prof[L]ing
PROTECTING CITIZENS' RIGHTS FIGHTING ILLICIT PROFILING

Working Paper

Evolution of Technologies in Profiling

DANIEL GUAGNIN, LEON HEMPEL (Technische Universität Berlin),
JUSTIN JUNG (Philipps-Universität Marburg)

Table of Contents

With financial support from the
"Fundamental Rights and Citizens Programme" of the European Union

Introduction

Profiling is not a new phenomenon but its character has been changing in the last decades. Many (if not most) of concerns that become evident these days have been anticipated since the invention of the computer, but recent technical developments have built the technical ground for the realization of automatic profiling.

On the one hand profiling technologies imply a huge potential of making life easier: the idea of Ambient Assisted Living envisions a world of things which can retrieve information and communicate to each other in order to provide a kind of self-configuring environment. Smart Meters promise to handle global challenges through optimization of resource usage and allocation. On the other hand the growing availability and quality of data implies unintended consequences such as changes in the self-determination of individuals, if decisions are increasingly based on data calculations, and changes in the traceability of individual behaviour. As the current discussions about PRISM[1] called show, data stored for individual benefit can be used in unintended ways and open up human behaviour to unforeseen scrutiny – a fact that has been discussed by privacy scholars and privacy advocates for many years. But now it turned out this is an actually used practice according to Edward Snowden's leak of confidential NSA information. Accordingly the technical evolution of profiling affects social values, the constitution of societies and even fundamental rights. In this paper we will focus on the technical characteristics of profiling and the implications of socio-technical constellations.

The structure of this paper is as follows. Firstly we will revisit the historic background of the cybernetic paradigm to underline that the recent developments in profiling, more specifically automatic profiling, have been anticipated and discussed decades ago. In the second section we will describe the digitalization of data collection and data processing as a technical precondition for, and main driver of automatic profiling. In the third section we will elaborate the changing impact of digital data processing for the analysis and the interpretation of data, which are the basic practices for profiling. We will examine the socio-technical implications underlying automated profiling technologies and discuss how the interplay between humans and machines is modified. Afterwards we will conclude how the changes in the socio-technical actor constellations support a changing notion of profiling from descriptive to predictive profiling. In the fifth section we will describe some scientific projects and current applications of profiling to illustrate how profiling is implemented in practice and to underline its actual relevance.

1. The roots of profiling

Before describing recent developments which enabled or enhanced the possibilities of automated profiling, it is useful to shortly revisit the historic background of profiling technologies. For this purpose, three developments in the last century will be revisited: (1) how the beginning of computation fostered concerns about machines controlling humans, (2) the pervasiveness of the cybernetic paradigm, promising to overcome human deficiencies

1 PRISM is a United States National Security Agency mass surveillance programme. It is only one of several NSA spying programmes, and names the cooperation between the NSA and internet companies, whereby the companies allow the NSA access to user data. The revelations about PRISM lead to a general discussion about justification of warrantless state surveillance. For detailed information see https://en.wikipedia.org/wiki/PRISM_%28surveillance_program%29 and https://en.wikipedia.org/wiki/2013_mass_surveillance_scandal [both accessed July 2013].

through technologies, and (3) the turn from retrospective fact analysis to probabilistic foresight in law enforcement.

First, the invention of the computer changed the way of how things are represented, how knowledge is transformed to information (Degele, 2000) and the scope of how information can be processed. Information storage is key to control resources of power which enable control (i.e. to store, retrieve, display) over information. These resources can be used to perpetuate social relations across time-space. (Giddens, 1984, p. 261) Obviously what computers can do best is to store and process information, and promise to provide efficient possibilities to cope with complexities. Accordingly the idea of computers as a great resource of power fostered concerns about increasing control through information availability. Other concerns starting with the very beginning of computer technologies were that a growing dominance of computers could change the character of human judgement (Johnson & Wayland, 2010, p. 23; Kling, 1994; Weizenbaum & McCarthy, 1977).

Second, inspired by the cybernetic paradigm in the postwar period the vision of a society unfolded where human deficiencies could be overcome by technology. Yet, this implied not merely the adoption of technical artifacts, but a transfer of military principles such as early detection, reconnaissance and identification of friend and foes. Remarkably the utopian potential of technology grounds on its conceptualization as being neutral. Technological rationality is conceived as a measure to control and reduce human subjectivity – at the same time technical rationality encapsulates its implicit subjective share through an objectifying functionalism. This builds the ground for an epistemology of governance through a universal calculus of political regulation, prediction and control which is objectified and concealed in technology. The central concern is to anticipate future developments in order to manage the uncertainties of the presence. (Hempel, 2012)

Third, as Gary T. Marx (1990) states, in the 70ies, after political scandals like Abscam and Watergate have been revealed, the focus of undercover policing activities shifted to small-scale crime policing, thus everyday activities became subject to undercover surveillance and even methods of intelligence operations started to be more and more applied in everyday law enforcement practices. Thus, the application of control and surveillance techniques and technologies found new fields of application after losing legitimacy through scandals. This shift of application is a prime example for function creep, the phenomenon that regulations or technologies established for a certain purpose will be used for other purposes afterwards, which is termed as “one of the most operational dynamics of contemporary surveillance”. (Ericson & Haggerty, 2006) Marx examines various technological surveillance tools and discusses the expansion of their use in everyday contexts. He states that the boundary of criminality becomes blurred as well as the boundary between governmental and social control: “Powerful new information-gathering technologies are extending ever deeper into the social fabric and to more features of the environment. […] People are in a sense turned inside out, and what was previously invisible or meaningless is made visible and meaningful.” (Marx, 1990, p. 206) Some of the basic novel characteristics of “new” surveillance technologies stated by Marx are that they overcome physical barriers and enable to scrutinize from remote. The character of surveillance technologies changes to anticipation and prevention of behaviour and is conducted without consent of its subjects – often even without noticing. Moreover the subjects of scrutiny become an active part in their own surveillance as in everyday life activities they are involved in technologized processes. This is even more true for new Internet-based applications as pointed out below. Finally the generation of suspicion is turned upside-down: “The new forms of control are helping to create a society where everyone is guilty until proven innocent.” (cf. Hempel, 2012; Marx, 1990, p. 219)

We will now describe some basic technologies underlying profiling, to describe how profiling works and how the evolution of these technologies corresponds to a tendency towards automatic profiling.

2. Digitized data generation and processing.

Considering that profiling is an old human practice of distinguishing different sorts of people, it is vital to first examine the grounds of data construction and the growing digitization of nearly all kinds of information. For that we will distinguish different data sources and describe the relevance of digitization. Afterwards we will describe basic aspects of digital data management, data bases and data processing, since the process of linking different data bases is used to build profiles and thus is essential for understanding profiling (cf. Rocco Bellanova et al., 2011, p. 60). Against the backdrop of the historic concerns about increased capabilities of control outlined above and in regard to the Fundamental Rights perspective of the PROFILING-project, we will discuss the social implications of the technologies explained.

2.1. Data sources

We distinguish three data sources – sensors, logs and user content – as three different ways of data construction and describe their changed characteristics caused by digitization. Most importantly digitization makes the information usable as data for digital processing which will be discussed in the next section.

i. Sensors

Generally speaking sensors are tools to measure physical quantities or chemical characteristics and to convert them to an electronic signal which can be interpreted by an observer or a machine. For example, a variety of sensors is applied on airports where luggage and people are searched for artifacts classified as dangerous. In the vision of ubiquitous computing and smart living, sensors are applied in many contexts from temperature control in the flat to car navigation. But also microphones and video cameras can be regarded as sensors.

The visualizations of sensors on monitors or the video and audio recordings require a human person to extract information. Once the signals are converted in digital signals, they can be processed digitally and computer programs can be used to analyse the signals and to enrich the material with additional information. For instance software can identify situations where people on a videotape are gathering, or distinguish conversations from arguing. Pictures can be enriched with location data or biometric characteristics of people can be extracted to identify persons on different materials. (cf. Marx, 2002)

ii. Logs – technical data traces

While sensor data is generated intentionally by the maintainer of the sensor, log files are generated through the mere use of digital infrastructures because logs are used by computer managed systems to provide information for system administrators. This makes huge amounts of data available – originally created for technical purposes – which can be used in many ways to analyse communications and online activities of users.

