Data Security on the Ground: Investigating Technical and Legal Requirements under the GDPR

Tina Marjanov*
tm794@cam.ac.uk
University of Cambridge
Cambridge, United Kingdom

Maria Konstantinou*
mariakonstantinou101@gmail.com
Vrije Universiteit Amsterdam &
Freshfields Bruckhaus Deringer
Frankfurt, Germany

Magdalena Jóźwiak
m.e.jozwiak@tilburguniversity.edu
Tilburg University
Tilburg, Netherlands

Dayana Spagnuelo
dayana.spagnuelo@tno.nl
Applied Cryptography and Quantum Algorithms, TNO
The Hague, Netherlands

ABSTRACT

The GDPR has been in force since 2018, but there is still uncertainty about how to comply with several of its provisions, including Article 32 which sets forth the requirements for data security. While scholars in this field have previously analysed the law or the industry standards, we use the fines imposed so far for violation of Article 32 as our primary data. We annotate and analyse technical and legal aspects of a representative subset of cases. Using clustering, four groups of cases with distinct characteristics emerge from our research. Three of the four groups of cases suffer from data incidents, but for different reasons: a targeted attack, non-technical human mistakes, or a combination of mistakes. The final group includes cases where no actual data incident happened, but fines were still imposed due to insufficient organisational measures and high risk or imminent harm to the data subjects. We uncover from the cases different measures that apply to each of the groups, ranging from compliance with the highest industry standards to organisational measures and enhanced internal privacy awareness.

KEYWORDS

GDPR, Article 32, security of processing, technical requirements, legal requirements, clustering

1 INTRODUCTION

The General Data Protection Regulation (GDPR)[1] was adopted in April 2016 and came into force in the European Union (EU) in May 2018. While it has been in force for more than four years now, there is still much uncertainty on how to meet its demands in practice. Most relevant from a technical point of view is Article 32 ("Security of processing") which stipulates that the data controller "shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk". The word *appropriate* is key; it is natural to expect higher security standards from a large corporation handling sensitive data (e.g., a hospital or a bank) than from a small business with a small number of employees and customers. The article gives some indication of the aspects that should drive the decision on appropriate measures, but it admits multiple interpretations. Article 32 is not only the most relevant, but also among the most commonly violated [3, 25]. Thus, when reading this provision, one question resonates: How to devise and put in practice concrete measures in order to guarantee the technical and legal demands for processing security?

If on one side we lack concrete guidelines on how to comply with the article's demands, on the other, information on what leads to non-compliance is already available: hundreds of fines have been imposed on the basis of violation of Article 32.[2] These fines contain a wealth of –yet underexplored– information about the legal interpretations of the Regulation in practice, frequent violation patterns, as well as suggestions by Data Protection Authorities (DPAs) on measures for compliance.

The goal of our research is to bridge the gap between the technical and legal interpretations and provide a more realistic view with practical recommendations for compliance with Art. 32. In the process, we also aim to uncover the relationships between the different types of data controllers and how the DPAs interpret appropriateness. To do so, we use DPA decisions (cases) as our primary data. This is a relatively novel starting point. We extract relevant information from a selected set of cases and analyse it from both technical and legal perspectives. By looking at cases from both a legal and computer science perspective, we provide an interdisciplinary framework. We identify the patterns and common pitfalls and systematise guidelines for compliance with the security of processing requirement set forth in Article 32 GDPR.

Our contribution to the literature is threefold. First, we provide an empirical answer to the key question of how DPAs interpret *appropriate* measures. We follow a data-driven approach, starting from the fines already imposed by DPAs for violation of Art. 32. This allows us to look at compliance with Art. 32 from a more practical point of view. Using clustering, we are able to group

[*] Both authors contributed equally to this research.

[1] 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), OJ 2016 L 119/1.

[2] As of May 2022, according to: *Privacy Affairs.*

Proceedings on Privacy Enhancing Technologies 2023(3), 405–417
© 2023 Copyright held by the owner/author(s).
https://doi.org/10.56553/popets-2023-0088

cases with similar characteristics. Among others, we find that how sensitive the data is, how vulnerable the data subjects are, and the type of mistake (human, organisational, technical) that lead to a breach, are all important factors that affect how the DPAs treat cases. This is our main contribution. Second, we contribute to the computational legal scholarship by providing a proof-of-concept on how to utilise existing cases, which can be used for further analysis by scholars assessing GDPR's effectiveness. Third, our results have useful implications for practitioners either operating in data handling or advising data controllers. More specifically, the danger points we discuss for each group provide a checklist that practitioners may use to prioritise implementing measures in order to achieve compliance.

The paper is structured as follows: section 2 provides some context on GDPR, particularly on the security of processing requirement under Art. 32, and section 3 presents related research in data security within the scope of GDPR. In section 4 we describe the case selection criteria as well as the methodology we use in our analysis. We define and explain the variables (characteristics) we codified to analyse the selected cases in section 5. In section 6 we then present the results of the case analysis and outlines the measures suggested throughout the cases. We conclude the paper with a brief discussion of challenges and directions for future research.

