PENN STATE LAW REVIEW

Articles

Governmental Data Mining and its Alternatives

Tal Z. Zarsky*

Abstract

Governments face new and serious risks when striving to protect their citizens. Of the various information technology tools discussed in the political and legal sphere, data mining applications for the analysis of personal information have probably generated the greatest interest. Data mining has captured the imagination as a tool which can potentially close the intelligence gap constantly deepening between governments and their targets. Data mining initiatives are popping up everywhere. The reaction to the data mining of personal information by governmental entities came to life in a flurry of reports, discussions, and academic papers. The general notion in these sources is that of fear and even awe. As this discourse unfolds, something is still missing. An important methodological step must be part of every one of these

* Hauser Research Fellow, New York University Law School, 2010-2011. Senior Lecturer, University of Haifa, Faculty of Law. Research for this paper was partially funded by an NWO (the Dutch Research Foundation) funded project "Data Mining without Discrimination" and I thank my co-researchers Bart Custers, Bart Schermer and Toon Calders for their insights. I also thank Chris Slobogin, Kathy Strandburg, Helen Nissenbaum, Ira Rubinstein, Richard Stewart, the participants of the Hauser Research Forum, the DePaul Law School CIPLIT presentation for their comments and Talya Ponchek for her assistance in research.

inquires mentioned above—the adequate consideration of alternatives. This article is devoted to bringing this step to the attention of academics and policymakers.

The article begins by explaining the term "data mining," its unique traits, and the roles of humans and machines. It then maps out, with a very broad brush, the various concerns raised by these practices. Thereafter, it introduces four central alternative strategies to achieve the governmental objectives of security and law enforcement without engaging in extensive data mining and an additional strategy which applies some data mining while striving to minimize several concerns. The article sharpens the distinctions between the central alternatives to promote a full understanding of their advantages and shortcomings. Finally, the article briefly demonstrates how an analysis that takes alternative measures into account can be carried out in two contexts. First, it addresses a legal perspective, while considering the detriments of data mining and other alternatives as overreaching "searches." Second, it tests the political process set in motion when contemplating these measures. This final analysis leads to an interesting conclusion—data mining (as opposed to other options) might indeed be disfavored by the public, but mandates the least scrutiny by courts. In addition, the majority's aversion from the use of data mining might result from the fact that data mining refrains from shifting risk and costs to weaker groups. This is yet one of the ways the methodology of examining alternatives can illuminate our understanding of data mining and its effects.

INTRODUCTION: THE LURE AND CONFUSION OF GOVERNMENTAL DATA MINING

Governments around the world are facing new and serious risks when striving to assure the security and safety of their citizens. Perhaps the greatest concern is the fear of terrorist attacks. Various technological tools are being used or considered as means to meet such challenges and curb these risks. Of the tools discussed in the political and legal sphere, data mining applications for the analysis of personal information have probably generated the greatest interest. The discovery of distinct behavior patterns linking several of the 9/11 terrorists to each other and known operatives[1] has led many to ask: What if data mining had been

1. *See* Kim Taipale, *Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy, and the Lessons of King Ludd*, 7 YALE J.L. & TECH. 123, 134 (2004).

applied in advance? Could the attacks and their devastating outcomes have been avoided?

Data mining has captured the imagination as a tool that can potentially close the intelligence gap constantly deepening between governments and their new targets—individuals posing a risk to security and the public's wellbeing.[2] Data mining is also generating interest in other governmental contexts, such as law enforcement and policing. In recent years, law enforcement has shifted to "Intelligence Led Policing"(ILP).[3] Rather than merely reacting to events and investigating them, law enforcement is trying to preempt crime. It does so by gathering intelligence, which includes personal information, closely analyzing it, and allocating police resources accordingly—all tasks which could be enhanced by data mining technology.[4] The growing appeal of data mining in all these contexts results from similar reasons and sources—the development of cutting edge technologies; advances in mathematics, statistics, and computer science; and the sinking costs of the hardware, software and manpower needed for their implementation.[5] Reports on the success of prediction through the use of data mining[6] in the commercial realm have strengthened the appeal of these models for governmental actions as well.[7]

It should therefore come as no surprise that in the United States, data mining initiatives are popping up everywhere. A recent U.S. General Accounting Office report indicates current data mining

2. For a countering view, see Jeff Jonas & Jim Harper, *Effective Counterterrorism and the Limited Role of Predictive Data Mining*, CATO INST. POL'Y ANALYSIS, Dec. 11, 2006; *see also* Bruce Schneier, *Why Data Mining Won't Stop Terror*, WIRED, Mar. 9, 2006, http://www.wired.com/politics/security/commentary/securitymatters/2006/03/70357.

3. Fred H. Cate, *Data Mining: The Need for a Legal Framework*, 43 HARV. C.R.-C.L. L. REV. 435 (2008).

4. *See, e.g.*, Press Release, IBM, Memphis Police Department Reduces Crime Rates with IBM Predictive Analytics Software (Jul. 21, 2010), *available at*http://www-03.ibm.com/press/us/en/pressrelease/32169.wss. For a paper discussing these initiatives in the Netherlands, see RCP van der Veer et al., *Data Mining for Intelligence Led Policing*, 15 ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING (2009), http://www.sentient.nl/docs/data_mining_for_intelligence_led_policing.pdf.

5. For a discussion of the building blocks of data mining, see Tal Z. Zarsky, *"Mine Your Own Business!": Making the Case for the Implications of the Data Mining of Personal Information in the Forum of Public Opinion*, 5 YALE J.L. & TECH. 4 (2002-2003) [hereinafter Zarsky, *MYOB*].

6. Such success has been recently detailed in several popular books. *See, e.g.*, STEPHAN BAKER, THE NUMERATI (2008); IAN AYRES, SUPER CRUNCHERS (2007).

7. DANIEL J. SOLOVE, NOTHING TO HIDE—THE FALSE TRADEOFF BETWEEN PRIVACY AND SECURITY 182 (Yale Univ. Press 2011) [hereinafter SOLOVE, NOTHING TO HIDE].

initiatives in a broad array of contexts.[8] The Defense Against Research Projects Agency (DARPA) has famously promoted the Total (later changed to "Terrorist") Information Awareness (TIA) program—an ambitious project which planned to analyze vast amounts of personal information from governmental and commercial sources. This project was catastrophically handled in terms of public relations. Public concern and outrage led to Congressional intervention and the project's quick demise.[9] It is broadly understood that similar projects are living on under different names and acronyms, however.

The reaction to the data mining of personal information by governmental entities came to life in a flurry of reports, discussions, and academic papers. The general notion in these sources, as well as the one in the public sphere,[10] is that of fear and even awe. Information privacy, which many sense is under constant attack in both the commercial and governmental realm, seems to be utterly compromised. The visceral feeling of many is that the outcome of data mining analyses, which allow the government to differentiate among individuals and groups in novel ways, is extremely problematic.

Understanding what stands behind this strong visceral response is a difficult task. Even though governmental data mining is extensively discussed in recent literature,[11] an overall sense of confusion is ever present. Given the fact that data mining will probably prove necessary (or a "necessary evil" for some), scholars have moved to examine how the problems it generates could be mitigated and how its risks and benefits should be balanced. While mapping out these matters, scholars, as well as policymakers, must further establish which paradigms of legal thought are most fitting to address these matters. Central paradigms are constitutional law, privacy law, and anti-discrimination—yet other fields

8. U.S. GEN. ACCOUNTING OFFICE,GAO-04-548, DATA MINING: FEDERAL EFFORTS OVER A WIDE RANGE OF USES 9-54 (2004). The General Accounting Office has since been renamed as the Government Accountability Office.

9. *See* Cate, *supra* note 3, at 441; *see also* SOLOVE, NOTHING TO HIDE, *supra* note 7, at 184-85.

10. This outcome is interesting, as stories related to privacy in general have generated limited interest, lest they involve an actual catastrophe—personal data about a judge blocks his nomination, information regarding the address of an actress leads to her murder, and many other examples. The data mining stories here addressed focus on potential harms, which have yet to materialize. This outcome tells an interesting story about the data mining risks.

11. *See, e.g.*, Cate, *supra* note 3; Christopher Slobogin, *Government Data Mining and the Fourth Amendment*, 75 U. OF CHIC. L. REV. 317 (2008); *see also* Anita Ramasastry, *Lost in Translation? Data Mining, National Security and the Adverse Inference Problem*, 22 SANTA CLARA COMPUTER & HIGH TECH. L.J. 757, 760 (2006), *and* Daniel J. Solove, *Data Mining and the Security-Liberty Debate*, 75 U. CHI. L. REV. 343 (2008).

will surely prove relevant. As this discourse unfolds, something is still missing. An important yet often overlooked methodological step must be part of every one of the inquires mentioned above—the adequate consideration of *alternatives*. Scholars and policymakers swiftly point out the troubles of data mining, as well as the dangers of ignoring it. Yet they are not equally quick to consider the alternatives which will surely be applied by governments setting data mining aside, with their many detriments and shortcomings. Understanding the importance of this analytical step follows from acknowledging that the challenges bringing data mining to the forefront of our discussion are not going away. Governments must confront new security and law enforcement challenges and pressure to take action. They must also address the challenges of optimally utilizing the vast volumes of personal information at their disposal. Moreover, considering alternatives is also helpful in sharpening our understanding of the benefits, detriments, traits and qualities of data mining itself.

This article strives to develop a methodology for examining alternatives to data mining and to bring it to the attention of academics and policymakers. It provides basic tools for engaging in this important analytic exercise and a brief demonstration as to how it could be carried out. To achieve its objective, this article proceeds as follows: in Part 1, it briefly demonstrates and explains the government's data mining initiatives. This is a crucial step, as the term "data mining" has almost taken on a life of its own, and is applied in several—at times contradictory—ways. The article also notes specific unique traits of these practices while focusing on the distinct roles of humans and machines. Part II maps the various concerns data mining generates while drawing from ongoing literature in legal journals and policy papers. Part III presents the center of the thesis and introduces four alternative strategies of data usage and management for achieving the governmental objectives of security and law enforcement. It also addresses an additional strategy (contemplated by policymakers and think tanks) for using a specific form of data mining while anonymizing the data and thus minimizing some of the mentioned concerns. In the second segment of this part, I sharpen the distinctions between the central alternatives to promote a full understanding of their advantages and shortcomings

In Part IV, I demonstrate how an analysis that takes alternative measures into account can be carried out in two contexts. First, from a legal perspective, while considering the detriments of data mining analysis as a "search" of personal information pertaining to specific individuals without their specific and informed consent. For that, the article briefly maps out three theories for understanding "searches" in these contexts and tests them for every alternative. I conclude that the

results are mixed; while generally data mining proves to be the most problematic option, the outcomes vary among the theories. Therefore, as results are complex and unpredictable, a full comparative analysis will be required at every juncture prior to setting relevant policy. Second, I briefly demonstrate how this methodology could be applied to studying the political process set in motion by adopting measures for selective law enforcement. Here I address the different social and political dynamics which will transpire under every alternative regime. This analysis leads to two interesting preliminary results. First, that data mining might indeed be disfavored by the public, but mandates the least scrutiny by courts. Second, that the majority's general discontent with data mining might result from the fact that data mining refrains from shifting risks and costs to weaker groups. Thus, the political process might not be leading to the selection of the most fair and efficient option. This comparative analysis provides an important insight that would enrich future discussions and court decisions.

The discussion of data mining and its alternatives goes beyond the actions of government. Private entities are applying similar techniques to distinguish among their actual or prospective clients/customers, while analyzing personal behavior. Advertisers, marketers, banks, credit card issuers, and insurance companies all engage in the data mining of personal information.[12] While the commercial context is of great importance, it is beyond our current scope. It is important to note, however, that the rationales and internal balances discussed in the governmental context cannot be applied directly to the private sector. With private firms, competitive forces (when these indeed exist) might play an important role in achieving some of the needed objectives.[13] These differences and their implications must be explored elsewhere.

Finally, although this article claims to merely make a methodological contribution, I confess to arguing a normative point between the lines. While I do not carry through a full analysis of the pros and cons of the data mining strategies, my sense is that when taking the full scope of alternatives into account, data mining is far less problematic than when it is considered at first blush. The problems data mining brings to mind persist, and with greater force, when applying

12. For a recent example, see Leslie Scism & Mark Maremont, *Insurers Test Data Profiles to Identify Risky Clients*, WALL ST. J., Nov. 19, 2010, http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html?mod=WSJ_hp_LEADNewsCollection.