While, for instance in the realm of telephony this kind of logging data has been produced to invoice the telephone connections for long, log data is growing along with the increased use of computer mediated activities (since any computer network data is transmitted via a telephone connection). Moreover any information received through Internet connections and any communication and interaction which is mediated by the Internet is logged on computers which process this information. (cf. Leenes & Koops, 2005, p. 332f) More and more activities are mediated through computer networks, namely the Internet. For instance smart phones are nearly constantly online, continuously submitting and synchronizing various data with the “cloud”. Consequently a growing part of daily interactions like mailing, banking, shopping and many more are digitally recorded and the data generated can be subject to digital processing (see below).

iii. User generated content

Another kind of data is user generated content. This usually includes personal data and information that document user's personal interests. Communication is more and more mediated digitally: Email and Voice over IP replace its analog ancestors telephone and postal delivery. New ways of expressing oneself are implemented: Blogs which are sometimes similar to publicly visible diaries. Microblogs – most prominently twitter – enable users to post short messages via SMS, finally SMS are replaced by Instant Messaging Programmes (e.g. Google-Talk and What'sApp).[2] Consequently personal communication becomes user generated digital information, which are digitally stored on the service providers' servers and can be retrieved for further processing.[3]

iv. Implications of digitized data sources

Observations made by human beings need to be written down to be made explicit. Sensors and Video and Audio recordings make it possible to externally store and exchange information. Storage used to be expensive and usually human interaction was needed to extract valuable information (e.g. video analysis). The written documentation of observations can be regarded as a first step to enable a generalized and objectified way of information exchange between individuals. Digitized sensors's data, however, can be processed and analysed digitally (see below) which is much easier and cheaper to store, process and analyse as we will discuss in the following. An illustrative example of how exhaustive and expansive the detailed documentation of people's activities and behaviour was, is the comparison recently drawn between digital data the NSA stored with the amounts of files the Stasi produced. The comparability is questionable but it is interesting that the print out of all the data captured by the NSA would need a filing cabinet nearly as big as the USA. (See figure 1)

The example also shows the efforts needed to collect data. While the Stasi needed to install microphones, hire staff to monitor and document people to gain information about

2 These diverse ways of communication are often centralized in online platforms provided by big companies like Facebook and Google. Specifically Facebook is a great example for an online platform providing diverse kinds of online interaction in one user interface: Messages can be sent as Instant Messages (Chat) or as a kind of e-mail. Pictures can be shared, Status Messages are an equivalent of micro-blogging. The centralization of different communication channels makes the data easier accessible because all the information is maintained by one company. Consequently a broader data set can be used for consumer profiling.

3 Of course use of such data may be limited through the providers' terms of services which have to be agreed by their users. But often these agreements are very long, hard to understand and accepted without reading – as the TOSDR project (“Terms of service – didn't read”) puts it: “'I have read and agree to the Terms' is the biggest lie on the web.” see http://tosdr.org [accessed July 2013]

their habits, attitudes and social networks, in a digitized world a lot of that information is stored on (service providers') servers and can be accessed by request. To use a less dark scenario, let us think of rebate marketing which – in pre-digital times – enabled companies to collect a broad range of information about consumers' interests and activities (and the terms of service allowed them to use the data in the way they wanted to). But with computer mediated information streams the scope and granularity of information is massively amplified and data can more effectively be used to optimize webpages, advertisements and product offers, or to select the information provided to a consumer.

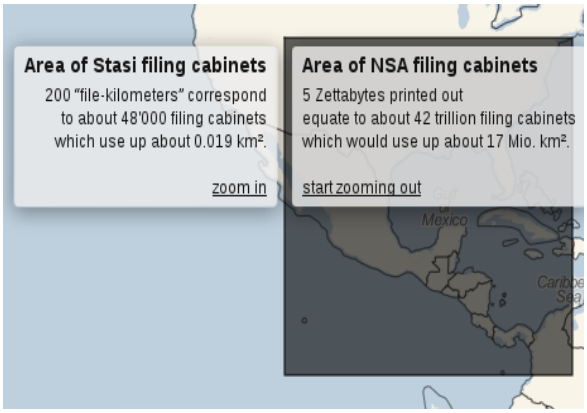

Figure 1: Area of filing cabinets, source: Opendatacity (2013)

For collecting consumer information logging data and user generated content are used. Notably the user is not necessarily aware that logging data exists and can be collected. Web browsers submit a lot of information about the current "session" and every request contains information about the current user: pages visited, durations of visits, which links were followed and using "cookies"[4], information can be maintained and reused over longer period.[5] Even IP-Numbers can be identified and collected to monitor surfing habits.[6] IPs are unique numbers assigned by the Internet Service Provider who can use them to identify every single household. (cf. Leenesand Koops 2005, 332f).

A crucial role for building profiles about individuals plays user generated content, which is a great innovation in information retrieval. In Online Social Networks (OSN) people provide information about their interests, friends, and much more, information formerly was not easy to get. One comfortable way to get access to this information are companies offering a login via established OSNs. The user can easily register without creating a new account while the company can acquire various information from his preferred social network platform.[7]

2.2. Processing

As outlined above information is increasingly digitized and thus becomes easy to store, access and analyse. We will go through some basic aspects of data processing such as data management, exchange and database coupling, data preprocessing and discuss the socio-technical implications of these methods.

i. Data management: Databases and tables

The large buildings containing big archives are replaced by server farms with huge amounts of computing power and digital storage. Digital or digitized data is usually stored in digital databases consisting of tables. These databases can be accessed via digital net-

4 Small files which are managed by the websites and used to write and read information browser.

5 Some of the informations are gained by intentionally generating logging data. For instance "tracking pixels" generate a HTTP-request to special servers which are logging these HTTP requests.

6 Usually the IP(v4) number changes from time to time due to number scarcity, but a new generation of IP numbers is in preparation which would make it possible to identify every single device (IPv6). See https://en.wikipedia.org/wiki/IPv6 for more information

7 See for instance http://janrain.com to learn how so-called social logins through established online social networks enable thrid party service providers to access personal information from its clients' preferred online social network.

works, thus every access can be made from a desktop through network cables. Cost of storage and access has become very low (comparing to analog archives) and efforts in time and space necessary to maintain and access information have reduced dramatically.

Moreover the organization of data is much more efficient. Data is usually fed into databases consisting of tables where data records are represented in rows enriched with a number of attributes. In the process of optimization redundancies are reduced by splitting tables in small tables representing subsets of relations, which can be re-linked by request.[8] To keep the data re-linkable it is vital to have unique identifiers (UIDs) for any records.

ii. Data exchange

Digital data bases and automated data processing enable the distribution of data on a large scale to a comparatively low price, e.g. instead of hard-copying loads of paper pages or carrying files, a whole archive can be stored on a USB stick or just remotely downloaded on Hard-Disk. The Internet as a global digital infrastructure makes it easy to exchange and widely distribute data in short time at low cost. This enables data controllers to easily exchange data and to integrate data collected in different contexts into new data bases for automated data analysis and knowledge construction.

iii. Database coupling

If databases contain overlapping information, more specifically if for some records of one database attributes from another database can be derived, the information can be put together – a process called database coupling and a crucial tool for data mining. With data base coupling the number of relations between data grows significantly. A difficulty is to correctly link the corresponding data sets (rows in the table), because a common identifier needs to be found. Usually the UIDs differ between databases, or values which are suitable to be used as a UID are stored in a different format and need to be converted first. For example it is not a trivial task to combine a database of students with a database of student jobs, if in the first UID is the student registration number and in the second the social security number. Given the fact that both informations are exclusively in the original data base, matching the names might be the only way to link the corresponding data rows by name. It may be difficult to match different conventions of representing the name: for instance if surnames shortened or not and if surnames are noted before or after the surname, if separators are comma or period, where second surnames and prefixes are put, etc. (Calders and Custers 2013, 31)

iv. Data preprocessing

Before data can efficiently be processed for analysis, data needs to be prepared and refined. Missing values must be substituted with valid values (Missing value imputation). Depending on the data it can be useful to divide values into non overlapping ranges (discretization) or to reduce dimensionality through summarizing attributes into groups. (Calders and Custers 2013, 39) New features can be extracted or constructed from data by derivation from existing values. (e.g. age can be derived from someone's birthday) (Canhoto and Backhouse 2008)