2 BACKGROUND

The GDPR constitutes a binding legislative act, directly applicable across all the EU Member States. It superseded –and refined– its predecessor, the Data Protection Directive (DPD), adopted in 1995. The new complex yet protective [15] regulatory framework aims at addressing the inadequacy of previously existing data protection frameworks, at enhancing the enforcement powers and mitigation measures against data mishandling and, ultimately, at safeguarding a high level of protection of personal data within the single European market and the data-driven economy of the 21st century.

The Regulation sets out both general principles for processing personal data and specific requirements or procedures. The former includes *e.g.,* Art. 5, which specifies that processing activities should adhere to the principles of lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability. The latter category refers, among others, to the lawfulness of processing in Art. 6, the conditions for consent in Art. 7, the stricter conditions for the processing of special categories of data or "sensitive data" in Art. 9, the requirement and conditions of maintaining records of processing activities in Art. 30, and the requirement to designate a data protection officer within the controller and/or processor in Art. 37.

A critical provision is set forth in Art. 32, which stipulates that appropriate technical and organisational measures should be imposed, in order to safeguard the security of processing. Expressly, it requires that all processing operations (activities) are executed in accordance with the principles of confidentiality (data is accessible only to authorised parties), integrity (prevention against data loss or manipulation) and availability (data can be accessed whenever required). Previous research and available statistics released by various DPAs show that Art. 32 has been among the most problematic articles in terms of compliance already from the beginning of its entry into force [3, 25, 32].

Based on the definitions provided in the GDPR, the guidelines of the European Data Protection Supervisor (EDPS) and the information available on the European Commission's website, we provide the reader with a short glossary of terms below. The glossary does not aim to be an exhaustive list of all relevant terms; we rather aim to familiarise the reader with the terminology used in the remainder of the paper.

- *Personal data*: refers to any information relating to an identified or identifiable individual (*i.e.*, that can be directly or indirectly identified based on the available data), now or in the future.
- *Data subject*: refers to the identified or identifiable natural person, to whom the personal data in question relate.
- *Processing*: refers to any individual operation or set of operations (*i.e.*, information activities) which are performed on personal data, for example collection, storage, retrieval, transmission, disclosure or erasure of data.
- *Data controller*: refers to the natural or legal person, public authority, agency, body or any other type of organisation or entity which decides on the purposes and the means of processing. Hence, controllership represents a functional concept which allocates responsibility to the controller based on its actual role.
- *Data processor*: refers to the natural or legal person, public authority, agency, body or any other type of organisation or entity which is a separate entity from the controller and processes personal data on the latter's behalf.
- *Data protection authorities (DPAs) / supervisory authorities*: these are independent public authorities (one in each Member State), which are responsible for monitoring the application of the GDPR in order to protect the rights and freedoms of data subjects in relation to the processing of their personal data. DPAs hold investigative and corrective powers, providing expert advice on data protection issues and handling complaints filed against data breaches or other violations of the GDPR.

Since the most common outcome of a security violation under Art. 32 is a data breach, the term and its nuances [1] –as they are used in our paper– should be clarified. Personal data breach refers to a breach of security leading to adverse effects on personal data, *i.e.* via accidental (non-malicious) or unlawful (malicious) destruction, loss, alteration, unauthorised disclosure of, or access to, the personal data processed. Pursuant to the definition provided in Art. 4, a data breach presupposes the existence of a security breach, that is an incident violating the technical and/or organisational measures for security under Art. 32, or a lack thereof. However, it should be noted that not all security breaches under Art. 32 lead to a data breach in the understanding of Art. 4. For instance, a security breach (an incident or a lack of appropriate security measures) may not (adversely) affect any personal data. In these cases, the DPAs might assess that there is an infringement of Art. 32, due to the security breach, but additionally determine that the (potential) risks to the rights and freedoms of data subjects are particularly high, and, therefore, in these cases we refer to a potential data breach. We provide the exact text of the provision in Table 1.

Table 1: Article 32 GDPR

Art. 32 GDPR - Security of Processing

1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing, as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:

(a) the pseudonymisation and encryption of personal data;
(b) the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;
(c) the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;
(d) a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

2. In assessing the appropriate level of security account shall be taken in particular of the risks that are presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed.

3. Adherence to an approved code of conduct as referred to in Article 40 or an approved certification mechanism as referred to in Article 42 may be used as an element by which to demonstrate compliance with the requirements set out in paragraph 1 of this Article.

4. The controller and processor shall take steps to ensure that any natural person acting under the authority of the controller or the processor who has access to personal data does not process them except on instructions from the controller, unless he or she is required to do so by Union or Member State law.

3 RELATED WORK

Previous research into the topic of GDPR compliance is diverse. Early research –constrained by the shortage of available real-world data– commonly uses the Regulation as the starting point. Researchers attempt to derive concrete or practical measures from the GDPR text itself using various methods such as interpretation of the law [18], concretisation of legal requirements [24], agile software development with user stories [4], and formal concept analysis [29].