13. In some instances, the services rendered are not essential, thus allowing for consumer choice—an option which requires rethinking many of the elements to be addressed below. Finally, the obligations and motivations of governmental entities are different than their commercial counterparts, thus altering the internal calculus leading to the final recommendations.

other options. Understanding this point might lead policymakers to reconsider the overall negative treatment data mining options receive.

PART I: DATA MINING: IN THEORY AND IN PRACTICE

I.1. Data Mining: Definitions, Processes and General Terms[14]

The term "data mining" has recently been used in several contexts by policymakers and legal scholars. For the discussion here, I revert to a somewhat technical definition of this term of art. Here, data mining is defined as the "nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data."[15] Even within this definition, there are several intricacies. The term "data mining" refers to both "subject based" and "pattern based" searches.[16] The former refers to database searches of and for specific individuals, events, and predetermined patterns. However, the core of this article focuses on the latter forms of analysis (also referred to as "event-based" data mining). These methods provide for a greater level of automation and the discovery of unintended and previously unknown information. Such methods can potentially generate great utility in the novel scenarios law enforcement and intelligence now face—where a vast amount of data is available, yet there is limited knowledge as to how it can be used and what insights it can provide.

With "pattern based analyses," the analysts engaging in data mining do not predetermine the specific factors the analytical process will use at the end of the day. They do, however, define the broader datasets which will be part of the analysis. Analysts also define general parameters for the patterns and results they are seeking and that thus could be accepted—such as their acceptable level of error. Thereafter, the analysts let the software sift through the data and point out trends within

14. Since the matters here addressed were drawn out elsewhere, the analysis is brief. For a more in-depth discussion, see Zarsky, *MYOB*, *supra* note 5. *See also* Kim A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 COLUM. SCI. & TECH. L. REV. 2 (2003), *available at*http://www.stlr.org/html/volume5/taipaleintro.php; MARY DEROSA, CTR. FOR STRATEGIC AND INT'L STUDIES, REPORT: DATA MINING AND DATA ANALYSIS FOR COUNTERTERRORISM 14, (2004), *available at* http://csis.org/files/media/csis/pubs/040301_data_mining_report.pdf.

15. This is the most common definition of data mining. For example, see U.M. Fayyad et al., *From Data Mining To Knowledge Discovery: An Overview*, *in* ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING 6 (1996).

16. For a discussion regarding the distinction among the two, see Cate, *supra* note 3, at 438, and Slobogin, *supra* note 11, at 323.

the relevant datasets, or ways in which the data could be effectively sorted.[17]

The data mining process could achieve both descriptive and predictive tasks. Descriptive data mining provides analysts with a better understanding of the information at their disposal, while uncovering hidden traits and trends within the dataset. When applied by law enforcement to vast databases of personal information, such analyses can uncover disturbing behavior patterns and assist in ongoing investigation to find criminals and terrorists already being sought. While this ability generates legal concerns, this paper focuses on the use of the data mining of personal information for predictive modeling and analysis—an issue that generates far more interest (and subsequent fear).[18]

In a predictive process, the analysts use data mining applications to generate rules based on preexisting data. Thereafter, these rules are applied to newer (while partial) data, which is constantly gathered and examined as the software constantly searches for previously encountered patterns and rules. Based on new information and previously established patterns, the analysts strive to predict outcomes prior to their occurrence (while assuming that the patterns revealed in the past pertain to the current data as well). In the law enforcement and national security context, such insights can prove quite helpful—at times allowing for sufficient reaction before it is too late.[19]

I.2. Data Mining, Automation, and the Human Touch

As mentioned above, one of data mining's unique traits is the high level of automation it provides. The scope of automation this process entails might be easily overestimated. Counter to what one might initially believe, even with predictive data mining, the role of the human analyst and his or her discretion is quite extensive. For example, the dataset must be actively constructed, at times by bringing together data from various sources. The analyst must also predefine the parameters of the search.[20] These actions directly affect the outcome of the process, and thus impact policy.

17. For a discussion as to how these data mining techniques are carried out, *see generally* Zarsky, *MYOB*, *supra* note 5.

18. For similar reflections on this dichotomy and its normative implications, see SOLOVE, NOTHING TO HIDE, *supra* note 7, at 195.

19. The data mining process includes other stages as well, such the preparation of the data, data warehousing, cleansing and sorting. For more on these stages, *see generally* Zarsky, *MYOB*, *supra* note 5.

20. This is done both in advance and after the fact by "weeding out" results she might consider as random, wrong, or insignificant.

The extent of human discretion involved in this process is not a factor set in stone. Rather, it is a result of various policy decisions. For instance, it is impacted by whether the process is *interpretable* or *non-interpretable*. With a non-interpretable process, the rationales for actions premised upon the predictions the data mining process provides are not necessarily explainable to humans; the software makes its decisions based upon multiple variables (even thousands!) that were learned throughout the data analysis.[21] This process is not easily reduced to words. Therefore, applying non-interpretable schemes affects the role and discretion of the analysts. With such processes in place, human discretion is minimized to setting the parameters for generating predictive algorithms ex ante. The subsequent process of sorting objects, events, or people is carried out automatically, with minimal human oversight. In addition, when a process is non-interpretable, it is very difficult to provide an answer as to *why* a specific result was reached beyond the fact that this is what the algorithm identified based on similar cases in the past.

The flip side of these processes would be a fully interpretable analysis: one which uses a limited number of factors which in turn could be reduced to a human-language explanation. With interpretable results, an additional stage could be added to the process in which the analyst works through the patterns and criteria set forth by the computer algorithms. These could be indications of higher risk associated with individuals of a certain height, age, specific credit or purchasing history—and the interaction of these factors. With an interpretation in hand, the analysts can track and set aside factors which they find offensive, ridiculous, or problematic. In addition, the analyst could provide a response after the fact to questions as to what initiated the special treatment of an event or individual. An interpretable process would be costly, both in terms of additional expenses for analysts and the efficiency and effectiveness lost in the process. Yet these costs are balanced by gains in accountability and transparency.

Providing for an interpretable process also enables an additional level of human scrutiny in the predictive data mining dynamic. If analysts have a good grasp of the elements used, they can further seek out a theory of causation. Such a theory would go beyond the mere correlation that data mining reveals and seek out explanations as to why

21. David Martens & Foster Provost, *Explaining Documents' Classification* 6 (N.Y.U. Stern School of Business, Working Paper No. CeDER-11-01, 2011), *available at* http://pages.stern.nyu.edu/~fprovost/Papers/martens-CeDER-11-01.pdf.

these are proper indicators[22] (as opposed to merely acknowledging that they "work"). This step can prove helpful in weeding out ridiculous and random findings as well as those which resemble problematic (or even illegal) discriminatory practices. The notion of "interpretability" and the causation/correlation distinction will be addressed throughout this article, as it analyzes the various alternatives to data mining.

To summarize, this segment provided an overview of the meaning and use of data mining when applied to the analysis of personal information by governments. It also briefly clarified the extent of human discretion and computer automation. The entire discussion is premised on an underlying assumption that the tools here discussed are effective in achieving their analytical objectives while maintaining an acceptably low level of false positives and negatives. Whether this is indeed true is currently hotly debated[23] and notoriously difficult to measure. The answer to these questions will depend on context as well as the costs, consequences, and chances of false positives and false negatives. Therefore, prior to engaging in data mining, a relevant authority must conduct an assessment of the effectiveness of the data mining process.[24] If such an analysis indicates that data mining schemes are doomed to technical and operational failure, data mining must be abandoned. The critiques presented below, however, will be premised upon the contrary assumption that data mining is effective and operational.

PART II:THE FEARS AND CHALLENGES OF GOVERNMENTAL DATA MINING

Data mining presents promising opportunities for bridging the gap between the government's informational needs and the vast datasets of information at its disposal. With data mining, such data could be transformed into knowledge. However, these practices generate a variety of concerns. These concerns, in turn, are now requiring policymakers and courts to engage in an extensive discussion and analysis. A discussion of these matters splinters quickly into a multitude of claims and counterclaims. Fully addressing all these issues is beyond the confines of this (or any) article. For that reason, this article focuses on a specific methodological point which must be applied in every one of the

22. oweverH, constructing a theoretical justification to a statistical correlation is usually easy and merely requires some imagination. Thus, one can easily question the extent of protection from arbitrary results this requirement will provide.

23. *See* sources cited *supra* note 2.

24. This is a well-accepted notion. *See* TECHNOLOGY AND PRIVACY ADVISORY COMMITTEE, SAFEGUARDING PRIVACY IN THE FIGHT AGAINST TERRORISM (2004) [hereinafter TAPAC REPORT]. For more on this point, see Cate, *supra* note 3, at 476.

data mining contexts: addressing alternatives.[25] To briefly demonstrate how that should be done, I focus below[26] on merely examining alternatives for the first segment of such an analysis.

In the interest of giving context to the critique of data mining and its alternatives, this segment maps out the specific analytical junctures where data mining is challenged. It is at these points where addressing alternatives is crucial. This analytic mapping relies upon scholarship and policy reports addressing this matter in the last few years. For the sake of clarity, I distinguish among the different steps of personal information flow: the collection and analysis stage and the usage of personal data.

The following description is mostly theoretical and normative, with only limited attention provided to positive law. The article takes this approach for several reasons. First, setting aside the positive analysis for now allows for quickly working through the relevant issues and leaving room for an in-depth discussion of the alternatives. Second, to a great extent, the legal and policy standing on these issues is still up for grabs. In the United States, most of these issues have not been decided upon in the courts, which are probably awaiting regulation or legislation. The governmental data mining initiatives usually do not amount to breaches of constitutional rights; as Daniel Solove succinctly summarized, "Data mining often falls between the crevices of constitutional doctrine."[27] These initiatives are also probably permitted according to current privacy laws in view of various exceptions and loopholes.[28] Yet public opinion and various policy groups do not approve of these practices.[29] Thus, change is inevitable.

II.1. Collection and Analysis

A data mining process inherently calls for automatically reviewing and analyzing profiles filled with personal information regarding many individuals. Such data was previously collected by either government or commercial entities. It is hard to imagine that individuals conceded to the data mining process here described at the time of collection or at a later stage. If the information was collected by the government, citizens

25. Transparency is an additional category that requires scrutiny and discussion, yet it calls for a very different form of analysis. For more on this issue, see Tal Zarsky, *Transparency in Prediction* in *Data Mining without Discrimination* (forthcoming2012). *See also* SOLOVE, NOTHING TO HIDE, *supra* note 7, at 193.

26. *See infra* Part IV.1.

27. Solove, *supra* note 11, at 355.

28. *See generally* Cate, *supra* note 3.

29. For an empirical study pointing in this direction, see CHRISTOPHER SLOBOGIN, PRIVACY AT RISK: THE NEW GOVERNMENT SURVEILLANCE AND THE FOURTH AMENDMENT 194 (2007).

might not have provided consent at the point of collection. Rather, they merely received a basic and vague notice of the collection and future uses.[30]

Engaging in personal data analysis without the direct consent of relevant data subjects run counter to several "privacy"-related legal concepts. First, such actions might constitute unreasonable searches.[31] If so, data mining will be considered an illegal search when carried out without sufficient judicial approval—approval which is not currently sought. According to other privacy theories, which are more central in European thought, data mining without prior consent constitutes a violation of the realm of control individuals have over their personal information.[32] The information is also analyzed and used outside the original context in which it was collected, thus violating the principles of "contextual integrity" recently set forth by Helen Nissenbaum to conceptualize proper information uses.[33] Currently, however, under American law at least, such practices are permitted if the data is collected legally and a very general and vague notice is provided.[34]

On a more pragmatic level, these vast analyses might cause a "chilling effect" with regard to many important activities and behaviors;

30. In the United States, such rights are governed by the Privacy Act of 1974, 5 U.S.C. § 552a (2010), which calls for the publication of System of Records Notices (SORNs) to notify the public of such uses. For more on this, see U.S. DEP'T OF JUSTICE, OVERVIEW OF THE PRIVACY ACT OF 1974 (2010), *available at*http://www.justice.gov/opcl/1974indrigacc.htm. For more on the role of the Privacy Act in this context, see Cate, *supra* note 3, at 464-65.

