data by mapping them into categories created during data analysis and determined by the data;

(c) provide a **summary** of the data, which is useful in a sense that the raw data are not;

(d) describe **dependencies** between variables;

(e) find **links** between data fields;

(f) use **regression** to predict future values of data and

(g) **model sequential patterns** in the data that may indicate revealing trends.

Data mining is most easily accomplished when the data are highly structured and available

---

4 See Usama Fayad, et al., *From Data Mining to Knowledge Discovery in Databases*, AI Magazine 17, Fall 1996, at 39 [hereinafter Fayad 1] (referring to "the application of specific algorithms for extracting patterns from data").

5 Id. (referring to work preparing the data so that it can be searched for patterns, and work done on the patterns to make them useful after they are found; i.e., pre-processing and post-processing).

6 See Usama Fayad, et al., *The KDD Process for Extracting Useful Knowledge From Volumes of Data*, COMM. OF THE ACM 39, November 1996, at 31 (noting that classification is the mapping of data “into one of several predefined categorical classes”) [hereinafter Fayad 2].

7 See id. (stating that clustering is mapping of data “into one of several categorical classes . . . in which the classes must be determined from the data . . . [and] are defined by finding natural groupings of data items based on similarity metrics or probability density models.”).

8 See id. at 32 (noting that summarization is a “compact description for a subset of data”).

9 See id. (noting that ‘dependency modeling’ is the description of significant dependencies among variables).

10 See id. (stating that 'link analysis' is the relation “between fields in the database . . . [to] deriv[e] multi-field correlations satisfying support and confidence thresholds”).

11 See Fayad 2, supra note 6, at 31 (noting that regression is the mapping of data items to real-valued prediction variables).

12 See id. at 32 (noting that the goal of ‘sequence analysis’ is “to model the states of the process generating the sequence or to extract and report deviation and trends over time”).
in many different forms and at many levels in what are known as “data warehouses.”\textsuperscript{13} A
data warehouse contains:

(a) \textit{integrated data}, which by allowing data to be compared and contrasted in different
form(at)s does away with much of the need for “data cleansing”;\textsuperscript{14}

(b) both \textit{detailed} and \textit{summarized} data. The former is important because certain patterns
can be detected only by examining “data in its most granular form,” while the latter is
important because some patterns become apparent only on higher-order data;\textsuperscript{15}

(c) \textit{historical data}, which if mined, can yield cyclic and seasonal activity as well as long-
term trends,\textsuperscript{16} and,

(d) \textit{metadata}, which provide the context of the data.\textsuperscript{17}

III. THE ISSUE

Much of the current concern about privacy arises because of data mining and, more
generally, because of knowledge discovery. In traditional computer-science terms, data
is uninterpreted, while knowledge has a semantics that gives it meaning. While the data
stored in databases is not truly uninterpreted, the old legal rule that anything put by a
person into the public domain (e.g., by purchasing an item in a public place of business)
is not legally protected served well when the data was not mined so as to produce
classifications, clustering, summaries and profiles, dependencies and links, and other
patterns.

\textsuperscript{13} See W. H. Inmon, \textit{The Data Warehouse and Data Mining}, COMM. OF THE ACM 39, November 1996, at 49.

\textsuperscript{14} See \textit{id.} at 49-50 (remarking that ‘integrated data’ allow a data mining system to “easily and quickly look across vistas
of data” by concentrating on mining data rather than cleansing and integrating it, and without such integrated data, the
data mining system would have to “spend inordinate amounts of time cleansing and conditioning the data before the
process of data mining could commence in an effective manner”).

\textsuperscript{15} See \textit{id.} at 50 (noting that detailed data are needed when a data mining system desires to “examine data in its most
granular form [as] [v]ery low levels of detail hide important patterns that can be discerned no other way than by carefully
scrutinizing the detail.” Summarized data are important as well, as it makes sure that if an analysis has already been
made, the process of data mining does not have to be repeated. Additionally, a data mining system can “build” from
previous work, instead of having to start from the beginning of the process.).

\textsuperscript{16} See \textit{id.} (noting that historical data are vital because it holds important information; information that is useful to track
patterns and trends that a data mining system utilizing only current data cannot find).