For instance, [29] presents a formalisation of the GDPR's text with the goal of gaining insights to support software designers and engineers in aligning their products to the regulation's requests. The author identifies four main practical design principles that lead to the observation that software should be re-designed in order to assist the duties of data protection officers, and to support multiple types of data subjects (*e.g.,* distinguishing between data from children and adults). With respect to security measures, the research recommends that software should be re-designed in order to make use of privacy-enhancing technologies as well as security enhancing approaches, but offers little practical guidance on how to conduct such re-design.

In contrast, the work in [24] offers an interpretation of the requirements set forth by the GDPR. The work derives from the regulation text's technical requirements that can serve as guidelines for its implementation in practice. The requirements are meant to be reusable and easily applied to multiple software, which is both a strength and a limitation of the work. In doing so, the requirements overlook details of the data processing practices that incur more risk, such as the presence of special categories of data, or data from vulnerable data subjects (*e.g.,* children) instead, leaving to the system designers the task to select the requirements which convey the appropriate level of security.

Another strand of early research starts from a more practical/technical point of view by evaluating existing industry standards (*e.g.,* ISO-family [8, 19]), practices (*e.g.,* Privacy by Design [21], Privacy Policy [22]) and existing technologies (*e.g.,* smartphones [13], blockchain [23, 30], Internet-of-Things [5]) to assess whether they can coexist with or ensure compliance with GDPR. However, such works tend to look at only specific technologies or discuss only the highest level of compliance, often overlooking the common cases where such requirements are not strictly necessary.

More recent research takes advantage of the available statistics, practitioners' experience and concrete decisions issued in case of non-compliance. For example, articles [32] and [3] provide a quantitative overview of the early fines imposed for non-compliance with GDPR and report on the heterogeneity of fines across GDPR articles as well across countries. Article [25] takes upon a more technical approach, and uses machine learning to extract features from a number of fines, which are later used (along with case metadata) to predict the size of future fines.

Finally, there are works which provide a qualitative overview of the GDPR compliance and its challenges through a series of interviews with relevant stakeholders [14, 26]. We draw attention to the work presented in [26], which fortifies the *raison d'être* of our work. In an attempt to gain understanding of the real challenges faced by companies when implementing the GDPR, semi-structured interviews with 12 practitioners from various companies were conducted. Among their findings, the authors highlight the fact that while larger and more technologically advanced companies experience compliance as achievable, and find the GDPR to have, we quote, *"a tone of flexibility and nuance"*, the same cannot be said about smaller companies, which perceive the regulation's requests as unclear.

While the works presented in this section provide valuable insights, they often only paint a bigger picture without digging deeper into specific articles, the challenges of their application, and potential solutions. Our work aims, instead, at the in-depth study of one provision of the GDPR - Article 32 - and the cases where it was infringed, in order to uncover the levels of implementation of security measures which are deemed *sufficient* for compliance.

4 METHODOLOGY

Our research was conducted in three steps. In order to capture the requirements set forth in Art. 32, we first identified the most important technical and legal aspects of data security under this provision. We did so based on the interpretations provided in related literature, and through preliminary analysis of a subset of cases. To eliminate personal bias and increase objectivity in the interpretation of the decisions, two researchers performed the initial annotations separately and cross referenced the results. The classification was then used as a blueprint to annotate a subset of available cases related to Art. 32. During this stage, we annotated each case with a number of technical and legal tags, which capture the case's defining characteristics and allow further analysis. In the final stage, the fines and their tags are quantitatively analysed in order to uncover the underlying patterns. In addition, we use the characteristics of the observed cases to bring light to the danger points of data processing where incidents commonly happen and discuss guidelines for compliance that are often highlighted by the DPAs in their decisions.

4.1 Case selection

At the moment of selection (August 2021), over 800 GDPR fines have been imposed, with roughly 200 of them related to the security of processing requirement under Art. 32.[3] Since GDPR fines are not yet reported in an organised manner on EU level [3], we rely instead on private or non-profit entities that track and collect them. In order to get a representative and –as much as possible– complete picture, we use three of such repositories as primary guidance when organising the full set of cases: GDPRHub[4], EnforcementTracker[5] and PRIVACYAffairs[6]. To the best of our knowledge, the three repositories contain the most complete publicly available set of cases and will be treated as ground-truth for the purpose of this study. In our sample, we only analyse cases where the full text (in quality that allows machine translation) or a sufficiently detailed (*i.e.* allowing the coding of a full set of tags) press release are available.[7] More specifically, we analysed 43 full decisions and 7 press releases. For annotation, we used the English machine translations of these cases –either already available or translated by us from the original text of the decision.[8]

We selected a subset of 50 cases for our analysis, which corresponds to roughly 25% of all Art. 32 cases available at the time. In choosing which cases to include in the sample, we considered cases covering a wide range of countries (19 EU countries and the UK) and fine sizes. Overall, our goal was to select a subset that was representative across country and year as well as fine sizes.