31. This is not the classic understanding of a "search," which does not pertain to searches of data already collected. However, newer theories reexamining the "search" terminology question such wisdom. Slobogin, for instance, believes the term should be used in the same way the public understands it. According to his empirical studies, that includes data mining. *See* Christopher Slogobin, *Is the Fourth Amendment Relevant in a Technological Age?* 13-14 (Vanderbilt Univ. Law Sch. Pub. Law & Legal Theory, Working Paper No. 10-56, 2010), *available at* http://www.brookings.edu/~/media/Files/rc/papers/2010/1208_4th_amendment_slobogin/1208_4th_amendment_slobogin.pdf. Mark Blitz is also examining whether searches within data or other sources the government obtained lawfully could be considered a "search" nonetheless, while focusing on DNA samples. Mark Blitz, *Warranting a Closer Look When Should the Government Need Probable Cause to Analyze Information It Has Already Acquired?*, PLSC 2011 Workshop (unpublished draft) (on file with author).For an in-depth normative discussion of data mining as "searches," *see infra* Part IV.1.

32. The notion of "privacy as control" was set forth by Alan Westin and implemented in various aspects of both the Organisation for Economic Co-Operation and Development (OECD) Principles and the EU Data Protection Directives. *See generally* ALAN WESTIN, PRIVACY AND FREEDOM (1967). For information on the EU Data Protection Directives, see DANIEL J. SOLOVE, PAUL M. SCHWARTZ, & MARC ROTENBURG, INFORMATION PRIVACY LAW (2006).

33. HELEN F. NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010).

34. *See* TAPAC REPORT, *supra* note 24, at viii-x.

if citizens fear that specific actions will generate additional governmental scrutiny, they will refrain from these actions—such as travel, communications, or consumption—even when they are legal and at times socially beneficial.[35] From a somewhat different perspective, knowledge of such actions impedes upon the citizen's autonomy, because it does not allow the citizen to develop his or her "self" to the greatest extent possible.[36]

Finally, even if these practices are justifiable in one context, such as that of homeland security, there is a fear that government and its agents will go beyond that singular context. For example, while data mining could be justified to protect citizens from risks that can lead to devastating outcomes, it most likely cannot be justified as a tool for locating deadbeat dads. This is the "project/function creep" concern which has many commentators and policymakers worrying.[37] This concern might even lead to recommendations that these projects should not be initiated in the first place, even in limited contexts.[38]

II.2. Usage

Using the knowledge derived from the data mining process for various governmental objectives generates an additional set of concerns. One such concern is that the outcomes will be used to unfairly discriminate among citizens.[39] Discrimination could prove problematic for a variety of reasons: it could be based (at times, tacitly) on unacceptable social factors (such as race and nationality). It could also be premised upon partial information or immutable factors individuals have no control over. In addition, some might object to distinguishing

35. For a discussion of this argument in the data mining context, *see generally* Cate, *supra* note 3 (noting it as perhaps the most powerful one in this context). Strandburg makes a similar argument, while pointing out that in some contexts data mining might impede on First Amendment rights, such as freedom of speech and association. Katherine J. Strandburg, *Freedom of Association in a Networked World: First Amendment Regulation of Relational Surveillance*, 49 B.C. L. REV. 741 (2008). This point is also reflected in Solove, *supra* note 11, at 358. For a general discussion of privacy and autonomy, see Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393 (2001). For a more cautious perspective of this concern, see Taipale, *supra* note 1, at 146.

36. For a discussion of the link between information privacy and autonomy, see Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 STAN. L. REV. 1373, 1424-28 (2000); *see also* Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1651-52 (1999).

37. *See generally* Cate, *supra* note 3. For an opinion that some of these concerns could be curbed through technology, see Taipale, *supra* note 1, at 149.

38. *See* Slobogin, *supra* note 11, at 326.

39. This process was perhaps first touched upon in the commercial context by Gandy. *See* OSCAR GANDY, THE PANOPTIC SORT: A POLITICAL ECONOMY OF PERSONAL INFORMATION (1993).

among individuals based on mere correlations with wrongdoers, as opposed to the specific actions of the relevant individual. This is the generalized/individualized suspicion distinction some scholars have already considered.[40] I am currently unaware of specific laws addressing discrimination by governmental[41] data mining in the United States. Note that in the European Union, specific rules governing automated searches may apply, providing individuals with additional rights to learn of the internal processes used.[42]

An additional concern often mentioned when addressing the data mining process is that it is ridden with *errors*. These errors can be of various forms and come at various stages of the process: they can result from errors in the initial data, errors in the aggregation process,[43] errors in the statistical modeling and computer programming, errors in the implementation of the system or errors in the system's ability to correctly define the risks and match them to the strategies on the ground. These errors can have devastating outcomes. First, they can render the entire process ineffective and inefficient—unable to identify real risks while leading law enforcement to follow bogus leads. Yet even when setting these concerns aside (and assuming that they can be tested), errors can have a detrimental effect on specific individuals by causing their subjection to discomfort, additional scrutiny, and even castigation and suspicion for no real reason.[44]

Finally, lack of knowledge and understanding of the internal data mining processes might also raise fears related to due

40. For a discussion and critique of this distinction, see SLOBOGIN, *supra* note 29, at 40.

41. I intentionally emphasize the lack of laws in the governmental realm. In the commercial realm there is some reference to this issue in the Fair Credit Reporting Act, 15 U.S.C. § 1681b (2010). For a critique of this situation and a call for a change, see BERNARD E. HARCOURT, AGAINST PREDICTION (2007). For a very different perspective, see FREDERICK SCHAUER, PROFILES, PROBABILITIES AND STEREOTYPES (2003).

42. For a full discussion of the nature of EU law (as well as the law in the various states), see Douwe Korff, *Data Protection Laws in the EU: The Difficulties in Meeting the Challenges Posed by Global Social and Technical Developments* (European Comm'n Directorate-Gen. Justice, Freedom and Sec., Working Paper No. 2, 2010), *available at*http://ec.europa.eu/justice/policies/privacy/docs/studies/new_privacy_challenges/final_report_working_paper_2_en.pdf.

43. For a discussion of errors in general and of this context in particular, see Anita Ramasastry, *Lost in Translation? Data Mining, National Security and the "Adverse Inference" Problem*, 22 SANTA CLARA COMPUTER & HIGH TECH. L.J. 757 (2006).

44. It should be noted, however, that data mining tools maintain the ability to self-correct the process. As the process rolls on, information regarding success rates, false positives and false negatives becomes available and is "fed" into the analyzing process. Analysts can use such data to fine-tune the algorithms they later apply. In addition, data mining techniques could be used to study the datasets and seek out information which does not fit other data patterns. Analysts could then examine whether such anomalies in the data result from errors and correct the database accordingly.

process.[45]Individuals might fear that adverse action was or will be taken against them without their ability to examine the reasons or challenge the allegations. The data mining process might be inherently opaque, or its inner workings hidden from the public for other reasons. Lacking an understanding of this internal process would encumber the individual's autonomy.[46]

PART III: ALTERNATIVES TO DATA MINING

> *Indeed, it has been said that democracy is the worst form of government except all those other forms that have been tried from time to time.*
>
> Winston Churchill

III.1.Mapping out Alternatives

As the previous section shows, a policy analysis of the data mining of personal information is an extremely complex matter. A comprehensive analysis calls for addressing all these elements and more. In addition, however, a policy study of data mining must consider the alternatives to applying data mining analyses. These are the policy strategies of choice, to be set in place if society refrains from applying data mining. As the quote above demonstrates, examining an issue without considering the alternatives is a futile exercise. In this section, I will briefly present the following five alternatives: (1) doing away with selective security and law enforcement all together, by treating all individuals and events equally; (2) differentiating among events and individuals randomly; (3) distinguishing among events and individuals while relying on the human discretion of field officers who examine personal information pertaining to the specific individual; (4) relying upon profiles and patterns constructed by experts; and (5) applying data mining only to anonymous or anonymized[47] data.

These alternatives are not without overlaps. Solutions might include elements from some or all of these options. Rather than alternatives, these are best understood as trajectories for various policy strategies which could be implemented—with every "alternative" pushing a different form of compromise. An understanding of the solutions' pros and cons along these lines prior to selecting one of them

45. *See generally* Daniel J. Steinbock, *Data Matching, Data Mining, and Due Process*, 40 GA. L. REV. 1 (2005).

46. The "due process" doctrine does not apply for various reasons. *See id.*

47. "Anonymized" data refers to data that went through an anonymization process—the process of removing identifying information and rendering the dataset anonymous.

for further implementation is imperative. I now move to draw out these alternatives.

(1) The Elimination of Selective Security and Law Enforcement

The first and most obvious alternative to government data mining initiatives is refraining from the analysis of personal information to identify individuals and events of higher risk and setting them aside for specific treatment. Generally, this is the alternative to data mining usually envisioned. Yet as I will here explain, it is probably the most unlikely strategy to follow.

Setting aside data mining technologies and the policies they enable will lead to treating *all* individuals (or events) as potentially risky and subjecting everyone to higher scrutiny. What will follow, however, is the transformation of risks of errors in the process into inefficiencies and discomfort, as well as excessive governmental costs.[48] These costs will no doubt be satisfied by using resources that could have otherwise been used to better society (or left in the taxpayers' pockets). A shift to such policy also leads to difficult legal questions as to the authority to subject all individuals to additional burdens when absolutely no evidence indicating elevated suspicion exists.[49] Finally, such a course of action could lead to substantial breaches in security. The fatigue resulting from applying higher security standards to individuals and events that are clearly of low risk might adversely impact the alertness of relevant officials. These officials, at the end of the day, might miss or react poorly to an actual threat.

Deciding whether to prefer this option, as opposed to using data mining, calls for a difficult balance of interests. It calls for tough decisions as to whether society should adopt an initiative which will risk the inconvenience, harm, and even liberty of specific individuals at several junctures. Or perhaps society might opt for this first alternative in which the public in its entirety is taxed, either financially, in terms of attention, or in some cases by raising risks of security (when broad initiatives compromise efficiency). Clearly, liberal and democratic societies should be willing to accept this first alternative and refrain from any data analysis if balancing indicates this is necessary.[50] Furthermore,

48. *See* SCHAUER, *supra* note 41, at 167 (explaining why this alternative is unpractical in the context of law enforcement).

49. For an in-depth discussion of this category of cases, see Christopher Slobogin, *Government Dragnets*, 73 LAW& CONTEMP. PROBS. 107 (2010).

50. It would mean that *all* individuals, for instance, would be required to arrive thirty minutes earlier at the airport to go through heightened security checks.

democratic societies are obligated to do so[51] when important interests of specific harmed groups are at stake. This is the case when discrimination is intentionally carried out on the basis of race or nationality. Yet in other instances which do not involve the risk of reinforcing problematic stereotypes, balancing becomes far more difficult and the results far less clear. In many instances, governments will decide that applying some form of selection and focused attention is prudent.

Yet beyond the normative balancing, this first option is, in many cases, politically unsustainable. As risk manifests and law enforcement resources are stretched, politicians and policymakers will face great pressures to "do something" with the vast datasets at their disposal. Thus, they will be pressured to move away from this alternative. Given the high risks and limited enforcement resources, a form of selection will transpire. The question is, of course, how the selection will take place. This is where data mining and the other options come into play.

(2) Random Selection

Refraining altogether from selective practices in the context of security or law enforcement is unreasonable and unfeasible; the costs would be too high[52] and the fatigue to the system too great. This leads to considering alternatives which call for *selective allocation of resources*. The second alternative applies randomness to meet the security risks at hand.[53] In other words, searches, stops, and other steps of enforcement would be carried out randomly.

Scholarship points to this option as either a strategy that complements data mining profiling or replaces it entirely.[54] Random allocation and testing is an important measure to be applied in conjunction to data mining analyses (or any other strategy). It is important for statistically monitoring the effectiveness of data mining initiatives and examining whether they are justifying the compromises these initiatives call for. Here, however, I am referring to a much broader implementation of random allocation and a much narrower role for data mining.

51. For instance, discrimination on the basis of "sensitive information" such as race is illegal, even when such discrimination is statistically justified. For a critique of this outcome, *see generally* SCHAUER, *supra* note 41.

52. One way to conceptualize such costs is by arguing that encumbering the ability of all individuals to travel when striving to provide for security might limit their freedom of movement. I will refrain from developing this notion. For more on this point, see SLOBOGIN, *supra* note 29, at 102.

53. An option promoted by Bernard Harcourt. *See generally* HARCOURT, *supra* note 41.

54. *Id*.