8 See for instance https://en.wikipedia.org/wiki/Database_normalization

v. Enhanced capabilities of digital data generation and processing

Figure 2: Manual analysis of fingerprintsat the federal armory during WWII, Source:FBI 1945

It becomes clear that the management and processing of data changes substantially through digitization. Storage and access of data is becoming independent from time and space. Accordingly it becomes much easier for institutions to exchange data. In pre-digital times data was stored in documents and tape recordings, which were stored in huge archives. Every access needed to be made physical, i.e. someone had to go into the archive and get the file. Accordingly storage as well as access was expensive. Thus it had to be considered well if or not to store or to access a file. More specifically the scope and extent of data retrieval and analysis has changed through this technological shift. At the same time it becomes more opaque how data is exchanged and thus how it is used and by whom. Also through data base coupling data of different areas can be linked to generate substantially new knowledge about individuals. Many personal data is unproblematic as a single data but becomes interesting – or confidential – when linked with other kinds of data.

Customers often cannot overlook the extent of “third party use” but are asked to agree with the dozens-of-pages terms of services , Data is shared between different local police departments, international institutions[9] up to transatlantic exchange between secret services[10], which is hard to grasp. Thes are just to examples how digitized data generation and processing decouple data from its local origin and its originally intended use. Data can easily be reused without users realizing which kind of data is collected and how it is used. Moreover new ways of analysis and interpretation generate new unforeseen (and for most citizen unimaginable) kinds of knowledge (cf. Hildebrandt 2009).

3. Socio-technical implications of digitized data analysis and interpretation

Data analysis and interpretation are key tools for profiling. Once data is digitally processed, it is a logical consequence that data is also analysed more and more automatically. Data is produced and stored on a large scale which implies problems of scalability. Humans alone cannot cope with the amount of data and even computer resources be-

9 e.g. on the European level: Occhipinti 2003; cf. De Hert and Gutwirth 2006; Hempel, Carius, and Ilten 2009, in USA data fusion centers were installed to gather information not only from government but also from private sources, see Monahan 2009)

10 As currently debated in the context of prism, see for instance http://www.tagesschau.de/inland/nsa-skandal100.html, http://www.spiegel.de/international/world/spiegel-reveals-cooperation-between-nsa-and-german-bnd-a-909954.html [both accessed 08/01/13]

come limited. It is a challenge how to select subsets of data and how to optimize algorithms to produce outcomes in a reasonable amount of time. (Calders and Custers 2013, 31) Han and Kamber (2006) define data mining as “automated or convenient extraction of patterns representing knowledge implicitly stored or catchable in large databases, data warehouses, the Web, other massive information repositories or data streams”. The “need for data mining [is] motivated by the challenges posed by the huge amounts of data available” (Calders and Custers 2013, 41)[11]

As will be shown, the data mining technologies support a shift to probabilistic and predictive approaches interpretation of data which fosters a changing understanding of profiling. This tendency is reflected in the term Knowledge Discovery in Databases (KDD) – often synonymously used for Data Mining: generated knowledge is not only generated but unintentionally “discovered” in data bases.

Data mining mainly consists of algorithms which are constructed or trained to find patterns in the data. Some basic data mining techniques will be outlined in this section to provide an understanding of how data mining basically works. Afterwards the ramifications for interpretation of data will be discussed. Finally CRISP-DM[12], a generic Data Mining approach will be revisited to illustrate how an exemplary Data Mining procedure is conducted.

3.1. Digital data analysis – basic data mining techniques

Clustering is used “to assign objects to groups or classes on the basis of criteria that can be adjusted on the basis of theory and experience. While classification into categories or groups may be the ultimate goal of these analytical efforts, a preliminary stage in the process may emphasize the discovery of patterns of association , or covariation”. (Gandy Jr 2006, 368) There are many other Data Mining algorithms, but most of them are based on one of the three types described in this section (Calders and Custers 2013, 32)[13]

i. Pattern mining

The purpose of this method is to find patterns – or relationships – in the data, which describe data or predict attributes. E.g. a regression analysis tries to find a mathematical function that describes points in the system of coordinates, correlation coefficients can then be used to measure how the functions represent the data. The computed pattern visualizes the data and can help to estimate future developments or to identify irregularities. In Data Mining, often the quality is not measured based on how the patterns together represent the complete dataset, but how surprising the patterns are, i.e. diverging from the global structure of the data. (Calders and Custers 2013, 31)

Like clustering, pattern mining is an unsupervised technique, that means algorithms are trained automatically, without human supervision, and no targets (labels) need to be previously defined. (Calders & Žliobaitė, 2013, p. 48)

ii. Clustering

Clustering does not need a pre-definition of classified examples and is thus – like pattern mining – a non supervised technique. (Oracle, 2008, Chapter 7) It is mainly used to iden-

11 Note that this technical view stands in contrast to Lyon's statement that it is not merely because of the capabilities but due to an increasing number of perceived risks and the desire to manage populations, cited in section 4.

12 Cross Industry Standard Process for Data Mining

13 A more detailed description of Data Mining algorithms can be found in Bernhard Anrig et al. (2008).

tify and describe groups through grouping similar sets of attributes. Specific types of clustering are partitional clustering – which generates disjoint groups – and hierarchical clustering – which builds a complete taxonomy. Because it does not need human interaction it is suitable for exploring data and to establish classes or groups which can be used for classification (Oracle 2013). At the same time it identifies/detects outliers – objects that are unlike many other objects (i.e. which do not belong to a large cluster, or forming a cluster themselves) and which can thus be interesting for further analysis. (Calders and Custers 2013, 35)

iii. Classification

Classification is a method to order data into predefined classes on the basis of their similarity (Bailey 1994). Classes need to be exhaustive and mutually exclusive, for that the boundary between classes is clearly defined and each object is classified in one of the classes. The classification is a supervised method, the algorithm needs to be trained to refine the selection mechanism. Different algorithms can be used to find different relationships in the data which are summarized in a model. (Oracle 2013, chap. 5)

3.2. Changing focus, changing facts? Ramifications of digital data analysis

In this section, we will be arguing that the increasing degree of automation in analysis procedures leads to changing perspectives on the objects of analysis.

Long before the process of analysis and interpretation, even the request of analog files for a data analyst – be it a marketing specialist or a policewoman – can be considered as totally different from accessing digital data. As argued above the cost and effort to get analog files on a desk is substantially different from downloading digital files on a PC desktop. The contents of the folders differ, since there are things which are easier to be put in analog files and other information is easier to be produced digitally. Thus the availability of certain kinds of data changes. Third, the access on the data at the desktop is very different. Analog folders can be touched and felt, digital files are browsed on a screen or can even be searched by keywords. Consequently the way of reasoning changes: more interaction of the analyst is oriented towards computer interfaces and thus influenced by the way user interfaces are designed, informations are presented and how searches can be conducted.

Nowadays big amounts of data are produced and made available and the availability leads to a need to use the data. Data mining is the preferred approach for analysing big amounts of data, but its methods focus on pattern matching and are thus discovery driven in contrast to statistical hypothesis building and testing. Accordingly the (preceding) reasoning about causes and relations is replaced by the need to interpret outcomes which were generated partially automatic.

