# Chapter 18
# Data Mining as Search: Theoretical Insights and Policy Responses

Tal Zarsky

**Abstract.** Data mining has captured the imagination as a tool which could potentially close the intelligence gap constantly deepening between governments and their new targets – terrorists and sophisticated criminals. It should therefore come as no surprise that data mining initiatives are popping up throughout the regulatory framework. The visceral feeling of many in response to the growing use of governmental data mining of personal data is that such practices are extremely problematic. Yet, framing the notions behind the visceral response in the form of legal theory is a difficult task.

This chapter strives to advance the theoretical discussion regarding the proper understanding of the problems data mining practices generate. It does so within the confines of privacy law and interests, which many sense are utterly compromised by the governmental data mining practices. Within this broader theoretical realm, the chapter focuses on examining the relevance of a related legal paradigm in privacy law – that of governmental searches. Data mining, the chapter explains, compromises some of same interests compromised by illegal governmental searches. Yet it does so in a unique and novel way. This chapter introduces three analytical paths for extending the well accepted notion of illegal searches into this novel setting. It also points to the important intricacies every path entails and the difficulties of applying the notion of search to this novel setting. Finally, the chapter briefly explains the policy implications of every theory. Indeed, the manner in which data mining practices are conceptualized directly effects the possible solutions which might be set in place to limit related concerns.

## 18.1 Introduction: Beyond the Visceral Response to 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

Tal Zarsky
Faculty of Law, University of Haifa, Israel
e-mail: tzarsky@law.haifa.ac.il

B. Custers et al. (Eds.): Discrimination & Privacy in the Information Society, SAPERE 3, pp. 325–338.
springerlink.com 

fear of terrorist attacks. Various technological tools are 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 other known operatives (Taipale, 2004) has led many to ask: What if data mining had been applied in advance? Could the attacks and their devastating outcomes been avoided?

Data mining has captured the imagination as a tool which could potentially close the intelligence gap constantly deepening between governments and their new targets – terrorists and sophisticated criminals. Data mining is also generating interest in other governmental contexts, such as law enforcement and policing. In recent years, law enforcement worldwide has shifted to "Intelligence Led Policing" (ILP) (Cate, 2008). 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 data mining could enhance. It should therefore come as no surprise that, at least in the United States, data mining initiatives are popping up throughout the regulatory framework (GAO, 2004).

The visceral feeling of many is that the outcome of data mining analyses, which enable the government to differentiate among individuals and groups in novel ways, is extremely problematic. Yet framing the notions behind this strong visceral response in the form of legal theory is a difficult task. Even though governmental data mining is extensively discussed in recent literature, an overall sense of confusion is ever present. Additional thought is still required to properly articulate the concerns these practices generate, and the context in which they arise. While mapping out these issues, scholars as well as policymakers must further establish which paradigms of legal thought are suitable for addressing these matters. Central potential paradigms are constitutional law, privacy law and anti-discrimination, yet other fields will surely prove relevant.

This chapter strives to advance the theoretical discussion regarding the understanding of the problems data mining practices generate. It does so within the confines of privacy law and interests, which many sense are utterly compromised by the governmental data mining practices. Within this broader theoretical realm, the chapter focuses on examining the relevance of a related legal paradigm in privacy law – that of governmental searches. Examining whether an intrusive act is a legal or illegal search is a common analytical query invoked when approaching various governmental actions which might compromise privacy interests. It is analytically helpful – this chapter will explain – to conceptualize the privacy harms data mining might cause by using paradigms of thought arising in "search" related analyses. To some extent and from some perspectives, data mining compromises the same interests affected by illegal governmental searches. Yet it does so in a unique and novel way. This uniqueness renders the discussion of data mining and its detriments difficult and complex. This chapter introduces three analytical paths for extending the well accepted notion of illegal searches

into this novel setting. It also points to the important intricacies every path entails and the difficulties of applying the notion of search to this novel setting.

Addressing this interesting comparison need not be a mere theoretical exercise. The theoretical concepts drawn out here will prove important in the future. Regulators will surely strive to move from theory to practice, approach data mining initiatives and establish which practices are to be allowed, and which must be prohibited. Therefore, this chapter would be of interest not only to readers interested in legal theory. It might also prove helpful to regulators and ractitioners seeking ways to ground the novel data mining practices in existing legal concepts.

