Yale Journal of Law and Technology

Volume 5
Issue 1 *Yale Journal of Law and Technology*

Article 1

1-1-2003

"MINE YOUR OWN BUSINESS!": MAKING THE CASE FOR THE IMPLICATIONS OF THE DATA MINING OF PERSONAL INFORMATION IN THE FORUM OF PUBLIC OPINION

TAL Z. ZARSKY

Follow this and additional works at: http://digitalcommons.law.yale.edu/yjolt

Part of the Computer Law Commons, Intellectual Property Commons, and the Science and Technology Commons

Recommended Citation
ZARSKY, TAL Z. (2003) ""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 and Technology*: Vol. 5: Iss. 1, Article 1.
Available at: http://digitalcommons.law.yale.edu/yjolt/vol5/iss1/1

This Article is brought to you for free and open access by Yale Law School Legal Scholarship Repository. It has been accepted for inclusion in Yale Journal of Law and Technology by an authorized administrator of Yale Law School Legal Scholarship Repository. For more information, please contact julian.aiken@yale.edu.

ARTICLE

"MINE YOUR OWN BUSINESS!": MAKING THE CASE FOR THE IMPLICATIONS OF THE DATA MINING OF PERSONAL INFORMATION IN THE FORUM OF PUBLIC OPINION

TAL Z. ZARSKY*

I. INTRODUCTION 2
II. A TECHNICAL DESCRIPTION OF THE DATA MINING PROCESS 6
A. A BRIEF INTRODUCTION TO DATA MINING 6
B. DATA WAREHOUSING AND DATA CLEANSING 7
C. DATA MINING: METHODS AND PRACTICES 9
1. CLUSTERING 9
2. ASSOCIATION RULES 11
(a) ASSOCIATION DISCOVERY 12
(b) SEQUENTIAL PATTERN DISCOVERY 13
(c) SIMILAR TIME SEQUENCE DISCOVERY 14
3. THE NEXT STEPS IN THE KDD PROCESS 15
D. FINAL POINTS AND INTERNET APPLICATIONS 16
III. THE SOCIAL IMPACT OF THE DATA MINING PRACTICES 18
A. GENERAL OVERVIEW 18
B. DISCRIMINATION 22
C. MANIPULATION AND THREATS TO AUTONOMY 35
1. EFFECTS ON THE INDIVIDUAL 38
2. EFFECTS ON SOCIETY 40
D. ABUSE & MISUSE 44
E. SECLUSION 45
F. "THE TRAGEDY OF ERRORS" 47
IV. DATA MINING AND PUBLIC OPINION 50
A. PREFACE AND NOTE OF CAUTION 50
B. THE DATA MINING CAMPAIGN 53
V. CONCLUSION 55

* J.S.D Candidate, Columbia Law School. The author would like to thank Eben Moglen, Lance Liebman, Paul Schwartz, and the members of the 2002 J.S.D Candidate workshop. The author also thanks Yochai Benkler and Eli Noam for providing additional insight and assistance regarding this paper. A prior version of this Article was presented at the 2002 S.J.D/J.S.D Conference at Harvard Law School.

"MINE YOUR OWN BUSINESS!": MAKING THE CASE FOR THE IMPLICATIONS OF THE DATA MINING OF PERSONAL INFORMATION IN THE FORUM OF PUBLIC OPINION

TAL Z. ZARSKY*

Today's world of constant surveillance and data collection allows for the gathering of vast amounts of personal information. In this reality, sophistication in the analysis of information is key. Data mining is probably the information collectors' only hope to close the sophistication gap, yet the use of advanced means of analysis is certain to impact individuals and society in various ways. This Article addresses the use of data mining applications in analyzing personal information and its impact upon society. It begins with a description of current data mining practices from a technical point of view, a perspective often overlooked in legal scholarship. The Article next describes the current privacy debate, highlighting the issues most relevant to the new reality data mining creates. Among others, it addresses issues such as discrimination, threats to autonomy, misuse of data and the consequences of erroneous information. The analysis is facilitated by several concrete "hypotheticals" that address some of the otherwise abstract concepts this debate presents in simple terms. The author asserts that in view of data mining tools, some traditional claims of privacy are rendered trivial or obsolete, while others are of particular importance. After focusing on the role of public opinion, the Article concludes by outlining a public opinion campaign which may prove useful in finding solutions to the legal problems data mining tools create.

I. INTRODUCTION

"Know what is above you: a watchful Eye, an attentive Ear, and all your deeds are recorded in a Book." (Values of our Fathers II,1)[1]

"We are all being surveyed. All the time. Even though it is not apparent to consumers shopping virtually online via e-commerce, purchasing in 'real world' supermarkets, or carrying out other mundane activities and transactions, our actions are being watched

---

* J.S.D Candidate, Columbia Law School. The author would like to thank Eben Moglen, Lance Liebman, Paul Schwartz, and the members of the 2002 J.S.D Candidate workshop. The author also thanks Yochai Benkler and Eli Noam for providing additional insight and assistance regarding this paper. A prior version of this Article was presented at the 2002 S.J.D/J.S.D Conference at Harvard Law School.

1 THE COMPLETE ARTSCROLL SIDDUR 551 (1994).

and examined." This mantra has been gaining popularity in recent literature concerning sociology, Internet law, and other social sciences. "The Death of Privacy" has been a recurring headline in many magazines and periodicals, bleakly portraying the manner in which modern society is closely watched and scrutinized. In short, surveillance has become the issue of the hour and "Big Brother" is back in vogue.

Mere surveillance, however, is not grounds for concern, at least not on its own. The fact that there are an eye watching and an ear listening is meaningless unless the collected information is *recorded and analyzed*.

Recording is easy. In the world of large corporations and with the use of today's high technology, nothing needs to be forgotten or lost in oblivion. Memory is cheap, and therefore such entities invest in the storage of trivial information, hoping to reap benefits in the future.

The source of difficulty they face at this time is the need to *analyze*. The number of terabytes gathered and stored is vastly greater than the quantity of information faced in the past.[2] Companies are learning quickly that when attempting to cope with mountains of accumulated information, sophistication is key.

The first, "classic," option for analyzing databases is the simple statistical "query."[3] Using relatively simple statistical tools on a neatly organized database created for this use, corporations can retrieve various types of information about the database as a whole and their individual customers by "presenting" the database with simple queries. Advanced practices include segmenting the database into groups and analyzing each sub-database both on its own and as compared to other sub-groups. These slightly improved procedures generate large benefits to their users, as they assist in locating disparity between store locations, seasons of the year, and so forth. But are these tools sufficient to overcome the current difficulties of immense databases and a competitive market?

"*Tell us something that we don't know,*" is the database holders' and analysts' response when offered the use of these tools. They mean

---

2 Some researchers estimate that only 7% of the information that is recorded is processed. Ann Cavoukian, Information and Privacy Commissioner/Ontario, *Data Mining: Staking a Claim on your Privacy*, 1, *at* http: www.ipc.on.ca/english/pubpres/papers/datamine.pdf (last visited November 21, 2002).

3 A "query" is a search in a database for all records satisfying some specified conditions.

both that they are familiar with these applications and also that they need tools that will allow them to reveal deeper, unknown connections. Breaking into groups has been applied for decades and has been very useful, but may not be adequate when facing databases of such proportions. Since it is hard to know what should be asked and how to subdivide a database, the use of these old practices is costly and must be followed by many eyes. Even so, such analysis may not discover important information; *if you don't know what you are looking for, you will not be able to find it!*

In order to cope with these difficulties, the sophisticated techniques of "knowledge discovery in databases" ("KDD"), also called "data mining" ("DM"),[4] have emerged. The building blocks of these techniques are complex algorithms, artificial intelligence, neural networks and even genetic-based modeling; they can discover previously unknown facts and phenomena about a database, answering questions users did not know to ask. They carry out the analysis without receiving a hypothesis from the human analysts, instead searching for hidden patterns on their own. Not only can the KDD tools describe the database as it is, they can also make predictions about future data. KDD can be embedded in the operating network of a business or organization and requires minimal intervention or supervision. KDD closes the sophistication gap.

However, the technical discipline of data mining is part of a larger social context. The descriptive and predictive information that KDD produces significantly affects those subject to the analysis and therefore should be the focus of legal scrutiny. When such KDD tools are linked to the ongoing online surveillance, the potential for adverse effects increases greatly, presenting a double threat compared by Jason Catlett, President of Junkbusters.com,[5] to "going hunting with nuclear weapons."[6] On the other hand, KDD may have positive effects as well. These effects are not captured by simple paradigms of privacy. This Article explains why.

I begin in Part I with an introduction to KDD, explaining its history, methods, techniques, and results. I also present examples of

---

4 Definitions are a problem in the dating mining field, as every writer uses the terms differently. The term "Data Mining" is used in two distinct ways: both to define the entire process (as KDD does) and to describe the specific stage in which the algorithms are applied. PETER CABENA ET AL., DISCOVERING DATA MINING – FROM CONCEPT TO IMPLEMENTATION 15 (1998).

5 Junkbusters.com (www.Junkbusters.com) is an advocacy group active in the field of privacy rights, personal information and mass marketing.

6 Patricia Odell, *Gotcha!,* DIRECT, Nov. 2000, *available at* http://directmag.com/ar/marketing_gotcha/index.htm (last visited Nov. 21, 2002).

recent KDD applications as they are employed in practice, particularly in connection with the Internet.

Subsequently in Part II, I leave the fields of computer science and statistics to examine the impact KDD tools are having on society, emphasizing present and future problems arising from the analysis of personal information. Only a deep understanding of the problems data mining presents will enable us to contemplate what legal solution these problems will require.[7] Here I focus on the interaction between private parties, rather than between individuals and the state.[8] In this Part, I address the major issues discussed in the ongoing debate regarding information privacy, as characterized in the existing literature of that field. But not all the issues that are raised in the literature are of equal relevance and importance when the discussion shifts to "data mining." Indeed, while the spread of KDD may render traditional claims of privacy trivial or obsolete, others are of particular importance when we adjust to the data-mining world, and therefore deserve special consideration.

In the interaction between KDD and traditional privacy claims, we should pay special attention to public opinion. Public concern over the privacy of personal information is rising rapidly for various reasons. However, the public debate concerning issues of personal information and privacy has had a tendency to stray from the most severe and direct issues. This tendency could be due to "innocent" error and the complexity of the matters at hand, but there is always the possibility that the public is being purposefully deluded. I therefore conclude this Article in Part III by giving reasons to believe that the role of public opinion is extremely important in bringing solutions to this field. I also outline the matters on which a public opinion campaign should focus to sufficiently address the problems arising from the use of KDD applications.

In what follows, for pragmatic reasons, I will not participate in the ongoing debate about rights.[9] Such discussion requires confronting

---

7 There are not many legal resources on the issues posed by KDD and DM. *See generally* Joseph S. Fulda, *Data Mining and Privacy*, 11 ALB. L.J. SCI. & TECH. 105 (2000).

8 The right of privacy towards the state and government has created a wide debate and raises various constitutional issues. *See generally* ROBERT ELLIS SMITH, BEN FRANKLIN'S WEB SITE 153, 309 (2000). In addition, several statutes are pertinent to this subject. *See* Privacy Act of 1974, 5 U.S.C. §552a (1994); Computer Matching and Privacy and Protection Act of 1988, 5 U.S.C. §552a(o) (1994).

9 The specific right of privacy has been mentioned as early as 1890, in the famous Warren & Brandeis article *The Right to Privacy*, 4 HARV. L. REV. 193 (1890) (for the background of this article, see Smith *supra* note 8 at 121-152). However, the existence of this right is a debated issue. For example, some

various theories of rights, both of individuals and of groups,[10] and the appropriate balance between them.[11] My focus, however, is on the practical relationships between private parties and on the power of public opinion.

II. A TECHNICAL DESCRIPTION OF THE DATA MINING PROCESS

A. A BRIEF INTRODUCTION TO DATA MINING

Before descending to definitions, we must keep in mind that though legal literature frequently refers to data mining, it often does so vaguely or in the wrong context. Data mining[12] is correctly defined as the "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data."[13] Each of the described components is important, but novelty is the most essential. It is novelty that distinguishes data mining from previous statistical practices. Data mining provides its users with answers to questions they did not know to ask.

Data mining's origins were in the 1990s, but it has come a long way since then. It has been described as one of the "top ten" emerging fields in today's technological world, with potential to dominate the

---

commentators point out that from the economics perspective, the creation of such rights is not recommended (or at the least should be regarded as "rules of the second order"), as they inhibit decision making. Richard Murphy, *Property Rights in Personal Information: an Economic Defense of Privacy*, 84 GEO. L. J. 2382 (1996).

10 Regarding the privacy rights of groups, see for example reference to the rights of Ashkenazi Jews in the information collected in DNA surveys. SIMON GARFINKEL, DATABASE NATION: THE DEATH OF PRIVACY IN THE 21ST CENTURY 190 (2000).

11 On these issues, see ETZIONI, THE LIMITS OF PRIVACY 183 (1999) discussing the importance of balancing among various issues of privacy.

12 Interestingly, it seems that the phrase "data mining" was originally derogative (compare the history of "democracy"). At first, it referred to the process of extracting ridiculous regressions with no hold on reality from large databases (such as the correlation between the stock market and the amount of milk cows produced in a certain area). DAVID HAND ET AL., PRINCIPLES OF DATA MINING 23 (2001).

13 This is the most common definition of data mining, offered by Fayyad himself. *See* U. M. FAYYAD ET AL., FROM DATA MINING TO KNOWLEDGE DISCOVERY: AN OVERVIEW, IN ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING 6 (1996). There are several other proposed definitions of the field; the discovery of new information from an existing database is their common denominator. E.g. "...True data mining software doesn't just change the presentation, but actually discovers previously unknown relationships among the data." *See* Webopedia.com, (http://www.webopedia.com/TERM/d/data_mining.html).

future as well.[14] Some have traced the development of KDD to research conducted by Professor Usama Fayyad in an attempt to identify latent defects in General Motors products. For this purpose, Fayyad constructed advanced algorithms to search GM's databases and retrieve useful information, answering questions the GM engineers did not know to pose.[15] The practice of data mining Fayyad pioneered has grown rapidly since then, and the annual KDD conferences started by Fayyad in 1995 have drawn large attendance and interest,[16] and account in part for the growing importance of the field.[17]

Even though the DM practices may be new, they constitute adaptations of statistical algorithms that have been used for a long time.[18] Nevertheless, recent advances in computer speed and the collecting of data by many businesses have inspired the improvement of software to achieve today's mining abilities.[19] As parallel processing[20] and the use of artificial intelligence[21] have met with improvements in software and growing business awareness of the benefits of database analysis, the interdisciplinary field of KDD, based on both statistical tools and computer science, has emerged.

But how is the actual "mining" process carried out?

B. DATA WAREHOUSING AND DATA CLEANSING

---

14 *10 Emerging Technologies that Will Change the World*, TECHNOLOGY REVIEW, Jan 1. 2001.

15 Id.

16 Information on the 1997 conference is found at http://www-aig.jpl.nasa.gov/public/kdd97/kdd_home.html.

17 An interesting indication of the emergence of the field: in 1999, the Library of Congress added a category of technical books, "QA 76.9 D343," devoted to Data Mining, to the various topics under the classification of "Computer Software" (I would like to thank the librarians of the Columbia University's Butler Library for their assistance regarding this remark).