Reconsidering Han and Kamber's (2006) previously cited definition of data mining as “automated or convenient extraction of patterns representing knowledge implicitly stored or catchable in large databases”, it is important to note that “automatic extraction” indicates the changing role of human actors: learning algorithms are trained on specific data sets to build categories or to find patterns in the data. Previously made assumptions or hypotheses about the issue itself play a minor role, hypotheses are rather derived from the material. At the same time implicit assumptions may lead the actors when defining the training data, selecting and preprocessing target data and choosing suitable algorithms. Also, search algorithms can be implemented differently and different sets of algorithms can be

offered to a user of a software program (thus interfaces play a role in constructing the outcomes). This considerably leads the search and influences the process of reasoning and finally the outcomes.

Consequently we can claim that there is a development from hypothesis checking to pattern mining in data analysis approaches. This strengthens a predictive focus of the observation. Subjective assumptions become hidden into the technology in the process of automatization while outcomes based on models which are computed on the basis of databases are often perceived as solid statistics and thus more objective than models constructed by humans. (Calders & Žliobaitė, 2013) This perception as objectified knowledge of computer generated models supports the thesis of hardening function of technologies in general and more specifically the thesis that social sorting becomes strengthened if mediated through technology. (Lianos and Douglas 2000; van Brakel and De Hert 2011)

3.3. Excursus: Steps and tasks of an idealized Data Mining process – the CRISP approach

The CRISP (Cross Industry Standard Process for Data Mining) Data Mining approach is conceived as a reference model and a user guide for data mining which shall help data analysts to take into account fundamental as well as crucial aspects in the data mining process. It was developed in March 1999 by the CRISP-DM consortium[14] with funding from the European Commission's Framing Programme 4[15]. Gasson and Browne (2008) state that such best practice models are important to "enable a reasonable level of assurance that the involved, complex and esoteric data mining process will ultimately render useful, repeatable and, most importantly, valid results." They stress that it is vital to keep in mind that profiling is merely an "attempt to deal with the diversity and complexity of reality, through categorisation" – it is never an exact science but an approximation.

The CRISP-DM model is thus helpful to foster reflexivity in the data mining process to reduce the gap between customers' expectations and realistic possibilities of data mining analysis. The guide helps to avoid overlooking important steps and to make the whole data mining process accountable. The CRISP approach separates the data mining process into six phases.(Chapman et al., 2000)

The first phase called "**Business understanding**" is mainly to get an understanding of the problems addressed and the goals that should be achieved taking into account the concrete business context, the resources, and the restrictions of the data mining endeavour. It can thus be seen as the initial translation of a business goal to a data analysis design.

The second phase "**Data understanding**" aims on getting familiar with the data to make sure that the data available is useful for the objectives and the data quality is sufficient.

"**Data preparation**" is necessary before any modeling procedure because each tool and method has its specific preconditions of data formats and data structures. Moreover relevant data for the respective analysis and the particular tool need to be selected, if necessary data sets need to be merged to get a richer data set etc. This step is likely to be performed multiple times, especially if several models are envisaged.

In the "**Modeling**" phase the concrete modeling techniques are chosen and the parameters are calibrated. Before the model building process, a test procedure should be gener-

14 The consortium included NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), Integral Solutions – later SPSS Inc. (USA, and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands)

15 http://www.cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_RCN=2923980

ated to measure the quality and validity of the model. Eventually the selection and configuration of data, techniques, and parameters need to be adjusted. After the model building processes are conducted, the success of the modeling and discovery techniques is assessed.

The "**Evaluation**" phase goes beyond the assessment of accuracy and generality of the model, but evaluates if the model meets the business objectives or if there are deficiencies. If possible the model can be tested in a "real application" context. If applicable, other data mining results which are related to the respective data mining project are taken into account in the evaluation. Finally a thorough review is useful to check if any important factor has been overlooked. If necessary the steps above are repeated to validate or improve the outcomes.

In the "**Deployment**" phase, the knowledge gained is organized and presented in a way the customer can use it. Possibly the procedures are generalized for future uses, sometimes a "live"-model needs to be generated which can be used for real-time analysis.

4. The changing notion of profiling

The digitization of sensors and the generation of data through logs and user engagement leads to a big amount of data which is available for digital processing and analysis. The amount of data available necessitates digital processing and an increased degree of automation to cope with the complexities of data management. At the same time data is more and more regarded as a resource that can be "mined", knowledge can be "discovered" in the grounds of Big Data: Frawley et al. (1992) describe data mining as the "nontrivial extraction of implicit, previously unknown and potentially useful information from data". With the changing capabilities of computers and the availability of data in searchable databases, the understanding and purpose of profiling has changed. Profiles can be seen as hypotheses. The hypotheses are not necessarily developed on the basis of a theory or a common sense expectation, but often emerge in the process of data mining. This can be regarded as a shift from more traditional assumption-driven approach to a discovery-driven approach. *(Hildebrandt, 2008)*This is not merely because of the new capabilities; as Lyon (2003, p. 20) argues, a central role play the increasing number of perceived risks and the desire to manage populations.

The use of profiling has developed from a case based deductive approach, which aims at developing characteristics that are incident specific, to an inductive generalizing procedure, deriving characteristics from criminal populations to follow pre-emptive and predictive strategies, for instance to find individuals that are most likely to commit crime. (van Brakel & De Hert, 2011) Clarke (1988) introduces the term Dataveillance and distinguishes between Personal Dataveillance, which focuses on identified individuals, and Mass Dataveillance, which implies a generalized suspicion. This means that data is more and more collected despite any suspicion in order to mine the data afterwards and find individuals "in need of attention" (ibid., 502).

The increasing automation of profiling is reflected in the terms machine profiling and automated profiling.[16] These sorts of profiling are based on data which is already stored. Raw data is generated through recording events and actions or the correspondent data. Data is translated into knowledge through interpretation. It is important to keep in mind that this

16 For a more detailed discussion of different definitions of profiling and the discourses about profiling see the PROFILING project's Working Paper "Defining Profiling", (Valeria Ferraris, Francesca Bosco, Gillian Cafiero, Elena D'Angelo, & Y. Suloyeva, 2013)

knowledge is not a knowledge of assured facts but a interpretation derived from data sets. Consequently changing data can quickly change the interpretation and thus the facts. Data bases need to be stored in a form where new interpretations can quickly be regenerated instead of storing variable facts. (M.Hildebrandt et al., 2005)

The manual analysis and interpretation of e.g. behavioural data and the classification of data subjects is getting replaced by highly automated computer processing, which is justified by increased amounts of data and the desideratum to reduce cost by increasing efficiency. (Canhoto & Backhouse, 2008) Automated profiling can be regarded as a powerful tool, “but there are consequences for using it inappropriately or for drawing ill-informed conclusions from it, and these issues serve to fan the flames of the technophobes.” (Gasson & Browne, 2008) But it is not only technophobes who criticisize actuarial methods. As stated in the introduction of this paper from the very beginning of computers critics warned of a generalization of technical rationality – especially because technology is too often understood as being neutral. Recently Ben Harcourt warned in his book “Against prediction” that actuarial methods in criminal law serve only to accentuate the ideological dimensions of the criminal law. (Harcourt, 2008, pp. 190–191; cited in Sapir, 2008, p. 261)

5. Current developments

In this section we will introduce some current applications of profiling, mainly in the security sector. This will give some practical examples of how automated profiling is actually used. Afterwards two ambitious research projects developing profiling techniques and applications at European level are revisited to give an impression of what is currently envisioned how profiling could be used in the future. Additional examples are given in the related PROFILING projects' working paper “Defining Profiling”. (Ferraris, Bosco, Cafiero, D’Angelo, & Suloyeva, 2013)

5.1. Practicing profiling – Examples of current applications of profiling technologies

The examples provided here are mainly focused on applications in the field of security and law enforcement, because data which is used for manifold services in the end can be easily accessed by governmental authorities as the current PRISM scandal shows. Moreover if it comes to fundamental rights concerns, for a democratic constitution of a society it is crucial to limit the governmental control over citizens' democratic engagement and activities which are increasingly mediated over the Internet which is a leading source of data – also for security services. Finally we will also shortly revisit profiling capabilities and practices over the Internet.