While applying a random scheme when specific personal information is available might seem a strange (to be polite) option, in some contexts it might suffice. When carried out in public, random checks might achieve sufficient deterrence of criminals and others fearing to be singled out randomly. Random schemes also allow the government to show it is doing *something*—or in other words, create a "security theater."[55] By doing so, governments will sidestep many of the problems data mining presents while also averting the problems of fatigue and the stretching of resources.

With randomness, as with almost any factor, there are several crucial details which must be established. First, there is the actual chance of being randomly selected. A very low chance due to limited law enforcement resources will probably fail to achieve deterrence.[56] A model featuring a very high chance of selection will begin generating the problems of alternative (1). Another issue is how to achieve "randomness." While this might sound trivial, in fact it is quite difficult for individuals in the field to engage people randomly. People are often affected by internal biases and external factors while trying to act randomly. This leads to unfair outcomes on the one hand and the fear of gaming[57] and ineffectiveness on the other hand. For a random search to be truly random, a randomizing tool must be applied—a computerized gadget that will indicate when someone would be selected, stopped, or questioned. Training field agents to ignore their judgment and succumb to a random number generator will not be simple. For all these reasons, administrating randomness might not be as easy as one would think (and it must be clearly distinguished from the use of discretion, addressed below).[58]

Yet even if these problems could be resolved, I believe this option usually is not feasible from a political/policy perspective. Engaging in mere random selection when a great deal of information which might be of relevance is available will be hard for the public to swallow. The notion of ignoring information on the one hand, and subjecting individuals who are clearly of a very low risk to a higher level of scrutiny

55. For a discussion of this concept, see discussion of this reference in Paul M. Schwartz, *Reviving Telecommunications Surveillance Law*, 75 U. CHI. L. REV. 287, 310-11 (2008).

56. When the chance of random selection is very low, such enforcement loses its teeth, as the penalties inflicted cannot be disproportionate to the specific transgression. For more on this point, see SCHAUER, *supra* note 41, at 161. Similar dynamics occurred in the music and film industry when right holders strived to enforce their rights online.

57. Cleary just selecting every tenth person or a similar strategy will allow easy gaming of the system by interested parties (as all they have to do is travel in pairs and one of them will surely be beyond suspicion).

58. *See infra* Part III.1(3).

on the other, would be difficult to accept politically. At times, the public must overcome its aversion to such a solution for important reasons, such as battling racial discrimination. Yet the broader context of this article—overall policy for identifying risks to security and law enforcement—does not provide strong justifications for randomizations to be applied in other contexts (which do not involve unlawful or unethical discrimination).

(3) Selection through Discretion

The third alternative already concedes to both the need for specific treatment of individuals and the use of personal information in this process. With this alternative, a decision-maker examines specific personal information about an individual and makes an informed, ad-hoc, decision. The decision-maker might rely on the information she directly collects at the time of a personal encounter (what the individual is carrying, doing, saying, etc.). Yet she might also rely upon information in the individual's governmental profile when making this decision (what has he done? where has she been?).[59] In most cases, the decisions made in this scheme involve a field officer or a lower level bureaucrat exercising their discretion. Possible examples are tax officers selecting a return for audit, security officers deciding which individuals to subject to additional questioning, or police officers deciding down which street to trek.[60]

To further explain the nature of this alternative, it is important to note what decision-makers at this juncture are *not* doing. First, they are not running analyses which involve the datasets of the entire public (and thus involve individuals entirely removed from the relevant context). Second, the process is not automated (in the computerized sense), although the decision maker might use a computer to view personal information about the subject in real time. Third, it does not involve the formulation of statistical groupings of factors indicating a higher or lower level of risk (at least not intentionally or explicitly). In addition, it is also interesting to point out that this alternative might have operational

59. For more on the practice of structuring the profile, see SCHAUER, *supra* note 41, at 155.

60. The discussion in the text is intentionally avoiding instances in which the actions resulting from the higher level of scrutiny constitute searches or other actions which directly impede upon the liberty of the subjects according to current doctrine (such as extensive stops and searches). I am doing so to sidestep the broader discussion about *Terry* stops and other such actions, where "reasonable cause" or other levels of scrutiny are mandated. For a mapping of these contexts, see SLOBOGIN, *supra* note 29, at 23.

advantages; this model requires officials to think on their feet.[61] Therefore, it differs from some data mining schemes which require individuals to merely apply an algorithm—a modest role which might adversely impact their motivation and performance (although the motivational problem could probably be resolved with alternative measures).

In its most basic form, however, this alternative is merely hypothetical; governments no longer operate in this fashion. Field officers never have full discretion, but are subject to protocols which are the result of central planning. Allowing full discretion and lack of any protocol is simply unthinkable given the inability to control and regulate the actions of these officers.[62] In addition, opting for this alternative will call for ignoring a great deal of knowledge within the system which one field officer cannot possibly integrate. When neglecting to make use of such additional information, existing threats will not be sufficiently met, and potential evildoers will easily circumvent security measures by hiding their intentions.

For these and other reasons, addressing and critiquing this alternative might seem to be like attacking a straw man. There is, however, still merit in examining this practice, even in its purest form. While this alternative is probably rarely exercised or even advocated, policy choices will no doubt reflect variations of this option. Actual policy will be somewhere along the continuum between this alternative and the next one to be discussed, alternative (4).[63] Or, in other cases,

61. This benefit of the discretion model, as opposed to the use of profiles, was emphasized by Justice Marshall's dissent in *United States v. Sokolow*, 490 U.S. 1, 13 (1989). *See* discussion in SCHAUER, *supra* note 41, at 172.

62. For a similar opinion, see SLOBOGIN, *supra* note 29, at123. It should be noted, however, that this issue was recently visited by the United States Supreme Court in a somewhat different context, and with different results. In *Walmart v. Dukes*, 131 S. Ct. 2541 (2011), the Court did not approve a discrimination class action against Walmart under Title VII. Walmart's official practice prohibited discrimination, but it provided a great deal of discretion to local managers. Plaintiffs argued that the class could include all workers who were subject to gender discrimination. The plaintiffs tried to approve a very broad class by arguing that Walmart's policy of local discretion in fact led to these forms of discrimination. The majority emphasized, while relying on previous cases (such as Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 990 (1988)), that individual discretion in itself is a common and reasonable way to conduct business. *Walmart*, 131 S. Ct. at 2554. The court explained that while a discretion-based policy could lead to disparate impact, this result in itself cannot lead to striking down such policy (of local manager discretion). To strike down policy, plaintiffs must challenge a specific discriminatory action (of Walmart, in this case). It is important to note that the specific context here is the approval of a class. In the context of specific claims, I would still believe that adapting a policy which provides a great deal of individualized discretion to local managers is ill-advised.

63. *See infra* Part III.1(4).

some balance between this alternative and a data mining-based system, which provides officers with recommendations, will be applied.

It is also important to point out that these practices are not as distinctively different from the use of profiles or even data mining as they purport to be. The difference is one of degree. On its face, applying discretion calls for treating every person individually. The decision making process this alternative entails is confined to reaching conclusions while only relying on data pertaining to the relevant subject. It is perhaps the most salient example of "individualized suspicion" (as opposed to generalized suspicion)—the ideal form of governmental selection. Every future-looking statement pertaining to one individual's risk and prospects, however, is premised upon statistical analyses (even if it is an unconscious one) of the behaviors of others.[64] In this case, the prediction is carried out within the minds of field officers. These officers generate their predictions on the basis of behavioral patterns they have previously witnessed or learned. In addition, the policy behind the law enforcement framework which leads to the field officers' final decisions is premised (at times, quite subtly) upon predictions, which in turn were premised on some form of statistical analysis. For instance, in some cases, field officers are instructed that relatively minor crimes or actions (such as carrying box cutters) are indicative of other, more serious crimes (such as commandeering aircrafts). This rule is in fact a prediction premised on previous findings and behaviors.[65] In other instances, field officers are required to present individuals with specific tests or questions and scrutinize the results they receive. Again, these questions and tests are structured with previous encounters in mind, and an assumption that similar behavior patterns will reoccur.[66] While these instances do not seem to be predictions and statistical grouping at first, they indeed must be considered as such after some thought.

To sum up our introduction to this alternative, let us examine two important factors which were previously introduced when providing an overview to data mining practices: interpretability and correlation/causation. At first blush, the process involving this alternative is inherently interpretable. It should be possible to inquire as to the reason leading to any specific decision simply by *asking* the decision maker (and steps could be taken to assure that decisions would be logged to assure effective retrieval). Intuitively, this aspect provides an important advantage to data mining practices which might lack

64. *See generally* SCHAUER, *supra* note 41. For instance, if the officer focuses on someone with a gun, it is because he created a profile with the category "people with guns" and is focusing his attention on those within that category.

65. *See id*. at 243.

66. *Id*. at 66.

interpretability at times. Yet the interpretability of this alternative could be called into question. The reasons officials or field officers report may not be their actual reasons (and there is almost no way to verify their claims). In addition, if the officer states that he or she relied on basic intuition or hunch, the decision is virtually non-interpretable.[67]

A similar observation could be made regarding the correlation/causation divide. When initially considering this alternative, one would assume it promotes the use of causation by field officers when applying various decisions and measures. Using causation will provide a safeguard against unfair or erroneous policies. However, law enforcement decisions might be opaque and rely upon intuition. In these instances, they may be premised merely on correlations the relevant official noted in the past which have yet to be backed by a relevant theory (or even authenticated empirically). Again, a closer look at this alternative shows that it is not as promising as we might have originally thought.

(4) Selection through Profiling

The fourth alternative to data mining requires law enforcement to rely upon predetermined profiles for the allocation of resources and risks among individuals and groups.[68] This profile is constructed by experts who apply their common sense, expertise, and experience to the task in a top-down process. For instance, experts will set up parameters for selecting tax returns, individuals at borders, or the location of police cars. They will do so while working through datasets of previous actions and perhaps other forms of behavioral trends addressed in the social sciences.

The differences between this model and data mining (as well as the former alternative) are set along three themes. First, the process does not call for "combing" through the entire dataset of personal information available to the government in the same way that data mining applications require. This difference will surely mitigate public concerns with data mining processes.[69] Note, however, that the profiling process does call for some examining of datasets pertaining to previous problematic acts. In addition, datasets will be reviewed as a whole to get

67. Solove makes a similar point. *See* SOLOVE, NOTHING TO HIDE, *supra* note 7, at 191.

68. *See* SCHAUER, *supra* note 41, at 166 (explaining that such practices are widespread and applied by customs, as well as by the IRS). Schauer also notes that the Supreme Court has upheld the use of profiles in the context of identifying drug couriers, referencing *United States v. Sokolow*, 490 U.S. 1 (1989). *See* SCHAUER, *supra* note 41 *at* 170.

69. *See* Taipale, *supra* note 1, at 180 (quoting Solove and Regan disapproving of such practices for this reason).

a sense of the "normal" levels of the parameters used so that a profile of deviations from the norm can be constructed.[70]

Second, the process will not be automated, but rather generated by human discretion. As opposed to the previous discussion, this process is triggered by the discretion of experts. Obviously, this option again calls for some use of technology: a system will provide the decision-maker with relevant facts, perhaps even with recommendations. Yet the expert is the one making the final decision.[71] Note here that the focus of discretion in this context is quite different than the one explored in the previous example; in this example, discretion is centralized as opposed to peripheral.

The third distinction is derived from the notion of relying on statistics and an "actuary model" which uses "generalizations" when making decisions regarding specific individuals. Clearly, this is the path employed by this alternative. Analysts create groups and subgroups of individuals based on set parameters. These groupings instruct law enforcement to treat those within them differently. Such modeling relies on specific assumptions regarding the ability to predict the future behavior of individuals, as well as to deduce it from others.[72] It also accepts the risk of wrongfully treating an innocent individual who happens to fit within a problematic group or profile.

I again conclude this segment by returning to interpretability and causation. With this alternative, the process will not only be inherently interpretable, but will usually rely on various theories of causation for explaining the elements it includes. This will arguably enhance the autonomy of those subject to the analysis; there will always be an understandable answer to explain the singling out of a specific individual.[73] This alternative will also promote procedural transparency. Relying on causation will, as explained above, provide a check against problematic forms of discrimination and errors.