18 An algorithm in this context has been defined as a "well defined procedure that takes data as input and produces output in the form of models or patterns." Hand, *supra* note 12 at 141.

19 Omar Faruk Alis et al., *Data Mining for Database Marketing at Garanti Bank*, *in* DATA MINING II 93 (N.F.F. Ebecken ed., 2000).

20 This technology is used to run the algorithms at the same time on several parts of the database, enabling faster and better results. *See* generally M. Holsheimer, *Data Surveyor Searching the Nuggets in Parallel*, *in* Fayyad, *supra* note 13 at 447.

21 A great amount of research has been devoted to this area. At this time, KDD applications are able to engage in "machine learning" – learning from one search to another and changing the process as it goes on. *See generally* C. Hsu, C.A. Knoblock, *Using Inductive Learning To Generate Rules For Semantic Query Optimization*, *in* Fayyad, *supra* note 13 at 427.

In order to commence "data mining" you first must have "data." The party intending to mine must decide what information to include in the database to be analyzed. One widespread practice is to organize business related information in a "data warehouse,"[22] an aggregation of several databases from multiple sources. For example, in a large banking organization, there are "snippets" of information pertinent to any specific customer scattered in various databases. Information on loans, certificates of deposits, checking accounts, and life insurance may each be in a different repository. The "data warehousing" process aggregates this information, combining it with personal information that has been submitted by the customer in various questionnaires and forms[23] and personal data purchased by the collector from third parties.

An efficient data warehouse[24] is a clean warehouse where unreliable information is thrown out. When it is clear that the information available is tainted, a special, "neutral," value can be used instead, which the algorithm can ignore during future analysis. This process of "data cleaning" (or "cleansing") is governed by a database manager who must constantly attend to the database in order to make data mining possible. Though data warehousing may seem trivial, it is thoroughly addressed in the technical and business literature,[25] as it is both a prerequisite to successful data mining and an important practice on its own. Even though the warehousing does not enable the user to produce new forms of clustering and predictions, the input the data warehouse provides is more than adequate for many businesses. Businesses that are not interested in the additional expense data mining would entail could utilize the data warehouse to carry out some of the simple tasks mentioned above, such as query-based searches and group comparisons.[26]

When the preparation of the data warehouse is concluded, the actual data mining can commence.

---

22 A data warehouse has been defined as "a subject oriented, integrated, time variant and nonvolatile collection of data in support of management's decisions." CABENA ET AL., *supra* note 4, at 19.

23 E.g., the customer's address, income, and family status.

24 BHAVANI THURAISINGHAM, DATA MINING – TECHNOLOGIES, TECHNIQUES, TOOLS, AND TRENDS 49 (1999).

25 *See, e.g.*, JILL DYCHÉ, E-DATA, TURNING DATA INTO INFORMATION WITH DATA-WAREHOUSING (2000)(addressing business aspects of data mining).

26 See Thuraisingham, supra 24, P.60.

C. DATA MINING: METHODS AND PRACTICES

Various methods and practices of data mining employ different algorithms, each suited to some required task. This Article focuses on two archetypes of such analysis: *clustering* and *association rules*. Every method has both descriptive and predictive applications.[27] While these two forms of applications may seem trivial from the technical perspective, from the lawyer's point of view, prediction and description present different legal and social problems to different classes of people.

1. CLUSTERING

First, let us take *clustering*. Here is an example of such an application loosely based on a paper presented in a recent KDD conference:[28]

An insurance firm decides to upgrade its marketing scheme. Instead of searching for clients in the dark and offering each a bulky package of papers destined for the trash, it adds sophistication to its marketing process by employing data mining. The goal of the mining is to determine which clients would most likely respond to the company's advertisements, and which insurance policy they would prefer. Learning the preferences of prospective clients is extremely important, as an advertisement directed to the specific interests of the recipient has a significantly higher chance of attracting attention and invoking a response.

Such a task relies on data available from several sources. The prime source of information is the clients themselves, who have provided personal information as part of the demanding policy application process. Additional information on the policyholders' histories is available from sources within the firm, including their methods and promptness of payment, their chosen policies, and their addresses and places of business. Further information could be provided by external sources: affiliates of the firm with whom the

---

27 Linda C. Smith, *Knowledge Discovery, Capture and Creation*, BULL. AM. SOC. FOR INFO. SCI., Dec. 1999 – Jan. 2000. *Available at* http:// www.asis.org. *See also* Neena Buck, *Eureka! Knowledge Discovery*, SOFTWARE MAG., Dec. 2000 – Jan. 2001. There are other applications of data mining, such as visualization. Visualization tools are designed to display the patterns found in the database and assists users in noticing such patterns and making the appropriate business decisions. *See* THURAISINGHAM, *supra* note 24 at 68. Our discussion will remain focused on clustering and association rules, and the social problems they create.

28 *See* G. Pedrazzi et al., *CRM in a Real-World Insurance Company*, in DATA MINING II, *supra* note 18 at 53.

current and future clients happen to be associated, or information brokers, who gather personal information for bulk sale.[29] Using the KDD tools, the firm could divide its database of existing clients into various segments according to several traits. By observing those clusters, the firm is able to determine which attributes had the greatest effect on policy choice among present clients and is able to make strong predictions about the preferences of potential clients, even though only a minimal amount of information about them may be available.

In general, the objective of the clustering task is to divide the database into several homogeneous sub-groups. Each grouping includes people or objects whose traits are relatively similar. This division is not carried out in accordance with any predefined criteria, but is based solely on patterns found in the data itself. The user can define what variables are taken into account, how many groups are being sought, and other statistical properties of the analyzing process. The mining process itself is usually performed in at least two tiers. First, the algorithms scan the dataset and search for similarities among variables in an attempt to "group" together variables that share a certain level of statistical affinity. The algorithms thus make several initial divisions into groups and inform the user of the strengths of correlations between the defined variables within every group and of possible overlaps between the suggested clusters. After examining these options, the miner can decide on the level of acceptable accuracy and request that the software divide the database into the final grouping.

In the insurance firm example provided above, the analysis is divided into two parts in this way. The database, once thoroughly warehoused and cleansed, is initially clustered. A variety of personal variables are used for the clustering, but *not* the variable indicating the policy chosen. Several options for the splitting of the customer pool emerge, and at this stage the "policy" variable is used to determine which of the possible "clustering" options should be preferred.[30] Having decided on a specific level of accuracy, the firm establishes seven clusters, each cluster reasonably homogenous with respect to both the traits of the policyholders, and the actual policy owned.[31]

---

29 Resources for such sales appear regularly in *The DM News. See, e.g.*, DM NEWS, *at* http://www.dmnews.com/cgi-bins/listdb.cgi (last visited Nov.21, 2002)(containing ads for the sale of lists created in the automobile tool and home brewery markets).

30 That is, the analysts chose a preliminary grouping that offered clusters that were relatively homogenous with regard to the policy held.

31 Another example from the same conference involves research done for a Turkish bank. The bank conducted a data mining study in order to try and improve its service. For this matter, it created a database using the following

After establishing the "clustering", both *descriptive* and *predictive* inquiries are possible. On the one hand, it is possible to *describe* the general pool of information in several distinct categories, and by marketing pinpoint every sub-group with a different strategy. *Predictions* as to the behavior of future variables (or prospective customers, for that matter) are also feasible. When confronted with a prospective customer providing only partial or initial information, it is possible to "place" her in the relevant cluster (using the information about her that is available) and predict that the factors that are unknown at this time will match results previously obtained from other individuals within the same cluster.

Thus, regarding the insurance firm mentioned above, the firm could establish which specific attributes led to a particular customer's propensity to purchase particular policies- a *description* that provides insight to its clients and could be used in future marketing initiatives. Yet *predictive* tasks are possible as well: pursuant to the clustering process, the firm analyzed an additional database consisting of prospective clients. Using partial personal information about these individuals, the firm "matched" every potential client to a relevant cluster of the existing policyholders. The results of such assignments indicated a propensity toward a specific policy type for every one of the prospective clients (which would be the one dominant in the assigned group). This was, in fact, a prediction as to the best policy the firm could offer every prospective customer. The firm was able to turn initial, sketchy information into a reliable marketing strategy regarding the insurance policy choices of prospective clients.

2. ASSOCIATION RULES

The association rules are another application of data mining we encounter very often, usually without noting their existence. Here are two examples:

(a) This first example is from a personal experience. While searching "Amazon.com" for books on the topic of information privacy and additional issues of interest, I decided to examine one of the interesting features provided by the site- "view your own web-page". When doing so, I was quite astounded by the results: not only

---

information: balance on accounts, loans, and insurance taken out by subsidiary and utility bills paid through the bank. The bank used the clustering method and seven groups evolved. Thereafter, the common traits of these groups were examined. The results of this analysis were also used to cross-sell products and services, which were predicted as popular within every group. ALIS ET AL., *supra* note 18.

did this page provide me with many books that were within my field of research, it offered me several music CDs and DVD movies I was interested in, that at least on their face are not connected to any privacy or technology law issues! I would have never assumed that Amazon.com has the ability to predict my taste in films and in music with such precision based on the personal information they were provided with!

(b) Let us take another example, based on current publications in the field: David has been a customer of "Bank East" for several years, holding both a checking and savings account. Recently, however, he has been unsatisfied with the service provided by the bank, and decided to move his active accounts to "Bank West". He never, however, discussed his dissatisfaction with any official at "Bank East". David was therefore very surprised to receive a personal phone call from the local manager of "Bank East", informing him of great new rates the bank could offer him. The fact that David was interested in terminating his account was never mentioned explicitly yet was an obvious undercurrent in the conversation.

How did Amazon obtain such knowledge regarding my tastes? How did "Bank East" know David was interested in terminating his account?

This is where the use of "association rules" comes into play. This KDD method (also referred to as "link analysis") uses algorithms in searching the database to reveal patterns of variables that typically associate with each other. As with "clustering", association applications do not require that the user define the form sought. The algorithms "check" whether there are any *rules* that could describe the relation between various variables in the examined databases. These "rules" (or patterns) refer to logical statements such as: *If A=1 and B=1 then C=1 with probability P*[32] or other elaborations of this rule including the use of multiple factors and other boolean symbols.[33]

There are three general methods for the searches of such "rules":

(a) ASSOCIATION DISCOVERY

This method, also referred to as "market basket analysis", involves observing which events happen at the same time, or which products tend to be purchased together.[34] Such analysis is usually

32 HAND, *supra* note 12, at 158.
33 *E.g.*, If A=1 and not B=1 then C=1 with probability P.
34 THURAISINGHAM, *supra* note 24 at 100.

carried out on retailers' databases[35] that record "baskets" of items purchased. This analysis uses algorithms to search these "baskets" in an efficient manner for rules describing the way sets of items are purchased (or not purchased) together. It could also be conducted in the transportation market, looking at passenger lists for answers to the questions: "Who travels together?" and "When?",[36] a search that could reveal interesting results especially in this age of airline (in)security. Every rule "discovered" by the algorithm is ranked according to its **support** – the relative occurrence of the rule within the overall database, and **confidence** – the degree of truth which the rule has across all the relevant transactions.[37] For efficiency reasons, only associations showing a sufficient level of support and confidence are examined, since any given dataset includes endless associations. Deciding on the necessary level of support and confidence is a challenging task that requires experience,[38] as setting the level too high leads to results that include only obvious rules that are of no value to the analyst. On the other hand, setting the threshold too low leads to the accumulation of an excessive amount of rules, many of which would be far-fetched and obscure.

Example (a) is a good example of a "market basket analysis". Amazon's ability to predict customer's preferences is a result of "mining" all the shopping carts used at the Amazon.com website, in a search for patterns of behavior. Such analysis probably revealed, with an ample level of support and confidence, that a search for information or merchandise in my field of interest would be followed by interest in music and films of a specific kind.

(b) SEQUENTIAL PATTERN DISCOVERY

This form of mining is aimed toward understanding the behavior of long-term customers. This is done by identifying associations across "related purchase transactions" carried out over time. Here, obviously, the "rules" formed are more complex, as the algorithms track data accumulated about *the same* objective over a certain period of time, rather than focusing on a single transaction (or "basket"). These applications are widely used by credit card companies

---

35 Such a database could be constructed from information collected from transactions occurring at an e-commerce web site, or at the cash register of any retailer.

36 Note however, that the information resulting in the first example is not specific to any person, as opposed to the transportation information, that may hold "explosive" private information. In addition, in this analysis, the "basket" is the list of passengers on various flights!

37 CABENA ET AL., *supra* note 4, at 81.

38 CABENA ET AL., *supra* note 4, at 57.

to monitor card use and detect fraud, as these companies assume that a credit card transaction that does not conform to previous patterns of behavior is more likely to be fraudulent. Example (b) above, however, demonstrates a use of these applications that does not focus on customer security. As illustrated, banks are applying data mining to search for patterns in their customers' behavior that eventually lead to the termination of their accounts. This practice is generally referred to as "customer retention and churn[39] analysis."[40] The banks study patterns of their current customer's behavior, focusing on clients that terminated their account, and search for preliminary signals to such termination (such as the lack of deposits in the months prior to termination).[41] With the information accumulated in this analysis, banks try to prevent the termination of the account by intervening every time such preliminary signals appear. These developments are very profitable to the banking industry, which is driven by the general conviction that it is considerably cheaper to retain an existing customer than to try and recruit a new one.[42]

(c) SIMILAR TIME SEQUENCE DISCOVERY[43]

This form of data mining 'searches' for links between two sets of data that are time dependent. Retailers use this technique to examine whether a product with a particular pattern of sales over time matches the sale's curve of other products (even if the pattern is "lagging" behind with regard to the time factor). The results of such analysis could be used in "grouping" together products with similar cycles.

In addition to finding associations, the applications must "prune" the results[44] and intelligently sort them. The software must

---

39 Defined in the business management field as the number of discontinuances or termination of service encountered. *Available at* http://www.babylon.com.

40 DYCHÉ, *supra* note 25, at 63 presents a lengthy discussion on this matter. Companies view this as an issue of grave importance – as it is "three to ten times cheaper" to retain a good customer – rather than find a new one. Many companies are now engaging in "churn" analysis and software vendors are adding "propensity to churn" models to the offered DM software packages.

41 Walter J. Trybula, *Data Mining and Knowledge Discovery*, Ann. Rev. Info. Sci. & Tech. 197, 215 (1997).

42 Obviously, this claim is accurate only with regard to "good clients", but the KDD procedures have the ability to assess this factor as well, and not try to retain "bad" clients.

43 CABENA ET AL, *supra* note 4, at 69.

44 "Prune"- narrowing down the results by aggregating similar rules ("interestingness"). This issue is addressed in ROBERT J. HILDERMAN & HOWARD J. HAMILTON, KNOWLEDGE DISCOVERY AND MEASURES OF INTEREST (2001).

"group" similar rules together, thus presenting the analyst with a limited number of results to be used and implemented. This proves to be a difficult task given the immense size of the databases and the vast number of redundant rules that are usually revealed.