Before proceeding, several analytical foundations must be set in place. Therefore, in section 18.2, the chapter briefly demonstrates and explains the meaning of data mining initiatives and what they might entail. 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. Data mining also presents specific unique traits, and sets distinct roles for humans and machines. Section 18.3 sets forth the central thesis of this chapter. It first explains why the chapter chose to import theoretical insights from "search" related interests in privacy law. It also explains why specific theories of search were selected for this discussion. It thereafter moves on to map out three ways in which the somewhat abstract notion of "search" could be conceptualized, and applies these notions to the data mining context. In doing so, the analysis addresses specific points where applying the relevant theory to the data mining context might face theoretical and practical obstacles, and discusses ways to overcome them. The chapter concludes in section 18.4, where it briefly explains the policy implications of applying every theory, both in terms of direct and ancillary policy measures which might be called for to minimize privacy related concerns. In these last two sections, the chapter demonstrates the importance of the theoretical analysis presented; indeed, the manner in which data mining practices are conceptualized directly effects the possible solutions which might be set in place to limit related concerns.

The chapter specifically focuses on the data mining practices of government, while purposefully neglecting similar initiatives carried out by commercial entities. This is not to say that the latter practices do not raise privacy concerns in general, and those related to the concepts of unacceptable searches in particular. Indeed, marketers, advertisers and insurers are all crunching away on the vast datasets of personal information at their disposal. In doing so, they open the door to a flurry of policy and legal problems regarding the permitted scope of using personal data and (among others) the form of consent data subjects must provide prior to such uses. This chapter, however, sets these issues aside for now. While the commercial-related issues are severe, governmental data mining leads to concerns of a far greater magnitude. The government has great datasets of personal information at its disposal and almost endless resources and opportunities to obtain many more. It can collect such information without the data subjects' consent (and in many cases without their knowledge). Perhaps most crucially, it can potentially use such information to impact the property, liberty and even life of the data subjects, given the government's almost limitless powers. For these

reasons (and others) I choose to focus on governmental data mining and leave a discussion of the actions of the commercial entities for a later day.

In addition, at this point it is useful to point out what this chapter will *not* discuss (even within the realm of governmental data mining) given the chapter's focus on privacy. The analysis here presented will be 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 (Harper & Jonas, 2006; Schneier, 2006) and notoriously difficult to measure and prove. Those opposing data mining can make a strong case that these predictive automated processes are, in general, inherently flawed and ineffective. In addition, they might argue they are particularly unfair to the individuals they implicate. This position has merit, and is no doubt true in specific contexts. The critiques presented below, however, will be premised upon the contrary assumption (which I believe is true in a variety of other settings), that data mining is effective and operational. Yet even so, such forms of analyses might prove problematic as they clashes with other important interests. In addition, data mining generates concerns related to the lack of transparency this practice entails, as well as discrimination it could generate. These too are important aspects which are addressed elsewhere within this volume (Chapter 17 and 19).

## 18.2 Governmental Data Mining: Definitions, Participants and Problems

The term "data mining" has recently been used in several contexts by policymakers and legal scholars. For this discussion, 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*" (Fayyad et. al, 1996). Within this broader topic, the core of this chapter focuses on data mining which enables "pattern based searches" (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 apply 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 which they are seeking and that 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 the relevant datasets, or ways in which the data could be effectively sorted (Zarsky, 2002-2003). The data mining process could achieve both descriptive and predictive tasks. In a predictive process (on which this

chapter is focused), 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. In doing so, software searches for the patterns and rules it previously established and encountered. 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 and environment as well).

A notion usually mentioned when considering data mining analyses is the level of automation this process facilitates. Data mining analyses indeed provide a higher level of automation than that available with other governmental alternatives; the predictive process somewhat limits the extent of human discretion in the process. Yet the level of automation this process entails might be easily overestimated. Analysts play important, yet at times hidden roles in the online process. Their actions (such as those mentioned in the previous paragraph) directly impact the outcome of the process and therefore affect actual governmental policy.

## 18.3 Governmental Data Mining and/as (Illegal) Searches?

### *18.3.1 Finding a Theory*

A governmental 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 any later stage. If the information was collected by the government, citizens might not have even provided consent at the point of collection. Rather, they merely received a basic and vague notice of the collection and future uses provided by the government.