(5) Selection by Anonymized Data Mining

The fifth and final alternative already accepts the ability of data mining to achieve the objectives at hand. It requires that the analysis be

70. For the process of constructing the profile, see HARCOURT, *supra* note 41, at 104.

71. On the role of experts in the profiling process, see SCHAUER, *supra* note 41, at 155.

72. For an extensive discussion of the "actuarial model," see HARCOURT, *supra* note 41. In Part I of his work, Harcourt draws out the rise of the use of the actuarial paradigm. Critiques of the paradigm are discussed in Part II. *Id.*

73. For the connection between the understanding of the predictive process and the notion of dignity, which is closely aligned to the notion of autonomy, see Steinbock, *supra* note 45, at 23.

conducted using anonymous (or anonymized) data sets; a recommendation set forth by several recent policy reports.[74] These reports call upon the government to engage in analysis through the use of several cryptographic tools which allow for data matching, warehousing, and even mining without providing the analyst with access to the personal information being mined. Such access could be provided at a later time if suspicion arises.

This alternative calls for a different form of balancing. It mitigates only some of the problems of data mining, while leaving others unaffected or even exacerbated. This strategy might reduce some forms of privacy and autonomy-related fears, as the public's concerns of being searched and tracked will be eased by knowing that the government cannot connect their personal data to their real identity.[75] This alternative, however, increases the chances of errors within the process and the lack of transparency. In addition, concerns regarding the practices that follow from data mining will persist. This alternative still allows for the generation of patterns, which could later be used to unfairly distinguish among individuals and events as parts of groups (rather than strictly being considered as individuals). Thus, applying this alternative comes with non-trivial costs (in terms of both real out-of-pocket costs as well as the costs of errors and engaging the system with additional process). It also appears to solve only few of the overall concerns.

Considering this alternative also requires some rethinking as to the actual protection anonymity provides. Recent studies have indicated that a massive anonymous database of personal information with a multitude of factors about every individual can be re-identified by sophisticated attackers.[76] This is especially true if another database of identifiable personal information is at these attackers' disposal.[77] Thus, rogue government analysts would probably be able to circumvent the protection measures of anonymization here mentioned should they choose to do so. These new findings weaken the attractiveness of this fifth alternative. However, in the governmental context at least, concerns of hacking and circumvention are probably manageable by applying internal security measures which would limit access and control

74. *See, e.g.*, *TAPAC Report*, *supra* note 24;MARKLE FOUNDATION TASK FORCE, CREATING A TRUSTED NETWORK FOR HOMELAND SECURITY (2003).

75. For empirical findings showing this point, see SLOBOGIN, *supra* note 29, at195.

76. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010).

77. This was the case in the Netflix/IMDb fiasco. Such multi-factored datasets are now at the disposal of many public and private entities. *Id.*at 1746-48.

the data.[78] Still, this fifth alternative requires a great deal of further review.

III.2.Distinguishing between the Field Officer, the Profiler, and the Data Miner

As the discussion above indicates, three key options enable government to engage in selective enforcement and scrutiny: data mining and alternatives (c) and (d). There are key differences between these options, with crucial policy implications. In this segment, I briefly examine these differences in greater depth. The first point—that of discretion and the model of decision making—is one which must be constantly revisited when examining data mining and its alternatives. In addition, I will address distinctions which are commonly considered and referred to—the use of statistical groupings and computerized automation—yet should not receive substantial weight in subsequent analysis of policy options.

Let us first examine the notion of human *discretion* and the different methods of decision-making these models employ. Selecting among alternatives leads to a choice between various forms of human discretion and a balance between human and automated discretion. Choosing between methods of discretion has several implications. The main implication is the forms of errors the alternative generates.[79] If one form of discretion generates predictable errors, the system would be easily gamed and manipulated, even if those errors are unsubstantial. If the errors are systematic, a specific population segment would be harmed (again, even if overall efficiency is maintained). If the errors are both systematic and detrimental towards segments of the population which are either weak or were singled out in the past, then those errors lead to an additional set of problems.

Preferring human discretion, as opposed to abiding by an output of a data mining-powered application, leads to at least two shortcomings (which pertain to almost all decisions premised on human cognition) that quickly transform to errors in the final outcome: human decisions tend to (a) make use of heuristics and (b) employ hidden biases. Both dynamics

78. This option still hold substantial benefits, as it minimizes the risk of illegal abuse of the information by government executives (such as the many stories occurring every year of tax officials sharing or selling personal information about citizens). Note, however, that this problem could also be mitigated through internal disciplinary actions.

79. If one form of discretion generates frequent errors, the entire process is compromised. Let us assume, however, that the threshold of a reasonable level of errors would be attended to as a preliminary matter—and if the level of errors will be unacceptably high, the project would be set aside. Yet as I demonstrated in the text, even with an overall acceptable level of errors, problems still remain.

are systematic and predictable. The latter shortcoming also generates errors detrimental to specific weaker and vulnerable population segments. I now take a closer look at both elements, explain how they generate differences between the models, and briefly note the implications of these differences.

A vast literature regarding heuristics clearly indicates that when dealing with complex tasks, the human brain applies various shortcuts, which allow it to overcome information overload.[80] These rules of thumb often lead to correct decisions. At times, however, heuristics leads to predictable errors. This occurs when individuals face the need for quick decisions with limited attention and vast information to consider. While some of these errors can be corrected through training and experience, many others cannot.

Considering the alternatives pointed out above quickly leads to recognizing flaws in alternative (3),[81] which relies heavily on the individual discretion of field officials. This alternative will lead to predictable cognitive traps where heuristics will be applied but lead to wrong results, which adversaries might abuse. Thus, for this reason alone, opting for alternative (3) will come at a high price in terms of efficiency and fairness. When opting for alternative (4) (expert-driven profiles), this concern might be somewhat mitigated.[82] Experts have greater awareness of these tendencies to err and focus more on empirical findings than intuition. They also need not make quick decisions under pressure. This process could be inflicted with heuristic-related errors as well, however, given the reliance on human-based discretion. On its face, data mining should face the least of these troubles. Computers have no need for shortcuts and heuristics when they have the capacity to address all relevant data. And when, for efficiency purposes, only segments of data are addressed or another analytic shortcut is applied, it is a shortcut of which operators are well aware and can take into consideration.

Furthermore, relying upon human discretion allows for the internal biases of the individual decision-makers to impact their actions and decisions, even if only inadvertently.[83] At times, behind the discrete decision of the experienced official or expert is a discriminatory notion

80. For a discussion of these dynamics, see Avishalom Tor, *The Methodology of the Behavioral Approach to Law*,4 HAIFA L. REV. 237 (2008), and Russell Korobkin, *Bounded Rationality, Standard Form Contracts, and Unconscionability*, 70 U. CHI. L. REV. 1203, 1206 (2003).

81. *See supra* Part III.1(3).

82. *See supra* Part III.1(4).

83. I have argued this point elsewhere. *See* Zarsky, *MYOB*, *supra* note 5. For a similar argument, see Taipale, *supra* note 14, at 33 n.118.

or belief premised upon (at times, subconscious) animosity towards specific segments of the population, or other forms of prejudice. These outcomes may be inefficient.[84] And far worse, these outcomes are unfair towards the weaker and "protected" segments of society.

Biases can transpire within the frameworks of both alternatives (3) and (4). Field officers are the most susceptible to generating these distortions. A recent review of studies addressing law enforcement field decisions and race shows an alarming and distorted picture of biased conduct.[85] For this reason, providing full discretion to field officers is unthinkable.[86] Yet even relying on expert decisions (as in alternative (4)) might lead to some of these concerns. Expert decisions might be plagued with internal biases. Rather than relying upon empirical findings and expertise, experts might be motivated by beliefs and prejudice. Note, however, that alternative (4) has the advantage of a central process. As opposed to a system where decisions are made at the periphery, the experts' profiles could be closely audited and studied in an attempt to identify arbitrary conduct that might lead to unfair discrimination.

With data mining, however, these problems are again mitigated. Applying an automated process allows the central planner to retain better control over the actions in the periphery. Yet data mining provides an additional, more substantial benefit in that computer modeling is not driven by human assumptions (which might be both hidden and biased) but, rather, by the data. Therefore, concerns regarding hidden biases premised on prejudice might be sidestepped by applying data mining.

Many will disagree with this last statement. Data mining, these dissenters will argue, allows for a flurry of biased decisions to be carried out undetected.[87] These biases might be put in place through the points of human interaction listed above (especially in the process of programming the relevant software), which in many cases are hidden from public scrutiny. Thus, data mining allows for the embedding of

84. According to Schauer, this was exactly the case in O'Hare airport, where it was revealed that the percentage of minorities made subject to intrusive cavity searches was disproportionately high. When such practices, which were no doubt motivated by racial animosity, were stopped, the success of such searches has increased. *See* Schauer, *supra* note 41, at 176-79.

85. Bernard E. Harcourt & Tracey L. Meares, *Randomization and the Fourth Amendment* (The Univ. of Chi. Law Sch., John M. Olin Law & Econ. Working Paper No. 530, 2010).

86. *See id.* at 4 (explaining that part of the role of the Fourth Amendment is to limit the discretion of law enforcement). Yet note in a somewhat different context the recent Supreme Court opinion in Wal-Mart Stores, Inc. v. Dukes, 131 S. Ct. 2541 (2011), which states that individualized discretion is a reasonable way to conduct business and should not necessarily be understood as a policy which promotes discrimination. *See* discussion *supra* note 62.

87. SOLOVE, NOTHING TO HIDE, *supra* note 7, at 191.

values as well. The difference between these options amounts to the ease of embedding values ex ante and the ability to identify these instances ex post. Those arguing against data mining will state that biases can be built into decision-making processes quite easily ex ante, and they could be very difficult to identify after the fact. For that reason, data mining runs high risks of generating biased conduct as well.

I believe, however, that the problems mentioned are not inherent features of data mining, and they certainly are not beyond repair. If the data mining process is sufficiently transparent, it can effectively overcome these challenges. Adding interpretability and even causation to the data mining process could allow policymakers to assure that biases are averted. In addition, analysts could keep a close eye on the forms of software used and the protocols applied when using it. Biases in a central computer code, once acknowledged, could be tackled with ease and identified effectively by external review. Managing and mitigating hidden biases in the actions of numerous field officers vested with a great deal of discretion is a much harder task. This would call for tracking, evaluating, and disciplining all actions carried out in the periphery.[88] Similarly, examining the actions of a group of central experts seems daunting, will generate numerous painful confrontations, and might prove ineffective.

A *second* issue related to the differences between these central alternatives pertains to the use of decisions premised on statistical groupings as opposed to individualized suspicion. Deciding on the basis of a statistical analysis of a group leads to a much broader debate in which some scholars show great resentment to the "actuary method."[89] This is the notion that individuals are treated as parts of groups which have specific predefined traits and parameters. Similar methods are broadly adopted in many features of modern life (especially in insurance), as opposed to actual clinical work to examine the relevant situation.[90]

While using this form of statistical analysis might generate negative sentiment, I believe categorically rejecting the "actuary method" is unwise. Relying merely on an individual's record not only is inefficient, but implicates (at times, subconscious) reliance on groupings as well.[91] In addition, a non-automated process (which fits within the confines of alternative (3)) usually includes several crucial detriments, especially the

88. I acknowledge that even when using a central system, auditing of the actions of the periphery operation is needed as well. Yet this would be substantially less than the level required in the alternative model.

89. HARCOURT, *supra* note 41, at 16-17.

90. For more on the clinical/actuary distinction, see *id*. at 106-07.

91. *See* discussion *supra* Part III.1(3).

lack of interpretability and transparency.[92] If the public or another auditing entity do not understand how assumptions about groups are used in an arguably "individualized" process (because grouping is carried out unintentionally or subconsciously), such assumptions can go untested and lead to the various errors mentioned above. Options that explicitly use statistical analysis (such as alternative (4) or data mining) could be rejected, but it should be for other, more specific reasons.

The *third* issue is that of automation. Deciding between the alternatives mapped out above is also a subset of a broader discussion concerning the role of computer-generated decision-making in society.[93] Those fearing computerized and automated decision-making show disdain for the tyranny of computers, believing that the process of analysis by computers lacks dignity, may lead to systematic errors, and is incapable of accounting for the delicacy of the human condition.[94] They also fear that society does not easily accept errors made by computers, as opposed to easily accepting that "to err is human."[95] These are all powerful arguments against the spreading use of data mining as well.[96]

We must always fear, however, that the arguments mentioned are in fact rationalizations of a fear of technology with a neo-Luddite flavor.[97] In other cases, a negative sentiment towards automation might result from a tendency to underestimate technology and its ability to match the analytical abilities and decision-making capabilities of humans. This final belief should be challenged, especially in the context of tedious tasks which call for difficult decisions premised upon multiple variables. Therefore, I do not find this issue on its own a sufficient factor to be considered in this discussion.

