Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR

Nils Gruschka[†], Vasileios Mavroeidis[†], Kamer Vishi[†], Meiko Jensen[*]
[†]Research Group of Information and Cyber Security, University of Oslo, Norway
Email(s): {nilsgrus, vasileim, kamerv}@ifi.uio.no
[*]Faculty of Computer Science and Electrical Engineering, Kiel University of Applied Science, Germany
Email: meiko.jensen@fh-kiel.de

Abstract—**Big data has become a great asset for many organizations, promising improved operations and new business opportunities. However, big data has increased access to sensitive information that when processed can directly jeopardize the privacy of individuals and violate data protection laws. As a consequence, data controllers and data processors may be imposed tough penalties for non-compliance that can result even to bankruptcy. In this paper, we discuss the current state of the legal regulations and analyse different data protection and privacy-preserving techniques in the context of big data analysis. In addition, we present and analyse two real-life research projects as case studies dealing with sensitive data and actions for complying with the data regulation laws. We show which types of information might become a privacy risk, the employed privacy-preserving techniques in accordance with the legal requirements, and the influence of these techniques on the data processing phase and the research results.**

Keywords—**big data; data analysis; privacy; data protection; GDPR; data anonymization; information security, biometric privacy**

I. Introduction

The term big data describes large or complex volumes of data, both structured and unstructured that can be analysed to bring value. The typical definitions (e.g., by NIST [1] or Gartner [2]) refer to big data by a number of *V*-properties, such as volume, velocity, and variety. Today, big data has become capital, with enterprises improving substantially their operations and customer relations, and the academia developing and enhancing research (e.g., in climate [3] or biology research [4]). In addition, the huge amount, generation speed, and diversity of data require special architectures for storage and processing (e.g., MapReduce [5] or Apache Hive[1]).

While the usefulness of processing big data is mainly unquestioned, it also comes with high privacy risks when operating on personal data. This is mainly due to two aspects of big data analysis. First, the larger the amount of data the higher the probability of re-identifying individuals even in datasets which seem not to have personal linking information. Second, big data analysis is able to infer from "harmless" personal data new information that is much more critical and was not intended to be revealed by the affected person. A famous example is the analysis of shopping patterns for creating customized (targeted) advertisements by a department store, where the algorithms correctly inferred that a teenage girl was pregnant [6]. There are areas where privacy threats may become even more critical, such as in medical treatment or research [7].

In order to protect individuals and their data a number of technical means and regulations for privacy-preserving data processing have been initiated and developed. However, implementing these methods in a data processing system obviously requires additional effort during the design phase, and in many cases such methods influence the performance of the system. As a result, in the past, enterprises and other organizations were not always willing to make this effort, but this tend to change due to the pressure applied from new privacy laws and regulations.

This paper describes privacy issues in big data analysis and elaborates on two case studies (government-funded projects[2,3]) in order to elucidate how legal privacy requirements can be met in research projects working on big data and highly sensitive personal information. Finally, it discusses resulted impacts on the processing of data and the results due to the employed privacy-preserving techniques.

The paper is organized as follows: Section II presents the current state of legal and technical aspects related to processing personal information. Section III presents and analyzes two research projects operating on large datasets containing personal information. In Section IV, we discuss the influence the utilized privacy-preserving techniques had on the data processing and results of the projects. Finally, Section V concludes the paper.

II. Privacy Issues in Big Data Analytics

A. Legal Regulations

From a legal point of view, in this paper we focus on the EU *General Data Protection Regulation* (GDPR) [8], which came into force in May 2018. It is relevant to all organizations inside the European Union (EU), the European Economic Area (EEA) and also to organizations from other countries, if they process data of European citizens. Thus, the GDPR has effect on most major companies worldwide.