i. Centralization and interplay of private actors and governmental access

Data is collected for various reasons through private companies in the first place. Data collection can serve the optimization of advertisments, risk assessments for invoicing or granting loans. Data is also necessarily generated in technical processes like telecommunication systems and is vital for these systems to work and to track and solve problems. Additional information is produced voluntarily by individuals to shape identities and exchange informations, what is often referred to as social software or Online Social Networks (OSN). The possibilities of data usage and analysis – and thus power over the individuals – grow with the access to a broad range of data. In the beginning the various kinds of data are usually spread over diverse companies and service providers. Accordingly data about individuals is only available in parts. However companies are often enabled to exchange information to get a more complete picture of their customers. Moreover some big companies might already have a very broad data basis, most prominently Google could potentially monitor from search terms over Mail and Scheduler informations up to web browsing habits and even location data from android devices. Finally data stored once, can be easily accessed by governmental authorities (cf. Gandy Jr, 2006; Soghoian, 2011). It was reported that those possibilites exposed political activists in serious trouble, for instance during the “arab spring”, but also the revelations about the USA's National Security Agency activities (PRISM) illustrated the extent to which individuals can be monitored. (see for instance Reporters Without Borders, 2012)

ii. Collaborative filtering

The ever increasing availability of information in the Internet makes filters necessary which can select the information relevant to its users. One filter-technique is collaborative filtering, also known as social filtering. Basic for that approach is the assumption that there are specific patterns and trends regarding the interests and tastes of persons and groups. Col-

laborative filtering collects an individual's information and product preferences and compares these to other individuals' preferences to provide more information to the respective user – based on the preferences of similar users. Unsolved problems of that approach include predictions which are based on wrong former decisions of users, the time exposure of user rating, wrong ratings and the need of a initial data set to enable effective outcomes.

The techniques of collaborative filtering existed before the World Wide Web through communication tools like the USENET and e-mail. With the increasing relevance of the WWW and the increasing availability of information, this technique was adopted in order to enable users to cope with the big amounts of information. Early examples were Mosaic, Firefly, Yahoo!, Point's Top 5%, PHOAKS und Fab; the most prominent might be today Amazon. Firefly even granted the users special options to control their data. The next step in collaborative filtering could be OpenFolders, a tool which compares which files users store on their computer desktop and how they use these files to derive patterns of relevance for generic computer files.[17]

iii. Cell phone monitoring

Funkzellenabfrage (GSM localization query/ query of radio cells) The GSM network is divided in cells and each cell has its own mast. Cell phones always register to the closest cell mast, what makes any cell phone easy to locate. For investigation purposes, the police can collect and analyse this data to find out which persons were near a specific location during a specified time. For example, in Dresden, Germany, 138,630 data signals of 65,645 cell phones were collected and analysed for further investigation during a protest march against neo-Nazism in February 2011 (Sächsisches Staatsministeriums, 2011). The fact that not only the data of suspicious persons, but all cell phones in the area were part of the query was highly debated because of proportionality concerns of data protection proponents. (Ibid.).

IMSI-Catcher – IMSI stands for “International Mobile Subscriber Identity” – is a technology that pretends to be a cell mast. It forces cell phones to connect with them instead of the regular cell mast. IMSI-Catcher collect IMSI and International Mobile Equipment Identity numbers (IMEI) of cell phones in range. With multiple measurements it is possible to spot the location of a specific person more exactly than analysing the data from an entire radio cell. Some IMSI-catcher have the possibility to force a cell phone to drop its encryption and even listen to phone calls and read text messages – if an encryption is used, what is not standard in every country.(Fox, 2002; Harnisch & Pohlmann, 2009). Activists report that this technology is used by police on protests.[18]

iv. No-fly, Selectee lists and Trusted Traveller programs

The best known watch lists and trusted traveller programs are constituted by the United States of America. The Terrorist Screening Centre was established under the US Homeland Security Presidential Directive 6 in 2003 in reaction to 9/11. Its main objective is to maintain the Terrorist Screening Database. It is created to have a single database of identifying people known or suspected of being involved in terrorist activities. Informations are collected and analysed from global sources on international threats (Krouse & Elias, 2009). The Terrorist Screening Database is used to compile the No-fly and Selectee lists (FBI, 2013). Individuals on the No-fly list “are considered a direct threat to U.S. civil aviation” (Krouse & Elias, 2009). The Transportation Security Administration collects the name,

17 See: http://www.moyak.com/papers/collaborative-filtering.html

18 Cf. “Your Phone May Not Be Safe at Protests.” http://privacysos.org/node/737 [accessed March 2013]

date of birth and gender of people who want to fly into, out of, within and over the continental United States for their Secure Flight program. Its main task is to identify high-risk passengers for appropriate security measures and prevent individuals on the No-fly list from boarding an aircraft, whereas low-risk passengers can be sort out for expedited screening.

Expedited screening for trusted travellers is possible for members of Global Entry, NEXUS and Secure Electronic Network for Travelers Rapid Inspection (SENTRI). To participate at the Global Entry program, a background check and an interview is mandatory, biometric informations, e.g., fingerprints, and a photo are taken[19]. Individuals who have been convicted of any criminal offence, who are subject of an ongoing investigation or cannot satisfy the U.S. Customs and Border Protection (CBO) are not allowed to participate[20]. The Transportation Security Administration incorporates random appropriate screening[21]. NEXUS is a trusted traveller program under similar conditions as Global Entry and is open for Canadian and US citizens. It allows participants a quicker entry to both Canada and the US when travelling between both states. (Canada Border Services Agency, 2012) SENTRI is a program for trusted travelling between the US and Mexico and is based on similar conditions. (CBP.gov, 2012)

v. Profiling database of violent hooligans in Germany

In 1994 German police created a database to coordinate nationwide preventive measures against football fans who are willing to resort to violence. The database includes about 15.000 individuals from Germany and other foreign countries. The database contains the data of individuals who are convicted or suspected for commission of a crime in connection to a sports event. Data is collected in Germany and foreign countries. Individuals are mostly banned from stadiums and are not allowed to approach a specified radius around the stadiums. During a sports event in a foreign country, these individuals are most likely not allowed to travel to the concerning country. The data includes personal information, stadium bans, the individual's favorite football club and exercised preventive measures (Polizei Nordrhein-Westfalen, n.d.).

In connection with the G8 summit 2007 in Germany, several politicians demanded a Europe-wide database to include left-wing autonomists, based on the model of the violent hooligan database (sueddeutsche.de, 2010).

vi. Web user profiling

There are several methods to reconstruct the session of a web user, such as log file analysis, session tracking, user and multi-server user tracking. One common method is log file analysis. Every time a user requests data from a server, several information will be sent to the server and stored in a log file. This data includes IP address of the user, requested data, HTTP-Header, user agent, operating system, the previously visited page (URL-referrer), return value (if the requested document is found or not) and size of the information. (E. Benoist, 2005) Usually this information is used to calculate statistics which can be used to advertisement optimization by generating user profiles. (Mattioli, 2012)

Cookies are information sent by the server and stored on the computer of the user. The cookie is usually set when the user visits a web site for the first time. During future website

19 http://www.globalentry.gov/howtoapply.html
20 http://www.globalentry.gov/eligibility.html
21 http://www.tsa.gov/tsa-pre%E2%9C%93%E2%84%A2/tsa-pre%E2%9C%93%E2%84%A2-faqs

visits, the server can retrieve information about former visits from the cookie and update the information stored in the cookie – if the user does not actively prevent it. Accordingly cookies facilitate the creation of individualized user profiles. (E. Benoist, 2005) Other methods to track user behaviour and build profiles are tracking pixels, java-script in general, HTTP-Request IDs (Inserting IDs in URLs) and browser fingerprints[22]. To a certain extent also IP addresses can be used for tracking online sessions. (ebd.)