Engaging in personal data analysis without the direct consent of relevant data subjects contradicts to several "privacy" related legal concepts. However, the precise meaning of privacy is elusive, and the privacy concerns arising in this context could be articulated in a variety of ways. In this chapter, I choose the salient paradigm of "searches" to try and illustrate the nature of privacy concerns data mining analyses generate. Of course, other paradigms of privacy might pertain to the data mining context as well. Yet this chapter focuses on a relatively specific privacy notion, which on its face is relevant and can prove insightful.

Applying the search paradigm to this context would imply that given various traits of the data mining process, this form of analysis should not be considered reasonable. Applying "search" related arguments to the data mining context has several implications. On the theoretical level, such a linkage will allow for "importing" well developed concepts of law into a novel context where they can potentially enrich a confused discourse. However, such linkage can have far reaching practical ramifications. In many cases, for a legal search to commence,

various forms of ex ante judicial approval and supervision are required. If data mining will be considered as a search, data mining analyses would be considered an illegal search when carried out without sufficient judicial approval – approval which is not currently sought.

The link between data mining practices and the concept of search can be made on several levels – only one of which would be examined in this chapter. It could be carried out on an intuitive level. It could also be carried out on a doctrinal level. Finally, it could be carried out on a theoretical level. This chapter merely focuses on the latter aspect. Yet before doing so, I hereby provide a few explanations about the former two realms, and explain why I chose to set them aside for now.

On an intuitive level, data mining seems to invoke the notion of "searching" and perhaps therefore, the legal implications of such terminology. The data mining process calls for the substantial analyses of personal information pertaining to specific individuals. 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. This is a process which will certainly be referred to as "searches" in laymen's terms (Slobogin, 2007). Yet intuition is a fickle prospect. In many instances it could be plainly wrong, as the public might be ill-informed regarding the true meaning and implications of data mining – including its vast benefits. For that reason, I set this discussion aside. Indeed, not all activities which are "searches" to laymen are or should be considered as searches in the eyes of the law.

Linking data mining and searches will have real world implications and therefore opens the door to an elaborate doctrinal analysis. When the law recognizes searches as such, it moves to regulate them, limits their scope, and sets systematic boundaries to assure the protection of rights. It is however unclear whether under current case law and the existing concept of "search" as articulated by the courts, data mining analyses constitute searches. In the US, for instance, these steps are commonly discussed in the Fourth Amendment context, which protects the people from unreasonable searches (Kerr, 2007). Whether current Fourth Amendment doctrine will find data mining to be a "search" is a difficult doctrinal question, which is beyond the scope of this chapter, but will probably be answered negatively (Cate, 2008). Therefore, the starting point for this discussion is that data mining analyses are not "searches." 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 either third parties and later passed on to the government or directly (yet lawfully) by the government itself. With this assumption in place, American law regarding searches generally assumes that individuals have a very limited subsequent privacy interest (at least in terms of "searching" and the Fourth Amendment) given the initial lawful collection of data (Kerr, 2010). The point of data collection is where data subjects relinquish control over the data and its future uses. To summarize, the governmental data mining initiatives usually do not amount to breaches of constitutional rights; or, as Daniel Solove succinctly states, "Data mining often falls between the crevices of

constitutional doctrine" (Solove, 2008). At least in the US, these initiatives are also probably permitted according to current privacy laws in view of various exceptions and loopholes (Cate, 2008).

As mentioned above, this chapter sets aside the doctrinal analysis and examines the issue at hand from a theoretical and normative perspective. Doing so allows for quickly working through the relevant issues, and leaving room for an in-depth discussion of various perspectives. Yet this discussion might not remain entirely theoretical for long. It should be noted that the doctrinal outcome mentioned is not set in stone. Data mining 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 abovementioned assumptions regarding privacy expectation in lawfully-collected datasets. Thus, courts might choose to change the existing doctrine in view of new theoretical understandings (which I strive to promote here), changes in public opinion, or other factors.

### *18.3.2 Data Mining as "Searches": Introducing Three Perspectives*

On a theoretical level, linking data mining concerns to search-related interests in the privacy context can be an illuminating exercise. This is because some of the underlying theories for articulating the interests compromised by illegal searches directly address the elusive privacy interests compromised by data mining initiatives. These nexuses between search interests and data mining practices are indeed the premise of this entire chapter. However, linking data mining and the notion of illegal searches in privacy law must be done with caution. This is due to the lack of consensus among scholars regarding the definition of illegal searches and the rationale behind their prohibition.