As with "clustering", the use of "association rules" makes both *descriptive* and *predictive* practices possible. Descriptive information is easily obtained by scanning the customer data now available. In addition, predictions can be made regarding future conduct, which will arguably conform to the patterns and rules revealed. In the examples mentioned above, it is hard to distinguish between descriptive and predictive practices, as in many events the required task calls for a mixture of both. However, pure descriptive or predictive tasks are at times required, as in the following example of a purely "descriptive" task conducted by the Health Insurance Commission (HIC) of Australia.[45] The goal of this analysis was to determine whether physicians had been prescribing needles and superfluous tests for their patients (and thereby spending HIC's funds). The data mining analysis revealed that the practice of ordering test B was associated with the use of test A with high probability. However, additional analysis found another minor trend, associating test C (which was much more expensive then test B) with the administration of test A. From these patterns, the analysts deduced it was very likely that the requests for the expensive tests (test C) were not required and represented an "unnecessary upcoding", since the ordering of test B should have sufficed. In this example, the KDD task was focused entirely on understanding the existing dataset, without trying to predict future conduct.

3. THE NEXT STEPS IN THE KDD PROCESS

Subsequent to the "mining" (be it "clustering", "association rules" or any other method), the analysts examine the results thoroughly in order to decide whether they are helpful to the corporation. If the results were found to be useful, appropriate action would be identified and implemented by the business managers of the relevant entity. Following the implementation, a "follow up" process would normally take place in order to evaluate the benefits created, introduce corrections to the analysis, and thereafter begin the next cycle of data mining and implementation. With time, the accuracy and efficiency of the various methods could be assessed and modified accordingly within this "feedback circle".

---

45 CABENA ET AL, *supra* note 4, at 106.

In today's software market, several commercial tools enable the use of the abovementioned analysis by any interested entity. These applications are not only extremely useful, but also user friendly. They are simple to operate, and accessible not only to the experienced computer analyst, but to the managerial executive as well (at least for the basic applications).[46]

D. FINAL POINTS AND INTERNET APPLICATIONS

At the present, KDD practices are used in a variety of areas, from fraud detection to the promotion of customer service. Several periodicals are published on these issues[47] and conferences are held frequently. From reviewing these periodicals and conference schedules, it is evident that a vast quantity of research and development is carried out in the area of Internet related applications. This is not surprising, as the Internet is a "data miner's paradise", presenting immense databases that are updated constantly. In addition, it is an appealing medium for the use of KDD, given that the analysis, implementation of results, and re-analysis of the feedback could be carried out automatically, without the knowledge of the consumer or even the direct intervention of an analyst or executive. Unlike the physical store, the e-commerce site could re-arrange the shelves, re-price the products and even dim the lights immediately. These data mining practices used for e-commerce sites are preformed both online (in real time while the customer is connected to the site) and offline, where immense amounts of gathered information are analyzed for patterns and clusters. Here are several concrete examples of some uses of data mining tools in the Internet setting: [48]

1. **Assessment of a Website:** Data Mining is used for assessing and evaluating the infrastructure of websites by analyzing the

---

46 *Software for Data Mining and Knowledge Discovery, at* http://www.kdnuggets.com/software/index.html, presents several tools that describe themselves as user-friendly. In addition, ISoft products present themselves as being user-friendly. *Isoft Data mining Technical Presentation, at* http://www.alice-soft.com/html/tech_dm.htm.

47 For example, the DATA MINING & KNOWLEDGE DISCOVERY, published by Kluwer Academic Publishers, a periodical from the computer science field. There are additional resources on the Internet from the business perspective. *See, e.g.*, CRM Daily, *at* http://www.crmdaily.com (last visited Nov. 21, 2002); Dstar *at* http://www.tgc.com/dsstar/dstitle.html (last visited Nov. 21, 2002).

48 Generally, I have been introduced to many of the mentioned applications at a NYU Business School course, "Data Mining and Knowledge Systems" taught by Professor R. Feldman in the Fall of 2001. For a list of the various applications, see Ron Kochavi et al., *Web Mining,* 6 DATA MINING AND KNOWLEDGE DISCOVERY 5-8 (2002).

"surfing" patterns of their patrons. The results of such analysis could be used to amend the sites architecture.

2. Recommendation Systems: Such systems offer the site's patrons recommendations that are fitted for the specific user.[49] The features presented on the Amazon.com website are an example of such applications.

3. Banner Targeting: These applications engage in tailoring the advertisements appearing on the user's screen when browsing the site, according to the specific profile of every user.[50]

4. "One on One" Marketing:[51] These applications are designed to create a different virtual store for every customer. "Customer Relationship Management" ("CRM") tools are the popular application for powering such marketing schemes. The objective of such applications is to construct a "relationship" with the customer, based on previously acquired information,[52] so that the customer would feel "at home" within the confines of the web site.

The future has much in store for KDD, as several fields that at this time are still undeveloped will no doubt make great progress. These fields include "text mining" and "multimedia mining" that facilitate the ability to search and find patterns and rules through text, audio and video. These subtopics present serious technological challenges, as usually text and media are not organized in databases and are therefore not easily searched.[53] However, if successful, such methods present great opportunities for mining and would enable

---

49 There are several ways in which recommending could be pursued: The customer could be provided with the "best sellers" list –an option that does not require any data mining. Another option could be providing recommendations according to the demographics of the customer, which would be carried out using "Clustering" tools. Finally, the recommendation could be based on past purchasing patterns. Such applications are available on the large e-commerce sites such as Amazon.com and CDnow.com. *See, e.g.*, J. Ben Schefer et al., *E-Commerce Recommendation Applications*, 5 DATA MINING & KNOWLEDGE DISCOVERY 115 (2001).

50 The issue of banner targeting has been subject to FTC scrutiny, leading to self-regulation that has been offered by the NAI (network advertising initiative). *See* http://www.networkadvertising.org/default.asp (last visited Nov. 21, 2002).

51 Companies such as ATG and BroadVision offer tools to promote and facilitate the use of such marketing schemes.

52 Axicom and Experian offer software tools for carrying out the above application, and in addition provide access to an extensive database of personal information.

53 One of the techniques used to overcome such problems is to scan the text, "tag" relevant parts and set them in a database that would be mined using the tools mentioned above (THURAISINGHAM, *supra* note 24, at 166).

searches of information appearing in these forms, on the Internet and elsewhere.

III. THE SOCIAL IMPACT OF THE DATA MINING PRACTICES

A. GENERAL OVERVIEW

Subsequent to the description of the current and to some extent future practices of data mining, we must now confront their implications, the problems they create, and the changes they bring in our lives. As above, emphasis is placed on problems stemming from the vast amounts of information made available through the Internet that will no doubt prove to be a source of future debate.

One obvious change the data mining tools bring about is a significant benefit to large corporations, now able to utilize the vast databases that are at their disposal. Using the tools mentioned in Part I, organizations are capable of gaining additional knowledge about themselves, their competitors and their clients/constituents/customers. Indeed, many corporations are rushing to implement such systems and extensive literature exists in the business management field regarding the pros and cons of such actions.[54]

I chose to put emphasis elsewhere and focus on the effects data mining has on society and on the individuals subject to its analysis, especially when the information analyzed is personal.[55] Clearly most of the information analyzed in the data mining procedures is not of personal nature, but is either general or a corporation's internal information.[56] Nevertheless, the data mining tools are bound to have a strong impact in the field of personal information analysis, even though such repercussions may not be apparent at first glance.

---

54 On these issues, see generally DYCHÉ, *supra* note 25. In addition, as part of the "Media Management" seminar in the Fall of 2001, Professor Eli Noam (Columbia Business School) described a business model that takes into account the expenses of installing and implementing the data-mining on the one hand, and the benefits that could be derived from the process on the other.

55 In this paper, I focus on the affects of data-mining on personal information. There is a great deal of writing on the general issues of personal information. To obtain a general perspective of the field, see *Symposium: Cyberspace and Privacy : A New Legal Paradigm? Information Privacy/Information Property*, 52 STAN. L. REV. 1351 (2000), as well as many of the other sources mentioned throughout the paper.

56 For example, the issue of the "assessment of the website" mentioned above, which could be carried out without the use of personal information at all.

To facilitate our discussion, let us examine several hypothetical individuals affected by the technical examples of the data mining practices discussed thus far:

1. Ms. Violet is a frequent shopper at an e-commerce "grocery" website. Upon returning to the site this weekend to conduct her weekly purchases, she is greeted by the following message:

"Welcome back Ms. Violet, I hope that you had a nice weekend. Please note that your favorite cheese is out of stock, we have a fresh supply of your favorite strawberries, and one of your preferred kinds of toilet paper is on sale. Just one more thing: Happy Birthday! (Click here for specials on birthday cakes)."

Other, less frequent customers did not receive similarly cheerful messages, and were not offered the mentioned discounts on toilet paper and birthday cakes.

2. Mr. Green is a minimum wage worker who lives in a poor neighborhood. Even though he is interested in taking out an insurance policy, he never receives any promotions, solicitations, or special offers for insurance. The situation is radically different for his identical twin, an executive with a high salary who receives many such offers.

3. Ms. Red is a member of a minority group, a fact well known to the e-commerce site she uses for shopping. After conducting marketing research, the company operating the site decided to charge members of this minority group higher rates for various products (a decision that may or may not have been motivated by bigotry).

4. Mr. Yellow is a philosophy student from a prominent family. Due to the fact that he is extremely busy writing his thesis on existentialism, he does all of his book shopping online. Upon entering an e-commerce website, he navigates straight to the book he is interested in, does not check for sales or discounts, and pays with his parents' credit card. The online vendor, observing this pattern, never informs him of any shipping discounts and has recently started to charge him a little more for every book purchased. In addition, the e-commerce site has determined the timing element of Mr. Yellow's purchasing patterns, and tends to charge higher prices towards the final weeks of the semester, when Mr. Yellow is in dire need of books.[57]

---

57 This example was inspired by Professor Eli Noam of the Columbia University Business School, who mentioned that such practices are at least possible (if not already occurring).

5. Mr. Orange often purchases through an e-commerce grocer and has recently stopped buying cigarettes. The grocer, anxious to cash in on potentially lucrative tobacco sales, notices that Mr. Orange has just purchased a "nicotine patch" and concludes that he is trying to quit smoking. Mr. Orange is then presented with cigarette ads at the websites he visits, and even receives a "complementary" cigarette pack in his most recent grocery shipment.

6. Mr. Black is a forty year old man who is paying a high premium for life insurance. The high rate is due to the fact that Mr. Black has suffered two heart attacks in the last five years (a fact that Mr. Black himself revealed to his insurer). An employee at the insurance firm with an entrepreneurial spirit launched a private data mining campaign, searching for people with physical problems who might be concealing such information from their employers. His search identifies Mr. Black as an individual who might fit such criteria, correctly as it turns out. Thereafter, the employee informs Mr. Black's employer of Mr. Black's heart condition (for a hefty fee, of course), leading to Mr. Black's immediate dismissal.

In addition, it has come to Mr. Black's attention that computer hackers have broken into the insurance firm's database, and downloaded some of his personal information.

7. Ms. White received advertising materials promoting an insurance policy for her mother who, according to the firm's records, has just moved to Ms. White's residence. Receiving the materials causes Ms. White distress, as in reality her mother has recently passed away (Ms. White had her late mother's mail forwarded to her address).[58] In addition, the insurance materials are taking up space in her mailbox and lengthening the time (and as we know, time is money) that Ms. White usually devotes to sorting her mail.

8. Mr. Blue, who is interested in taking out an insurance policy, received notice from his insurance firm indicating that he would be charged a high premium for coverage. Mr. Blue, a man of good mental and physical health, is bewildered. He does not know that the firm has erroneously categorized him as a "high-risk person" due to the fact that he takes Prozac regularly. However, the insurance company has made an error in its records—it is not Mr. Blue who is taking the Prozac, but Mr. Blues, his next-door neighbor.

---

58 This example has been inspired by GARFINKEL, *supra* note 10, at 156.

9. Ms. Gray has also received notice indicating that she would be charged a high premium for insurance. In her case however, the facts accumulated by the company are true: Ms. Gray subscribes to *Scuba Magazine*, visits Internet sites discussing bungi jumping, and travels each year to the Himalayas. Given these facts, the insurance firm concluded that Ms. Gray is a "risk-taker" and priced her policy accordingly. However, this conclusion is far from accurate, as Ms. Gray's idea of risk-taking is buying blue chip stocks and boarding the subway after 6 p.m. She is currently writing an article about the dangers of extreme sports, and travels to Tibet to visit her son.

Every one of the stories in this rainbow of hypotheticals represents a current concern in the field of information privacy. Together, they give a contemporary snapshot of the information privacy debate. Though they differ in concept and complexity, all these perspectives support a conviction that the current state of affairs in the personal information field is not acceptable, and that changes must be made immediately. These concepts have been the subject of widespread discussion and argument prior to the emergence of data mining practices, and they continue to be relevant today, when privacy experts draw on the example of data mining to strengthen their arguments about the increasing dangers to privacy.

However, not all of the privacy arguments and perspectives carry equal strength, merit, and potential for influencing the public. While some present strong claims, others have weaker analytical backing. I will address these strengths and weaknesses in the analysis below. Furthermore, I will examine the interaction between these perspectives and the implications of the growing use of data mining. Given these new technologies, some of these problems deepen, while others remain unchanged, are resolved, or are even rendered obsolete. Therefore, it is necessary to separate the crucial issues from the weaker ones in view of a changing reality. The identification of these central problems is essential not only for a comprehensive understanding of the issues themselves, but also for the later challenge of constructing solutions.

Throughout the analysis of the interaction between privacy perspectives and examples, and the use of data mining techniques, I will address the distinction between descriptive and the predictive tasks. This distinction involves two different groups of individuals whose privacy might be affected: persons whose information is present in a current database, already subject to scrutiny and analysis by the database holders, and the prospective clients/customers whose future behavior is predicted on the basis of partial information and the previous conduct of others. Each of these groups has different

expectations of privacy, is facing separate problems and therefore might require different solutions.

B. DISCRIMINATION

We first examine Examples 1-4, all of which are related to the concept of discrimination[59] that can have several manifestations in the privacy debate. These examples also demonstrate the use of "profiling" as a tool to facilitate the practice of discrimination. The ability to create profiles is part of a new reality in which vendors are able to collect vast amounts of information about present and prospective customers, including data regarding their traits and behavior, both at the present and in the future. Such information is gathered by surveying customers' conduct in the virtual or physical store and elsewhere. The gathered information is analyzed, possibly with the addition of data purchased on the now-vibrant data market, to create a profile of the individual, or of a group of individuals fitting certain criteria. After such analyses, vendors have the ability to discriminate[60] between consumers based on this profile. The discrimination could include creating a pricing scheme tailored to each customer by offering a different basket of services to distinctive groups of clients, or by avoiding certain customers based on their purchase histories.[61] Advertisers also practice profiling by promoting various products in accordance with the specific consumption traits of a consumer.[62] In Example 1, the vendor has learned the shopper's name and shopping habits, as well as her date of birth. Based on such information, Ms. Violet, as a preferred customer, received discounts that others did not, as well as personally tailored promotions.[63]

The DM perspective: The profiling described above does not require the use of data mining, yet such tools could greatly aid the vendors both through building better profiles and in reaching a wider clientele through an automated process. Using both clustering and

---

59 In this article, I discuss issues relating exclusively to discrimination within the private sector, as opposed to the public sector where there are stronger rationales for regulation.