[1] https://hive.apache.org/

[2] https://www.mn.uio.no/ifi/english/research/projects/oslo-analytics/

[3] https://www.ntnu.edu/iik/swan

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

The GDPR regulates the collection, storage, and processing of personal data. Personal data are any data that can be linked to a specific natural person. This includes not only direct personal identifiers (e.g., full name, national ID number) but also indirect identifiers like phone numbers, IP addresses, or photos with identifiable people. Data that do not include such identifiers are commonly regarded as *anonymous* and are outside the scope of GDPR (Recital 26). The results of big data analysis are very often statistical findings without direct links to specific individuals. Hence, a simple method to conform to all requirements of GDPR is to process only anonymous data. However, the definition of anonymity is not trivial. Even if directly identifiable parameters are removed from a dataset, it might be possible to *re-identify* single individuals by combining the dataset with other information. This approach for de-anonymization is called *background knowledge attack* [9], [10].

A famous example of re-identification is the Netflix challenge in 2006. As part of a competition for finding more accurate movie recommendation methods, Netflix released a dataset containing movie ratings of 500,000 customers. In the dataset, any personally identifiable information (PII) was removed and only subscriber IDs (without any connection to the actual identity) and movie ratings (score, movie info, date) were published. However, researchers combined these data with other publicly available information (e.g., IMDB ratings) and were able to identify individual customers with a high probability [11]. Other well-known cases include identification of individuals from internet search terms [12], anonymized DNA [13] and mobility data [14].

There are numerous formal metrics for measuring the degree of anonymity of a dataset (see next section). GDPR without giving a precise or concrete definition of anonymity considers a dataset anonymous when re-identification is only possible with high effort or unlikely means.

For processing personal data the GDPR defines a number of legal, organizational and technical requirements, and proposes different methods. The most relevant principles are described here. First of all, in most cases, processing of personal data is allowed only if the data subject has given its *consent* (Article 6)[4]. Exceptions apply when the data processing is explicitly allowed by a law or regulation, or ensures ”vital interests of the data subject”. Additionally, the consent given must be limited to a specific purpose for data processing (Article 5)[5]. The data controller (the entity that is responsible for collecting the data) can neither define a too generic data processing purpose nor change the purpose later arbitrarily (see Figure 1).

Another data processing principle is *data minimization* (Article 5) which refers to limiting personal data collection, storage, and usage to data that are relevant, adequate, and more importantly necessary for carrying out the purpose for which the data are processed. Worthy of noting is that *pseudonymization* is explicitly mentioned as a data minimization measure. In pseudonymized data, identifiable parameters are replaced by other (randomly) generated identifiers. This usually does not have any negative impact on the data mining process and preferably should be initiated by the data controller before transferring the data to the data processor. If the data processing results require to be linked this can be achieved by the data controller, as it holds the mapping (also known as *a pseudo-lookup table*) pseudonyms to the identifiable parameters. In addition to the anonymized data, storage at the data controller must conform to the GDPR by employing techniques that protect the data in rest (e.g, encryption and tight access control). Furthermore, the GDPR requires ”*appropriate technical and organizational measures to ensure a level of security appropriate to the risk*” (Article 32)[6], which commonly includes the application of techniques like data encryption, access control, physical protection and (again) pseudonymization.

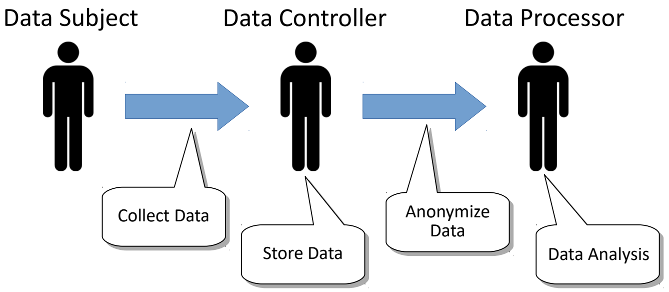

Fig. 1. Personal data handling process

An extension to the data minimization principle is the *storage limitation* principle, which restricts the duration of data storage to a specified (necessary) period.

In the context of data processing, it must be further taken into account that automatic decision-making processes with impact on individuals (Article 22)[7] as well as processing of extremely sensitive data, such as biometric data (Article 9)[8], requires “*explicit*” consent from the data subject.