This chapter sets forth three normative theories, which are especially helpful in understanding concerns related to governmental data mining. These theories are drawn from the existing literature and case law examining searches in the technological age in general, and in the context of data mining in particular. With these theories in mind, it is easy to see how privacy concerns in the context of data mining could be articulated using the terminology and concepts of illegal searches.

As presented below, not all of these theories are of equal strength. Some (the first) are weaker than others in explaining the privacy concerns arising in this context. Every theory however addresses a different aspect of the harms of privacy. The first focuses on the individuals and their state of mind while the second on the government and its unchecked powerful force. The third theory presents somewhat of a combination of both elements, and calls for limiting the government's ability to engage in "fishing expeditions." I now move to present these theories, how they might apply to the data mining context and what analytical obstacles might arise when doing so.

***Searches as Psychological Intrusions***

The *first* theory for distinguishing between legal and illegal searches looks to their *intrusive* nature. In other words, the government should meet a higher threshold of scrutiny if its actions are understood to be intrusive.[1] While intrusion is usually understood to be one that is physical, it has a psychological aspect as well (it should be noted, that this theory was *not* yet accepted in US courts).[2] The notion of psychological intrusion can be easily identified when government searches the home and self of citizens. Yet it need not be limited to these instances. Intrusiveness of various forms is the mirror image of key privacy interests, such as the right to solitude and "to be left alone."

Examining whether mere psychological (as opposed to physical) intrusions are afoot can lead this normative theory to the data mining context. It is fair to assume that many will feel intruded when confronted with the existence of data mining practices carried out with regard to their personal data. This aggravated sense of intrusion (as opposed to any other form of review of personal information on file with the authorities) could be derived from two key unique elements of the data mining process (which distinguish it from other governmental practices). First, the process's *automated* nature might generate additional anxiety. Second, data mining's ability to *predict* future behaviors could cause greater 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. Yet now they are in the hands of the government. Empirical data gathered regarding the public attitude towards searches upholds this theory, while showing indications of anxiety towards these novel and (assumedly) "intrusive" practices (Slobogin, 2007).

The "psychological intrusion" theory provides an interesting perspective for examining the extent of privacy concerns arising from governmental data mining analyses. However, when rigorously applying this theory to the data mining context, it does not provide a conclusive response as to the intrusiveness of these governmental practices. This should come as no surprise, as psychological

[1] In the US, the test for the legality of searches is one which is premised upon the "reasonable expectation of privacy." Such expectation has two elements – subjective and objective/normative. Clearly, this discussion pertains to the subjective element – and a search might indeed be found to be subjectively unreasonable if considered intrusive – even merely on a psychological level. Indeed, wiretapping which does not involve a physical intrusion is considered unreasonable as well. However, the test includes an important objective/normative layer. Here, justices decide which form of subjectively unreasonable conduct is objectively unacceptable as well. As mentioned in the text, the courts have yet to find that psychological intrusions in the form of governmental searches throughout legally obtained data are unreasonable. For more on the theoretical analysis of the Fourth Amendment, 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 his first model – the *Probabilistic* Mode – a descriptive model which is premised about expectations based on current social norms. *Id.* at 508-13.

[2] 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).

intrusion is a complicated notion. Some individuals might be greatly troubled by the automated nature of the data mining process, and the lack of human decision-making and discretion. Yet others might have a very different set of preferences when it comes to governmental analyses of personal data. To properly assess the notion of psychological intrusion at this specific juncture, one must remember the alternative strategy to governmental data mining. This would call for broader roles for experts and field officers in the law enforcement decision making process; in such a non-automated process, actual humans are those sifting and considering the individual's files. For some individuals, data mining generates greater anxiety than this latter options given concerns with automated and computerized decision-making processes. For others, however, the opposite would be true.[3] These persons would not be alarmed by the faceless computer searching their data (Tokson, 2011). They would, however, be gravely concerned with actual individuals looking through their information.