The fear of technology in general and data mining in particular is in many cases difficult to articulate. This fear has strong and meaningful explanations which we must diligently seek—the fear of errors, the loss

92. *See* discussion *supra* Part III.1(3).

93. For a discussion of this matter in the Corporate Risk Management setting, see Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669 (2010).

94. SOLOVE, NOTHING TO HIDE, *supra* note 7, at 186-88.

95. *See generally* Korff, *supra* note 42.

96. *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1305-06 (2008) (explaining ways to "debias" this notion).

97. *See* Taipale, *supra* note 1, at 126 n.3 (discussing the Luddite Movement). *See also id.* at 137-38 (discussing the illogical fear of technologies and its foundations). For a very different perspective, see SOLOVE, NOTHING TO HIDE, *supra* note 7, at 201. Solove notes that those embracing technology too quickly might be vain and unwilling to take needed precautions. *Id.* at 203. He refers to such conduct as the "Titanic Phenomenon" where the planners of the ship were so confident it would not sink they saw no reason to provide a sufficient number of lifeboats. *Id.* He notes such dynamics might also be in play in the data mining this context. *Id.*

of autonomy, the lack of transparency, and others. Yet these concerns must be tackled directly while setting the broader, somewhat vague, fear of automation aside.[98]

PART IV: CONSIDERING ALTERNATIVES—EXAMPLES OF A NORMATIVE AND POLITICAL ANALYSIS

Data mining practices generate several forms and layers of concerns. They are also here to stay. It is fair to assume that the policies governments finally implement will be a compromise between data mining and the five alternatives mentioned. As part of establishing these compromises, various alternative courses of action must be taken into account. This part briefly demonstrates the importance and impact comparative analysis can have in two contexts. The first is normative. When contemplating the implementation of data mining processes, policymakers and courts will examine whether the measures used are the least harmful and intrusive. This should be done with regard to all the concerns raised in this context.[99] At every juncture, the framework adopted above could prove helpful in exploring other measures and establishing their suitability. This segment will merely launch such a process. It will demonstrate the complexity of seeking out the least intrusive measures while examining the potential harms of data mining analyses when conceptualized as forms of "search."

The second context is that of the political process theory. According to this theory, at some points, the outcomes of the political process should receive limited judicial review. This segment separately reviews data mining and its alternatives from this theoretical perspective and strives to understand the differences among these options and their implications. It takes an additional step and proactively shows that the political process might reject data mining for a very problematic reason—data mining provides a *balanced* treatment of minorities.

IV.1. A Normative Example: Data Mining as "Search"

Let us begin with a normative analysis which focuses on the "search" data mining entails. This issue pertains to the first step in the personal data flow. It is suited for further elaboration because it presents some of the greatest challenges to data mining. The comparative analysis illuminates these challenges, which scholars and citizens might sense viscerally but find difficult to clearly articulate.

98. For a similar view, see Taipale, *supra* note 1, at 178.
99. *See supra* Part II.

The data mining process calls for the substantial analyses of personal information which might be referred to as "searches" in layman's terms. In this process, computer programs work through a broad array of datasets on their way to developing clusters, links, and other outputs. Thereafter, the programs examine specific sets of personal data in real time in an effort to establish whether they fit the predictive models previously constructed.

While the public might refer to these actions as searches, they are not clearly "searches" in the eyes of the law.[100] The law regulates "searches," limits their scope, and sets systematic boundaries to assure the protection of rights. These steps are commonly discussed in the Fourth Amendment context, which protects the people from unreasonable searches.[101] Whether current Fourth Amendment doctrine will find data mining to be a "search" is a difficult question, which is beyond the scope of this article (but, as mentioned above, will probably be answered negatively).[102] Fourth Amendment jurisprudence has focused on the gathering of information as opposed to its subsequent analysis. Therefore, deriving an answer from the existing case law is difficult.

In this inquiry, I choose to examine whether the interests (as opposed to the legal doctrine) commonly associated with the Fourth Amendment and its protection from unreasonable searches calls for limitations to data mining. If that is the case, we will assume that the law, generally speaking, will move to limit these forms of searches, either through legislation or, in the less likely case, the rulings of courts. Yet, at least in the context of regulation and legislation, policymakers must inquire which alternatives will surely follow if data mining is ruled out and whether they will prove to be more harmful to the specific interests here explored.[103]

100. For a discussion of the disparity between "search" in law and in layman's terms, see SLOBOGIN, *supra* note 29, at 33-35 (citing Christopher Slobogin & Joseph E. Schumacher, *Reasonable Expectations of Privacy and Autonomy in Fourth Amendment Cases: An Empirical Look at "Understandings Recognized and Permitted by Society,"* 42 DUKE L.J. 727, 743–51 (1993)).*See also* Stephen Henderson, *Nothing New Under the Sun?: A Technologically Rational Doctrine of Fourth Amendment Search*, 56 MERCER L. REV. 507, 544 (2005).

101. U.S. CONST. amend. IV.

102. *See generally* Cate, *supra* note 3. For additional discussion, see TAPAC REPORT, *supra* note 24. I have argued this point elsewhere. *See* Zarsky, *MYOB*, *supra* note 5. For a similar argument, see Taipale, *supra* note 14, at 33 n.118.

103. Framing the discussion in this way is crucial for two reasons. First, it allows us to break away from the current Supreme Court case law, which probably leads to a simple response to these questions: that data mining does not constitute a search. The approach stated in the text allows for further considering whether the important interests which might be beyond the constitutional language and doctrine could be attended to

This inquiry is complex due to the lack of consensus among scholars regarding the Fourth Amendment's central principles and interests. To overcome this analytical challenge, this part sets forth three normative theories. These theories demonstrate the way data mining practices compromise interests addressed in other discussions of "searches." The theories set forth are inspired by the existing literature examining the Fourth Amendment in a technological age in general, and in the context of data mining in particular. This section's goal is to merely familiarize the reader with the basis of these theories so to facilitate the limited discussion that follows. Furthermore, the list of theories presented is not necessarily an exhaustive one, nor need it be. If other theories are set forth, this methodology could be applied to them as well.

Before beginning, a few words regarding the problematic starting point of this discussion are required. The starting point of this analysis is that data mining analyses are not "searches" (according to existing Fourth Amendment doctrine). The analysis set forth assumes that data mining (or other forms of data analysis) is carried out while relying upon data which was initially collected lawfully by third parties and later passed on to the government, or directly by the government from the data subjects. With this assumption in place, American law regarding searches[104] generally assumes that individuals have no subsequent privacy interest (at least in terms of "searching" and the Fourth Amendment).[105] The point of data collection is where data subjects relinquish control over the data and its future uses. Yet data mining

through a legislative response. Orin Kerr, for instance, has recommended in several contexts that the correct response to situations that are probably beyond the doctrinal reach of the Fourth Amendment, but which are still troubling, is a legislative one. Orin Kerr, *The Fourth Amendment and New Technologies: Constitutional Myths and the case for Caution*, 102 MICH. L. REV. 801, 806-07 (2004). For a recent critique of this opinion, see SOLOVE, NOTHING TO HIDE, *supra* note 7, at 164, stating the statutes have not been able to keep up with new technologies or draw out comprehensive rules and are unclear. Rather, surveillance should also be regulated by courts, which can rely on broad concepts of the Fourth Amendment. *Id.*

In addition, framing the discussion as one that involves normative values and concepts that are central to the constitutional discourse, as opposed to constitutional analysis, allows for a discussion of balancing among options. If the analysis would show that data mining practices constitute illegal searches, the fact that other alternatives are just as harmful would have only limited importance—as such a finding would probably lead to striking down those practices as well, rather than legitimizing data mining practices.

104. Note that EU law and even US law in other contexts (such as the Privacy Act) set out a different balance. *See supra* note 30 and accompanying text. For more on this point, see Cate, *supra* note 3, at 453.

105. The Privacy Act provides limited rights in this instance. *See supra* note 105 and accompanying text.

allows the government to add additional layers of knowledge after further analyzing the data—knowledge previously undiscovered by either side. This novel development might lead to changing the underlying assumptions (and thus the law) regarding "searches", data mining, and the applicability of the Fourth Amendment on the basis of the following theories.[106] It might find that in these instances as well, "search"-related interests are compromised. Yet making this case is an uphill battle.

The first theoretical response ("A") states that the prospect of data mining analysis psychologically intrudes[107] upon individuals and their rights. Therefore, data mining analyses resemble physical searches to the home and self and should be regulated accordingly. This sense of intrusion is primarily derived from two key elements of the data mining process. First, the process's automated nature might generate anxiety. Second, data mining's ability to predict future behaviors may cause worry. These predicted behaviors might be premised upon thoughts and traits that relevant individuals have strived to keep secret or perhaps did not fully grasp.[108] Empirical data gathered regarding the public attitude towards searches upholds this theory, while showing indications of anxiety towards these novel practices.[109]

A second theory ("B") is similar to the first but looks beyond the psychological intrusions and takes an objective and normative approach.[110] Theory B notes that as a policy matter, the government should not be entrusted with a powerful tool which can turn even seemingly benign factors into a powerful mapping of an individual's

106. For a discussion of this issue, see Blitz, *supra* note 31.

107. This notion of "psychological intrusion" in computer searches (as a notion that would provide Fourth Amendment protection) was not accepted by the Sixth Circuit Court of Appeals in *United States v. Ellison*, 462 F.3d 557 (6th Cir. 2006). It was, however, noted by the dissent. *Ellison*, 462 F.3d at 568 (Moore, J., dissenting). For further discussion, see SOLOVE et al., *supra* note 32 at 207.

108. *See* Orin Kerr, *Four Models of Fourth Amendment Protection*, 60 STAN. L. REV. 503 (2007) (mapping out four theoretical models to understand and analyze the Fourth Amendment which are used interchangeably by courts). The theory presented in this segment coincides with Kerr's first model—the *Probabilistic* Model—a descriptive model premised on expectations based on current social norms. *Id.* at 508-13.

109. SLOBOGIN, *supra* note 29, at 184-85.

110. Returning to Orin Kerr's "Four Model" framework, this theory would be part of the "Policy Model" which engages in a normative balance between the needs of police and the threat to civil liberties. Kerr, *supra* note 108, at 519-22. However, it might also be classified under his second model—the "Private Facts" Model, noting that the predictions here disclosed should be considered as private and therefore protected. *Id.* at 512-16. As the predictions cannot necessarily be considered as "facts" but as mere correlations or predispositions, it is unclear whether theory "B" could indeed fit within this model.

persona and insights[111] (or, if the government is allowed to do so, such power should at least be closely scrutinized). In other words, when it comes to data mining, the results of the analysis amount to more than the sum of the parts of the dataset that were previously collected. Therefore, the fact that the governmental actions were reviewed by courts at the data collection stage is insufficient given the additional layer of knowledge which the subsequent mining process can provide. Thus, additional scrutiny is required at the data mining "search" stage as well.

The notion that many seemingly innocuous bits of information, which were collected lawfully, should be treated differently when aggregated (as opposed to when the government contemplates the collection of every bit on its own) is generating traction in the ongoing Fourth Amendment discourse. It is currently fiercely debated in the context of location-based data—especially that which can be collected by mobile phone operators. For instance, in a controversial opinion, the D.C. Circuit chose to restrict governmental collection of location-based data over a long time period while promoting the "Mosaic Theory."[112] This theory argues that small bits of innocuous information, when brought together, can provide a full mosaic of an individual's persona, and therefore such practices should be further scrutinized. Data mining arguably should be considered as a "search" and thereafter restricted as part of a similar analytical move. Data mining transforms small segments of information into an overall "mosaic" of human behavior. Such a force should not be left unchecked in the hands of government.[113]

The third theory ("C") focuses on another novel attribute of data mining—the fact that this analysis calls for actively examining and analyzing personal datasets pertaining to a very broad segment of the population, including those whom are substantially removed from the matter at hand. Data mining applications might at times work through very broad segments of the governmental dataset when formulating patterns, trends, and clusters. According to this theory, such practices

111. This second theory also fits well with those who see the role of the Fourth Amendment as a tool to limit the permeating of government. *See* Marc Blitz, *Stanley in Cyberspace*, 62 HASTINGS L.J. 357, 380 (2010).