60 At this part of the analysis, discrimination should be taken literally as "treating differently." Later, I will address legal and economic definitions of the term.

61 This could be affected by the use of direct mail, telemarketing, or the Internet.

62 *See* DIV. OF FIN. PRACTICES, FTC, ONLINE PROFILING: A REPORT TO CONGRESS (June 2000) [hereinafter FTC REPORT]. *See also NAI Initiative, supra* note 49.

63 Receiving a discount amounts to discrimination against the other customers, when compared to Ms. Violet, who receives preferred treatment.

association rules in the creation of consumer profiles, vendors will have the ability to offer customers targeted products with ease.

In order to correctly identify the source of the problem, we must ask: "What is wrong with this picture?" Does the situation described in Example 1 create a social harm, or is it a matter that requires legal intervention? Arguably, it does not. The above practices are not prohibited *prima facie*[64]—and moreover, have been an integral part of business relations for centuries. Storeowners often treated familiar, good customers with a smile. Merchants at a market set their starting prices in a negotiation in accordance with the appearance of the client, aiming for the highest price possible. Door-to-door salesmen chose to offer and advertise different products, depending on the appearance of the prospective client and home. Moreover, from the economist's perspective, in many events the practices of price discrimination are beneficial to all parties to the transaction. When a large, non-homogenous class of people is charged a single price, clearly for some the price will be high, while for others it will be excessively low. This inevitably leads to redistribution within the class. By the

64 It should be noted that price discrimination is prohibited according to the Robinson Patman Act, 15 U.S.C. §13 (a) (1982). The elements of a prima facie case are (1) a price difference (2) between sales to two buyers (3) of commodities (4) of like grade and quality (5) that creates a reasonable possibility or probability of competitive injury.

The major legislative purpose behind the Robinson Patman Act was to provide some measure of protection to small independent retailers and their independent suppliers from what was thought to be unfair competition from vertically integrated, multi-location chain stores. Two types of possible injury are most commonly alleged. The first is often referred to as "primary line injury," because the actual or threatened injury is to competition between the seller granting the discriminatory discount and other sellers. The second type of injury is often referred to as "secondary line injury," because the actual or threatened injury is to competition between the favored customer of the seller who receives the discriminatory price and the seller's disfavored customers. *See* FTC Secretary Donald S. Clark, The Robinson-Patman Act: General Principles, Commission Proceedings, and Selected Issues, Address Before The Ambit Group Retail Channel Conference for the Computer Industry (June 7, 1995)(transcript available at http://www.ftc.gov/speeches/other/patman.htm).

In this article, we are concerned with the detriments price discrimination might cause potential buyers, rather than other sellers. The practices described in this article are not intended to undercut competitors and competition and are focused on retail sales. Therefore, this Act is not applicable to most of our discussion as the problems we address are neither primary nor secondary line injuries. On the inapplicability of the Robinson-Patman Act (as well as its general flaws), see Mark Klock, *Unconscionability and Price Discrimination*, 69 TENN. L. REV. 317, 360 & 370 (2002).

segmentation of a clientele database into homogenous groups, such redistribution can be avoided.[65]

To find a substantial social harm emerging from these profiling and discrimination practices, we must look beyond the simple Ms. Violet example to more elaborate hypotheticals, starting with Mr. Green, the minimum wage worker. Example 2 shows a use of profiling methods that result in harm to poorer sectors of the population, which might be neglected or avoided based upon their personal information.[66] These situations arise out of popular business strategies, which instruct corporate entities to focus on prominent clients and to neglect individuals who are not predicted to generate large revenues.

The DM perspective: The use of KDD tools would surely assist in carrying out this form of discrimination, by enabling vendors to predict with a high level of accuracy the spending ability of each client, as well as the probability of default on payments, and to conduct business accordingly. For example, a company can use data mining to locate groups with the strongest buying power, and solicit these groups exclusively (in one case, a company recognized that by using these tools, it could send out fifty percent fewer advertisements, and still reach ninety percent of its best prospects)[67]. The result of such practices would be that less well off individuals will receive fewer discounts, solicitations, and sales information, and possibly an inferior variety of goods. The identification and seclusion of the individuals forming this group could be carried out effortlessly.

Even though the described behavior and misfortune befalling Mr. Green seems infuriating, it does not seem to present a legal issue, or one that is contrary to public policy. In a capitalist regime, such problems should be resolved on their own, through the powers of the market. In a well-functioning market, other companies would step into the place of those that are neglecting a certain segment of the population (provided that there is no shortage of goods) and supply that segment with its needs. The government might choose to intervene should these practices amount to covert discrimination, or by providing welfare assistance,[68] but there is arguably no reason to

---

65 *See* Murphy, *supra* note 9, at 2385. Regarding the benefits of such price discrimination, see Jonathan Weinberg, *Hardware Based ID, Rights Management, and Trusted Systems*, 52 STAN. L. REV. 1251, 1274 (2000).

66 The sectors of the population in this example are defined by strict economic criteria. Racial and other problematic criteria will be discussed with regard to another example.

67 CABENA ET AL., *supra* note 4 at 123.

68 Some issues are of more concern than others—for instance, the FTC remarked on the difficulty encountered by certain classes of society in receiving life

interfere with the actions of vendors who choose to focus their marketing initiatives on a specific segment of the public.[69] In conclusion, the claim that data mining causes discrimination against the weak is weak itself, unless accompanied by additional evidence, such as claims of welfare, unfairness, or market failure (which could amount to sufficient cause for legislation, as discussed below).[70]

We therefore proceed to examine the case of Ms. Red and Example 3, in which the discrimination is based on certain criteria, the use of which is considered to be prohibited by, or adverse to, public policy (such as gender, race, or nationality).[71] In today's market, corporations are able to obtain information regarding their customers' personal traits and apply these kinds of problematic criteria in their pricing schemes.[72] Stereotypes could be used to decide which clients should be actively pursued and which ones snubbed. The motivation for such forms of discrimination could be either subconscious or overt bigotry,[73] yet might also result from the vendor's efforts to increase revenue or profits from specific transactions, using stereotype modeling merely as a tool to pursue these objectives.[74] These practices

---

insurance at a decent rate. FTC REPORT, *supra* note 62. This might be an area where government intervention is required.

69 Professor Lessig does not agree with this view, and finds this situation problematic, as it interferes with an important principle of equality. L. LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE 154 (1999).

70 Note that customers who state that they have been wrongfully designated as "bad customers" have a different claim. *See infra* pp. 45.

71 For the affect of cyberspace on racial issues, see J. Kang, *Cyber Race*, 113 HARV. L. REV. 1131 (2000).

72 According to *Weblining*, BUSINESSWEEK, Apr. 3 2000, Axicom now offers a new service called the "InfoBase Ethnicity System," which provides ethnic and racial information about millions of American households.

73 Regarding these issues, the work carried out by Ian Ayres regarding the sale of cars is extremely relevant. *See, e.g.*, Ian Ayres, *Fair Driving: Gender and Race Discrimination in Retail Car Negotiations*, 104 Harv. L. Rev. 817 (1991) [hereinafter Ayres, *Harvard Article*]; Ian Ayres, *Further Evidence of Discrimination in New Car Negotiations and Estimates of its Cause*, 94 Mich. L. Rev. 109 (1995) [hereinafter Ayres, *Michigan Article*]. In these papers, Ayres described the used cars market and the problematic negotiations that facilitate sales to minorities and women. Discrimination that is motivated by "pure bigotry" is referred to as (1) "associational animus" when the sellers dislike the buyers because of their race, and wish to compensate themselves for the time spent with the buyers by charging a higher price; or (2) "consequential animus," when the discrimination is due to the seller's special desire to disadvantage a certain group. Ayres, *Michigan Article*, *supra* at 125. The article indicates that such forms of discrimination would not survive in a well-functioning market (referring to GARY BECKER, THE ECONOMICS OF DISCRIMINATION (University of Chicago Press, 2d ed. 1971)).

74 Ayres mentions that such discrimination should be described as (1) cost-based statistical discrimination, which takes place when the seller believes a certain group is more averse to bargaining; (2) revenue-based statistical discrimination, when there is a belief that there is a difference in the distribution of

are at times referred to as "Weblining,"[75] due to their similarity to the "redlining" process through which segments of the population have historically been denied services, or offered a lower level of products, based on their geographic place of residence that was used as an indicator of race. The process of "Weblining" can be carried out at a higher level of sophistication through the use of "one-on-one marketing"[76] that is gaining popularity in the Internet society. By using this marketing method, discriminatory conduct can be concealed, and racial and other prohibited profiling methods can be carried out, camouflaged as economic analysis. Arguably, such means of discrimination are either illegal or contrary to public policy that aims to treat all people as equals and regards traits such as race, gender and nationality as irrelevant, or unsuitable for differentiating between individuals.

The DM perspective: At this point, several perspectives can be seen. One perspective would assert that the use of KDD tools would assist in these discriminatory practices and perhaps bring them to an even more effective level, enabling the structuring of profiles for every individual customer, or clustering individuals into racial groups while providing a fertile opportunity for discrimination. The discrimination can be carried out with greater ease, thanks to the automated characteristics of KDD tools. The customer might never suspect such discrimination is taking place and would be unable to ascertain the breadth of the profile constructed nor determine how certain assumptions are being formed. Therefore, since this view considers data mining as the promoter of such virtual discrimination, this aspect

---

the reservation prices between different. Ayres, *Michigan Article*, *supra* note 71, at 137. Ayres also states that discrimination can result from an assumption that specific groups have higher opportunity costs, and therefore less ability to compare prices. Ayres, *Harvard Article*, *supra* note 71, at 849. Competition would not deter such discrimination, as the sellers stand to gain greatly from just a few profitable "sucker" transactions. *Id.* at 853.

75 *See* Paul M. Schwartz, *Beyond Lessig's Code for the Internet Privacy: Cyberspace Filters, Privacy-Control, and Fair Information Practices*, 2000 WIS. L. REV 743, at 757 (2000); *see also Weblining*, *supra* note 72; FTC REPORT, *supra* note 62. The use of the zip code, or area of residence factor, was initially used as a racial indicator. An interesting interaction between this old and new "redlining": Wells Fargo was sued for having a web site that was used to assist in finding housing—and would automatically refer customers to apartments based on their current zip code. Dee DePass, *Wells Fargo Pulls Criticized Data; Housing Information on Web site is Labeled Racist*, STAR TRIB., June 23, 2000, at D1. The result of this application was that people of specific backgrounds were continuously referred to specific areas rather than others. The resulting suit was based on the violation of housing laws that prohibited such conduct. For more on this incident, see Mike Hatch, *The Privatization of Big Brother: Protecting Sensitive Personal Information from Commercial Interests in the 21st Century*, 27 WM. MITCHELL L. REV 1457, 1485 at fn 173 (2001).

76 Generally, the ability to market to a specific customer, based on specific needs and traits.

of discrimination should be dominant in the public debate over the permitted breadth of the data mining practices. It should figure as a cornerstone in any attempt to construct solutions for the various privacy problems arising from the data mining of personal information.

I believe in an opposing perspective, though, contrary to the above argument. I assert that the importance and relevance of these problems to public debate is minimal, and the effects of discrimination of the kind discussed above should not be a major factor when constructing a solution regarding the permitted data mining practices. I base my assertion on the following arguments:

First, discriminatory practices motivated by a determination to increase profits or revenue are already serious issues, regardless of the future use of data mining applications that would have only a marginal effect on the legal and social landscape. These issues should be solved by specific regulation that would leave the greater concerns of the data mining phenomena extant, and therefore should not be presented as central in the data mining debate.[77]

Moreover, it is possible to view the effect KDD tools have on this matter of discrimination in a different light, and possibly find that the use of such applications should be encouraged, as they may be of great assistance in diminishing the negative effects of such discrimination and "Weblining." The support for such statements rests in the salient traits of the data mining process: automation, and the fact that human-generated hypotheses are not required for the analysis.[78]

Here are two examples to assist in understanding the benefits of these applications:

1. As a result of ongoing video surveillance, the "unequal gaze" problem has arisen.[79] It has been claimed that when surveillance tools are controlled manually, examining their recordings leads at

---

77 Such regulation could include a restriction on the use of specific criteria in data mining analyses, eliminating the issue at hand, yet leaving many of the larger issues to be discussed unsolved. Note that some of these practices are adopted independently by the practitioners. *See* CABENA ET AL., *supra* note 4.

78 These claims are more relevant to discrimination motivated by bigotry, or subconscious resentment. Even though it is claimed above that such conduct would not hold up in a competitive market, the effects of such conduct still pose a problem that should be addressed. Discrimination should be avoided in the interim period, as the dynamics of the competitive market do not always reflect the model.

79 *See* GARFINKEL, *supra* note 10, at 116.

times to the finding that the surveying device is not gazing evenly, but tends to focus on minorities, even when these individuals are not exhibiting suspicious behavior. The obvious result of such unequal gazing is the finding of a higher rate of events involving minorities than non-minorities, as they are the people who are constantly being surveyed. This example depicts how manual intervention may lead to a discriminatory result.

2. Unlike the above example, which focused on the information collection stage of data mining, this example focuses on the data analysis stage that follows. In the process of creating a marketing strategy, an analyst might commence an analysis using biased hypotheses rooted in racial or other discriminatory beliefs. In the event that one of these hypotheses proves to be of statistical merit, it is possible that such a hypothesis would be implemented into the company's strategy, leading to an adverse effect on minorities. Other hypotheses that are not racially based would not be tested, due to the concentration on bias-based assumptions.

When data mining tools are used for these tasks, the problems addressed in these examples can be mitigated. By using KDD, the entire process is carried out via computer algorithms that present the final result without being manually focused on one group or another by a human eye or arm[80] and after taking into account *all* available information.[81] When applying data mining, the results of database analysis are balanced, displaying patterns drawn from the population in general that were chosen according to objective criteria and not subjectively driven.[82] The difference between the two forms of analysis is the inherent difference between the data mining tasks of *clustering*, and the hypothesis driven practice of *classification*.[83] Using classification, the analyst decides what hypothesis to examine, chooses what query to post, and offers the taxonomy for the grouping of the factors to be compared. If the analyst experiments with classifications

---

80 As mentioned above, most data mining procedures *do* require manual intervention, yet it is of the technical sort and does not necessarily affect the substance of the result.

81 Since it is possible to analyze all information, there is no need to "focus" the collection and narrow down the amount of information accumulated.

82 As opposed to computers, humans tend to err when forced to make decisions regarding large datasets. They tend to rely on heuristics that are true only part of the time, rather than solid logic. That is why I believe that minimization of the human component is advised in this context. For more on the issues of heuristics, see Amos Tversky and Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1124-31 (1974).