The terms big data or data analysis are not addressed by the GDPR directly. However, it is clear from the description above that big data and the GDPR are not always compatible [15]. For example, big data mining relies on the analysis of large amounts of data, which often contradicts the principle of data minimization. In addition, in data analysis very often new hypotheses for testing are introduced after the data were collected. However, the data subjects from which the data were collected have given consent initially for a different purpose. Thus, from a legal perspective data processing should be done—if possible—on anonymized data, otherwise great care must be taken that the GDPR is respected. This for example might require a *data protection impact assessment* (DPIA); a privacy-related impact assessment whose objective

[4] Article 6 of the GDPR regulates the lawfulness, fairness and transparency of collecting and processing personal data.

[5] Article 5 of the GDPR regulates principles relating to processing of personal data.

[6] Article 32 regulates the security of data processing.

[7] Article 22 regulates automated decision-making and profiling.

[8] Article 9 regulates the processing of special categories of personal data.

is to identify and analyse how data privacy might be affected by certain actions or activities (Article 35)[9].

The fines that can be levied as penalties for non-compliance are based on the specific articles of the Regulation that the organization has breached. The GDPR also gives individuals the right to compensation for material and/or non-material damage resulting from an infringement of the Regulation. Specifically, data controllers and processors face administrative fines of

- up to 10 million Euros or 2% of annual global turnover (whichever amount is higher) for infringements of articles: 8 (conditions for childrens consent), 11 (processing that does not require identification), 25-39 (general obligations of processors and controllers), 42 (certification), and 43 (certification bodies).
- up to 20 million Euros or 4% of annual global turnover (whichever amount is higher) for infringements of articles: 5 (data processing principles), 6 (lawful bases for processing), 7 (conditions for consent), 9 (processing of special categories of data), 12-22 (data subjects rights), and 44-49 (data transfers to third countries).

B. Technical Aspects

The legal requirements presented in the previous section must be implemented by technical means. This section presents some methods for privacy-preserving data mining. Well-known early approaches in this area are the works of Agrawal and Skrikant [16], and Lindell and Pinkas [17]. In the first one, data were anonymized by distortion, and a special decision tree classification analysis was performed on the anonymized data. In the second one, the data were split over two separate databases (which can be seen as a type of pseudonymization) and a special multi-party computation algorithm was developed for analysing the dataset. This shows the typical parts of privacy-preserving data analysis: anonymization (as effective as possible; at least pseudonymization) and potentially mining algorithms adapted to this kind of modified data.

To support the anonymization process the attributes in a dataset are usually divided into four distinct categories [18]:

- *Explicit identifiers*: attributes that each directly link to a single individual, like social security number or email address.
- *Quasi-identifiers*: attributes that do not directly link a person, but can re-identify an individual when the values of multiple attributes are combined. Examples are: date of birth, ZIP code or profession.
- *Sensitive information*: attributes containing information the data subject does not want to be revealed or at least not be linked with its person. Examples might be: diseases, financial situation, sexual orientation, current position.
- *Non-sensitive information*: attributes that do not fall in any of the aforementioned categories (e.g., weather data).

Unfortunately, this categorization is not always obvious: a (rare) disease might also identify a person. Also inside the categories there are large differences: the place of residence as a quasi-identifier can refer to millions of people if referring to a large city, but can also point to only a handful of individuals for small villages. To quantify the rate of anonymity and thereby the threat regarding re-identification, a number of anonymity models exist. The most common approaches are *k-anonymity* [19], *l-diversity* [9], *t-closeness* [20] and *differential privacy* [21].

It is a fact that most of the times sensitive information is of high value for data mining, but also for adversaries. If the linkage between explicit identifiers and sensitive information is the goal of an analysis, obviously anonymization is not possible and the GDPR must be regarded. In this case, at least pseudonymization should be applied. Very often, however, data mining is looking for connections between quasi-identifiers and sensitive attributes, which allows anonymization of data. Common anonymization methods are [22]:

- *Suppression*: removing the values of an attribute completely or replacing them with a dummy value (typically an asterisk "*"). This operation is usually performed on explicit identifiers.
- *Generalization*: replacing values with more general or more abstract values inside the attribute taxonomy, for example, date of birth → age (in years); age (in years) → a range of years; ZIP code → first two digits of the ZIP code. This operation is usually performed on quasi-identifiers.
- *Permutation*: partitioning the data into groups and shuffling the sensitive values within each group. As a consequence, the relationship between quasi-identifiers and sensitive data is eliminated.
- *Perturbation*: replacing values in a way that linkage to the original data is removed, but keeping the statistical properties similar. A typical method for perturbation is adding noise [11].