Cookies are not only used to track the session on a single web page, beyond that they provide possibilities to track the whole online sessions of users, logging the surfing behaviour over different web pages, called multi-server tracking. An element, e.g. a picture of a web page can be stored on a third-party server that is different from the server the original web page is hosted on. When the user requests a file from the web page, there is also a request to the server with the mentioned element and therefore, two different cookies with two different user IDs are assigned for the user. This element or several elements belonging to the same server can be displayed on various web pages on other servers. This information are fundamental for generating personalised and group profiles (ebd.) which is also the ground for targeted advertising (Toubiana et al., 2010)

vii. IATA Checkpoint of the future

The Checkpoint of the Future describes the vision of an airport and aviation security program developed by the International Air Transport Association (IATA). In contrast to the current “one-size-fits-all” security screening, the endeavour focuses on a risk based approach with based on future profiling technologies. The major goals of the program are “strengthened security, increased operational efficiency and improved passenger experience. (IATA, 2012)

Risk assessments base on the assumption that the majority of travellers are of a lower risk. The assessments are made by using travel data like Passenger Name Records, Advanced Passenger Information and information from Known Traveller programs. These measures include a wide range of data and inputs from national and international agencies. Additional behavioural analysis is conducted. For direct questioning and behavioural observation it is envisaged to engage specialists. Automated behaviour detection and behavioural characteristic observation on the entire airport is planned to be supported up from 2017. Factors taken into account for risk assessments are type of flight (business or tourism), traveller type, passenger data and traveller program membership. Preceding to be accepted for a traveller program, measures like backup checks, associated passengers, checks against several watch lists (e.g. No-fly lists) and behavioural analysis are obligatory. (IATA, 2012)

5.2. Two exemplary scientific projects Research projects

Out of a variety of EU research projects, we selected two projects, which have a extraordinary futuristic focus of Big Data profiling: the INDECT and VIRTUOSO project. The INDECT project has been strongly debated because of its ambitious plans of behavioural prediction and the linkage of various kinds of data sources. The second, VIRTUOSO, intends to provide a tool for using a broad range of so called “Open Source Data”, i.e. data available in the Internet.[23]

22 See :https://panopticlick.eff.org/ for more information about browser identification possibilites.

23 This term has also repeatedly used during the European Police Congress in Berlin, February 19 & 20, http://www.european-police.eu/

i. INDECT

INDECT is a research project, involving European scientists and researchers, it was initiated by the Polish Platform for Homeland Security. The consortium consists of several European universities, surveillance technology companies, the Police Service of Northern Ireland and the Polish General Headquarters of Police[24]. The project started in 2009 and will end on December 31, 2013[25]. Its main objective is "to develop advanced and innovative algorithms for human decision support in combating terrorism and other criminal activities"[26]. INDECT is developing algorithms for automatic threat detection, recognition of serious criminal behaviour and recognise danger events.

In the category Intelligent Monitoring for Threat Detection, the project develops novel algorithms for automatic threat monitoring and detection. Therefore, "novel algorithms" to detect dangerous events, based on automatic object detection, object classification, analysis of interactions between objects are developed (Cetnarowicz, Dąbrowski, Pleva, Juhar, & Ondas, 2012). Moreover algorithms for collecting and processing open source information, e. g. news reports, blogs and chats, and algorithms based for pattern mining to detect suspicious websites are developed (Klapaftis, Manandhar, & Pandey, 2009; Pandey & Dorosz, 2012). Further algorithms are intended to apply machine learning on existing **behavioural profiling methods**, biometric and non-biometric, such as a novel algorithm for identifying sexual predatory conversations on public chatrooms (Pandey, 2012). For that data from sources like chats, blogs and other social networks are collected and modified for further processing. Afterwards it is analysed to identify whether a conversations involve illegal activity or not. Natural language processing tools are applied to extract sentences and named entities. To detect illegal behaviour discriminative behavioural profiling features are used. Content and stylistic features of texts and authors are analysed, based on the methodology of **authorship profiling**. For instance conversation partners' characteristics like age, habits, mental state etc. are derived. **Relationship mining** is a method based on pattern mining to automatically learn if there exist a relationships between entities, websites and documents, to extract named entities and to constitute the relationship between those entities (Klapaftis, 2012).

INDECT develops digital watermarking and cryptographic algorithms what is conceived as Data and Privacy Protection tools. Digital watermarking is applied to conceal the identities of monitored persons or objects for the case that unauthorised persons get access to the data, cryptographic algorithms support a secure data transmission and storage. Accordingly the project seems to use a very limited privacy notion which is mainly focused on data security but does not take into account general data protection principles. At least a user access management is intended with detailed logging of access (date, time and person), the effectiveness will depend on the concrete implementation.[27].

ii. VIRTUOSO

VIRTUOSO, Versatile InfoRmation Toolkit for end-Users oriented Open-Sources explOitation, is a research project developing a toolkit for European Security stakeholders. The project duration is from May 1, 2010 to April 30, 2013. It is coordinated by the Commissariat à l'Energie Atomique (CEA) in France, other participants include European and

24 http://www.indect-project.eu/faq

25 http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_LANG=EN&PJ_RCN=10374914&pid=15&q=FD8A9BBC079BD5FCECD584ADBD3CE6A7&type=adv

26 http://www.indect-project.eu/

27 http://www.indect-project.eu/faq

Swiss Universities and surveillance technology companies[28]. The project goals for the VIRTUOSO tool-kit are functionalities for extracting and exploiting “open source” information from the Internet for decision support. The components of the tool-kit are categorised in infrastructural and functional components. Infrastructural components ensure interactivity and collaboration of the functional elements, functional components are developed for finding, selecting and acquiring “open source” information. The main functional components consist of “information gathering components (acquisition), information extraction and structuring components (processing), knowledge acquisition components (knowledge management), decision support and visualization components”[29]. The project defines “Open source data” as “sources that are freely available to the public, i.e., information sources that have no access or other types of restrictions nor payment requirements.” (Koops, Cuijpers, & Schellekens, 2011)

A possible end-user scenario for the use of the VIRTUOSO platform envisaged is the detection of illegal cross-border migration into the EU via Greece (ibid.). VIRTUOSO could be applied to collect and analyse information from websites, blogs, media, academia and to link these information with components for facial recognition and translation. The data would then be added to the knowledge base and sent via email or sms to authorities involved in preventing illegal cross-border migration.

In regard to privacy, ethical and legal aspects, the VIRTUOSO project refers to 'Code as Code' or 'Code as Law'. A concept aiming on translation of regulations into the software of technology design. This approach is considered as effectively reducing the potential of abuse of the VIRTUOSO tool-kit, because in software code the ways of use can be variously constrained. With this approach, developers acknowledge that not only the end-user is responsible for possible violations of ethical and human rights, but that it is also the task of technology designers to minimize the potential of such violations. (Ibid.)

6. Conclusion

We revisited the evolution of profiling technologies starting from the digitization of data and information exchange which leads to a growing availability of digital data. Especially the increasing relevance of networked communication produces a new quality of data, because data is generated for administration purposes (logging data) or for interpersonal communication (user generated data) but is accessible for not-intended uses, most importantly for secondary data analysis. Accordingly nearly all interactions mediated through computers become at least traceable (metadata: e.g. who communicates with whom) if not accessible in more detail (content: e.g. what information is exchanged). Moreover digital data can be easily exchanged over computer networks and it has become difficult to keep control over data flows and to oversee which knowledge can be gained through the data one produces on a daily basis.

The availability of big amounts of data leads to an understanding of data as a resource which can be mined what is reflected in wording and the implementation of data analysis methods. For Knowledge Discovery and Data Mining key technologies used are pattern mining, classification and clustering. These approaches imply that irregularities are to be found which strengthens societal norms and the classification of deviants, and consequently entails a changing generation of suspicion. At the same time the norms underlying

28 http://cordis.europa.eu/search/index.cfm?fuseaction=proj.document&PJ_LANG=EN&PJ_RCN=11316357&pid=42&q=FD8A9BBC079BD5FCECD584ADBD3CE6A7&type=adv

29 http://www.virtuoso.eu/

the analysis are increasingly incorporated in data mining tools and algorithms. The norms and leading values potentially are hidden into the technologies but are conceived as objectified due to the neutral perception of technology and hard mathematics. Finally, this builds the ground for a changing understanding of profiling from assumption-driven to discovery-driven data analysis.