To illustrate, out of all cases, the five most prevalent countries are Romania (18%), Spain (15%), Italy (13%), Germany (7%) and Norway (5%). In our selected subset of cases, we follow a similar distribution with Spain (16%), Italy (10%), Romania (10%), Germany (6%), and Norway (6%). Similarly, in the original sample, roughly 37% of cases were from 2021, 43% from 2020, and 18% from 2019. In our selected sample, roughly 28% of cases were from 2021, 52% from 2020, and 16% from 2019. Given that fine sizes are numerical, we can formally check that our subset is representative via a Kolmorogorov–Smirnov test [27]. The hypothesis of equality of distributions of fine sizes between the full set and our selected subset cannot be rejected ($p = 0.248$), suggesting that our subset is indeed representative with respect to fine sizes.

4.2 Case clustering

For obtaining groups of similar data breach cases, we apply clustering, using one-hot encoded categorical features as the input. We explored two commonly used methods of clustering, namely hierarchical clustering and k-modes clustering, and obtained similar results. For conciseness, we present here results from k-modes as this is the most apt model to handle categorical data –which is the nature of the data in our study.

Classical clustering methods, such as k-means, compute distances between all variables, and use two criteria to group observations; (i) minimising the total distance *within* groups, and (ii) maximising the total distance *between* groups. With categorical data, the notion of distance needs modification. The method of *k-modes* [16] circumvents this issue by using the Hamming distance (also known as dissimilarity measure [12]) and groups data based on how similar they are. We apply k-modes to our data in order to discover groups of cases with similar characteristics.

An important element of clustering analysis is determining the number of clusters. To do so, we use the elbow method. The elbow method is based on a graph of percentage of variation explained plotted against the number of clusters. Naturally, the more clusters added, the better the model fit, but the higher the likelihood of overfitting. The graph essentially visualises the diminishing returns of adding new clusters. Initially, each additional cluster improves model fit substantially. After a while, an additional cluster does not improve model fit by much, and the curve flattens. Hence, the curve obtains an elbow shape with a kink at the optimal number of clusters.[9]

5 THE CLASSIFICATION

In this section, we present and precisely define the variables used for the clustering analysis. All variables were categorical with their values showed below in Table 2. Our initial coding included a larger set of variables. Those variables were not included in the analysis because they either contained textual information that we only used to better describe the emerging cluster groups or had too little variation to be suitable for analysis. They are presented in more detail in the Appendix.

[3] Whether a data breach proceeds into an investigation and/or court proceedings is not necessarily random. DPAs may selectively go after certain types of cases, and some of them may be negotiated before reaching the court. Both those factors may induce survivorship bias which in turn may influence the representativeness of the initial 800 cases.

[4] https://gdprhub.eu/

[5] https://www.enforcementtracker.com/

[6] https://www.privacyaffairs.com/gdpr-fines/

[7] The lack of actual ground truth and limited availability of original case text (due to translations or redactions by the DPAs) are potential threats to the external validity of our research. In the paper we operate under the assumption that those limitations are random, which implies that the available cases are representative.

[8] The annotated data is made public at https://github.com/tau200/gdpr_master.

[9] We proceed in our clustering analysis under the assumption that the sample size and the number of features are sufficient to accurately distinguish the clusters, while acknowledging that a larger sample size may be necessary for full scale implementation of our methodology.

Table 2: The classification codebook

Variable	Possible values
Descriptive only	
Country	
Year	
Fine size (in €)	
Compliance	Yes, No
For clustering	
Data incident	Yes, No
Origin of threat	External, Internal
Maliciousness	Yes, No
Mistake type	Human, Organisational, Technical
Data subject	Vulnerable, Non-vulnerable
Nature of data	Sensitive, Non-sensitive
Controller status	Private, Public

Descriptive characteristics include the country of the DPA handling the case, the decision's publication date, the amount of the fine (converted to € if necessary), and whether the data controller was found to be compliant with Article 32 or not after an investigation. We note that for cases where compliance was established, we cannot code the origin of threat and type of mistake as none was established. However, given that our random sample included compliant cases, we later discuss them as a separate group. In other words, only cases finding a violation of Article 32 are included in the clustering analysis. Additionally, there are some cases where no fine was given. It is important to note that this does not mean that no breach of GDPR was established, but rather that the warning was given instead, despite the breach.

Data incident refers to the presence of an incident where personal data was wrongly processed (*e.g.,* accessed, modified) by an unauthorised party. Note that this is not the same as a (legal) breach of the GDPR -non-compliance-, as determined by a DPA (*e.g.,* a data controller might not suffer a data incident, but could still be in breach of the GDPR for insufficient organisational measures).[10] When such an event happens, we can distinguish between a *malicious* (*i.e.* intentional act that led to a breach) and *non-malicious* (*i.e.* accidental) incident. *Threat origin* refers to the origin of a threat that leads or may lead to a data breach. We distinguish internal (*e.g.,* employees) and external (*e.g.,* attackers/intruders) threats. Since external threat includes internal by definition (*e.g.,* if a third party can access an unprotected database, so can the employees of a company), we only tag the wider of the two. *Mistake type* refers to the nature of mistake that leads or may lead to a data breach. We distinguish between human, technical, and organisational mistakes. As the lines are sometimes blurry and inter-dependent (*e.g.,* a human mistake can cause a technical problem), we focus on the main mistake(s). Note that any combination of mistakes can coexist.