The models and anonymization techniques presented are not just of academic interest, but are used in practical privacy guidelines (e.g., the Norwegian data protection authority) [23]. The aforementioned anonymization operations obviously cause a loss of information and reduce the *utility* of the data [24]. Using the metrics for utility and anonymity (see above), one can evaluate different anonymization approaches and find the trade-off between privacy and utility.

Data mining on anonymized datasets sometimes requires specially adapted mining algorithms. Typical examples of classification and clustering algorithms for which privacy-preserving versions exist are decision trees, Bayesian classification, support vector machines (SVM) and secure multi-party computation [22].

[9] Article 35 regulates the data protection impact assessment (DPIA).

III. Case studies: GDPR and Real-life Research Projects

A. Oslo Analytics

The *Operable Subjective Logic Analysis Technology for Intelligence in Cybersecurity*[2] is a research project (project number: 247648) funded under the ICT and Digital Innovation program of the Research Council of Norway for the University of Oslo for the period of 2016 – 2019. Oslo Analytics develops advanced analytical methods based on big data analysis, machine learning and subjective logic [25] to gain a deep situational awareness and understanding of security incidents. The project is organized in collaboration with national and international institutions, organizations and security vendors, such as mnemonic, the Norwegian Computing Center (NR), the Norwegian National Security Authority (NSM), The Defence Intelligence College, the US Army Research Labs, and the Technische Universität Darmstadt.

Oslo Analytics needs to conform to the GDPR and the Personal Data Act of 2000 (managed by the Norwegian Data Protection Authority - *Datatilsynet*). The Norwegian Centre for Research Data (NSD) is responsible for implementing the statutory data privacy requirements in the research community, and thus requires notification from every research project processing personal data that are not fully anonymized. Fully anonymous data are information that cannot in any way identify an individual either directly through name and national identity number, or indirectly through background variables, a name list, scrambling key, encryption formula or code.

1) Handling Sysmon Data - End Point Security: Data of particular research importance for Oslo Analytics are Sysmon logs. Sysmon is a Windows system service and device driver that monitors and logs system activity of Windows workstations, servers and domain controllers. Sysmon provides some of the most effective events needed to trace attacker activity and increase host visibility. For example, Sysmon event class ”Process Create” with ID 1 can detect initial infection and malware child processes by capturing hashes. Sysmon event class ”File Creation Time Changed” with ID 2 can detect anti-forensic activities, such as changes in the file creation time of a backdoor to make it look like it was installed with the operating system. Sysmon event class ”Network Connection” with ID 3 can be used to identify network activity, such as connections to command and control servers (C&C) or even download encryption keys from ransomware servers. Research on Sysmon aims to reduce the cumbersome process of investigative analyses (threat hunting with NoSQL database systems or graph databases) by providing new complementary means based on Artificial Intelligence (e.g., ontologies [26]) and specifically machine learning.

Like many other datasets, Sysmon contains multiple privacy-sensitive identifiers (Windows account usernames, computer names, static internal IPs) and user-behaviour (running processes, internet activity) that Oslo Analytics has to deal with prior processing. For example, sensitive fields in events with ID 1 include *computer name*, *command line*, *current directory*, *user*, *parent image*, *parent command line*. All the aforementioned fields can reveal the identity of the user either directly or indirectly. In addition, complete removal of the aforementioned information (fields) would disallow researching and experimenting with technologies such as natural language processing narrowing down our options to more simplistic and less effective approaches.

A fallacy identified in the very early stage was the hashing of computer names in the dataset to keep the mapping between parent and child processes, as well as the time-sequenced activity of computers. This approach (hashing computer names) could allow re-identification of the original computer names and consequently the users operating the computers and their activity by re-hashing the computer names found in new Sysmon data [27]. Thus, for keeping the computer activity linkability in the dataset we generated unique integer identifiers that replaced the computer names, without keeping any mapping between them.

