MO' DATA, MO' PROBLEMS? PERSONAL DATA MINING AND THE CHALLENGE TO THE DATA MINIMIZATION PRINCIPLE

Liane Colonna

1. Introduction

Data minimization is a bedrock principle of data protection law. It is enshrined in privacy regulations all around the world including the OECD Guidelines, the EU Data Protection Directive, the APEC Privacy Framework and even the recent US Consumer Privacy Bill of Rights. The principle requires that the only personal data that should be collected and stored is that data, which is necessary to obtain certain specified and legitimate goals. It further requires that the personal data should be destroyed as soon as it is no longer relevant to the achievement of these goals.

Data minimization is a rather intuitive and common sense practice: do not arbitrarily collect and store data because this will only lead down a road of trouble consisting of such things as privacy and security breaches. It's the analogue to "mo' money, mo' problems."[1] The predicament is, however, because of recent advances in software development and computer processing power, "mo' data" often means "mo' knowledge" which, like money, can be used to solve many of life's problems.

This paper is about how the concept of personal data mining, a term used to explain the individual use of dynamic data processing techniques to find hidden patterns and trends in large amounts of personal data, challenges the concept of data minimization. It is an attempt to demonstrate that fair information principles like data minimization, while providing a useful starting point for data protection laws, must give way to more nuanced legal rules and models. It stresses that a shift of paradigms from the current paternalistic[2] approach to handling personal towards an empowered-user approach is needed in order to better protect privacy in light of recent advancements in technology.

The outline is as follows. First, the notion of "data minimization" will be commented upon. Second, the technology of data mining will be explained, paying particular attention to a subset of the field that has been dubbed "personalized data mining." Finally, the paper will reflect upon how an unyielding commitment to the principle of data minimization is problematic in a world where the indiscriminate collection and the ad hoc retention of data can lead to many benefits for individuals and society alike.

2. Data minimization

The principle of data minimization first emerged during the 1970s at a time when there was great concern over the large-scale collection and processing of personal

[1] "Mo' Money, Mo' Problems" is a song by the legendary rapper and hip hop artist Notorious B.I.G. (also known as Biggie Smalls). It is the second single from his album Life After Death. The single was released posthumously and topped the Billboard Hot 100 for two weeks in 1997. *For more, see* Wikipedia, *Mo Money Mo Problems* retrieved at http://en.wikipedia.org/wiki/Mo_Money_Mo_Problems

[2] I use the word "paternalistic" because fair information principles are designed to protect individuals from, for example, too much data collection but maybe that's not what individuals want or need?

data in centralized, stand-alone, governmental computer databases. The idea was simple: limit the collection and storage of personal data in order to prevent powerful organizations from building giant dossiers of innocent people which could be used for purposes such as manipulation, profiling and discrimination. That is, minimizing data collection and storage times, would help protect the individual against privacy intrusions by the State or other puissant organizations. After all, data cannot be lost, stolen or misused if it does not exist.

At that time the concept of data minimization was first formulated individuals did not have the software or the processing power to handle large amounts of data themselves. Nor was there a way for ordinary people to collect and distribute limitless amounts of data via an international super network. In other words, while the concern to protect individuals from Big Brother's exploitation of large-scale personal data repositories was palpable, there certainly was little regard for the fact that individuals could somehow benefit from an amassment of their personal data. This is, however, no longer the case.

3. The technology of data mining

3.1. Data mining in general

Data mining is often thought to be the most essential step in the process of "knowledge discovery in databases", which denotes the entire process of using data to generate information that is easy to use in a decision-making context.[3] The data-mining step itself consists of the application of particular techniques to a large set of cleansed data in order to identify certain previously unknown characteristics of the data set.[4] Data mining techniques can include, for example, classification analysis (takes data and places it into an existing structure[5]), cluster analysis (clumps together similar things, events or people in order to create meaningful subgroups[6]) or association analysis (captures the co-occurrence of items or events in large volumes of data[7]).

A key feature of data mining is that, unlike earlier forms of data processing, it is usually conducted on huge volumes of complex data and it can extract value from such volume.[8] Data mining is also highly automated, sometimes relying on "black boxes."[9] Another interesting feature of data mining is that it creates "new

[3] Han, J. and Micheline Kamber, *Data Mining: Concepts and Techniques (Second Edition)(*San Francisco: Morgan Kaufmann Publishers, 2006).