A similar complication will follow when considering the psychological intrusion resulting from fears of powerful revelations made by a computer algorithm. While this is the perspective of some, others might have greater fears of the other practices government might apply if data mining is set aside. When relying on experts and field officers, the process might be ridden with errors and biases which result from the cognitive limitations and opinions of humans (Zarsky, 2012). These are concerns that the computer analysis could avoid with greater success.

The last few paragraphs set out arguments which explain that data mining processes might generate a sense of psychological intrusion for some, yet might be comforting to others. The latter are individuals whom believe that this process is preferable to its inevitable alternatives. Both arguments and points of view seem acceptable, even reasonable. The differences of opinion people will have regarding the intrusiveness of data mining will result from differences in their understanding of the data mining technology, its benefits, and its detriments. A possible measure to overcome the analytical obstacle this theory faces might be through conducting surveys to establish the public's position. Yet administering such surveys would be a very difficult, perhaps near-impossible task (Solove, 2010).

To conclude, the "psychological intrusion" perspective to the law of searches can easily be applied to the context of governmental data mining. It can easily explain why, for some, the governmental actions breach privacy rights. However, this perspective – if ever accepted and applied by law – will face problems when moving from theory to practice. Establishing whether data mining is indeed intrusive will depend on a variety of unpredictable factors. Thus, this theory will probably fail to provide clear-cut policy.

### *Limiting Searches to Limit the Force of Government*

A *second* theory distinguishing between legal and illegal searches which can illuminate the privacy-in-data mining debate looks to the normative reasons (as

[3] For instance, see Goldman, *Data Mining and Attention Consumption,* 225, 228, as discussed by SOLOVE, NOTHING TO HIDE, *supra* note 7, at 183.

opposed to visceral feelings) for limiting governmental power. This theory notes that searches are found to be illegal when they are a *powerful tool government should not be entrusted with* (at least without various forms of judicial supervision). Again, this rationale applies naturally to searches of the home and self, as well as wiretapping of communications.

When considering the use of data mining for automated predictive modeling, one can easily argue that government should not be entrusted with such a powerful tool without being closely scrutinized. Data mining can potentially turn even seemingly benign factors into a powerful mapping of an individual's persona and insights. For that reason, ex-ante judicial (or other forms of scrutiny) must be applied.

The challenge of applying this theory to the data mining context and finding that a privacy interest was compromised is that the analysis here discussed uses information which was collected lawfully by government. Therefore, the power of government was already examined and limited when information was collected. Accepting that a search-related interest might have been compromised in the data mining context calls for accepting a non-trivial argument: at times the knowledge provided by the analysis of the sum of the dataset goes beyond the value of the parts of the dataset previously collected, when viewed on their own. If this is indeed true, then the fact that the governmental actions were reviewed by courts at the data collection stage is insufficient. Additional scrutiny is required at the data mining "search" stage. Given the enhanced ability of data mining tools to engage in broad, automated and predictive tasks, this argument seems quite convincing. Data mining transforms small segments of information into an overall "mosaic" of human behavior.

The provocative notion that many, seemingly innocuous, bits of information which were collected lawfully should be treated differently in the aggregate is slowly gaining recognition in US courts which examine the notion of "search" (although it has yet to be accepted into Fourth Amendment doctrine). Most famously, this issue is fiercely debated in the context of location-based data (which is currently easily collected by mobile phone operators and other GPS devices), while questioning whether there is a difference between collecting limited and vast amounts of such data. For instance, in a controversial opinion, the Federal Court of the D.C Circuit chose to restrict governmental collection of location-based data over an extensive time period while promoting the "Mosaic Theory."[4] This finding contradicted previous cases which found that individuals have no privacy in GPS information which pertained to their actions in the open.

[4] *U.S. vs. 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).

The "Mosaic Theory" argues that small bits of innocuous information, when brought together, can provide a full mosaic of an individual's persona. Therefore such practices of aggregation should be further scrutinized. It should be noted, that very recently the US Supreme Court addressed this case on appeal (United States v. Jones, 2012). It unanimously found the governmental search to be unconstitutional, yet the majority relied on other grounds and left the acceptance of the "mosaic theory" into the law for another day.