Data subject refers to the distinction between vulnerable and non-vulnerable data subjects. We consider as vulnerable data subjects those persons that due to their physical or mental state (*e.g.,* patients), their age (*e.g.,* children or elderly) or the position of dependence (*e.g.,* employees under a *de facto* power imbalance towards the employer) [20] are in need of higher protection because of the higher imminent risks. *Nature of data* refers to the distinction between sensitive and non-sensitive personal data. In accordance with Recital 10 and Art. 9, sensitive data constitutes "special categories" of personal data, which require higher protection and stricter processing requirements, as it could involve significant risks for the rights and freedoms of data subjects. Sensitive data relevant to our research include those concerning the health, political views, racial or ethnic and sexual orientation of the data subjects. Conversely, all other types of data are classified as non-sensitive. *Controller status* refers to the type of organisation, under which the controller operates, for example: public (*e.g.,* hospital, university) or private (*e.g.,* financial institution, e-commerce company or telecommunications company).

6 FINDINGS AND DISCUSSION

6.1 Overview of cases and general observations

Our analysis reveals that the majority of cases relate to digital data (44) with only a few cases including physical data (6). Such distribution reflects the current reality, where the majority of data is processed digitally (automatically), rather than in physical (analog) form. The data is most commonly in text format, with only few exceptions related to video or photographic data. Only one of the analysed cases is related to audio data, more specifically recordings of phone calls.

Regarding the stages of processing, perhaps unsurprisingly, most security incidents happen during the handling (36) or storage of data (32). Only a small number of incidents are related to the disposal of data (7), with four incidents related to the collection of data. We find that most threats are external, but there is also a non-negligible number of fines (14) where the data controller did not sufficiently protect the data from internal sources, namely its employees. The majority of incidents were non-malicious (36).

Regarding the risk factors, we find some distinct patterns of cases in terms of data controller type. We notice that private data controllers are more often fined (27) than public ones (18). Additionally, the noticeable number of fines stemming from the processing of non-sensitive personal data (26) and the classification of data subjects as non-vulnerable (25) underlines the observation that lower-risk cases still pose a challenge to security compliance [2], that the appropriate level of security is not easier to attain in these cases, and that, even when a processing activity entails lower (or no) risk, the data controller should assess that strong, state-of-the-art measures are necessary to safeguard data security [1]. In fact, there are cumulatively (32) cases in our analysis involving either of the aforementioned lower-risk factors or both.

However, it should be noted that the majority of cases where no explicit data incident happened, concern public data controllers, *i.e.* public authorities, agencies or bodies. There is no definitive explanation yet on why this is the case. This may denote that these data controllers are more diligent in making sure they are compliant with all the laws applicable to them, because of their position as state-related authorities, thus being subject to more scrutiny. It may also be the case that this finding is accidental, and a possible examination of a larger set of cases might lead to a diverging result.

[10] In technical literature and practice, the incident is often referred to as data breach, but since we want to differentiate between a technical/actual breach and a breach of GDPR in the legal sense, we make this distinction.

Delving into a cross-risk factor analysis, we should underline that even though whenever sensitive personal data are processed, the respective data subjects are classified as vulnerable, the opposite is not always true. That is, there are cases where the data subjects affected by the security incident or the data breach are considered vulnerable, but the personal data processed are non-sensitive. Consequently, the sensitivity of the personal data and the vulnerability of the data subject seem to be non-bidirectional risk factors. Even though this was a foreseeable pair-pattern of risk factors that we expected to observe, the non-bidirectionality was not expected.

6.2 Case groups

The clustering analysis was implemented in Python using the kmodes library [7]. Figure 1 shows the model fit as a function of the number of clusters (i.e. the elbow method). The vertical axis measures the cost of the model, defined as the sum of all the dissimilarities between the clusters. We see that the optimal number of clusters for our case should be four (where the curve changes angle more drastically - the elbow).

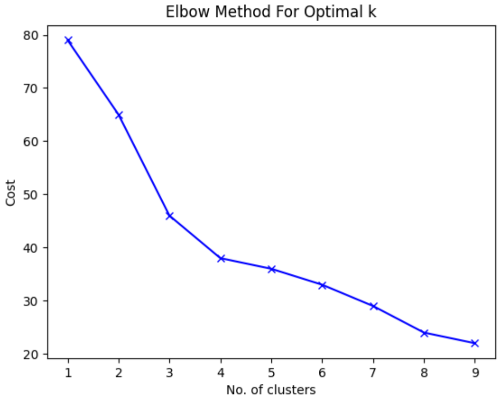

Figure 1: Elbow Method for optimal number of clusters

Figure 2 is a co-occurrence network representing the relationships between the case annotations (tags), with more strongly related (co-occurring) concepts appearing closer to each other. The size of nodes and the thickness of edges represent the number of annotations per tag and the number of co-occurrences respectively. As seen in Figure 2, there is a number of central characteristics that most cases contain: the most commonly occurring problem is a lack of or insufficient organisational measures; most cases indeed involve a data incident –and consequently a breach–, while most threats are external, but non-malicious.