\textsuperscript{17} See \textit{id.} (remarking that metadata are used by a data mining system to describe the context, rather than the content,
of information, because “[w]hen information is being examined over time, context becomes as relevant as content”).
Thus, for example, it is unnerving if a database of film, garnered from a bank ATM video camera, showing a pregnant woman with her ATM card is linked through the bank’s database to her home address and subsequently triggers an avalanche of circulars, advertisements, and e-mail spam for products for newborns sold by another subsidiary of the bank’s corporate holding company. Yet, in this scenario, it is not the pregnancy that is private, much less the address; it is the linkage that is disturbing. This example brings into relief the difficult philosophical question that data mining and knowledge discovery have created: Is it possible for data that does not in itself deserve legal protection to contain implicit knowledge that does deserve legal protection? If legal protection is needed for such data, what balance must be struck between the freedom to use whatever knowledge one has at one’s disposal to further one’s own ends and the freedom not to have one’s personal data mined into knowledge that will be used as a means to someone else’s ends?  

IV. ANALYSIS

I propose to address this issue based on my analysis presenting a theory of privacy rights, in turn based generally on tort law updated for the information age. Because data need not be encapsulated in property any longer, protection of information through (at least tangible) property rights is inadequate.

In my previous work, a standard was developed that privacy is invaded, when any means are used that bypass the subject’s consent as manifested by the subject’s observable (i.e., objective) behavior, reasonably interpreted. This rule, which might be termed an anti-circumvention rule, mirrors that recently passed by Congress for intellectual property, the Digital Millennium Copyright Act [“Act”]. But while the Act’s restrictions on technological progress (as opposed to its specific use to take property) are in this author’s view misguided, the thrust of the law protecting abstract property from circumvention measures is philosophically well motivated. Such circumvention in the realm of privacy constitutes the tort of appropriation refigured for the information age. In the case of a government actor acting without a search warrant (or in some cases merely probable cause), it is, under our theory, a violation of the Fourth Amendment of the United States

---

18 This is to put the dilemma in Kantian terms.


20 Id. at 316.

Constitution.\textsuperscript{22} Indeed, reputation, i.e., that which privacy protects, should be regarded as property, and is a fundamental part of what a person owns in himself, and this author has given a full account of this elsewhere.\textsuperscript{23}

The process we will use to perform this analysis is known as “eduction,”\textsuperscript{24} here finding hypothetical cases where two (conventional) data about an individual, each innocuous in itself, are combined, but together they produce new (conventional) knowledge about the individual. This is akin to data mining. The rule that emerges from the hypothetical cases is then formulated and applied, in turn, to (technological) data mining.

Thus, for example, when I have a guest over, I am often asked, ‘Can I read this magazine?’ The answer is always, ‘Well, it is published; nothing published can be considered private.’ While this is a good rule of thumb, there are some exceptions. For example, suppose I have a stack of magazines in my living room, and buried among them is a ‘men’s magazine.’ Wouldn’t it be wrong for someone to go through the stack of magazines, even though it is readily apparent that it contains only published material? I think it would, for although the magazine is not private and the stack showing that I keep magazines is not private, the fact that the former is in the latter—the association—is private.

Consider a second case: I have a small archives that contains everything I have had published. Can a guest simply take the liberty of browsing through my archives? It would seem so, based on the rule of thumb on the lack of privacy due published materials, until one learns that I have pseudonymously authored articles on two occasions. It is not that the articles themselves are private; I wanted them published for reasons that I thought were good. But for equally good reasons, I did not want my name associated with these articles. In isolation, the guest could read the articles, as they were openly published, but yet he could not examine the particular copies of the articles found in the archives. If I had the periodicals in which they appeared on my stack of magazines (see above), no problem would present itself. It is only the association with my authorship signified by their (mere) presence in the archives that presents a privacy issue.

Consider a third case: Suppose people generally considered their home addresses matters

\textsuperscript{22} Fulda, \textit{supra} note 19.

\textsuperscript{23} See Joseph S. Fulda, \textit{Reputation As Property}, 33 ST. CROIX REV. 30, 30 (April 2000) (noting that “reputation is based on our abilities, capacities, and even physiognomy as modified over the years by every action we take, every behavior we display. Thus, like personal property, reputation is formed by taking natural resources and mixing our labor with it.”).