To conclude, this search-related theory of privacy can explain why data mining must be limited, and when this must be done: in instances in which the tools used by government prove extremely effective! The theory here presented is premised on an interesting insight; data mining's analytical strength is the key to its normative disadvantage. The public has learned to live and accept decision making processes involving experts and field officers with their limited abilities. These existing alternatives strike an acceptable balance between law enforcement needs and civil liberty interests, even though they might compromise overall effectiveness. Data mining presents a challenge which law must now answer to, and a force which the law might find to be excessive if not properly checked. However, this theory has clear limits – if the data mining process is not found to be more powerful and insightful than other acceptable practices, this argument loses its analytical force.

### *Limiting Searches to Limit "Fishing Expeditions"*

A *third theory* which can prove helpful in articulating privacy-related concerns from the "search" perspective in the context of data mining analyses pertains to their very broad scope. Usually, when considering invasive searches, laws and courts find that they must be carried out narrowly, while limiting the gaze of government as much as possible. Searches which fail to do so 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. Curbing "fishing expedition" by governments is one of the central roles of judicial review (Solove, 2002). Thus, this theory finds a normative flaw with very broad searches, which impact to non-suspects.

Data mining initiatives famously call for actively examining and analyzing datasets pertaining to a very broad realm of individuals, including those whom are substantially removed from the matter at hand. The software does so while striving to formulate patterns, trends, and clusters. Thus, data mining generates a massive "fishing expedition" which resembles the most feared practices of government – searching datasets in mass, while hoping to locate relevant evidence (as opposed to initiating a search based on suspicion). On its face, this paradigm of thought might be extremely helpful in grasping the concerns data mining generates.

Yet again a theoretical obstacle blocks the application of this perspective in the data mining context. If, under existing doctrine (and as explained above) the government may review and analyze information which was lawfully collected in

any way it deems fit, data mining cannot be considered a "fishing expedition."[5] In other words, no "search-related" interest is compromised by the analysis (or, to carry through the metaphor, there is no "expedition" in these actions), as the government is clearly operating within its mandate, rather than intruding on the rights of the innocent. Therefore, this perspective does not prove helpful in mapping the boundaries of legitimate and excessive data mining practices.

This theoretical obstacle follows from today's understanding 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 (in which case no constitutional form of protection is called for). The harsh implications (for government) of actions being considered as "searches" have led courts to limit the breadth of this term. Yet this dichotomous perspective of the concept of searches is not set in stone and the problem not beyond repair. Recent scholarship argues that rather than a dichotomy, searches should be viewed on a sliding scale. In other words, the legitimacy of search should be established through a proportionality-based analysis (Slobogin, 2007); different forms of intrusions will be met by different forms of legal thresholds to protect search-related interests. Every such intrusion will call for a proportionate level of protection and standard of review.

The proposed shift to a proportionality based analysis of search interests will force policy makers to address the "fishing expedition" problem data mining practices set forth. Data mining analysis could be considered as a minute intrusion on its own (rather than a process which is not a "search" at all), when examing the impact on a single citizen. Yet when considering the aggregated impact on a broad segment of the population subjected to the data mining analysis, the result might be quite different. Indeed, in cases where the benefits of the data mining analysis are limited or unsure, and the population segment extremely broad, such practices might be found to be a disproportionate measure (Slobogin, 2007). Therefore, this specific theoretical perspective of "searches" can provide a different form of "calculus" for configuring whether a governmental data mining is acceptable – and a balance which is quite different than the one called for under the previously mentioned theories.

## 18.4 Conclusion: Novel Practices, Classic Concepts and Policy Proposals

This chapter draws out a basic conceptual framework for "importing" theoretical concepts used in the "search" discourse to properly understand concerns associated with governmental data mining practices. Yet the discussion need not stay in the realm of theory. This conceptual framework can also 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. In

[5] Note that most recently, in *US v. Jones* (1.2012), the count's concurring opinion questioned the wisdom of the third-party doctrine.

doing so, they must figure out ways in which "search" related interests could be answered within the governmental data mining analysis process.

As explained above, every such theory calls for a different set of balances and findings. However, the implications of these theories – if they are indeed accepted in the data mining contexts – run deeper. Every one of the theories mentioned above might point regulators in a different regulatory direction when considering ancillary privacy rights to overcome the concerns at hand. The *first* theory points to the sense of intrusion data mining generates. If this is the privacy-based theory which generates concerns in the data mining context, then this concern could be partially mitigated by a greater degree of transparency in the data mining process. With additional knowledge as to the process, the public aversion might be limited. Therefore, accepting this theory should promote this ancillary right.