The layout of the nodes (along with the colour coding) also allows us to see that certain tags commonly appear together (*e.g.* non-malicious internal threat - employees making honest mistakes), while some others rarely do (*e.g.* malicious internal actors). This suggests a number of distinct groups with specific common characteristics which is confirmed by k-modes clustering analysis. The remaining section and Table 3 present the the four groups identified along with their respective characteristics and two representative cases from each group. We also include the fifth group of cases where no breach of GDPR was identified. For a full list of cases and their characteristics, see Table A2 in the Appendix.

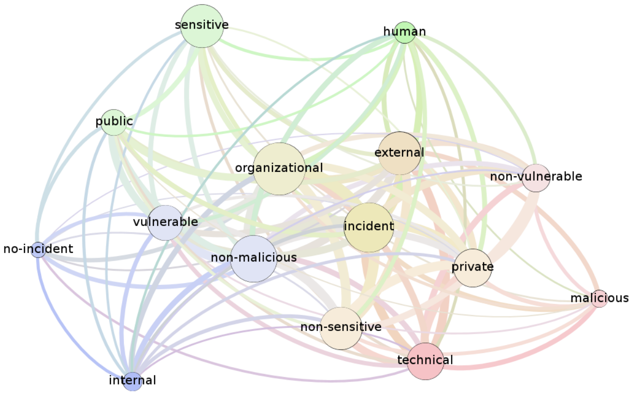

Figure 2: Co-occurrences of case characteristics

Group 1: Insufficient organisational measures.
Group 1 represents the cases whose main characteristic is that no actual data incident occurred, but a fine was still imposed for infringement of the security requirement under Art. 32. In these cases, the mistake leading to the fine is typically of organisational nature. More specifically, the DPAs found insufficient access control, auditing or protocols, too broad authorisation accounts, no logging of accesses to personal data *etc.* Subsequently, the immediate threat is typically of internal and non-malicious nature –*i.e.* employees having access to a wider-than-required range of data, without necessarily or explicitly bad intentions. Moreover, the data controller is consistently a public organisation (*e.g.*, university hospital), processing sensitive data of vulnerable data subjects, leading to enhanced scrutiny and, consequently, a fine even before an actual data incident might occur or before the rights of data subjects are adversely affected. In such cases, however, the DPAs considered that high risk existed for the data subjects, especially since both risk factors of sensitivity of data and vulnerability of subjects are present.

In particular, the group contains seven cases of such nature, with fines ranging from 4,750€ up to 440,000€. A characteristic case from this group is **Sweden (2020)**, where the data controller –a university hospital– gave the healthcare personnel access to patient journals based on whether they were doctors or nurses (with no additional distinction), effectively enabling access to almost all the medical care records regardless of necessity. The Swedish DPA held that the hospital had not taken appropriate organisational measures to limit access to personal data, therefore failing to ensure the appropriate security of personal data.

Since the main security issue of the cases in Group 1 is of organisational nature, and typically no data incident occurred, the suggestions for improvements and Art. 32 compliance are accordingly of organisational nature. Concrete DPA suggestions for such cases include: separate authorisation profiles depending on needs, implementation of a single profile per person, regular testing and evaluation of systems, logging access to personal data (*i.e.* who and when), security risk assessment and mitigation, implementation

Table 3: Groups with sample cases

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	organisation
Norway	2020	73,000€	-	Internal	-	O,T	-	✓	✓	Public
Sweden	2020	394,000€	-	Internal	-	O	-	✓	✓	Public

(a) Group 1: Insufficient organisational measures.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	organisation
Romania	2021	2,000€	-	Internal	-	H,O	✓	✓	-	Private
Spain	2021	3,000€	-	External	-	H,O	✓	-	✓	Private

(b) Group 2: Non-technical mistake.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	organisation
Poland	2021	22,000€	-	External	-	O,T	✓	✓	-	Public
Italy	2020	80,000€	-	External	-	O,T	✓	-	✓	Public

(c) Group 3: General breach.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	organisation
Germany	2020	20,000€	-	External	✓	O,T	✓	-	-	Private
Poland	2020	235,000€	-	External	✓	H,T	✓	-	✓	Private

(d) Group 4: Targeted attack.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	organisation
Spain	2021	-	✓	-	-	-	-	-	-	Private
Denmark	2019	-	✓	-	-	-	-	-	✓	Private

(e) Group 5: GDPR compliant.

Notes: The full table is available in the Appendix; Mistake type: H=Human, O=organisational, T=Technical

and adherence to data handling protocols, routines and incident response plans, regular staff training regarding privacy policy and protocols.

Group 2: Non-technical mistakes.

Group 2 consists of cases where a data controller, often a small business (*e.g.,* a hairdresser, self-employed lawyer), suffered a data breach due to a non-malicious human or organisational mistake. Interestingly, only a single case in this group suffered a breach due to a technical mistake. In contrast to the previous group of cases, this group (and the remaining groups) included cases where an actual data incident occurred. The reason for the incident is a low-tech mistake, such as employees mistakenly sharing data, wrong email attachments, leaving computers unattended, unlawful disposal or loss of data, mass emails without BCC (Blind Carbon Copy). The threat in these cases can be both internal (unlawful access to personal data within the company –by the employees) and external (loss or disclosure of data outside of the company).