\textsuperscript{24} See \textit{Merriam Webster’s Collegiate Dictionary} 367 (10th ed. 1993) (noting that the word, “educe,” “implies the bringing out of something potential or latent”).
of public record. The mailman comes by and drops off a stack of letters. Is it then appropriate for a guest to go through the stack of mail without opening any of the letters? Certainly not, for although the return addresses are, *arguendo*, not private, their association with my address and home is private. I may not want my guest to know with whom I correspond, even though the correspondents’ addresses, taken in isolation, are no more private than the fact that I received correspondence. However, it is the association between the two data, both relating in some way to me, that is private.

None of these cases involve technology, but sifting through a stack of magazines, an archives, or a stack of letters to find associations between two data and an individual are all pre-technological forms of data mining. Moreover, they are all improper. Technology cannot make right what is otherwise wrong. Consequently, under our theory, if data about an individual is mined and implicit knowledge about him is discovered, an appropriation has occurred, and further disclosure should not be permitted.

Unfortunately, this is not what the case law says. The governing case is *Smith v. Maryland*, in which it was held that the Fourth Amendment provides no protection against the use of pen registers, devices that record the telephone numbers dialed from a particular telephone line without recording the conversation itself. Clearly, neither the calling line nor the numbers called are private, but the association between them and the party in question is. Pen registers were, at the time of the ruling, still a new technology. In *United States v. New York Telephone Co.*, a case that was decided two years earlier, a lower court assumed that Fourth Amendment protections did apply to pen registers, even though the Supreme Court expressly withheld judgment on that issue. *New York Telephone* revolved around whether a telephone company could be compelled “to furnish the FBI ‘all information, facilities and technical assistance’ necessary to employ the pen registers.” In other words, the technology was new enough that the FBI needed outside technical expertise. This may perhaps explain why the Court reached what this author

---


26 *Id.* at 741.


28 *See id.* at 166 n. 7 (stating, in relevant part, “The Court of Appeals held that pen register surveillance was subject to the requirements of the Fourth Amendment. This conclusion is not challenged by either party, and we find it unnecessary to consider the matter.”).

29 *Id.* at 161.

30 *See id.* (noting that the Federal Bureau of Investigation was also ordered to fully compensate New York Telephone at the prevailing rates for any assistance it had furnished).
considers the wrong result: At the time, the Court possessed no intuition about pen registers, and failed to make appropriate comparisons to older technologies, such as print materials, as we have done here.\textsuperscript{31}

If the courts will not protect private data from being associated with each other and forming new information about an individual, there is a long-standing set of methodologies, based on cryptographic protocols, which can provide such protection. Devised by computer scientist David Chaum,\textsuperscript{32} these techniques prevent the “dossier society, in which computers could be used to infer individuals’ life-styles, habits, whereabouts, and associations from data collected in ordinary consumer transactions [and which] can have a ‘chilling effect,’ causing people to alter their observable activities . . .”\textsuperscript{33} while answering the need “for organizations . . . to devise more pervasive, efficient, and interlinked computerized record-keeping systems . . .”\textsuperscript{34}, so that everything from consumer credit to social services is not abused. Unfortunately, there is no real incentive for organizations to implement these rather simple methodologies, despite consumer concern with data collection and mining. Thus, as long as the law is silent on the subject, it appears that existing non-intrusive technological means of solving this problem will remain merely a unit in various advanced courses in the computer science curriculum.

\textsuperscript{31} More recent case law has upheld the anti-circumvention rule of the Digital Millennium Copyright Act, see supra note 21. In Universal City Studios, Inc. v. Reimerdes, 111 F. Supp. 2d 294 (S.D.N.Y. 2000), the district court found that communication of decryption source code for unlocking DVD’s was not protected by the First Amendment, a ruling with which I disagree. But that ruling is a far cry from a ruling that finds circumvention itself a constitutional violation of the Fourth Amendment (as against a state actor and in the case of privacy, not intellectual property). This non-binding precedent only suggests that were the States to pass legislation protecting privacy from circumvention, such legislation might not itself run afoul of the Constitution. Such a suggestion, however, is weak, since a long line of precedents allow invasion of privacy in the interests of free speech. (See, e.g., New York Magazine v. The Metropolitan Transp. Auth., 136 F.3d 123 (2nd Cir. 1998), \textit{cert. den.}, 525 U.S. 824 (1998).) As the judge himself noted, his holding was “very narrow” (111 F. Supp. 2d at 333) and based on the dual nature of computer source code as both expressive and functional, which is of no moment to our concern here with privacy.


\textsuperscript{33} \textit{Id.} at 1030.

\textsuperscript{34} \textit{Id.}