[4] *Id.*

[5] Alexander Furnasapr, *Everything You Wanted to Know About Data Mining but Were Afraid to Ask*, The Atlantic (April 3, 2012).

[6] Amit Kumar Patnaik, *Data Mining and Its Current Research Directions*, a paper presented at International Conference on Frontiers of Intelligent Computing retrieved at http://ficta.in/attachments/article/55/07%20Data%20Mining%20and%20Its%20Current%20Research%20Directions.pdf

[7] Gary M. Weiss and Brian Davison, *Data Mining*, In Handbook of Technology Management H. Bidgoli (ed.), Volume 2 (John Wiley and Sons, 2010), 542-555.

[8] M. Lloyd-Williams, *Discovering the Hidden Secrets in Your Data - the Data Mining Approach to Information*, Information Research: An International Electronic Journal (1997) retrieved at http://informationr.net/ir/3-2/paper36.html

[9] Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 Penn State Law Review 285 (*2011*).

knowledge" such as an abstract description or a useful prediction that did not exist *a priori.*[10] A final important feature about data mining is that it is not necessarily limited by the creativity of humans to create hypotheses because data mining can be used to explore the dataset and generate hypotheses automatically.[11]

In some respect, data mining can be thought of as voodoo science. According to the conventional scientific method, a hypothesis is built and then the data is carefully collected to test the hypothesis. Unlike with the conventional scientific method, the data-mining method involves an exploration of a dataset without a hypothesis in order to discover hidden patterns from data. Instead of being driven by a hypothesis, the process is driven by the data itself and therefore, the results are unanticipated and serendipitous.[12] Here, the concern is that scientific proposals that are derived without a preconceived hypothesis about the data are not valuable, reliable or significant because correlations that appear in the data could be totally random.[13] As such, it is important that data miners understand the risk in the approach and take steps to evaluate the reliability of their findings.[14]

3.2. Personalized data mining

Individuals today collect and retain large amounts of personal data through a multiplicity of different channels. Through, for example, participating in the so-called Web 2.0, a massive amount of personal data is stored in emails, blogs, Wikis, web browsing history and so on. Social media, a Web 2.0 innovation that introduced web-based sharing with the click of a button, also provides for rich sources of personal data. The information that a user puts onto Twitter and Facebook, for example, can reveal a tremendous amount about a person such as individual's speech patterns, the topics an individual obsesses over and the identity of an individual's "real" friends.[15]

Likewise, individuals are generating a huge amount of data about themselves through using technologies that are embedded in everyday objects that interact with the physical world. Here, there is no need to press any buttons or to self-report: the information is raw and unfiltered. For example, an individual's mobile phone can be used to automatically track location data or Nike+ can be used to record every mile an individual runs.

One way of understanding all of these data is to use the information for personal data mining. That is, this information can be mined to cluster, classify and discover rules in order to assist individuals to extract important insights about themselves and their

[10] K.A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 Columbia Science & Technology Law Review 2 (2003).

[11] Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, In Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases (Bart Custers, Tal Zarsky, Bart Schermer, Toon Calders)(eds.))(Springer 2013).

[12] J.A. McCarty, *Database Marketing*. Wiley International Encyclopedia of Marketing (Wiley 2010).

[13] D.B. Kell, S.G. Oliver, *Here Is the Evidence, Now What is the Hypothesis? The Complementary Roles of Inductive and Hypothesis-Driven Science in the Post-Genomic Era*, 26 Bioessays 99–105 (Wiley 2004).

[14] J.A. McCarty, *Database Marketing*. Wiley International Encyclopedia of Marketing (Wiley 2010).

[15] Christopher Mims, *How to Use Twitter for Personal Data Mining*, MIT Technology Review, (October 13, 2010) retrieved at http://www.technologyreview.com/view/421201/how-to-use-twitter-for-personal-data-mining/

worlds that might be hidden within these large datasets. For example, if an individual gets frequent headaches then he/she could use data mining to look for patterns that suggest what food or activity that seems to bring the headaches on.[16] Another example is using personal data mining to identify factors that influence weight.[17]

An interesting feature about personal data mining is that the data can be mined either alone or in conjunction with the data of others, possibly collected on multiple platforms, in order to reveal hidden information among the data and the associated users.[18] The question of how precisely an individual shall gain access to this "third-party data" is not straightforward or obvious. For example, in some circumstances, the individual may be able to purchase the data from third parties and in other circumstances the individual may be given free access to the data in the interest of the collective good. The individual is also likely to encounter denial of access to data due to the nature and the value of the information.