83 Unlike clustering, classification is the practice of breaking data into groups according to set criteria. This practice is at times referred to as "data mining", yet it lacks in the needed factor of "discovery" to be considered as such. *See* Thuraisingham, *supra* note 24, at 106.

based on his or her own racial or discriminatory views, he or she might be affecting the outcome. When using *clustering,* however, non-hypothesis grouping is not a priori based on any racial, economic, or gender factor, but rather on the aggregation of several variables that were chosen by the computer for their affinity.[84]

This pro-data mining argument would surely meet a strong opposition claiming that the above description contains in its essence the greatest flaw of the data mining process—that it is entirely automated and lacks human influence.[85] Yet I do not believe that the human touch on its own holds special merit, so that we should prefer it to data mining applications that do not require a person generate the original hypothesis. Moreover, the situations described above prove that better results will arise, especially for minorities, with less human intervention. In conclusion, the use of data mining may have an overall positive effect with regard to this sub-issue of discrimination, forcing us to continue our quest for the key social problem data mining might create.

We turn now to our last discrimination example, the one of Mr. Yellow (Example 4). In this hypothetical, the vendor charges Mr. Yellow a "marked-up" price using the information collected and obtained. This example differs from Example 1, as here the seller does not use personal information to segment the market and avoid redistribution between customers, but rather uses it to the detriment of specific customers in an attempt to enrich the firm.[86] The seller does this by manipulating the buyer and overcharging at a time of need, probably without the customer being aware of such overpricing. The sellers are now able to carry out such manipulations due to their ability

---

84 I concede that this claim is somewhat problematic. There is always the fear that discriminatory conduct could be hidden behind the façade of so-called scientific evidence indicating that even with no a priori presumption certain groups are created -- groups that have racial or other problematic characteristics. A plausible solution to this problem could be restricting the use of such variables in a KDD analysis employed in the marketing setting (a practice that has been adopted by some voluntarily—*see supra* text accompanying note 77).

85 These views are popular in the EU. *See* Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive* 17 COMPUTER L. & SEC. REP. 17 (2001).

86 I am assuming that the book vender in this example indeed is not using the excess profit from the transaction with Mr. Yellow to subsidize transactions with other clients. Still, it could be claimed that the fact that the allocation of such a sale's surplus to the seller is a plausible outcome, as, at the time of the transaction, the seller is willing to pay the price charged. I do not agree with this claim, and believe that a firm's use of personal information in such instances (especially information regarding the timing of the transaction) is unfair, manipulative, and in many cases, allows them to maintain virtual monopolies. On this issue, see Klock, *supra* note 64 at 327.

to treat every client differently, their ease of collecting vast amounts of personal information, and their ability to analyze this data and obtain from it powerful insights to the customers' preferences.

To understand the dynamics of Example 4, we must first readdress the "one-on-one marketing" schemes, in which every customer is treated separately and possibly differently. Using such methods, the vendors are able to offer every customer a different price and layout of products and services. Such schemes are increasingly popular in the Internet setting, where the page the specific consumer views could be easily individualized.[87] However, this marketing mechanism may spread with the use of new technologies to other mediums of retail in the near future. When using such applications, several potential outcomes exist: One possibility would be every customer easily finding the best "deal" without being troubled with products clearly outside his or her scope of interest or price range (as in Example 1). There is another possibility; Companies could use the "market of one" scheme to the financial detriment of the consumer while taking advantage of the personal information they have gathered – information usually conceded by the customer himself or herself. The sellers could use the personal data they obtained to manipulate the prices, discounts, and services offered in various ways. Products could be overcharged when it is evident from the information at hand that the customer is uninformed, unaware of competing products, hurried, or simply indifferent to the price of the product.[88] In our example, the e-commerce site was able to predict Mr. Yellow's needs, financial ability, and attention span, as well as the *time* the site's products were at highest demand, and then was able to exploit this knowledge to its advantage. From the economist's perspective, the information obtained by the sellers provides them with a strong indication of the customer's demand curve[89] and reserve price, while the customers have no idea of the vendors' reserve values, therefore placing them at a

87 Regarding these claims one could counterclaim that any online customer could easily counter these practices by logging on anonymously to the vendor's site (or other competing sites) and inquiring whether the products are sold to others for a different price. To this I would respond as follows: (1) not all are aware of such practices; (2) in the future (and perhaps at the present, with the use of cookies and webworms) such anonymous visits would not be possible, especially with the use of biometric identification; and (3) the discrimination could take place by providing some customers (who fit a specific profile) with discounts and special services. Therefore, a visit to another website would not be helpful.

88 In these cases, the sellers can present higher prices without the fear of competition, as they can identify the situation as one of a "virtual monopoly" at that given time frame.

89 Thus, from the economic point of view, the seller has the ability to predict how much the customer would be *willing to pay* for the product.

considerable disadvantage.[90] The practical consequences of this phenomenon are poor and misinformed people paying higher prices for products due to ignorance of information market dynamics. The wealthy and uninformed, on the other hand, could also be constantly overcharged without their knowledge, if classified as big spenders rather than thrifty and careful consumers.

Yet another aspect of such discriminatory practices is "online advertising". The content of "banners" is tailored to the various Internet users according to information collected about every user. The banners offer a variety of products and discounts to different types of customers, setting prices and discounts according to the relevant customer profile. This issue has received significant attention during recent FTC hearings and reports[91] citing concern regarding possible "price discrimination" practices. Nevertheless, the matter of online advertising should be of secondary concern when compared to the possible retail dynamic articulated above.[92]

The vendors' ability to collect vast amounts of personal information is key to their ability to employ the described pricing scheme. Surprisingly, the customers themselves usually provide such information. This reality is difficult to comprehend, as we would assume that had the customers known of such practices, they would have protected their personal information vigorously and certainly would not have disclosed information voluntarily. Alternatively, since many transactions in this day and age could be viewed as trades of goods for information, the rational customer should be bargaining diligently, prior to conceding personal information, to assure appropriate compensation. Returning to our example, we might say that had Mr. Yellow known of pricing schemes used by the e-commerce site, he surely would have concealed his real identity or alternatively would have demanded proper compensation prior to surrendering such data. In actuality, however, the concealment of personal information or the rigorous bargaining described above is not occurring. Somehow, consumers are unaware or are generally unconcerned about the implications of surrendering personal information that is thereafter easily gathered. This fact could be due to:

---

90 Marc Rotenberg, *Fair Information Practices and the Architecture of Privacy (What Larry Doesn't Get)*, 2001 STAN. TECH. L. REV. 1, § 104 (2001).

91 See FTC REPORT, *supra* note 62, at 763; and the NAI that followed, *see supra* note 50.

92 As the actions of the vendors are those that cause actual loss. In contrast, advertising plays a major role in the "autonomy trap" claim (discussed below).

1. **Unfair and Deceitful Conduct of the Vendors:** The vendors are collecting the personal information of their customers for future use without explicitly disclosing that such information is gathered and what its future uses might be. The most popular form of such collection is the tracking by the use of "cookies" or "web bugs" that are installed on the users' computers, often without their consent or knowledge.[93] Such conduct could be considered deceitful and unfair.

Note however, that the "unfairness" stated above, though widely addressed,[94] is based on a *moral* standard (one that is subjective and malleable), rather than a legal one.[95] In a set of recent federal cases, courts in several circuits failed to find the use of "cookies" for information gathering unfair after reviewing various federal statutes.[96] The court responded to such claims of unfairness by providing the following circular argument: "It is simply implausible that the entire business plan [involving the use of 'cookies'] of one of the country's largest Internet media companies would be 'primary motivated' by a tortuous or criminal purpose."[97]

---

93 However, such collection is carried out in the "real world" as well. One example is the collection of personal information by pubs when swiping the driver's license of a patron. *See* Jennifer Lee, *Finding Pay Dirt in Scannable Driver's Licenses*, N. Y. TIMES, March 21, 2002.

94 As for such unfairness, The FTC finds the authority to intervene with regard to this matter in its mandate to address issues of "unfair" practice granted in Section 5 of the Federal Trade Commission Act (Steven Hetcher, *Changing the Social Meaning of Privacy in Cyberspace*, 15 HARV. J. L. & TECH. 149, 172 n.81 (2001)).

95 Note, however, that rendering such conduct "unfair" is problematic, as technically the use of "cookies" could be blocked by the user, who is conceding to surveillance by agreeing to visit the site.

96 The salient example to such a case is *In re DoubleClick Inc. Privacy Litigation* 154 F. Supp. 2d 497 (S.D.N.Y. 2001); 2001 U.S. Dist LEXIS 3498. In this case, plaintiffs sued DoubleClick for the collection of various forms of information through the use of "cookies." The information collected was described as (1) Get—the parts of the URL that contains a "?," which indicates a query posted (2) Post—the places where blanks have been filled within a web page; and (3) GIFs—areas providing information regarding movements within the affiliated websites. At first, the court found that such practices are not prohibited according to Electronic Communications Privacy Act (ECPA), 18 U.S.C. 2701, an anti-hacking law; in fact, DoubleClick's conduct has been authorized as well as the relevance of several exceptions within the said Act (*Id.* at 506). The Wire Tap Act was also found not to be applicable (*Id.* at 513), as the plaintiff could not prove that DoubleClick was primarily motivated to commit a crime or a tortuous act. Lastly, the court confronts the Computer Fraud and Abuse Act. Here as well the court found that the statute is insufficient, as "damages" and "losses" of over $5,000 have not been proven. An additional case presenting similar facts and conclusions is *In re Toys-R-Us Inc. Privacy Litigation*, No. M-00-1381 MMC, 2001 U.S. Dist. Lexis 16947 (Oct. 9, 2001).

97 *Dane Chase v. Avenue A, Inc.* 165 F. Supp. 2d 1153, 1163 (W.D. Wash. 2001), which presented facts similar to those addressed in the *DoubleClick*

Additional acts that are claimed by some to be "unfair," and would probably fail to pass legal muster at this time are (1) the *sale and purchase* of personal information by vendors and data collectors from and to third parties, and (2) *secondary use*—the use of personal information for a purpose other than the one intended by the submitting party.[98]

This moral "unfairness" may develop into a legal matter when the vendor's actions (such as secondary use or sale to a third party) are contrary to their own privacy policy or specific representation,[99] yet such situations are uncommon as the representations made are, in many cases, vague and cloaked in heavy "legalese."[100]

2. Market Failure: Several commentators note the personal information market's inherent flaws, as customers lack the power, the information, and the understanding to negotiate effectively for a good bargain. This result occurs, among other reasons, due to the high transactional costs entailed in reaching such informed decisions,[101] the practical inability to price personal information[102] and public ignorance regarding these matters. It is further assisted by the pro-collector default rules, which provide collectors with the right to use the information without offering the consumer an "opt in"

---

litigation mentioned above. This statement was made with regard to the relevance of the Wire Tap Act when the court examined whether it is plausible that the defendant's actions have been purposefully illegal or tortuous. Even though this statement could be viewed narrowly, it indicates the court's general attitude toward such practices of gathering information.

98 The application of personal information for uses other than the purpose for which it was collected is a problematic practice to say the least. The European Union (EU) Directive regarding the process of personal information prohibits such practices, and laws along these lines have been enacted throughout the continent (*See* Fred Cate, *The EU Data Protection Directive, Information Privacy and the Public Interest*, 80 IOWA L. REV. 431, 433 (1995)). However, in the U.S., such protection depends on self-regulation, not a grant by the legislator. *See id.* at 437. *See also* J. Reidenberg, *Resolving Conflicting International Data Privacy Rules in Cyberspace*, 52 STAN. L. REV. 1315, 1331 (2000).

99 For example, the FTC chose to intervene when Toysmart.com attempted to sell its customers list, contrary to its explicit presentations and privacy statements. *See FTC Announces Settlement with Bankrupt Website, Toysmart.com, Regarding Alleged Privacy Policy Violations* (July 21, 2000) at http://www.ftc.gov/opa/2000/07/toysmart2.htm.

100 For an analysis of the vagueness of privacy policies and their effectiveness, see Hetcher, *supra* note 94, at 176.

101 J. Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193, 1253 (1998).

102 Lessig, *supra* note 69, at 116.

choice.[103] Therefore, collectors are able to obtain large amounts of information for a very low price.

The third development (in addition to the collection of personal information and the "marketing of one" environment) that enables vendors to implement such price discrimination schemes is their ability to analyze such vast amount of information. Through the use of powerful computers and skilled analysts, vendors are enhancing their ability to analyze the personal data they have now obtained. Here, clearly, the data mining perspective is of great relevance.

The DM Perspective: The introduction of the data mining tools to the scene only exacerbates this form of price discrimination. First, the KDD tools can assist the sellers and other data miners in improving their pricing and marketing methods. By means of predictive models (both using clustering and association rules), vendors can obtain a strong indication of the future buying habits of prospective clients, thereafter avoiding discount offers when certain transactions occur with high probability. Moreover, using KDD, vendors would have the capability of gaining additional insight into the proper *timing* of their actions and would use such information to their benefit (as with Mr. Yellow and his book purchase). By using KDD, sellers receive strong indications regarding the customer's demand curve as a function of time, enabling them to deduce at what instant the customer's demand would be at its peak.

Furthermore, the data-mining environment would enlarge the disparity between the value customers assign to their personal information, its actual value to the vendor, and the possible detriment it may cause the information provider, contributing to the severity of the market failure addressed previously.[104] By using KDD, the database holder has the ability to deduce a great amount of important information from only "scraps" of data disclosed by the customer. Therefore, it is extremely difficult (if not impossible) for consumers to evaluate how a specific bit of information would contribute to their overall profile. Any "snippet" of data could turn out to be of no importance to the collectors, or the missing piece of data needed in order to place the consumer in a specific cluster. As a result, it would be even harder for customers to decide whether to conceal or disclose personal facts as part of a transaction, or to determine what value to

---

103 On the issue of market failure, see P. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1607, 1682 (1999) [hereinafter *Privacy and Democracy*], for a discussion of the four reasons for the mentioned market failure.

104 Using the terminology stated above, data mining could create greater transactional costs for the consumers, since revealing the actual value of the information grows more difficult.

attach to every bit of information. Eventually, consumers would end up providing vendors with a greater advantage than before.

Therefore, unlike the examples previously discussed and in view of these powerful arguments, we could state with conviction that the availability of data mining tools and processes contribute directly to the severity and depth of this problem of price discrimination. This issue should be a dominant claim in any discussion concerning the regulation of the collection, analysis, use of personal information, and data mining especially.

I conclude our discussion regarding the discrimination examples with a linguistic remark regarding the term *weblining*. This term is often used to describe the discrimination issue as a whole, yet it is unsuitable for this purpose. "Weblining" focuses the attention of the privacy debate on a specific discriminatory aspect (the one of racial discrimination described in Example 3), a feeble argument prone to refutation. The use of the phrase "weblining" insinuates the dominance of this specific aspect of the greater problem, when emphasis should be placed elsewhere.[105] It is not in our interest, as consumers seeking "protection" in the ongoing data mining battle to use this term, since the word choice is important and the use of the term "weblining" is not sufficient in describing the greater problems at hand.

C. MANIPULATION AND THREATS TO AUTONOMY

Leaving our discussion regarding discrimination, we turn to our fifth example of Mr. Orange and his attempt to quit smoking. Unbeknownst to Mr. Orange, he is not alone in this battle to quit. Others know information about his personal habits and behavior, giving them the ability to impede his efforts to quit through the use of various manipulations aimed at affecting his free will. I will refer to these manipulations as the *autonomy trap*,[106] which may have implications on the autonomy of the individual[107] and on society as a

---

105 "Price discrimination" might be a better choice of words to describe this issue.