112. United States v. Maynard, 615 F.3d 544, 562 (D.C. Cir. 2010), *cert. granted*, 131 S. Ct. 3064 (2011). For a critique, see Orin Kerr, *Applying the Mosaic Theory of the Fourth Amendment to Disclosure of Stored Records*, THE VOLOKH CONSPIRACY (Apr. 5, 2011, 4:54 pm), http://volokh.com/2011/04/05/applying-the-mosaic-theory-of-the-fourthamendment-to-disclosure-of-stored-records. Several courts have taken the opposite position and allowed for these forms of surveillance. *Cf.* United States v. Hernandez, 647 F.3d 216 (5th Cir. 2011) (holding that government's use of hidden GPS to track defendant's movements was not an unconstitutional warrantless search); United States v. Cuevas–Perez, 640 F.3d 272 (7th Cir. 2011) (holding that placement of GPS tracking unit on defendant's vehicle did not violate Fourth Amendment).

113. I thank Kathy Strandburg for this insight.

amount to a "fishing expedition" on behalf of the state—the practice of looking through the files and personal effects of individuals who raise no suspicion while striving to build a case on the basis of information they might recover.[114] Curbing "fishing expeditions" by governments is one of the central roles of judicial review.[115]

Promoting theory "C" faces a serious analytical challenge. If the government may review and analyze information which was already lawfully collected in any way it deems fit (without such actions being considered a "search"), data mining cannot be considered a "fishing expedition." No "search-related" interest is compromised by the analysis. This result follows from today's acceptance of "searches" as an almost dichotomous variable; actions are either a search (and thus lead to a harsh legal analysis usually calling for the finding of "probable cause") or they are not.

Yet recent scholarship argues that this understanding of "searches" in the Fourth Amendment context should be abandoned for a proportionality-based analysis.[116] With proportionality, search-related interests could be compromised by various levels of intrusions. Every such intrusion will call for a proportionate level of protection and standard of review. Data mining analysis could be considered as a minute intrusion at times, but even so, engaging in the analysis of a very broad segment of the population would be rendered a disproportionate measure.

After briefly introducing these theories which present the clash between data mining analyses and search related interests, we must examine whether these theories generate similar concerns when applied to alternative strategies for selective law enforcement. A comparative analysis quickly shows that alternatives (1) (which called for equal treatment for all) and (2) (random selection) generate none of these concerns because there are no searches (of the form here discussed) with these options. As explained above, however, following these alternatives is usually not politically or practically feasible. Therefore, the policy debate must move to explore the next set of options. Introducing

114. For a critique of this aspect of data mining, see Taipale, *supra* note 1, at 180.

115. *See* Daniel J. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S. CAL. L. REV. 1083, 1106-07 (2002).

116. This theory is promoted by Slobogin, who applies it to data mining as well. SLOBOGIN, *supra* note 29 at 206-11. Slobogin concludes that, based on current findings and with the exception of instances related to homeland security, data mining practices almost always constitute disproportional measures. *See also id.* at194 (discussing data mining). Note that Slobogin asserts that some data mining processes are far from minute and call for a more rigorous standard of review. Slobogin also notes that political process theory (to be discussed below) might lead to accepting data mining nonetheless if specific requirements are met. *See infra* note 129.

alternatives (3) and (4) into this analysis calls for working through every one of the three theories stated above and closely examining the differences between these options.[117] Table I below summarizes the results of such an analysis.

Examining the relevance of theory "A" to alternatives (3) and (4) leads to inconclusive results. Theory "A" is premised on the fickle notion of psychological intrusion; the mental discomfort associated with learning of the data mining process. As mentioned, such mental discomfort might result from several elements. For one thing, the data mining processes are automated with a limited role for human decision-making and discretion. Alternatives (3) and (4) call for a very different process in which actual people look through the individual files. For some, data mining generates greater anxiety given concerns with automated and computerized decision-making processes. Thus, opting for data mining would not be advisable. For others, however, the opposite would be true.[118] These persons would not be alarmed by the faceless computer searching their data.[119] They would, however, be gravely concerned with actual individuals looking through their information, even if carried out to a lesser extent. A similar complication will follow when considering the psychological intrusion resulting from fears of powerful revelations made by a computer algorithm. While this might be the perspective of some, others might have greater fears of the main concerns alternatives (3) and (4) generate: errors and biases which the computer analysis could limit.

Both arguments and points of view seem acceptable, even reasonable. The differences of opinion people will have regarding this point will result from disparities in their understanding of the data mining technology, its benefits, and its detriments. Thus, ranking these alternatives on the basis of a theory of psychological intrusion (theory "A") is almost impossible at this juncture. A possible measure to overcome such difficulty might be through conducting surveys to establish the public's position.[120] Yet administering such surveys would

117. *See infra* Table I for a brief summary of this segment.

118. For instance, *see* SOLOVE, NOTHING TO HIDE, *supra* note 7, at 183 (citing Eric Goldman, *Data Mining and Attention Consumption, in* PRIVACY AND TECHNOLOGIES OF IDENTITY: A CROSS-DISCIPLINARY CONVERSATION 225, 228 (Katherine Strandburg & Daniela Stan Raicu eds., 2005).

119. A similar point was recently made by Mathew Tokson. *See* Mathew Tokson, *Automation and the Fourth Amendment*, 96 IOWA L. REV. 581, 602-09 (2011). On the other hand Bruce Boyden recently explored whether entirely computerized searches can be considered a "search." Bruce Boyden, *Can a Computer Intercept Your Email?*, 4 PRIVACY L. SCHOLARS CONF., June 2, 2011.

120. *See generally* Slobogin& Schumacher, *supra* note 100.

be a very difficult, perhaps near-impossible task.[121] Therefore, when accounting for alternatives, theory "A" seems of limited use in establishing which alternative leads to optimal outcomes.

A look at theory "B" illustrates how alternatives (c) and (d) prove superior to data mining from this specific perspective. The key to the concern voiced by this theory is data mining's ability to discover hidden behavioral trends while bringing many bits together to form a meaningful whole. Data mining's analytical strength is the key to its normative disadvantage. Policymakers might indeed find it unacceptable to provide the government with the unchecked ability to carry out such powerful searches and reach meaningful insights. Analysts would, however, be fine with the options we already know quite well: relying on the work of experts and field officers with their limited abilities. These alternatives strike an acceptable balance between law enforcement needs and civil liberty interests, even though they might compromise overall effectiveness.

A similar conclusion, which views data mining as a greater threat than its alternatives, will follow from theory "C." When experts (alternative (4)) or officials (alternative (3)) carry out their duties, they will confine their analysis to datasets of known criminals, terrorists, and harmful events. In these limited instances, a specific suspicion of wrongdoing exists or there is some relevance of the examined data to the issue at hand. Thus, the data analysis process in these two alternatives would not compromise the rights of a broader segment of the public. The analysis processes in alternatives (3) and (4) refrain from gazing, either electronically or manually, at the records of broad segments of the population.[122] Alternative (3) seems to provide the greatest protection from "fishing expeditions" by merely addressing specific individual datasets. This individual usually is *not* a suspect and has not committed any wrongdoings.[123] Yet, when considering theory "C," this alternative is superior to scanning the records of thousands of individuals with no connection to the event under investigation. Even with alternative (4), the concern voiced by theory "C" is mitigated in comparison to data mining. Experts cannot possibly engage in the extensive study of the

121. *See* Daniel J. Solove, *Fourth Amendment Pragmatism*, 51 B.C. L. REV. 1511, 1522-24 (2010) (noting the difficulty of collecting information regarding the public's true preferences, as opposed to inaccurate responses resulting from errors in judgment).

122. One can easily counter this argument by stating that even in the process of carrying out analyses according to alternatives (3) and (4), analysts must look to a dataset of the general public to see whether the pattern or protocol they constructed is overbroad and generates vast false positives. A possible response could be that this step of the analysis could be carried out with anonymous data.

123. He or she might be a traveler, an individual submitting a tax form, or even a pedestrian—however the context of the relevant inquiry pertains directly to them.

public's personal data to the same extent the new data mining applications allow.

I conclude this discussion by addressing our fifth and final alternative: the data mining of anonymous or anonymized data in the context of these three theories.[124] At first blush, this alternative should shield individuals from "search" related concerns on all dimensions. If the risks mentioned above[125] of the re-identification of personal information in the dataset can be limited through technological and disciplinary measures, then the "search" interests of the individual in the anonymous data are minimized. Working through the three theories stated above leads to interesting results, which at times counter this intuition. The results of this analysis are summarized below in Table I as well.

In terms of theory "A," it indeed is reasonable that psychological intrusion would be mitigated if the data analyzed were anonymous. It is questionable, however, whether "search" related concerns would be mitigated under theory "B." Even if data were rendered anonymous, the data mining analysis would be able to generate a whole which surpasses the sum of its parts. Anonymization is not intended to substantially curb the governments' ability to derive insights if carried out optimally. Therefore, when shifting the focus of the theoretical analysis to the outcome of the data mining process, this measure appears to achieve little to mitigate this specific concern.

Examining theory "C" in the context of alternative (5) probably raises the most difficult theoretical issues. It is unclear whether the "fishing expedition" concern is mitigated by the fact that the initial information used is anonymous. After all, should the analysis lead to suspicion, the cloak of anonymity would quickly be removed. Thus, individuals still might be fearful of these blanket anonymous searches, even if their names are not apparent to the searchers at the time of the analysis. On the other hand, at least at the preliminary stage, the data subjects' identity is concealed and thus the broad searches might not seem as intrusive. Additional work is required to establish whether anonymization mitigates these specific concerns.

124. *See infra* Table I for a summary of the analysis and discussion of alternative (5) as well.

125. *See supra* notes 76-77 and accompanying text.

	Alternatives		
	3 (Field Officer Discretion)	4 (Expert Profile)	5 (Anonymized)
A	Might limit concern (a countering argument could be made if individuals find comfort in the lack of human contact with data)	Might limit concern (a countering argument could be made if individuals find comfort in the lack of human contact with data)	Limits concern
B	Limits concern (ability of officers is limited)	Limits concern (ability of experts is limited)	No effect (yet generates errors)
C	Limits concern (only examine data pertaining to relevant individual)	Somewhat limits concern (examine relevant individuals and parts of the dataset)	Limited effect (?)

Table I

To summarize, examining the legal aspects of data mining through the prism of its alternatives leads to interesting insights which must be incorporated into policy decisions. In this section, the article merely demonstrates how an alternative-based study must be carried out across the various issues detailed above.[126] This segment also shows that the balance among alternatives depends upon the relevant theory adopted to explain the most basic factors of data mining analysis. Yet above all, this segment demonstrates the complexity of the question as to whether data mining analysis should be adopted or abandoned. A similar analysis, which takes all alternatives into account, should be carried out at every juncture and in every context prior to making final rulings and policy decisions concerning the implementation or abandonment of data mining-based initiatives for selective law enforcement.

IV.2. A Political Process Analysis: Data Mining vs. the Tyranny of the Majority?

Beyond the legal debate, the process of deciding whether and how data mining initiatives should be implemented is public and political. The public's actual and predicted reactions to the various policies set forth will affect politicians who will change their positions accordingly.

126. *See supra* Part II.

These positions will be reflected in the laws and policies ultimately set in place. Examining the fairness and effectiveness of this specific political discourse is of great importance and has recently been visited by several scholars.[127] Framing the analysis of data mining as a political process analysis might even prove to be the most suitable context for the discussion of optimal public policy. A legal analysis of the intrusiveness, effectiveness, and legality of data mining requires a close analysis of its alternatives. The previous discussion shows that this is an extremely difficult task. Given these difficulties, perhaps rather than having courts second guess the legislature and engage in the difficult task of balancing, the judiciary should provide deference to the political process when it can assume that the process was a fair one.[128] Therefore, political process analysis is fertile ground for further research of data mining and its alternatives. I outline these aspects here and hope that they will be developed in future works.

(1) Data Mining, Political Process, and Deference

The first aspect to be explored pertains to the role of the courts in the process of examining (and indirectly regulating) data mining initiatives. Courts have often been called upon to examine whether the balances set at these junctures are fair and respect the rights of the people. Some commentators[129] have called for limiting judicial review when the governmental action is said to impact groups.[130] In these cases, the groups should have sufficient power to attend to their own interests in the political arena, and deference should be given to the legislative process. Courts will intervene to assure that the rules are implemented even-handedly and are not patently irrational.[131]

127. *See* Slobogin, *supra* note 31, at 19 (citing John Hart Ely, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW (1980)). Note that Ely did not address these contexts specifically in his work. *See generally* ELY, DEMOCRACY AND DISTRUST. However, this thesis was set forth by Richard Worf. *See* Richard C. Worf, *The Case for Rational Basis Review of General Suspicionless Searches and Seizures*, 23 TOURO L. REV. 93 (2007).