The *second* theory points in a different direction. Addressing this concern should call for various measure for assuring that data mining analysis are only used for the specific tasks they are needed for the most. In other words, steps must be taken to assure that the use of these methods does not "creep" into other realms. This could be achieved by both technological measures (which safeguard the use of these tools) and a regulatory structure which closely supervises these uses. Again, the theoretical perspective can point policymakers (if convinced by this argument in the relevant context) in the direction of relevant (yet different) ancillary rights.

The *third* theory might call for yet a different regulatory trajectory in terms of regulatory steps. The concern it addresses relates to the very nature of the data mining practice – one that finds its broad scope problematic. Therefore, this theory might indeed lead to limiting data mining analysis. Another possible option might call for engaging in the mining of anonymized data – a practice which might somewhat mitigate these concerns (yet raises others) (Zarsky, 2012). Arguably, if the search is of anonymized data, the interests of the many subjected to it are not compromised by the vast net the data mining practice apply. Therefore the use of this measure would be found proportionate.

As there is probably a kernel of truth in every one of the theories, it would be wise to take all these proposals under consideration. However, at some points they might prove contradicting. Therefore, additional analytical work must prioritize among them, while relying on social norms and the balancing of other rights. This analysis of course must be context-specific, as in different contexts the relative force of every theory will vary.

In conclusion, existing risks call for analyzing and using personal information in an effort to preempt possible harms and attacks. Society will be forced to decide among several non-ideal options. At the end of the day, the solution selected will no doubt be a compromise, taking into account some of the elements here set forth. The theoretical analysis here introduced strives to assist in the process of establishing such a compromise, while acknowledging that there is still a great deal of work to be done.

## References

Cate, F.H.: Government, data mining: The need for a legal framework. Harvard Civil Rights-Civil Liberties Law Review 43(2), 435–489 (2008)

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Cambridge, Mass. (1996)

Henderson, S.: Nothing new under the sun?: A technologically rational doctrine of fourth amendment search. Mercer Law Review 56, 507–563, 544 (2005)

Jonas, J., Harper, J.: Effective counterterrorism and the limited role of predictive data mining. CATO Institute Policy Analysis 584, 1–12 (2006)

Kerr, O.: Four models of fourth amendment protection. Stanford Law Review 60, 503 (2007)

Kerr, O.: Applying the fourth amendment to the internet: a general approach. Stanford Law Review 62, 1005 (2010)

Schneier, B.: Why data mining won't stop terror. Wired (September 3, 2006), http://www.wired.com/politics/security/commentary/securitymatters/2006/03/70357?currentPage=all (accessed December 30, 2011)

Slobogin, C., Schumacher, J.E.: Reasonable expectations of privacy and autonomy in fourth amendment cases: An empirical look at "understandings recognized and permitted by society." Duke Law Journal 42(4), 727–775, 743–751 (1993)

Slobogin, C.: Privacy at risk: The new government surveillance and the fourth amendment, p. 194. The University of Chicago Press, Chicago and London (2007)

Solove, D.J.: Digital dossiers and the dissipation of fourth amendment privacy. Southern California Law Review 75, 1083–1167, 1106–1107 (2002)

Solove, D.J.: Data mining and the security-liberty debate. University of Chicago Law Review 74, 343–362 (2008)

Solove, D.J.: Fourth amendment pragmatism. Boston College Law Review 51, 1511–1538, 1522–1524 (2010)

Taipale, K.: Technology, security and privacy: The fear of frankenstein, the mythology of privacy, and the lessons of king ludd. Yale Journal of Law and Technology 7, 123–221, 134 (2004)

Tokson, M.: Automation and the fourth amendment. Iowa Law Review 96, 581–647, 602–609 (2011)

United States General Accounting Office. Data Mining: Federal Efforts over a Wide Range of Uses GAO-04-548, pp. 1–64 (2004), http://www.gao.gov/new.items/d04548.pdf (accessed December 30, 2011)

Zarsky, T.Z.: Mine your own business!: Making the case for the implications of the data mining of personal information in the forum of public opinion. Yale Journal of Law & Technology 5, 1–56 (2002-2003)

Zarsky, T.Z.: Governmental data mining and its alternatives. Penn State Law Review 116(2), 285–330 (2012)