2) Data Storage and Accessibility: The data are stored on a secure server with access restricted to authorized researchers working on Oslo Analytics under a very tight access control list adopting the principle of least privilege. Processing of the data can only occur on the server. Access to the secure server is only allowed from inside the organizational network and this is restricted to specific computers filtered by their MAC addresses, their internal static IP, and user account. In addition, a firewall has been configured to allow only incoming connections to the server on port 22 (SSH). Any other network activity is denied and consequently dropped. In this respect, the network restrictions disallowed us to personally install any extra programming libraries needed for processing the data after setting up the server. Thus, we had to inform the security team that is responsible for the security of the server and the data stored. Finally, the user accounts for processing the data on the secure server are only valid for the duration of the project (account expiration), meaning that the accounts will be disabled on a specific date. The same principle applies to the Sysmon data which restricts the duration of the data storage to the active period of the project.

3) Trade-off between Security, Reproducibility and Dataset Availability: Reproducibility provides transparency to data analyses and allows the transfer of knowledge to others who could learn from your data and methods. Reproducible research demands that data analyses and scientific claims are published with their raw data and software code so others interested may verify the findings and build upon them.

Even anonymized Sysmon datasets can be a great source of information for any malicious actor interested to harm an organization. It can be used to identify vulnerable and unpatched applications running on workstations and servers, to determine the version of Windows operating systems running in the organizational environment, to process network activity, to identify file names, etc. Network activity even anonymized can be used with various success for phishing attacks since insights for the most visited domains can be obtained. In case of re-identification network activity can be used successfully for crafting more targeted phishing attacks. For the aforementioned reasons Oslo Analytics could not make publicly available the anonymized Sysmon dataset used in the research.

B. SWAN

The *Secure Access Control over Wide Area Network (SWAN)*[3] is a research project funded by the Research Council of Norway (Grant number: IKTPLUSS 248030/070) for the Norwegian University of Science and Technology (NTNU) for the period of 2015 – 2019. SWAN is composed of the following six partners; NTNU as the coordinator of the project (Norway), the University of Oslo (Norway), the IDIAP Research Institute (Switzerland), the Association of German Banks (Germany), IDEMIA (France), and Zwipe (Norway).

The SWAN project develops authentication technologies for banking and other services by using biometric identifiers. Biometric references (also known as templates) are stored, controlled and verified locally (e.g., smartphones) based on a pre-shared secret, which can be used to seal and authenticate transaction data. This overcomes the need of centralized storage for the biometric data [28]. The SWAN biometric authentication solutions are designed to be privacy compliant and align with existing and emerging biometric standards.

SWAN needs to conform to the GDPR and the Norwegian Personal Data Act in the same way as Oslo Analytics. The creation of the biometric dataset has been permitted by the Data Protection Official for Research (NSD).

1) Data Collection, Processing, and Storage: Clause 1 of Article 9 of GDPR states that biometric data are to be considered a *"special category of personal data"* and are prohibited from being used for identifying individuals, unless the data subjects have given explicit consent. In the first phase SWAN had to collect biometric data from 200 people (data subjects).

Biometric Data Collection Consent: The SWAN team created a dedicated Biometric Information Privacy Policy to comply with the Privacy Act and lawsuits that were into force in 2015 (before GDPR came into effect). The policy includes the following sections and clauses:

- Definition of "biometric identifier" and "biometric information"
- Consent
- Disclosure
- Storage
- Retention Schedule

The data subjects were asked to aid in the construction of a biometric dataset which will be used for research purposes related to biometrics recognition and presentation attack detection (PAD) for face, voice, eye and fingerprint biometric characteristics. Prior to handing over any biometric data all participants (data subjects) signed a consent form and were informed both orally and written about the purpose of the collection.

It is worth mentioning that in the consent form it is clearly stated that in case the data will be used for new purposes the data subjects will be asked to assess and sign a new consent form.