While, at first blush, individuals may not appear to have the processing power or computer software that is available to governments and private companies, there are services being offered, which would allow individuals to search for novel and implicit information in large datasets. For example, Google offers a data mining service called Correlate that allows individuals to search for trends by combining individual data with Google's computing power.[19] Likewise, Microsoft has been granted a patent for personal data mining[20] and is currently offering Lifebrowers as a tool "to assist individuals to explore their own sets of personal data including e-mails, Web browsing and search history, calendar events, and other documents stored on a person's computer."[21]

4. Personal data mining and the challenge to the notion of data minimization

Reconciling the principle of data minimization and the notion of personal data mining is difficult. This is because a perquisite to personal data mining is the amassment of huge amounts of data. It is also because the potential benefits of mining this data are unpredictable and can grow exponentially with time, which means there is an interest in storing the data for an indefinite period.

One way of addressing this reality is to focus away from fair information principles such as data minimization towards a misuse model of data protection.[22] That is,

[16] Kevin Maney, *Download Net on Your Laptop? Maybe Someday Way storage is Growing, Who knows?* USA Today (July 12, 2006).

[17] Kuner Patal, *Personal Data Mining*, Creativity Online (April 28, 2009) retrieved at http://creativity-online.com/news/personal-data-mining/136077

[18] See *Google's Patent for Personal Data Mining US 20080082467 A1* retrieved at http://www.google.com/patents/US20080082467

[19] Douglas Perry, *Google Releases Data Mining Engine* (May 26, 2011) retrieved at http://www.tomsguide.com/us/google-org-data-mining-correlate-serach-engine,news-11343.html

[20] *Google's patent for Personal data mining US 20080082467 A1* retrieved at http://www.google.com/patents/US20080082467

[21] Tom Simonite, *Microsoft Builds a Browser for Your Past*, MIT Technology Review (March 15, 2012) retrieved at http://www.technologyreview.com/news/427233/microsoft-builds-a-browser-for-your-past/

[22] *See generally*, Fred H. Cate, *The Failure of Fair Information Practice Principles* In Consumer Protection In The Age Of The Information Economy (Jane K. Winn (ed.))(Surry, UK: Ashgate 2006);

instead of the placing the emphasis on limiting data collection, the emphasis could be placed on limiting the misuse of data. This, however, would require a more substantive approach to data protection where individuals can rely upon explicit remedies for the misuse of their personal data.

The main point here is that it matters who uses data and how they use the data and in what context.[23] The individual's mining of personal records in order to fulfill certain personal goals such as becoming more efficient, healthy or knowledgeable about his/her strengths and weaknesses in the context of self-discovery, requires a different reaction from, for example, Facebook mining an individual's personal data to reveal his/her credit worthiness in the context of a mortgage application.[24] While the personal data clearly has value to both the different controllers, it is the way that the data is used where it becomes obvious whether there has been a privacy infraction.

5. Conclusion

It is true that limiting the collection and storage of data could help safeguard privacy in certain contexts by, for example, guarding against security breaches. It is also true that the unlimited collection and storage of data can give rise to many individual and societal benefits. Consequently, the current mechanical approach to data protection that presupposes that the haphazard collection of data is always bad for individuals must give way to a more nuanced, relevant and organic model that reflects the recent and dramatic advancements in dynamic data processing and data storage techniques.

It is time to recognize that "mo' data" does not always mean "mo' problems" and to create an environment where individuals – and not just governments and big business – are able to benefit from the analysis of large repositories of personal data. It is time to pay closer attention to the ways that individuals can be empowered with tools to manage and understand their personal data. The current paradigm of "data protection" should be shifted towards "data empowerment" to exhibit greater connection with the technological reality.

Peter Seipel, *Privacy and Freedom of Information in Sweden in Nordic Data Protection* Law (First Edition)(Peter Blume (ed..)(Copenhagen: DJØF Publishing, 2001).

[23] Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford Law Books (Stanford, California 2010).

[24] Martha C. White, *Could That Facebook 'Like' Hurt Your Credit Score?*, Time Magazine (June 14, 2012).