106 The "autonomy trap" is addressed in P. Schwartz, *Internet Privacy and the State*, 32 CONN. L. REV. 815, 821 (2000) [hereinafter *Internet Privacy*]. Schwartz describes one of the traits of said trap as the "reduced sense of the possible." *Id.* at 825. An additional discussion of this concept is found in *Privacy and Democracy*, *supra* note 103, at 1662. However, the use of this important term in this paper is somewhat different, as it concentrates on the loss of autonomy due to the reception of specifically tailored information.

107 In this context, autonomy is best defined as "*second order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or to attempt to change these in light of higher-order preferences and values.*"

whole. This claim differs from the above mentioned price discrimination claims since it does not pertain specifically to marketers and e-commerce sites but, rather, is relevant to the media in general (with emphasis on content providers).[108] The claim is also one of higher complexity, dealing with a philosophical concept, rather than a mundane phenomenon.

The roots of this argument are set in the cultural background of the information flow at the beginning of the 21st century. At this time, much of our knowledge, lifestyle, beliefs, and consumer behavior are dictated to us by the media.[109] Fashion taste and related ideas are projected to us by "experts," celebrities, and other public icons. Their message is transferred to us by mass media tools,[110] which constantly bombard our senses with mantras and affect our ability to engage in autonomic thought. Such practices continue even though the media market is competitive and other forms of media are free to compete for the public's attention by providing other forms of content and information. In practice, however, the sum of voices that reach us is limited. This paucity is due to the high entrance barriers to the mass-media market and the common trend of mergers and consolidations, which together contribute to the diminishing of independent voices.[111] Moreover, some scholars claim that even a competitive media market will not eliminate the media's ability to impede on the public's autonomy for various reasons.[112]

G. DWORKIN, THE THEORY AND PRACTICE OF AUTONOMY 20 (1988). *Privacy and Democracy*, *supra* note 102, at 1655, uses a similar definition: "*...the vision of people controlling to some degree, their own destiny, fashioning it through successive decisions throughout their lives*" (citing JOSEPH RAZ, THE MORALITY OF FREEDOM 369 (1986)).

108 In "content providers," I refer to sites that provide the users with information, such as news sites and major Internet portals.

109 The strength of mass marketing has been discussed and debated by sociologists for years. For one view, see generally VANCE PACKARD, THE HIDDEN PERSUADERS (1957). In this book, Packard described how advertisers mold our inner thoughts and dreams in their attempt to convince us to purchase specific brands. The various tools of persuasion have been discussed in a recent article by Emily Eakin, *Penetrating the Mind by Metaphor*, N. Y. TIMES, Feb. 23, 2002.

110 Regarding the ability of the mass-media to "corrupt segmentation" within society, and the effects of advertising on this process, see C.E. Baker, *Giving the Audience What It Wants*, 58 OHIO ST. L. J. 311, 336 (1997), and C.E. Baker, *The Media that Citizens Need*, 147 U. PA. L. REV. 317, 375 (1998).

111 *See* Y. Benkler, *Free as the Air to Common Use: First Amendment Constraints on Enclosure of the Public Domain*, 74 N.Y.U. L. Rev. 354, 369 (1999), regarding the concentration of the broadcasting market.

112 *See* Y. Benkler, *Siren Songs and Amish Children: Autonomy, Information and Law*, 76 N.Y.U. L. Rev. 23, 68 (2001). Benkler mentions several reasons for the failure of the "market will solve it" hunch regarding this issue, including high transactional costs and negotiation costs for the individuals.

The rise of the Internet was thought to reduce the success of such persuasive means and to present serious challenges to those attempting to control public tastes and opinions, as the availability and accessibility of information via the World Wide Web has greatly changed for the better. The Internet presents lower entry costs, enabling almost anyone with an interest to start a website and share his or her thoughts with the world. Yet the promise of the Internet regarding these issues is yet to be fulfilled. With time it has become clear that obtaining public recognition and reputation on the Internet is hard given the number of voices. In addition, it is evident that Internet content markets are extremely difficult to penetrate, especially ones with strong incumbents.[113] The result is that many of the independent and non-mainstream content providers remained unheard, while the majority of Internet traffic is focused on the major content providers.

The ability to persuade and influence by the means of the media market have been greatly enhanced due to recent technological progress, which changed the market in two ways. First, content providers are accumulating vast amounts of personal information regarding the tastes and fields of interest of their customers. Such information is collected not only by Internet content providers[114] using the methods mentioned above, but also by conventional broadcasters.[115] This data could be later added to other data resources purchased from third parties to create a complete data profile on every content-receiving individual. Secondly, content providers can now provide every customer with specifically tailored content, which will differ from one customer to the next—thus perfectly segmenting the market. The mixed effect of these phenomena brings the "autonomy trap" claim to life. According to this claim, the following vicious circle could be created:

---

113 The "pioneers" such as Amazon.com, Ebay, and Yahoo!, as well as "real world" content providers such as CNN, have "taken over" a great amount of the Internet traffic. Such "first movers" have gained great advantages in the Internet environment (among others, due to a network effect taking place online).

114 The personal information would also accumulate in the hands of other Internet "players" such as the ISPs or other entities that have access to surfing habits or cookies. In the Internet environment, several companies constructed business plans that include the collection and use of personal information of Internet users. For example, Hotbar.com, a site that provides a recommending system to similar sites, and in return provides the site operators with the users entire browsing history. In addition, it has been revealed that several "free applications" such as Media Player and RealPlayer have tracked customer's surfing habits. Regarding the MS Media Player, see *Technology's Threats to Privacy*, N. Y. TIMES (Feb. 24, 2002).

115 Such information could be collected by means of the new services that provide selected programming to customers ("pay per view"), thus allowing the broadcasters insight to the individual's taste and schedule.

(a) Individuals inform the information providers which types of knowledge and information they are interested in and provide (both implicitly and explicitly) personal information such as their traits and interests;

(b) The content providers supply individuals with specific information "tailored" to the needs of every person, according to each provider's specific strategy, and chosen on the basis of the personal information previously collected;

(c) The individuals require additional information. This time, however, the request is affected by the information previously provided;

(d) Again, the information providers supply information, in accordance with their policies and discretion;

And so on. This "vicious circle" could effect both (1) the individual, and (2) society as a whole, as follows:

1. EFFECTS ON THE INDIVIDUAL

The autonomy trap will be sure to have a strong effect on individuals. By creating this "vicious cycle" it will "push" individuals towards certain products or services in which they initially were not interested (as with Mr. Orange's attempt to quit smoking). This is achievable by narrowing down the options they receive and offering persuasive arguments at exactly the right time, thus impeding their autonomy. The resistance the product "pushers" encounter would be substantially lower than before, since they are capable of "pinpointing" their campaign for the "target" customer (using special content for every type of individual), and receiving almost immediate and specific feedback as to the campaign's success or failure (thus enabling them to tinker with the future feedback until proven successful). An example of such practices is evident in the world of *online advertising*.

Advertising has been practiced for centuries, and has been known for its ability to influence the public through the use of various manipulations. The overpowering ability of advertisements in certain settings has been acknowledged, and the investigation and enforcement of "unfair advertisements" in the United States is one of the duties of the Federal Trade Commission ('FTC').[116] Recently, the

116 An example of the FTC intervention regarding "unfair advertising" is the FTC's complaint against R.J. Reynolds regarding the use of the "Joe Camel" figure in cigarette ads. The FTC investigated allegations that such use was targeting young children in an attempt to convince them to adopt smoking habits – an unfair means of advertising. *See* FTC Press release *Joe Camel Advertising Campaign Violates*

FTC investigated whether the use Internet advertisers make of 'customer profiles' in the distribution of advertisements constitutes an "unfair" practice. One way to understand the source of such "unfairness" is through the use of the "autonomy trap" paradigm: the fact that the "banners" are "tailored" for every customer according to his specific profile, thus enabling the advertisers to use an unfair advantage in weakening and undermining resistance.[117]

Since this point seems abstract and theoretical, here is a concrete (yet fictitious) example of the possible implementations of these practices:

> *John does not know it yet, but there is a 67.4% chance that he would become a full-scale vegetarian in the next 18-24 months. His purchases of beef have stopped completely and his visits to steak houses are becoming scarce. He has subscribed to the "National Geographic" and additional nature magazines, and has begun to purchase Tofu products. Even his surfing habits show some indications, in visits to sites discussing animal experimentation. But things were not always like this, as in the past John was a real beefeater – eating steaks and red meat regularly. However, people tend inevitably to change over time.*
>
> *Inevitably? Not if it were up to Meat-A-Mine, a new joint venture of the Beef, Pork and Poultry industries. This organization has been tracking the behavior of vegetarians for years, breaking them down into groups and trying to establish what are the early "symptoms" and more importantly, how "crossing over" can be avoided. John's name has been flashing on their screen for a few days now, as someone that might be soon passing the point of no return, so they better act fast.*
>
> *They start out by sending coupons to his favorite steakhouse from the past. They are able to locate his "cookie numbers", and make sure that he views commercials for some fast food restaurants and others that mention the importance of eating meat and protein. Using an affiliated supermarket, they make sure that he never receives any promotions for tofu products, and just occasionally is overcharged for his vegetables. When he visits his on-line supermarket store – the shelves are stocked heavily with beef products, yet the dairy products receive smaller icons on the screen and are barely presented.*

Federal Law, FTC says, (May 28,1997), *available at* http://www.ftc.gov/opa/1997/9705/joecamel.htm.

117 Several Internet experts voiced their opinion as to the severity of online advertising based on consumer profiling, equating them to the use of "subliminal advertising." *See Profiling Said To Be Worse Than Subliminal Advertising, available at* http://www.privacyplace.com/news/99news/12_dec/news_12099/profiling10.pdf.

Meat-a-Mine knows that it's a long shot – and only 50 out of every 1,000 John-like individuals are affected by this campaign, yet due to the full automation of the process, such a low success rate is worth their while and covers their costs!

On the other hand, there is "Vegi-data", a similar joint venture formed by entities holding the opposite interests. They too have had John's name popping on their screen recently as a person that might be switching to being a vegetarian, and as they are promoting various food labels that cater to the vegetarian crowd, they are very interested that he complete the transformation. They therefore launch their specially tailored campaign (that holds a better success rate than ones they've tried previously) that includes mailings to his current address about the dangers of eating meat, and offering coupons for the purchase of organic vegetables, etc.

John, in the background (or front stage, depending on the perspective), has no idea that he has become such a celebrity. He is too busy with his daily chores to give thought to these matters of dietary preference.

In this example, John is being influenced by the use of minor and brief "interventions" carried out at the right time, which are probably unnoticeable. These interventions powerfully interfere with John's autonomy and his ability to make decisions regarding the outcome of his own life, as such outcome is to some extent dictated by an external intervention. It is true that due to low probabilities and large deviations, it would be extremely hard to predict whether John "himself" would be influenced by the interventions described above. Yet surely many "John-like" personalities would be influenced. Therefore, the argument claiming that every person reacts to these signals differently and is unpredictable does not undermine the stated problem; the objective is not influencing John specifically, but having a successful overall effect.

The "autonomy trap" therefore, could have an adverse effect on the individual by altering his preference in the consumer market. Yet this is only the beginning of the problem. The real problem is the effect on society and democracy.

2. EFFECTS ON SOCIETY

In the broader perspective, the individual's freedom as well as society's autonomy of thought could be impaired. Thoughts and beliefs would be directed by pre-sorted information chosen by others,

rather than by the full breadth of ideas and information independently gathered and sought by the individuals themselves. In the world of the "autonomy trap", the individuals' ideas and thoughts would be molded by external entities that are controlling the input of information. Using the words of Lawrence Lessig,[118] the "observer will effect the observed", and the public would be "normalized" by the effects of the vicious cycle described above. People would not be open to new ideas and concepts, but rather change their concept of the "ought" to the "is" by ceasing to seek out new concepts and ideas and settling for the content provided to them.[119]

An archetype of an Internet application that is of relevance to our discussion is "The Daily Me".[120] This application consists of a service that provides listed users a newspaper made up of content links and pages that are focused on specific pre-selected fields of interest (as indicated by the specific user). As the "Daily Me" grows in popularity, it's possible future effects have become the subject of debate.[121] Several scholars claim that such sites and applications will have an adverse effect on the democratic process and free speech, creating rapid segmentation within society and forming splintered groups whom have interest (and therefore knowledge) only of their internal issues, while losing interest and contact with society as a whole. Others disagree with such negative prophecies and do not believe that such consequences are to be anticipated.

But beyond the problems of a fractured society, the "Daily Me" might contain additional problems with regard to the "autonomy trap" dynamic described above. The use of such filtering applications will give content providers excessive power in their ability to shape information channels according to the user's profile on the one hand, and the content provider's policy, whim or commercial interest on the other- resulting in the "Daily Them"![122] In this scenario, content providers gain control over the knowledge the public obtains, and are able to promote certain ideas while ignoring others, generally focusing the public's knowledge and opinions on issues and views they decide upon. This claim adds a "fatal twist" to the "Daily Me" debate, as it

---

118 Lessig, supra note 69 at 153.

119 P.Schwartz, *supra* note 106 at 825.

120 A discussion regarding this application is to be found in the Boston Review, Summer 2001. The leading article is by Cass Sunstein "The Daily We", that is based on Cass Sunstein, Republic.com (2001). However, the discussion is mostly focused on issues of the First Amendment rather than privacy. The concept of the "Daily Me" was coined by Nicholas Negroponte in his book BEING DIGITAL (1995).

121 *Id.* (including several articles by scholars as to the effects of the "Daily Me").

122 "Them" describing the content providers that control the real, hidden content and agenda of the "Daily Me."

supplements a scent of tyranny to a general discussion on the weakening of democracy.

The DM perspective: It is apparent that the use of KDD tools will cause an escalation of the above stated problems. Using data mining applications, advertisers, vendors and content providers will have the ability to enrich their knowledge regarding their customers (both present and prospective), target them with specific content in accordance with this information, and influence them with greater success. Furthermore, these tools could be used to facilitate experimentation on behalf of vendors interested in launching campaigns to promote various products. The vendors will seek out the traits of those who reacted in the strongest fashion to their solicitation by using "clustering" methods to break the population into groups. They will thereafter amend the campaign in order to reach other groups with a similar level of effectiveness, using "prediction" to approach new 'clients' who match a certain profile.

Beyond marketing matters, data mining can be a threat to democracy as well. Content providers may use KDD tools to assess individuals according to their demographics and behavior, and will try to influence their opinions and beliefs by sending every person specific content (as with the "Daily Them" example). This content will be tailored to have the strongest impact on the recipients, according to the cluster to which they belong. These actions will enable the content providers to mold the public's opinion regarding specific issues to their satisfaction.

Many of the tasks related to the "autonomy trap" issue were possible before data mining became available. However, the KDD tools enable the management and use of vast amounts of information without the necessity of employing a large number of analysts. Data mining facilitates the automation of the processes described above, and therefore enables control over a large number of individuals with Orwellian precision, using private information and personalized access.