The creation of the SWAN database is in accordance with the aforementioned GDPR data processing techniques, such as pseudonymization, meaning that the personal data could allow the re-identification of individuals when required. There are three main reasons for using pseudonymization measures for constructing this biometric database. First, the pseudo ID can be used to facilitate the destruction of data in the case of participation withdrawal from the project. (Article 7)[10] clause 3 of the GDPR specifies that *"consent can be withdrawn at any time"*. In such cases, all personal data including biometric identifiers related to the data subject are permanently deleted. Secondly, if the database holding pseudonymous data together with biometric characteristics is compromised, the attackers would not have the ability to look up the pseudo value and identify the data subjects. Thirdly, pseudonymization enables big data analysis without access to the raw data that contains sensitive personal information (biometric characteristics in this case). Since each data controller (project partner) have its "own" unique key, data cannot easily be linked among different data controllers, thus, further reducing the risk of re-identification, while affording the sharing of dedicated pseudonymous datasets (dedicated to processing for a specific purpose by an identified data controller).

- *Data collection:* is performed from all partners participating in the project. All the data subjects have used a purpose-built application on a smartphone to capture images and video recordings of their face, eyes and fingers, and audio recordings of their voice. In addition, the participant's name, email, gender, and age will be stored along with a pseudo ID, linking to the biometric data.
  It is worthwhile to mention that during the biometric data collection (voice data collection phase) the participants had to say four sentences: 1) "My name is $\underline{A}$, and I live in $\underline{B}$", 2) "My bank account number is $\underline{C}$", 3) "The limit of my card is 5000 Euros", and 4) "My PIN code is 9, 8, 7, 6, 5, 4, 3, 2, 1, 0". Where $\underline{A}$ indicates a fictitious name, $\underline{B}$ indicates a fictitious address, and $\underline{C}$ indicates a fictitious bank account number. Name, address, and the bank account number are considered to be PII (personal identifying information) according to the GDPR. Therefore, these fields contain pseudo-identifiers (fiction data) generated by a random data generator in order to comply with the GDPR pseudonymization methods.
- *Data storage:* the collected biometric data (pseudonymized) are shared among the SWAN partners, stored securely, and raw data are only accessible to researchers participating in the project from the aforementioned project partners.
- *Data processing:* all the partners working on the project are able to process the SWAN database based on their

[10] Article 7 of the GDPR regulates the conditions for consent.

needs for specific work packages defined in the project description.

All project partners that collected biometric data are responsible for their collected sub-dataset (data controllers), in terms of processing and storing the biometric data. Additionally, NTNU as a project leader serves as the main data controller.

Biometric data may also be shared through the BEAT platform[11], a research platform facilitating open research without compromising security and privacy of data as no access to raw data is given. The SWAN project is scheduled for completion during the 4th quarter of 2019, however in the SWAN project's consent form is specified that the collected data may be stored after the completion of the project for an additional maximum period. This would require the data subjects to sign a new consent form.

In compliance with the GDPR, personal data must be kept *”no longer than is necessary for the purposes for which the personal data are processed”* (Article 5 clause 1(e)). However, Article 5 also provides an exception to this rule allowing extensive data retention insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes subject to implementation of the appropriate technical and organizational measures required by the GDPR in order to safeguard (Article 89)[12] the rights and freedoms of individuals.

2) Privacy-Preserving Biometrics: Since biometric data are highly sensitive and cannot be easily changed, there is a need for privacy-preserving solutions to avoid misuse, loss or theft. The SWAN project applies biometric template protection methods to secure biometric references stored locally on smartphone devices. This can prevent misuse of biometric data in case of data theft. Biometric template protection can also prevent linking a user's biometric characteristics between different databases (cross matching), thereby preserving the privacy of the user. In addition, the SWAN project applies novel *cancellable biometric* techniques (biometric template protection using Bloom Filters [29]). Cancellable biometrics provide an intentional, systematic and repeatable distortion of biometric features in order to protect user's sensitive data. For example, if a ”cancellable” characteristic is stolen, the distortions provided are modified and remapped to a new template which will replace the one that has been compromised.