Coming back to the introductory section about historic concerns about the power of emerging computer systems it becomes clear that the spread of computer based technologies lead to an increasing control over subjects through information. The ideas of overcoming human deficiencies are still relevant as well as the idea of a technology as a neutral objectifiying tool.

The reduction of complexity through automated processing goes in line with a generation of complexity through the scope of possibilities to gain knowledge and the number of actors involved in the usage of data. For citizens it has become hard to understand which data is produced, who has the means and rights to store it and in which ways it can be analysed, more specifically what kind of knowledge can be derived from that data and how it affects future agency.

Beyond that, automated processing technologies imply intransparency. Human attitudes and assumptions, which always have been leading decision making, are translated into technology and become invisible and encapsulated in a conceived neutral objectivity of technology. Code as a new form of law (Hildebrandt, 2009; Lessig, 1999) regulates socio-technically mediated societies, and is maintained by a self-proclaimed technocrat-elite. Leading principles of societies become inscribed in code, e.g. democratic racism becomes translated in to profiling algorithms (Tator & Frances, 2006) as well as privacy is translated into technology (Cavoukian, 2009)

Of course, other things potentially do become more visible through technology: risk probabilities become computable, decisions based on data analysing algorithms become reproducible, deviant behaviour becomes traceable. Inequalities manifest in indirect discrimination measures or discriminatory algorithms and receive renewed attention. Societal principles like just punishment (Harcourt, 2008) and initial suspicion (Lianos & Douglas, 2000) are challenged.

Profiling technologies provide means of control which can be exercised for care and protection or coercion and repression. Control may be a necessary and inevitable function of social regulation, but it is vital to consider how to keep it democratic and empowering for citizen. (Monahan, 2010)

7. References

Bailey, Kenneth D. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques*. 102. Sage Publications, Incorporated. http://books.google.com/books?hl=de&lr=&id=1TaYulGjhLYC&oi=fnd&pg=PR5&dq=bailey+typologies+and+taxonomies&ots=LQ4NXF9hkP&sig=HHo_02waS6ww49rGdv80xCL7ES8.

Bernhard Anrig, Will Browne, and Mark Gasson. 2008. “The Role of Algorithms in Profiling.” In *Profiling the European Citizen*, edited by Mireille Hildebrandt and Serge Gutwirth. Springer Netherlands.

Van Brakel, Rosamunde, and Paul De Hert. 2011. “Policing, Surveillance and Law in a

Pre-crime Society: Understanding the Consequences of Technology Based Strategies." *Technology-Led Policing: Journal of Police Studies, Volume 2011-3* 20: 165.

Brown, Lesley. 2007. *Shorter Oxford English Dictionary*. OUP Oxford.

Calders, Toon, and Indrė Žliobaitė. 2013. "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures." In *Discrimination and Privacy in the Information Society*, edited by Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky, 43–57. Studies in Applied Philosophy, Epistemology and Rational Ethics 3. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-30487-3_3.

Calders, Toon, and Bart Custers. 2013. "What Is Data Mining and How Does It Work?" *Discrimination and Privacy in the Information Society*: 27–42.

Canada Border Services Agency. 2012. "Frequently Asked Questions about the NEXUS Program." http://www.cbsa-asfc.gc.ca/prog/nexus/faq-eng.html#a1.

Canhoto, Ana, and James Backhouse. 2008. "General Description of the Process of Behavioural Profiling." In *Profiling the European Citizen*, edited by Mireille Hildebrandt and Serge Gutwirth. Springer Netherlands.

Cavoukian, A. 2009. *Privacy by Design... Take the Challenge*. Information and Privacy Commissioner of Ontario, Canada.

CBP.gov. 2012. "SENTRI Program Description." http://www.cbp.gov/xp/cgov/travel/trusted_traveler/sentri/sentri.xml.

Cetnarowicz, Damian, Adam Dąbrowski, Matus Pleva, Jozef Juhar, and Stanislav Ondas. 2012. "D7.2 Creation of Event Model in Order to Detect Dangerous Events." INDECT.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. 2000. "CRISP-DM 1.0 Step-by-step Data Mining Guide." http://www.crisp-dm.org/CRISPWP-0800.pdf.

Clarke, Roger. 1988. "Information Technology and Dataveillance." *Communications of the ACM* 31 (5): 498–512.

Degele, Nina. 2000. *Informiertes Wissen: Eine Wissenssoziologie Der Computerisierten Gesellschaft*. Frankfurt am Main u.a.: Campus-Verl.

Ericson, Richard Victor, and Kevin D. Haggerty. 2006. *The New Politics of Surveillance and Visibility*. University of Toronto Press.

FBI. 2013. "Terrorist Screening Center - Vision & Mission." *FBI*. Accessed March 26. http://www.fbi.gov/about-us/nsb/tsc/tsc_mission.

FIDIS. 2005. "D7.2: Descriptive Analysis and Inventory of Profiling Practices."

http://www.fidis.net/fileadmin/fidis/deliverables/fidis-wp7-del7.2.profiling_practices.pdf.

Fox, Dirk. 2002. “Der IMSI-Catcher.” *Datenschutz Und Datensicherheit* (26).

Frawley, William J., Gregory Piatetsky-Shapiro, and Christopher J. Matheus. 1992. “Knowledge Discovery in Databases: An Overview.” *AI Magazine* 13 (3): 57.

Gandy Jr, Oscar. 2006. “Data Mining, Surveillance, and Discrimination in the Post-9/11 Environment.” *The New Politics of Surveillance and Visibility* 363: 363–64.

Gasson, Mark, and Will Browne. 2008. “Reply: Towards a Data Mining De Facto Standard.” In *Profiling the European Citizen*, edited by Mireille Hildebrandt and Serge Gutwirth. Springer Netherlands.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. University of California Press.

———. 1990. *The Consequences of Modernity*. Stanford University Press.

Haller, Stephan, Stamatis Karnouskos, and Christoph Schroth. 2009. “The Internet of Things in an Enterprise Context.” In *Future Internet – FIS 2008*, edited by John Domingue, Dieter Fensel, and Paolo Traverso, 14–28. Lecture Notes in Computer Science 5468. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-00985-3_2.

Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann. http://books.google.de/books?hl=de&lr=&id=AfL0t-YzOrEC&oi=fnd&pg=PP2&dq=han+and+kamber+%22data+mining:+concepts+and+techniques%22&ots=UvYTuTekF1&sig=qIcMH728bl_sLpAHeRGl8PfaUbI.

Hand, David J., Heikki Mannila, and Padhraic Smyth. 2001. *Principles of Data Mining*. MIT press. http://books.google.de/books?hl=de&lr=&id=SdZ-bhVhZGYC&oi=fnd&pg=PR17&dq=mannila+and+smyth+%22principles+of+data+mining%22&ots=ywQ3wpwmn2&sig=I1O1mKzKoaXmp3XG7FiskWff09I.

Harcourt, Bernard E. 2008. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press. http://books.google.com/books?hl=de&lr=&id=sqryh6ol1LwC&oi=fnd&pg=PR7&dq=Against+Prediction+harcourt&ots=vjCRj-ukh5&sig=0Sc-NwWsA6MKJiMV85dHmVAeVQw.

Harnisch, Stefanie, and Martin Pohlmann. 2009. “Strafprozessuale Maßnahmen Bei Mobilfunkendgeräten.” *HRRS - Onlinezeitschrift Für Höchstrichtertliche Rechtsprechung Zum Strafrecht* (5): 202–217.

Hempel, Leon. 2012. “Surveillance Studies.” In *Kultur*, 151–175. Bielefeld: Transcript.

Hempel, Leon, Michael Carius, and Carla Ilten. 2009. “Exchange of Information and

Data Between Law Enforcement Agencies Within the European Union". 29/09. ZTG Discussion Paper.

De Hert, Paul, and Serge Gutwirth. 2006. "Interoperability of Police Databases Within the EU: An Accountable Political Choice?" *International Review of Law Computers & Technology* 20 (1-2): 21–35.