We find twelve cases of such nature, with fines ranging between 0€ (warning) and 4,500,000€. Two characteristic cases are **Romania (2021)** where an employee's resignation letter was shared in a WhatsApp group, and **Spain (2021)** where envelopes with sensitive data of 29 employees were found abandoned in an industrial complex. There is one outlier case within this group, with a fine of 4,500,000€ . For this particular case the annotation does not reflect the full picture, as Art. 32 infringements only correspond to a small portion of the fine with the largest portion being imposed for violations of other GDPR Articles (not reflected in the annotations and therefore indistinguishable to the clustering method). Notably, fewer cases in this group relate to the processing of sensitive data or vulnerable groups.

Considering the nature of mistakes and the low risk for data subjects, cases from Group 2 most commonly occur in small to medium-sized companies. When the mistake is of merely organisational nature, the suggestions from Group 1 apply, with particular focus on enhancing staff training and raising awareness of proper data handling practices, especially when the data controller is a legal entity run by a single individual. The most important measures that could be implemented by such controllers at risk of low-tech mistakes include: the use of BCC within mass emails, data backup, password-protected hardware and software/files (if applicable), lawful - and timely - disposal of data (shredding, deleting), storage of data in a physically secure location, and use of secure platforms for data transmission and storage. For slightly larger data controllers, detailed data-handling internal protocols can catch small human mistakes and protect against a security incident and/or personal data breach. Devising and following internal protocols also allows the data processor to prove compliance with the GDPR when a single employee acts out of line.

Group 3: General breach of personal data.

Group 3 is the biggest and most diverse one and captures the most "general" of cases. It concerns both public and private data controllers, vulnerable and non-vulnerable data subjects and all three types of mistakes (human, organisational and technical). It predominantly consists of cases where a data incident occurred. Distinct characteristics are the mostly *external* origin of the threat and the *non-maliciousness* of the mistake. Such incidents most commonly happen to small or medium-sized entities with little or no expert IT or security staff, processing large quantities of data (extensive processing scope) which necessitate some system or automation (*e.g.,* small e-commerce businesses, schools and local chains). Examples of incidents for cases in this group include data leaks, unauthorised disclosure or transmission of data, non-encrypted data on websites or employee mistakes due to deficient organisational measures or internal protocols.

We find eighteen cases belonging to this group, with fines ranging from 0€ (warning) up to 27,800,000€. Two characteristic cases from this group are **Poland (2021)**, where personal data of over 50,000 students were exposed on the internet, likely due to technical mistakes made during a migration to a new platform, and **Italy (2020)**, where a technical problem in the infrastructure (managed by a third party) led to health data of competition participants being published on the controller's website.

Regarding the risk factors taken into consideration when determining the amount of the fines, the DPAs reasonably examine the lack of deliberate (malicious) mistake along with the processing of mostly non-sensitive data and the relatively non-extensive scope of processing. The broadness of cases and the inclusiveness of this Group are reflected in the variety of fines imposed for the infringement, ranging from a warning to a few million Euros but with the majority of them standing in the middle, at around a few dozen thousand Euros. As before, the two outliers in terms of the fine size include cases where the processor was fined for breach of multiple GDPR articles.

Taking into account the broadness of Group 3 cases and the existence of all three types of mistakes, compliance suggestions from all other groups apply. Given the nature of the data controller –little/no dedicated IT or security staff, but often the need to create own systems for data handling– DPAs suggestions for technical measures are especially relevant: sufficient encryption (AES, RSA, https, *etc*) and passwords, multi-factor authentication, use of dummy data in test databases, sufficient entropy when security relies on randomisation, use of firewall and anti-virus, proper testing of environments, databases and additional features of the systems, logging of operations, and regular updates and checks of the systems. Advisable organisational measures for this Group include appropriate certification mechanisms, codes of conduct, thorough internal privacy policies, as well as clear accountability frameworks and specified contractual obligations -when external data processors are involved.

Group 4: Targeted attack.

The final set includes cases where a data breach occurred in a *malicious* manner, *i.e.* an *external* attacker willfully attempting to access personal data. Such a breach commonly occurs due to a technical mistake on the data controller's part that allows an attack using malware, SQL injections, cross-site scripting, sometimes together with social engineering, threats or false promises to the data controller. The target of such an attack is usually a private organisation, with the attacker profiting either from exploiting the data itself or through post-attack extortion and threats to the controller.

The risk factors aggravating the fines are not only the maliciousness of the security incident, aiming for example at financial gains by materially harming data subjects via online fraud and financial loss, but also the existence of a personal data breach in every one of these cases, as severe harm was inflicted upon the rights and freedoms of data subjects. It is worth noting, however, that in the majority of cases no sensitive data or vulnerable data subjects were affected.

We identify eight such cases, with fines ranging from 0€ (warning) up to 1,434,000€. A characteristic case from this group is **Germany (2018)** where an online chat platform was hacked and 1.8 million data records of 330,000 users, including passwords (stored in plain text), were stolen and later posted online.