IV. Lessons Learned

We have described two rather different projects dealing with sensitive large datasets. The SWAN project processes biometric data. This kind of data cannot be anonymized as the biometric samples are personally identifiable information and, thus, the GDPR applies. The project applied the following privacy protection methods:

- *Explicit consent*: the participants were informed in detail about the processing steps executed on their biometric data and had to explicitly agree on this.
- *Security of the processing system*: the data are protected from unauthorized access using access control and encryption technologies.
- *Pseudonymization*: to impede re-identification the mapping between biometric data and the owner (directly identifying information) is replaced by a pseudonym.
- *Processing biometric templates*: SWAN applies different techniques to protect the biometric templates, such as cancellable biometrics which allow the revocation of a compromised biometric template. This technique does not remove the link to the data subject completely, but makes re-identification much harder.
- *Limited storage duration*: SWAN defines a maximum period for storing the biometric data that extends beyond the duration of the project in case additional research needed to be conducted. The dataset will be stored only for this period (and maximum up to the predefined period stated in the signed consent form) and then will be deleted. In addition, if the research deviates even slightly from the original purpose this would require the data subjects to assess and sign a new consent form.

The aforementioned methods allow SWAN to comply with the GDPR, and consequently utilize the collected data for research purposes. It is worth mentioning that under GDPR projects that collect and/or process biometric data should carry out a *data protection impact assessment (DPIA)*.

The Oslo Analytics project does not operate on data that were collected explicitly for research purposes, but diversified data used in security operations. Thus, for the data processed the subjects have only given consent for purposes which are required for monitoring and protecting the network from major disruptions and attacks. As explained before, this is a typical situation in projects dealing with big data. For conforming to the legal requirements Oslo Analytics applied the following privacy protection methods:

- *Anonymization*: all data fields which allow easy re-identification of the subject have been removed (*suppression*) or abstracted (*generalization*). This does not only apply to directly identifying data like usernames but also to fields like internal IP addresses.
- *Security of the processing system*: like in the SWAN project, access to the processing system was strictly controlled and restricted.

Like mentioned before, re-identification might be possible in large anonymized datasets. Nonetheless, the datasets used for processing should fulfill the GDPR with its rather weak (not a strong formal definition) anonymization requirement.

Like in the SWAN project, projects that operate on data which are sensitive, processed on a large scale, and fall under the special categories referred to in Article 9, clause 1 of the GDPR should carry out a data protection impact assessment (DPIA). In addition, it is the case that anonymizing methods

[11] https://www.beat-eu.org/platform/

[12] Article 89 sets out safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

can prohibit the use of specific technologies, such as natural language processing which would be beneficial for improving the successfulness and potentially the results of the research.

V. Conclusion

This paper presented the implications of data protection laws on projects dealing with big data, and by using two case studies analysed how privacy-preserving techniques can be applied. The results were quite different. In one project, for mitigating privacy concerns regarding biometric data collection and processing the participants were asked to give consent. In addition, no problems were faced during the data analysis phase. In the second project, data from an existing data source were used. Here, anonymization of many data fields was required, making the data analysis more challenging and in many cases limited. It is of great importance to remark that for projects and technologies dealing with sensitive data a data protection impact assessment should be conducted at the very early stages of the project to identify potential privacy challenges, and to adapt the analysis methods taking into consideration privacy-preserving techniques.

Acknowledgement

This research was partially supported by the Research Council of Norway under the Grant No.: IKTPLUSS 247648 and 248030/O70 for Oslo Analytics and SWAN projects, respectively.

References

[1] NIST Big Data Public Working Group Definitions and Taxonomies Subgroup, "NIST Big Data Interoperability Framework: Volume 1, Definitions," National Institute of Standards and Technology, Tech. Rep. NIST SP 1500-1r1, Jun. 2018. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1.pdf

[2] Gartner IT Glossary, "What Is Big Data?" 2018. [Online]. Available: https://www.gartner.com/it-glossary/big-data

[3] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate Data Challenges in the 21st Century," *Science*, vol. 331, no. 6018, pp. 700–702, Feb. 2011.

[4] V. Marx, "Biology: The big challenges of big data," Jun. 2013. [Online]. Available: https://www.nature.com/articles/498255a

[5] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[6] K. Hill, "How Target figured out a teen girl was pregnant before her father did," *Forbes, Inc*, 2012.