Comparison to Price Discrimination: In conclusion, the "autonomy trap" is a scary concept, portraying a frightening picture of a dysfunctional society. However, it is a difficult argument to present, using amorphous and vague concepts and introducing ideas that are philosophical and somewhat far-fetched. Therefore, it will probably be difficult to convince legislators to regulate and change current laws based on these concepts, prior to showing concrete damage or loss.[123]

123 In *In re Doubleclick Inc. Privacy Litigation*, 154 F.Supp. 2d 497 (S.D.N.Y. 2001), and *In re Toys-R-Us Inc. Privacy Litigation*, 2001 U.S. Dist. LEXIS

Nonetheless, the problem described above is gravely important due to its serious results and hidden consequences. It should therefore dominate any discussion regarding the effects of data mining, or possible solutions to the problems of personal information in the Internet age. Moreover, understanding this claim might provide us with the main motivation for data accumulation on behalf of large corporations, as the personal information not only provides it's holders with an option to make money, but the possibility to gain power and control – an objective stronger than any monetary desire. This desire for power is also a strong distinguishing characteristic between the autonomy trap rational, and the previous discussion of "price discrimination", which focused on financial gain.

Another distinguishing characteristic between the two important issues of autonomy and price discrimination is found in the different results they lead to. The "autonomy trap" creates a setting in which the individual is "forced" into a new market and is purchasing a product she has no initial interest in, as opposed to being overpriced for a product she initially wanted to purchase in the "price discrimination" situation. The line between these two issues is quite blurred and this distinction may not always be as clear as presented here.

Generally, the concerns of monetary gain and of public control display the essence of the entire "privacy dilemma" and highlight the importance of the 21st century's most important resource, information, and the dangers that might result from its misuse.

Even though this paper has set out to discuss the problems created by the data mining practices, it is essential at this point to mention a partial *solution*. In view of the analysis above, it is clear that to avoid entering the autonomy trap, emphasis should be put on the regulation of diversification in the media market, and the need for stronger rules against the ongoing phenomena of convergence[124] (yet clearly, such steps are the tip of the iceberg and would be elaborated elsewhere).

16947 (N.D. Cal. Oct. 9, 2001), the courts had extreme difficulty acknowledging the existence and magnitude of such damage.

124 A recent article describing the latest media merger and limiting FCC rule to be struck down discusses the adverse effects of such mergers on privacy, noting the content providers' growing ability to "really . . . know who you really are". *See* J. Schwartz, *Bigger is Always Way Better*, N.Y. TIMES, Feb. 24, 2002, at "Week in Review" 5. It is also mentioned that such control could be achieved on the Internet as well, as more people turn to high-speed service provided by large companies in the field (and thus providing them with the personal information).

D. ABUSE & MISUSE

We now return to our long list of examples: it is time to consider the matter of Mr. Black, who lost his job pursuant to misuse of information. The problem illustrated by this example could be presented as the abuse and misuse of information (with a touch of security problems). Pursuant to the above discussion of complex problems arising from the use of personal information, we must remind ourselves that a data collector could upset the information provider in very uncomplicated ways. Personal information could be (1) published counter to the will of the relevant person; (2) be used as an instrument in blackmail; or (3) be used by the holder to (3) cause embarrassment.[125] The current legal regime provides partial protection against such practices, yet does not always prove effective.[126]

The information with the potential of misuse is usually the 'raw' outcome of mundane sources such as workplace surveillance or privileged relationships. It could also be information accumulated through the use of various surveillance techniques such as cameras placed secretly on public or private property. Alternatively, the surveying tool may be placed out in the open while its possible implications are not always obvious to the public. A common example for such surveillance is the E-Z Pass payment system,[127] that when used by any vehicle reveals its (and its owners) movements to the system operator, a fact that is not always fully comprehended by the general public.[128] These forms of surveillance may provide collectors with potentially "explosive" data that can be used to the detriment of certain individuals (with no additional analysis required).

The DM perspective: When focusing on the implications of data mining, it is obvious that such applications would have a minor effect

---

125 *See* Kang, *supra* note 101 at 1212.

126 For a discussion on the limits of privacy law, see Daniel J. Solove, *Privacy and Power, Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1430 (2001). *See also* SMITH, *supra* note 8 at 224.

127 The use of E-Z Pass technology is spreading to additional areas (such as payment systems at off-highway fast food restaurants and gas stations), and the possible privacy implications spread as well. *See* Jeffrey Selingo, *It's the Cars, not the Tires, that Squeal,* N.Y. TIMES, Oct. 25, 2001. A possible use of the E-Z Pass information has been addressed in a *Law & Order* episode, where such information was used to track criminals and "destroy" alibis. *Law & Order: Dissonance* (NBC television broadcast, Nov. 1, 2000).

128 In the aftermath of the events of Sept. 11th, a new "twist" on the E-Z pass system is in planning – a card that would be used at all airports and border crossing. Obviously, the use of such a system (that would be operated by a private entity) would provide additional sensitive information as to the movements of individuals. *See* John H. Cushman, Jr. *Airlines Seek an E-Z Pass for Fast Security Checks*, N.Y. TIMES, Jan. 13, 2002, at "Travel" 3.

on this issue, as the source of the information in question is not from the discussed KDD tools. Instead, it usually exists in the relevant database without any need to conduct complicated analysis. The abuse of information rarely involves the use of elaborate tools and procedures since it is mostly predicated on "raw data" rather than patterns or clusters. Usually, it is based on data resulting from a specific query and seldom requires an automated search that was not hypothetically driven. Therefore, the issue of misuse and abuse of information would not be drastically changed or challenged due to the spreading use of data mining; Mr. Black could have suffered the same consequences in a world that did not have data mining.

It should be noted however, that possible abuses of information could stem from the use of KDD in the following setting. At the first stage, data-mining tools could search for patterns that describe certain minority groups (such as groups of a specific religion, nationality, and sexual preference). At the second stage, through the use of prediction models, it could be publicized (contrary to the will of the members of the "predictive" group) that others who answer to the same criteria are part of those groups, even though the specific member of the group chose to exercise privacy. It remains to be seen whether these practices will be utilized, and specific legislation could probably eliminate such issues.

Our last point regarding this example concerns security. In our example, Mr. Black suffers from breaches in the security of the database containing his personal information. Security is a field that is separate from privacy, though sometimes the two are confused. Security breaches involve actions carried out by external entities (though they could be enabled by internal negligence), contrary to the data mining process addressed throughout this Article that is conducted internally by the database holders. Since the focus of our discussion is data mining, the issue of security, which is one of great importance, is beyond the reach of this Article.[129]

E. SECLUSION

Moving down our long list of examples, we arrive at Example 7, describing Ms. White, a recipient of some bothersome mail. The matter at hand is the desire for seclusion[130] and comfort, which is

---

129 To learn about FTC initiatives in the area of security, see *Advisory Committee on Online Access and Security*, *at* http://www.ftc.gov/acoas/index.htm.

130 This perspective addresses the discomfort such practices cause. These claims also have a normative background of various rights to seclusion and dignity. *See, e.g.*, Reidenberg, *supra* note 98 at 1341.

confronted by the ongoing practices of telemarketing, direct marketing, and to some extent email spamming.

In this age of data accumulation, many shrewd retailers anxious to cash in on the personal information they gathered use such data to target prospective customers.[131] This results in an increasing load of junk mail arriving at our physical and virtual doorsteps, causing an ongoing nuisance and in some cases, actual loss.[132] At times, these practices force individuals to confront issues that they would rather avoid or ignore (such as a chronic illness or condition) or invoke unpleasant memories (as in our example regarding Ms. White and her late mother).[133] The retailers send out such mass mailings using the personal information they gathered themselves, or (in most cases) consumer lists now sold on the vibrant secondary information market.

The DM Perspective: Clearly, the use of KDD would have an impact on this issue, as it would surely assist the direct marketers in their practices. Data mining tools present opportunities to pinpoint the traits of every customer group and address them directly with specific offers based on their grouping or previous purchase patterns. Moreover, during the data mining process, the mining entity could discover new facts about prospective clients, and use this information to its benefit. In addition, the data mining process enables database holders to manage the vast databases in their possession by *automating*[134] the process. Data mining allows the retailer to search for groups of a certain affinity, find their common traits, and respond with a relevant marketing scheme that would be automatically evaluated

---

131 This practice has been growing due to the slowdown in the economy. The most recent market practice is to "rent" out the use of such lists. *See* Saul Hansell, *Seeking Profits, Internet Companies Alter Privacy Policy*, N.Y. TIMES, Apr. 11, 2002, at A1; *See also* SMITH, *supra* note 8 at 323 (regarding the use credit bureaus made of the personal information accumulated and the sale of such information to mass marketers).

132 In a recent *New York Times* article, it has been mentioned that the use of spam e-mail is on the rise. Besides the obvious nuisances spam mail takes time to delete and makes important messages harder to find. Spam also forces the recipients to confront and consider personal flaws when receiving mail that pertains to such shortcomings. *See You've Got Mail, Lots of It, and It's Mostly Junk*, N.Y. TIMES, Dec. 24, 2001, at A1. Spamming is gaining popularity as spammers adopt sophisticated means to locate addresses, such as random mailings based on various words from the dictionary. *Id.*

133 An additional example, taken from DAVID BRIN, THE TRANSPARENT SOCIETY, (1998), is a mother that had a painful miscarriage, yet continued to receive junkmail regarding products targeting a young mother with a newborn baby.

134 The "automization" of the process is one of KDD's main strengths, as it enables large corporations to avoid being smothered by the volume of available information.

thereafter. By using KDD, the marketers obtain an enhanced ability to discover hidden traits of their customers, and to potentially cause them additional distress.

Despite these seemingly solid claims, direct marketers have the ability to present powerful counterclaims mitigating the strength of these arguments. Direct marketers will assert that the use of data mining merely minimizes the inconvenience and intrusion caused to individuals, as the new processing tools would allow everyone to receive less advertising material compared to the quantities they received before.[135] Moreover, the advertisers would actually be delivering offers that the public wants to receive, as they would be specifically tailored for the relevant customer.[136] These claims may appeal to consumers, and more importantly, to legislators interested in satisfying this lucrative industry.

In view of these strong counterclaims, the troubles Ms. White encountered would, on their own, be insufficient to undermine the data mining technology used by the direct marketing industry that employs thousands of Americans and generates billions of dollars annually. The problems that mass mail creates should not be addressed using the claims presented in this hypothetical, but rather through those of price discrimination and the diminishing of autonomy.[137]

F. "THE TRAGEDY OF ERRORS"

Last, and perhaps least, we confront the troubles of Mr. Blue and Ms. Gray (Examples 8 and 9). Both of their concerns could be categorized as part of a *tragedy of errors*: the fear that database analyzers err in the conceptions they form about specific individuals. However, there is a fine line distinguishing both examples.

First, Mr. Blue, due to an error in the firm's database, has been charged a higher premium. Here we are facing the problem of incorrect information existing in the databases and the damage such inconsistencies cause. The dependency of many systems solely on

---

135 Clearly, spamming will continue to be a problem because the expense of sending out an e-mail message is so low that there is no reason not to send out a mass mailing. These problems, however, should be tackled by specific legislation, and are beyond the scope of this Article. For an example of the most recent legislation in this field, see Rob Gavin, *E-Commerce (A Special Report): The Rules*, WALL ST. J., Oct. 21, 2002, at R9.

136 As some direct marketers put it, "[T]here is no such thing as junk mail – only junk people!" GARFINKEL, *supra* note 10, at 155).

137 Another way to approach these problems is by arguing that such mailings violate the right to privacy of individuals. This Article does not address this approach.

computer-stored data tends to shift the burden to the individual,[138] who must try and prove that the computer and the database are wrong, and that the individual is right; contrary to the general assumption that the database is infallible. In today's reality, however, computer errors are actually widespread. Personal credit reports, for example, have been frequently reported to contain various flaws and misrepresentations. The existence of errors in such a report could have serious implications to the relevant individuals, and could result in the inability to secure a loan, open a bank account, or receive a charge card.[139] This concern is the basis for the ongoing demand that database holders provide individuals with the right of "access" to database records in order to examine, and, if necessary, to correct inaccurate information.[140]

The DM Perspective: when superimposing this argument on the data mining process, we reach a surprising conclusion: the use of data mining (and it's prerequisite, data warehousing) will not exacerbate this problem, but may help diminish it. KDD practices may facilitate the statistical verification of information through multiple sources,[141] assuring a smaller probability of mistakes. Information that does not fit into the patterns of the other data collected would stand out and be re-examined, diminishing the chance of errors. KDD applications would have little effect, if any, on the troubles of Mr. Blue and this issue besides promoting the uses of data mining.[142] Thus, claims regarding

---

138 LESSIG, *supra* note 69, at 152, speaks of shifting the burden of proving innocence to the individual.

139 GARFINKEL, *supra* note 10, at 25 discusses common errors made by the credit bureaus and the damage that they cause.

140 Access is one of the five "Fair Information Practices," which the FTC indicates as "widely accepted principles concerning fair information practices" throughout the world with regard to information privacy (the others are notice, choice, security and enforcement). *See* http://www.ftc.gov/reports/privacy3/fairinfo.htm; *see also Privacy Online: Fair Information Practices in the Electronic Marketplace: A Federal Trade Commission Report to Congress* (May, 2000) *available at* http://www.ftc.gov/os/2000/05/index.htm#22. This principle is also one of the foundations of the rights individuals have in relation to the credit bureaus. *See* SMITH, *supra* note 8, at 319-21 (giving the background of the drafting of the Fair Credit Reporting Act (1971)).

141 One KDD technique, in addition to those mentioned above, is the use of algorithms which search for deviations from a regular data pattern; if one factor within the database was incorrect, the system would track it down and re-examine why it did not fit into the larger dataset. Specific tools for this objective are offered by Axicom and Experian. These are referred to as CDI customer data integration applications. They create links between possibly redundant names by using the entire corporate database and external databases, as discussed *supra* note 52.

142 Clearly a counterargument can be made: the fact that KDD practices are available would further encourage the use and creation of databases by any entity that has access to personal information. Along with this growth in databases would come an increased number of errors and mistakes, many of which would not be picked up by the new and advanced algorithms. While such arguments may have

such errors in databases should not dominate a discussion regarding the regulation of data mining practices.