Hildebrandt, Mireille. 2008. "Defining Profiling: A New Type of Knowledge?" In *Profiling the European Citizen*, edited by Mireille Hildebrandt and Serge Gutwirth, 17–45. Springer Netherlands. http://www.springerlink.com/content/r70n22p620k62301/abstract/.

———. 2009. "Technology and the End of Law." In *Facing the Limits of the Law*, edited by Bert Keirsbilck, Wouter Devroe, and Erik Claes, 1–22. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-79856-9_23.

IATA. 2012. "Checkpoint of the Future Executive Summary." http://www.iata.org/whatwedo/security/Documents/COF-Concept-Definition-Executive-Summary.pdf.

Johnson, Deborah, and Kent A. Wayland. 2010. "Surveillance and Transparency as Sociotechnical Systems of Accountability." In *Surveillance and Democracy*, edited by Kevin D. Haggerty and Minas Samatas. Taylor & Francis.

Klapaftis, Ioannis. 2012. "D4.5 Novel Algorithms for Relationship Mining Including Comparison with Existing Methods Indicated in D4.2." INDECT.

Klapaftis, Ioannis, Suresh Manandhar, and Shailesh Pandey. 2009. "D4.1 XML Data Corpus: Report on Methodology for Collection, Cleaning and Unified Representation of Large Textual Data from Various Sources: News Reports, Weblogs, Chat." INDECT.

Kling, Rob. 1994. "Reading 'All About' Computerization: How Genre Conventions Shape Nonfiction Social Analysis." *The Information Society* 10 (3): 147–172. doi:10.1080/01972243.1994.9960166.

Koops, B. J, Colette Cuijpers, and Maurice Schellekens. 2011. "D 3.2 Analysis of the Legal and Ethical Framework in Open Source Intelligence." VIRTUOSO.

Koops, Bert-Jaap. 2011. "Forgetting Footprints, Shunning Shadows: A Critical Analysis of the'Right to Be Forgotten'in Big Data Practice." http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1986719.

Krouse, William J., and Bart Elias. 2009. "Terrorist Watchlist Checks and Air Passenger Prescreening." CRS Report for Congress. Congressional Research Service.

Leenes, Ronald, and Bert-Jaap Koops. 2005. "'Code': Privacy's Death or Saviour?" *International Review of Law Computers & Technology* 19 (3): 329–340.

Lessig, Lawrence. 1999. *Code and Other Laws of Cyberspace*. New York: Basic Books.

Lianos, Michaelis, and Mary Douglas. 2000. “Dangerization and the End of Deviance The Institutional Environment.” *British Journal of Criminology* 40 (2): 261–278.

Lyon, David. 2001. *Surveillance Society*. Open University Press Buckingham. http://www.festivaldeldiritto.it/2008/pdf/interventi/david_lyon.pdf.

———. 2003. “Surveillance as Social Sorting. Computer Codes and Mobile Bodies.” In *Surveillance As Social Sorting: Privacy, Risk, and Digital Discrimination*. Psychology Press.

Marx, Gary T. 2002. “What’s New About the ‘New Surveillance’? Classifying for Change and Continuity.” *Surveillance & Society* 1 (1) (September 1): 9–29.

Marx, Gary T. 1990. *Undercover: Police Surveillance in America*. University of California Press.

Mattioli, Dana. 2012. “On Orbitz, Mac Users Steered to Pricier Hotels.” *The Wall Street Journal*, June 26.

Monahan, Torin. 2009. “The Murky World of ‘Fusion Centres’ Torin Monahan Critiques the Emergence of Data-sharing ‘Fusion Centres’ Intended to Reduce Crime and Prevent Terrorism.” *Criminal Justice Matters* 75 (1): 20–21.

———. 2010. “Surveillance as Governance: Social Inequality and the Pursuit of Democratic Surveillance.” *Surveillance and Democracy*: 91–110.

Occhipinti, Jonh D. 2003. *The Politics of Eu Police Cooperation: Toward a European Fbi?* Lynne Rienner Publishers.

Oracle. 2008. *Data Mining Concepts*. Accessed March 27, 2013. http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/toc.htm.

Pandey, Suraj Jung. 2012. “D9.26 Novel Algorithms for Behavioural Profiling and Comparison with Baseline Systems Developed in 4.7.” INDECT.

Pandey, Suraj Jung, and Krzysztof Dorosz. 2012. “D4.11 Specification of Methods for Mining and Detecting Suspicious Websites.” INDECT.

Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin. http://books.google.com/books?hl=de&lr=&id=-FWO0puw3nYC&oi=fnd&pg=PT3&dq=The+Filter+Bubble&ots=g2MrCrsVOX&sig=YGhX3HqwRBiYMJ2eN028L-8gBZA.

Polizei Nordrhein-Westfalen. 2013. “Häufig Gestellte Fragen Zur Datei Gewalttäter Sport.” Accessed March 27. http://www.polizei-nrw.de/artikel__4596.html.

Raabe, Oliver, Peter Georgieff, Daniel J. Koch, and Peter Neuhäusler. 2010. *Ubiquitäres Computing: Das “Internet der Dinge” - Grundlagen, Anwendungen, Folgen*. edition sigma.

Rocco Bellanova, Matthias Vermeulen, Serge Gutwirth, Rachel Finn, Paul McCarthy, David Wright, Kush Wadhwa, et al. 2011. “Deliverable 1.1 - Smart Surveillance - State of the Art.” SAPIENT. FP7 Sapient Project, Brussels. http://www.sapientproject.eu/docs/D1.1-State-of-the-Art-submitted-21-January-2012.pdf.

Sächsisches Staatsministeriums. 2011. “Bericht Über Die Erhebung Und Verwendung Der Gemäß § 100g Strafprozessordnung I. V. M. § 96 Telekommunikations-Gesetz Vorliegenden Datenbestände Im Zusammenhang Mit Dem Ermittlungsverfahren Zur Verfolgung Der Am 19. Februar 2011 in Dresden Begangenen Straftaten”. Sächsisches Staatsministerium der Justiz und für Europa und Sächsisches Staatsministeriums des Innern. http://www.sachsen.de/download/Gemeinsamer_Bericht_zur_Funkzellenauswertung.pdf.

Sapir, Yoav. 2008. “Against Prevention? A Response to Harcourt’s Against Prediction on Actuarial and Clinical Predictions and the Faults of Incapacitation.” *Law & Social Inquiry* 33 (1): 253–264.

sosadmin. 2013. “Your Phone May Not Be Safe at Protests.” Accessed March 26. http://privacysos.org/node/737.

Stytz, Martin R., and Roland L. Trope. 2008. “Digital Rights Management and Individualized Pricing.” http://www.stttelkom.ac.id/staf/faz/DRM/papers/msp2008030076.pdf.

sueddeutsche.de. 2010. “Staatsmacht gegen Militante: Ruf nach Autonomen-Datei wird lauter.” *sueddeutsche.de*, May 17, sec. politik. http://www.sueddeutsche.de/politik/staatsmacht-gegen-militante-ruf-nach-autonomen-datei-wird-lauter-1.414542.

Swan, Melanie. 2012. “Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0.” *Journal of Sensor and Actuator Networks* 1 (3): 217–253.

Tator, Carol, and Henry Frances. 2006. *Racial Profiling in Canada: Challenging the Myth of “a Few Bad Apples”*. University of Toronto Press.

Toubiana et al. 2010. “Adnostic: Privacy Preserving Targeted Advertising.” *Proceedings Network and Distributed System Symposium* (March).

Weizenbaum, Joseph, and John McCarthy. 1977. “Computer Power and Human Reason: From Judgment to Calculation.” *Physics Today* 30: 68.

7.1. Figures:

Opendatacity 2013: http://apps.opendatacity.de/stasi-vs-nsa/english.html [accessed July 2013], Creative Commons License CC-BY 3.0

FBI 1945, retrieved from

http://commons.wikimedia.org/wiki/File:Fingerprinting_at_the_federal_armory_during_WWII_%E2%80%94_National_Guard_Amory,_Fingerprinting_Division,_92nd_street,_Washington,_D.C._-_1945.jpg [accessed July 2013], Public Domain