[7] M. Mostert, A. Bredenoord, M. Biesaart, and J. J. Delden, "Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach," *European Journal of Human Genetics*, vol. 24, no. 7, pp. 956–960, Jul. 2016. [Online]. Available: https://www.nature.com/articles/ejhg2015239

[8] European Parliament and Council, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)," May 2016. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj/eng

[9] A. Machanavajjhala, M. Venkitasubramaniam, D. Kifer, and J. Gehrke, "L-Diversity: Privacy Beyond k-Anonymity," in *22nd International Conference on Data Engineering (ICDE'06)(ICDE)*, 2006, p. 24. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICDE.2006.1

[10] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '11. New York, NY, USA: ACM, 2011, pp. 193–204. [Online]. Available: http://doi.acm.org/10.1145/1989323.1989345

[11] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, May 2008, pp. 111–125.

[12] M. Barbaro and T. Z. Jr, "A Face Is Exposed for AOL Searcher No. 4417749," *The New York Times*, Aug. 2006. [Online]. Available: https://www.nytimes.com/2006/08/09/technology/09aol.html

[13] J. Bohannon, "Genealogy Databases Enable Naming of Anonymous DNA Donors," *Science*, vol. 339, no. 6117, pp. 262–262, Jan. 2013.

[14] Y.-A. d. Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports*, vol. 3, p. 1376, Mar. 2013. [Online]. Available: https://www.nature.com/articles/srep01376

[15] T. Z. Zarsky, "Incompatible: The GDPR in the Age of Big Data," *Seton Hall L. Rev.*, vol. 47, p. 995, 2016.

[16] R. Agrawal and R. Srikant, "Privacy-preserving Data Mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 439–450. [Online]. Available: http://doi.acm.org/10.1145/342009.335438

[17] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," *Journal of Cryptology*, vol. 15, no. 3, pp. 177–206, Jun. 2002. [Online]. Available: https://doi.org/10.1007/s00145-001-0019-2

[18] M. B. Malik, M. A. Ghazi, and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects," in *2012 3rd International Conference on Computer and Communication Technology (ICCCT 2012)*, Nov. 2013, pp. 26–32. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICCCT.2012.15

[19] L. Sweeney, "k-Anonimity: A Model For Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, Oct. 2002. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218488502001648

[20] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, Apr. 2007, pp. 106–115.

[21] C. Dwork, "Differential Privacy: A Survey of Results," in *Theory and Applications of Models of Computation*, ser. Lecture Notes in Computer Science, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Springer Berlin Heidelberg, 2008, pp. 1–19.

[22] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," *IEEE Access*, vol. 2, pp. 1149–1176, 2014.

[23] Norwegian Data Protection Authority, "The anonymisation of personal data," 2017. [Online]. Available: https://www.datatilsynet.no/en/regulations-and-tools/guidelines/anonymisation/

[24] S. Yu, "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data," *IEEE Access*, vol. 4, pp. 2751–2763, 2016.

[25] A. Jøsang, R. Hayward, and S. Pope, "Trust network analysis with subjective logic," in *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*. Australian Computer Society, Inc., 2006, pp. 85–94.

[26] V. Mavroeidis and A. Jøsang, "Data-driven threat hunting using sysmon," in *In ICCSP 2018: 2018 the 2nd International Conference on Cryptography, Security and Privacy*. ACM, 2018. [Online]. Available: https://doi.org/10.1145/3199478.3199490

[27] M. Marx, E. Zimmer, T. Mueller, M. Blochberger, and H. Federrath, *Hashing of personally identifiable information is not sufficient*. Gesellschaft fr Informatik e.V., 2018. [Online]. Available: http://dl.gi.de/handle/20.500.12116/16294

[28] E. Kindt, *Privacy and Data Protection Issues of Biometric Applications: A Comparative Legal Analysis*, ser. Law, Governance and Technology Series. Springer Netherlands, 2013.

[29] M. Stokkenes, R. Ramachandra, K. B. Raja, M. Sigaard, M. Gomez-Barrero, and C. Busch, "Multi-biometric template protection on smartphones: An approach based on binarized statistical features and bloom filters," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2016, pp. 385–392.