Ms. Gray's example (Example 9), however, forces us to confront a second claim of error that is of a higher level of complexity. In this example, all the information collected about her is true, yet the conclusion reached is nevertheless wrong. Such examples lead to the claim that drawing conclusions regarding specific individuals on the basis of database analysis (while using statistics and probabilities) is inappropriate. These arguments are usually based on the assumption that people are too complex to be captured by data alone[143] and therefore are portrayed in a distorted manner within the database context.[144] Some commentators point to the inherent difference between the "messy" humans and the neat sets of databases as the source of this problem. [145]

The DM perspective: Clearly the growing use of data mining will enhance the problem described in this example. In the world of KDD, not only are people addressed through their "electronic shadows",[146] they are now placed in clusters and patterns on the basis of probabilities and rules of thumb, thus creating many opportunities for the distortion of the actual reality. Corporations will assess their customers solely by viewing the results of the data mining analysis, which might be unable to present the complex reality properly. This problematic aspect of the data mining practices is especially relevant when considering the use of predictive tools that try to predict future actions of the consumer. Using these applications, it is possible that decisions of great magnitude (such as the granting of credit approval) would be based on partial personal information linked to existing patterns and clusters that were constructed from information provided by other sources. The outcome of such analysis could be that credit

---

merit, they would not compare to the benefits that can be reaped from data mining. Another possible effect of the data mining process should be mentioned. One of the methods mentioned to fight breaches of privacy is the use of bogus names: switching digits in Social Security Numbers and using different spellings of your name when prompted at various web sites. These forms of "defense strategies" would be severely impaired by the use of data mining tools that would have the ability to detect these artificially created errors and correct them (These were the facts of the Avrahami cases mentioned by GARFINKEL, *supra* note 10, at 178).

143 Some discussion of this issue is available in Solove, *supra* note 126, at 22.

144 The key difference between the two concepts is that the second does not involve errors in the facts, but instead claims that the process itself is flawed.

145 REG WHITAKER, THE END OF PRIVACY: HOW TOTAL SURVEILLANCE IS BECOMING A REALITY 137 (1999).

146 The "electronic shadow" or Doppelganger is referred to often, and is borrowed from a Germanic myth of a creature, which walks in the shadows of others. *See* GARFINKEL, *supra* note 10, at 248.

would not be granted to an individual solely because he has comparable traits to others that tend to default on their payments. Furthermore, such crucial decisions would be made on the basis of this comparison alone, without meeting with the client, or even evaluating his or her data on its own.

This claim against data mining practices is analytically weak. There is no convincing reason to suppose that decisions made by software are inferior to the ones made by humans (and as mentioned above, there are several occasions where the opposite is true). The complications experienced by Ms. Gray could have occurred just the same had her information been reviewed by a human auditor, who is prone to make mistakes and misjudgments when encountering vast amounts of information. These fears might originate in conservative points of view toward technology or are perhaps part of a neo-Luddite wave of thought.[147] Such arguments would prove unconvincing in the search for strong claims in the data mining debate.

IV. DATA MINING AND PUBLIC OPINION

A. PREFACE AND NOTE OF CAUTION

"*Fortunate is the man who is always afraid*" (Proverbs 28,14).[148]

It is time for us to meet our last friend for the day – Mr. Purple. Mr. Purple did not receive solicitations for insurance policies, nor is he interested in such. He does not shop online, and actually rarely uses his email as a means of communication. On the other hand, he is a person with many friends, including all of the characters introduced in our examples, which often confer with him regarding their troubles. He is well read and up-to-date with regard to the issues of ongoing surveillance and analysis of personal information, and has just recently learned of yet another troublesome issue: the new aviation security scheme and the immense database constructed to enable its operation.[149] This scheme uses personal information accumulated from previous transactions and other resources to create predictive models of behavior. Authorities, working closely with private entities,

147 The Neo-Luddite wave with regard to database technology is mentioned in Kirsten Wahlstrom & John F. Roddick, *On the Impact of Knowledge Discovery and Data Mining, in* 1 Conference in Research and Practice in Information Technology, (J. Weckert ed., 2000). *Available at* http://www.jrpit.flinders.edu.au/confpapers/CRPITV1Wahlstrom.pdf.

148 PROVERBS: A NEW ENGLISH TRANSLATION 176 (A.J. Rosenberg, trans., 1988).

149 Robert O'Harrow, Jr., *Intricate Screening of Fliers in Works*, WASH. POST, Feb. 1, 2002, at A1..

use the revealed patterns to decide whether or not a specific passenger poses a risk and should be held for additional questioning. In addition, Mr. Purple is an enthusiastic reader of editorials, privacy magazines, and law review articles where he learns of other surveillance practices and their repercussions. With this garnered information, Mr. Purple has reached a specific state of mind with regard to the situation of his personal information: He is intimidated and afraid.

And he is not alone. In ongoing surveys conducted by Alan Westin[150] and others,[151] it is apparent that there has been a growing public interest in personal information. According to Westin, such concerns focus on issues of intrusion (unwanted mail and telemarketing), manipulation (profiling that allows "hidden persuader" marketing), and discrimination. Consumers want companies to implement good privacy policies, and are even interested in protective legislation and government intervention regarding several matters.[152]

The results of these surveys are quoted often, and rightly so. Even though public opinion may be wrong, or lack proper theoretical backing, it is a strong force in today's democratic society. An illustration of these forces can be found in recent FTC publications, where the results of such surveys have been mentioned as the motivation for the Commission's interventions and actions in the field.[153] However, when approaching these results and public opinion as a whole, some notes of caution should be added. These issues lack objectivity; one person's may be another's convenience.[154] Moreover, the use of surveys as an indication of the public's opinion should be performed with care, as it is only a one-time snapshot, which provides

---

150 Dr. Westin is Professor Emeritus at Columbia University and currently the president of "Privacy and American Business". In his prepared witness testimony before the Committee on Energy and Commerce, Westin sets out the public opinion, as viewed by him based on surveys conducted between the years 1979-2001 (he had been the academic advisor for 45 surveys and analyzed a sum of 120 surveys). *See Opinion Surveys: What Consumers Have to Say About Information* Privacy, 107th Congress (2000)(statement of Dr. Alan Westin), *available at* http://energycommerce.house.gov/107/hearings/05082001Hearing209/Westin309.htm.

151 *See* Hatch, *supra* note 75, at 1477-81 (listing 9 additional polls).

152 The polls also indicated that fears of consumers are now focused on private entities, in addition to the ongoing fear of Big Brother

153 FTC REPORT, *supra* note 62, at 14-17 (regarding the concern of the public with "profiling" practices).

154 For example, some people might be distressed by the fact that they are targeted by ads that indicate that the promoters are aware of their current actions while others would find receiving such ads useful. Many vendors choose to ignore the public fear and believe that targeted ads are important, as they help build the clients' loyalty and create intimacy in the relationship with the customer. *See, e.g.*, DYCHÉ, *supra* note 25, at 45.

partial information. It is also important to closely examine the questions included in such surveys prior to reaching any conclusions regarding the results, as they might have been drafted in a biased manner.[155] Many of the survey results should be taken with a grain of salt. Public opinion, as determined from these studies, should not be regarded as a sufficient reason for legislation and regulation but as a signal worthy of our attention, and a consideration when searching for the sources as well as solutions for privacy problems.

The DM perspective: With regard to this issue, we must inquire as to the impact of the KDD applications on the public opinion and on Mr. Purple's peace of mind. At first glance it is reasonable to assume that the KDD tools would cause additional public fear and wariness; they present many possibilities for various entities to enhance the practice of intrusion, manipulation, and discrimination, which, as mentioned above, are the three leading reasons for today's public anxiety. Public concern could also rise in view of the public's recognition of predictive modeling; The notion of corporate entities predicting - with a high level of precision - the future actions of individuals, is indeed daunting.

On the other hand, adopting a pragmatic perspective leads to a different conclusion. Up until now, the issues that fueled public interest and concern regarding information privacy were stories of others affected by problematic practices; people that felt 'violated' when they realized that their personal information was known to their adversaries or that it was used to their disadvantage.[156] Upon reading such stories, the public grew concerned - could the same thing happen to them?[157] The anecdotes that drew public interest did not usually involve data mining practices, but simple applications, such as using information extracted from a relevant database,[158] workplace

---

155 *See, e.g.*, FTC REPORT, *supra* note 62 (Commissioner Swindle's dissent).

156 One common resource for such stories is the Privacy Foundation website, published by Richard E. Smith, *at* http://www.privacyfoundation.org.

157 The sources of public fear might be deduced from the following brilliant TV commercial (and probably one of the first "privacy based campaigns") provided by "Earthlink", an Internet provider: In this commercial, a well-dressed man and two sleazy looking guys are sitting at the bar – with a girl. The girl gives her phone number to the well dressed man – who offers it to the other sleazy men – in her presence, for $10 a person. In this commercial, Earthlink attempts to encourage the public to consider the implications of Internet privacy through an example of a simple breach of trust. Obviously, a commercial discussing the problems of the autonomy trap and price discrimination would not have been as convincing.

158 A good example to the above would be the following:

"…a Los Angeles Man, Robert Rivera, says that after he sued Vons markets when he fell in the store and injured his leg the store looked up his record, discovered that he likes to buy a lot of liquor, and said it would use the information to defend itself in the lawsuit. The

surveillance by a preying employer or co-worker, abuse of information received as part of a privileged relationship, or even the misfortunate consequences of errors in a database. The availability of complex data mining tools, whose uses are not easily comprehended, will probably not have a strong impact on public opinion to the same extent the simple and mundane privacy issues have had. Therefore, in terms of the examples introduced above, Mr. Purple would probably be preoccupied with the misfortunate Mr. Black, who had his personal information sold to his employer, and with Mr. Blue, who was confused with his depressed neighbor. He would probably be enraged by the discrimination towards Ms. Red, and saddened for the conduct towards Mr. Green, the minimum wage employee. But would he be moved, or even acknowledge the problems of Mr. Yellow, the overpaying philosophy student? Or of Mr. Orange, who was unsuccessful in his attempt to quit smoking? I suspect not. These matters are too complex to have a strong public impact.

Moreover, I fear that the "autonomy trap" would move to trap the public in a state of ignorance, steering it away from important issues toward decoy problems, using the methods discussed.[159] Therefore, public opinion would not prove a dominant force regarding the use of data mining, as it would not focus on the strongest issues, due to their minimal public appeal. A strong campaign should be launched to educate the public otherwise, and guide it in the right direction.

B. THE DATA MINING CAMPAIGN

Up until now we have focused on the problems, yet the public opinion claim also allows us to address solutions to some extent. Public opinion might assist us in mitigating the major problems caused by data mining in several ways.

First, an effective change in the public opinion could result in public awareness toward the central issues stated above. Such

---

implication was that Rivera may have been impaired when he fell." PRIVACY J., Mar. 1999, at 5. Another good example for a "horror story" that grasps the public attention and focuses the public opinion, is the MetroMail case, in which a woman received a harassing letter with intimate details regarding her life from an inmate she did not know. These details were available inside the prison, as the inmates were employed in transferring personal information from various sources to electronic databanks. *See* Nina Bernstein, *The Erosion of Privacy*, N.Y. TIMES, June 12, 1997, at A1.

159 For example, the "Daily Me" editions would stop presenting discussions on privacy matters, focusing on security issues instead. Prof. Schwartz mentions that the autonomy trap might lock us into a lower level of privacy. See *Privacy and Democracy*, *supra* note 102, at 1660-65.

awareness could do wonders for solving the problems mentioned. The public, when aware of the privacy concerns, could reduce the amount of personal data it provides collectors with, and insist on proper compensation when they choose to submit such information. In addition, people might apply general caution towards any feedback they receive from various content providers and advertisers, knowing that it might have been tailored especially for them.[160]

Secondly, public awareness created by an effective public campaign could be constructive as a force in leading government and industry to change the common problematic practices. In the past, public opinion and pressure proved to be a dominant force on several occasions in the Internet setting, causing corporations to amend plans and terminate programs that had adverse effects on information privacy.[161] The Internet's traits of connectability and the ease of transferring information, which are part of the reasons for the severity of privacy issues in this environment, contribute to the effectiveness of public opinion.[162]

Such a campaign must focus on the crucial issues mentioned above: price discrimination and the autonomy trap. Failure to do so – i.e. broadening the arguments to include additional claims - would cause the downfall of the entire campaign. An unfocused campaign would lead to strong counter claims and mitigating arguments from the adversaries in this debate (the various data-mining entities), resulting in the loss of momentum, and an eventual halt of this initiative. This may cause public opinion to wander elsewhere and allow the data miners to maintain the status quo.

---

160 Note that regarding the problem described in the Ayres' articles, *supra* notes 73 & 74, it has been claimed that the situation has improved after the publishing of the articles, as the public has learned of the conduct of the car salesmen, and acted accordingly. *See Michigan Article*, *supra* note 73, at 143.

161 For more on this issue, see LAURA J. GURAK, PERSUASION AND PRIVACY IN CYBERSPACE (1997). This book mentions and analyzes several cases in which public opinion helped trump several corporate schemes. For example, the Lotus initiative to sell products that include personal information to retailers, the Lexis-Nexis's attempt to sell Social Security numbers, American Express's attempt to sell merchants transactional information, and the Intel Processor Serial number issue.

162 *See, e.g.*, Oscar Gandy, *Exploring Identity and Identification in Cyberspace,* 14 NOTRE DAME J. L. ETHICS & PUB. POL'Y 1104 (suggesting that through the use of the Internet, aggrieved individuals may find others affected in a similar way).

Therefore, a focused public campaign should be launched.[163] It should be dominated not by tales of collection and surveillance, but by the inherent unfairness of certain price discrimination practices.[164] It should also stress the damage to autonomic thought that might follow the revealing of personal information, and the ways in which such problems could be avoided. Price discrimination and the autonomy trap, the "failures" forming in the market of goods and ideas, should be the winning slogans in the personal information campaign.

V. CONCLUSION

Data mining, a powerful tool capable of impressive descriptive and predicative tasks, will no doubt have a great impact in various areas, above all in the field of personal information. After examining the various schools of thought regarding the ongoing privacy debate and its interaction with the practices and uses of data mining tools, we concluded that some of the mentioned perspectives and examples are of greater relevance in this debate, while others appear irrelevant or neutral. We also found that the claims central to this discussion are *price discrimination* and the *autonomy trap*.

When moving towards the establishment of solutions for the described problems, it is important to focus on the issues central to the discussion, rather than straying to solutions that may seem simple to implement, yet will solve little. Whatever solutions are implemented, they will have to be backed by a strong public opinion that would force legislators, regulators and courts to acknowledge the need for such actions.

To a certain degree, public opinion on it's own could serve as a partial and intermediate solution, since recognition and awareness could go a long way in solving the problems at hand. Moreover, as many of the problems addressed above occur in cyberspace, it is important to remember that the virtual public arena is a battleground of special rules of conduct and engagement, where strong claims spread at lightning speed through the use of Internet communities and mass emailing. When carried out effectively, the results are striking, as

---

163 SMITH, *supra* note 8, at 146, mentions that Brandeis, when contemplating a right of privacy and working on his famous law article, wrote to Samuel Warren about campaigning for a right of privacy, stating that: "All law is dead letter without public opinion behind it".

164 In some cases, price discrimination through the use of "club cards" at supermarkets has ceased after customers began to boycott the stores practicing said discrimination – thus demonstrating the power of public opinion. *See, e.g., Weblining*, *supra* note 71.

such online campaigns have killed or stalled several business initiatives in a very short time. We must construct such campaigns with care, and make strong and relevant claims. The data mining issue is an important one – it should not be exhausted on trivial matters, or on questions that are severe regardless of this phenomena. We must save the data mining card to address the matters that it affects directly.

It might be up to us – the legal readers, students, and scholars – to assist the public in focusing on the strongest arguments. We should focus on the arguments that cannot be rebutted easily, or solved in a pinpoint manner. The arguments we choose should be those that present the most severe problems and eventually lead to suitable solution.