128. I thank Chris Slobogin for articulating this point for me.

129. Note that this theory examines deference to legislation, where all the political voices and forces would be played out, as opposed to mere actions of the executive branch. Slobogin, *supra* note 31, at 20. Another prerequisite for this political process analysis is that the intrusion discussed is not so extensive as to harm individual rights and thus call for protection; and indeed one of the underlying assumptions of the article's overall thesis is that this requirement is met. *See generally* Christopher Slobogin, *Government Dragnets*, 73 LAW & CONTEMP. PROBS. 107 (2010).

130. For instance, in cases where policy was introduced which affected any person who entered an international airport, travelled on a ferry or crossed a border. *Id.* at 133.

131. *Id.* I again thank Christopher Slobogin for this observation.

The exception to this rule is when there is an apparent "defect in the democratic processes."[132] This would be the case when the relevant law discriminates against weaker segments of society, such as insular minorities or other groups that are not adequately represented. Thus, there is a fear that the majority will select a rule that will overburden the minorities while abusing the majority's political power. In these cases, the democratic system fails and it is the duty of the courts to protect these groups from the majority's tyranny.

In accordance with this line of thought, it is important to analyze the democratic process that might lead to selecting any one of the alternatives mentioned. If a specific alternative might be selected in view of a biased and defective discourse and process, it must receive closer scrutiny by courts. The first two alternatives can indeed easily meet the criterion calling for judicial deference. The alternatives calling for equal burden for all and even random selection will affect the entire public as a group with no additional burden to a weaker party or an insulated minority.

Further examination of the political process that might lead to adopting data mining and its two major alternatives provides interesting insights. Data mining has generated animosity and concern by the general public.[133] Such fear might indicate, however, that if data mining is ultimately selected, it will be a balanced solution beyond judicial review. While it might point to a specific group and require additional scrutiny, the process would not be arbitrary and the selected group would not be insular. The fact that anybody could be selected might indeed indicate that judicial deference would be prudent if the process is ultimately selected by the legislature.

This last argument follows from the assumption that the political process implementing data mining is not affected by failures which generate enhanced judicial scrutiny. With data mining, the chance of being subjected to additional scrutiny appears to be equally spread across the entire public. This assumption takes into account the existence of errors in the process but assumes that those errors are non-systematic (or in other words, may pertain to any individual with equal risk). A "veil of ignorance" separates citizens from knowing who might be adversely impacted by the data mining schemes.[134] Data mining will try and point

132. *See* S*lobogin*, *supra* note 31, at 19.

133. *See supra* Part II.

134. *See* JOHN RAWLS, A THEORY OF JUSTICE 118 (rev. ed. 1971)("[B]ehind a veil of ignorance . . . [citizens] do not know how the various alternatives will affect their own particular case and they are obliged to evaluate principles solely on the basis of general considerations."). In other contexts, examples exist regarding famous and prominent people being singled out, such as the late senator Kennedy, who was marked as a risk

out individuals and groups that are higher risks, yet there is no *a priori* assumption that these results will end up referring to minorities or weaker groups. In the context of this argument, there is no substantial difference between using anonymized personal information (alternative (5)) for data mining tasks and "open" data, as the political dynamics should be quite similar.[135]

Can the same be said for alternatives (3) and (4)? Do these processes lead to pointing towards non-insular groups in a process that is nonbiased? The previous analysis of the intricacies of these processes casts serious doubt on this proposition. As already mentioned,[136] these alternatives will have a tendency to generate biased outcomes given the prejudice of either field officials or experts. Therefore, there is a greater chance that the results of these processes will lead to the focusing of governmental attention on a specific and predictable segment of the population – one that constitutes at times a protected minority. It is also fair to assume that many individuals understand that such an outcome will follow and thus advocate policies along the lines of alternatives (3) and (4) as an attempt to shift the burden of additional scrutiny away from the powerful majority. Therefore, if these alternatives are implemented, courts should still scrutinize the laws and the structures that are set in place. These structures might be inherently unfair towards the minority who cannot assert its position in the political discourse.[137]

Many will argue that this distinction between data mining and alternatives (3) and (4) is misplaced. They will argue that data mining initiatives are tilted and tainted as well. Data mining analyses, these critics will assert, do not generate a risk equal to all, but rather focus on specific weak subgroups. This will follow from either biases in the automated process, problems with the learning datasets (which would be biased in view of historical dynamics of ill-treatment), or the human interactions the process involves.[138]

I believe the previous discussion as to the actual data mining process leads to setting such criticism aside.[139] Data mining might

prior to boarding a flight. *See, e.g.*, Sara Kehaulani Goo, *Sen. Kennedy Flagged by No-Fly List*, WASH. POST, Aug. 20, 2004, at A01.

135. Again, with the fifth alternative, the margin of error might be higher, but if these errors are reasonable, systematic, and unbiased, it should lead to a similar outcome.

136. *See supra* Part III.2.

137. For a discussion of a different opinion on this point, see *supra* note 62.

138. This is an argument made by Solove, who calls for limiting the role of data mining in this context. *See* SOLOVE, NOTHING TO HIDE, *supra* note 7, at 190-91.

139. Another critique might note that data mining initiatives still require judicial scrutiny because the individuals adversely affected by this process are unable to exercise political power and counter the unfair treatment they are facing. Therefore, the entire political process thesis cannot pertain to this situation, which focuses on individuals, rather than groups. Here, the group is dispersed and suffers from "collective action"

include hidden biases, but the process is data driven, automated and centralized. All of these characteristics provide some insulation from the biases mentioned. If data mining analysis is interpretable and audited to assure that bureaucrats are not successfully manipulating the process, this might be the best shot at an unbiased analysis which relies on datasets[140] rather than prejudice. Thus, judicial deference might indeed be called for at this specific juncture.

(2) Political Process and the Majority Data Mining Aversion

Our analysis thus far has concluded that if data mining is accepted by the legislature, it might only require limited judicial review. This is as opposed to the use of profiles and field officer discretion, which calls for greater scrutiny. Here, the comparative analysis does not necessarily steer us towards one option or the other, but rather elucidates the hidden traits of every alternative and the role of courts in every context. Introducing the analysis of the political process into our discussion of selecting among alternatives, however, might lead to even more provocative insights and assertions. It might explain why the majority is averse to data mining even when strong arguments could be made for its implementation.

The methodological tool of examining alternatives can be applied to try and predict the public's response to various policy proposals. Note, of course, that the "public" is not homogenous, but a heterogeneous mass which includes multiple sub-groups and minorities. For this segment of the analysis, I assume that the individuals in general would be interested in limiting their costs and burdens while shifting those burdens to a smaller, specific group which has no political power (and thus no ability

problems (as opposed to being an insular and inherently weakened group). For more on this critique, see Slobogin, *supra* note 129, at 139.

This critique may have merit and calls for additional analytical discussion and empirical testing. When approaching this question, one must distinguish between systematic and non-systematic errors. If the data mining process will lead to erroneously singling out individuals, but in a non-systematic manner, which does not focus on a specific sub-group, and continuously shifts its focus to different individuals based on the changes in risks and models, deference might still be the best policy by courts (beyond generally looking at the reasonability of the project). The "group" affected by such policy would be the broader population. However, if there is a systematic error which constantly refers to specific individuals or sub-groups, the political process model will not provide sufficient protection to this diffused group, and court intervention is indeed justified.

140. While the training data might be biased, as mentioned above, techniques could be developed (and are being developed) to overcome this problem and try to produce neutral results.

to shift these costs and burdens onward).[141] With this perspective in mind, a reason for the aversion towards data mining surfaces, and it is quite problematic. As opposed to the other feasible alternatives, data mining equally spreads the risk of error among individuals. This is not an outcome a strong majority will welcome. Such a majority would rather opt for alternatives (3) and (4) that allow for shifting the burden to a politically weak segment. This might be the reason why data mining initiatives are constantly condemned in the political sphere. Indeed, the quick political moves taken to shut down the Total Information Awareness (TIA) initiative[142] demonstrate that politicians are attuned to the public's thoughts (or perhaps merely the majority of it), even though its motivations might be problematic.

Table II below sets out the analytical process members of the powerful majority work through, which ends with the rejection of data mining solutions. The majority will reject alternatives (1) and (2). Its overall discontent with these policies is understandable. Any shift from the first alternative would be welcomed, at least by those whose burden is eased by a shift from segmenting the entire population. The second alternative (which uses randomization) will probably lead to similar results in terms of the majority's discontent. Here, the entire public is still subjected to the *risks* of additional burdens (even though those manifest only part of the time).

When forced to choose between alternatives (3), (4), and data mining (or alternative (5)), the majority will reject data mining. The majority will seek a solution that would allow it to pass the burden on to the weaker minority as much as possible. By advocating alternatives that provide for human discretion ((3) and (4)), members of the majority will be effectively taking steps to insulate themselves from the disturbance and discomfort of the selection process. They will do so as they acknowledge that options premised upon human discretion will lead to selecting weaker minority groups and sparing the majority.

Data mining processes, therefore, will not be the majority's favorite. These processes rely on facts and computerized analyses rather than discretion, and they cannot provide the majority with assurances that weaker minorities will be the focus of future inquiry. The majority's members will be equally exposed. The majority's rejection of data mining might not be a favorable outcome to minorities, however. In a

141. For more on the interplay between the political economy, data mining, and courts, *see generally* Slobogin, *supra* note 129. Of course there is an assumption that the public only cares of its own interests, and not that of minorities within society that might be unfairly treated. This of course is a difficult assertion, which could be easily challenged and requires an additional study.

142. *See generally* Cate, *supra* note 3;SOLOVE, NOTHING TO HIDE, *supra* note 7.

way, rejecting data mining by the majority might even constitute an abuse of power.

Alternative	General Social Impact	Explanation
(1) No Selection	General discontent	Everyone similarly impacted
(2) Random Selection	Same as (1)	Equal chance to be impacted
(3) Field Office Discretion	Greater chance of focused harm Weaker general discontent	Result of discretion, hidden bias, and previous experiences
(4) Expert Profile	Same as (3)	Same as (3)
Data Mining & (5) Anonymized Data Mining	Same as (2)?	If analysis is neutral, and errors equally spread, same as (2)

Table II

Examining this political process introduces at least two interesting insights. First, data mining, not the other, more popular solutions, should in many cases receive the greatest deference from courts examining the nature of the legislative structure. Second, the political process might lead to the abandonment of data mining initiatives for the wrong reasons. Therefore, policymakers should be cautious when rejecting data mining solutions in view of popular and political pressure.

Understanding these specific insights of the political process illuminates an important role for courts and academics; these influential players must assure that even when specific forces strive to shift burdens to weaker groups using various strategies, society must counter these forces and assure that the policy it applies is fair to all. This solution, in some cases and with adequate safeguards in place, may involve the use of data mining.

CONCLUSION: GOVERNMENTAL DATA MINING AND ALTERNATIVES: FROM A LEGAL QUESTION TO A POLITICAL BATTLE

In this article, I strived to draw out a methodology that will assist policymakers searching for a balance in today's world of global insecurity. These policymakers are now challenged with the structuring of schemes striving to use databases of personal information to promote law enforcement and stability. They are also struggling with methods to engage in law enforcement in a way that is both selective and fair. The

methodology here introduced calls for the close examination of alternative routes and the way these routes will impact the concerns voiced in the data mining debate. Comparing among alternatives will provide for a better sense of the balances and realistic compromises required at every juncture when applying selection regimes. Therefore, comparing among all options must be carried while taking into account all legal concerns. Yet it must not stop there. The comparison must account for political trends as well, while trying to understand the powerful social forces this discussion seems to awaken.

Existing risks call for the use of personal information in an effort to preempt possible harms and attacks. Society is forced to decide among several non-ideal options. At the end of the day, the solution to be applied will no doubt be a compromise. The methodological steps presented in this article strive to assist in the process of establishing such a compromise, while acknowledging that there is still a great deal of work to be done. I hope this article's small contribution promotes this broader objective.