Given the nature of these attacks, DPAs sometimes refrain from giving specific suggestions for compliance to data controllers and instead refer to accepted industry standards. For incidents where the main factor facilitating an attack is a human or technical mistake, the majority of suggestions from Group 1 or Group 3 may apply. Despite that, the most comprehensive and complete form of data security compliance is achieved by hiring trained IT staff, imposing regular auditing, stress-testing and evaluation of the systems, in order to ensure the effectiveness of the existent security measures at all times, and most importantly the application of relevant industry standards, such as ISO/IEC 27001 [17].

Group 5: GDPR compliant.

A separate group of five cases also emerged, where no fine was imposed since no severe insufficiency in technical or organisational security measures was identified.[11] In these cases, the data controller was investigated after a complaint or as part of routine (random) inspection. In four of the cases the DPAs held that the controllers had implemented sufficient measures by conducting a risk assessment, considering technical and organisational security measures and introduced measures to minimise the processing in question. Specifically, the DPAs are satisfied when the controllers use encryption in transit and at rest, secure servers (ideally located in the EEA), implement sufficient access control and have procedures in place for detection and mitigation of a potential data breach. Interestingly, in one of the cases the data controller actually suffered an incident, but was found to be compliant with the GDPR.

In the final case the Belgian DPA held that receiving an erroneous email with personal information does not necessarily constitute processing and, therefore, no data breach was identified. In addition, the decision was informed by the fact that no security measure was found to be inappropriate, but the mistake was purely of human nature. However, our research exposed cases with similar characteristics where a breach was indeed determined and a fine imposed. While further investigation of such observations is outside the scope of current research, it shows some level of variability between DPAs and their interpretations of Art. 32.

6.3 Magnitude of fines

Figure 3 presents the full set of Art. 32 cases ordered by fine size (as a black line), along with our selected subset of cases (as shape-marked individual fines). This figure serves two purposes. First, it displays the distribution of fine size in both the full set and our selected subset (mentioned in subsection 4.1). Second, it visualises the fines imposed across the different groups. Overall, we see that the cases range across the full spectrum and there is no clear correlation between the groups and the magnitude of the fines. Our sample

[11] The group was manually separated before the clustering was performed since some of the features cannot be determined (e.g. threat origin).

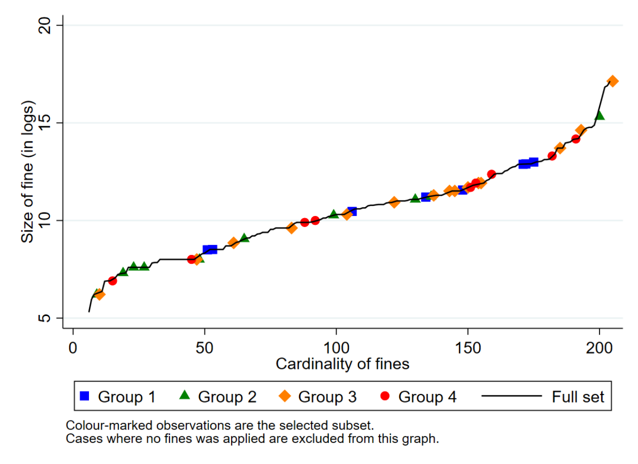

Figure 3: Fine sizes (in increasing order) with cluster groups

size does not allow for statistical inference, however, we can make some informal observations.

First, we notice that cases where no data incident occurred (Group 1) mostly appear on the higher part of the spectrum, rather close to each other. The relatively higher fines reflect the fact that Group 1 consists of large public institutions (*e.g.,* hospitals) that are processing sensitive data of vulnerable subjects (*e.g.,* patients) and possibly receive higher scrutiny. Even though no incident occurred, the material or moral harm to data subjects could have been severe, which is reflected in the size of the fine given.

On the opposite end of the spectrum is Group 2, which includes low-tech mistakes. These relatively lower fines correspond to the non-maliciousness of the mistakes and the lack of both personal data sensitivity and data subject vulnerability. Lack of technical mistakes is largely the consequence of manual processing of data, which also results in a small number of data subjects involved in a breach - often a single person.

The outliers belonging to any of the groups where fines reach several million euros demonstrate that processing an extensive quantity of data (belonging to millions of data subjects) intensifies the potential risk and harm to the rights and freedoms of data subjects, aggravates the severity of a breach, and, subsequently, exacerbates the amount of the fine. It is important to keep in mind that the issues in such cases relate to a number of additional GDPR articles along with Article 32.

Fines for Groups 3 (general breach) and 4 (targeted attack) cases appear across the spectrum without any noticeable patterns. Perhaps the most interesting general observation is the high size of fines given when the measures were deemed insufficient following an inspection (Group 1), rather than an actual incident (remaining groups). We speculate this is related to the public nature of data controller, vulnerability of data subjects and the sensitivity of personal data, but further investigation is needed to confirm such speculations.

In terms of the mistake type, we notice that human mistakes result in smaller fines (median=3000€), followed by technical mistakes (median=56,000€), and organisational mistakes resulting in the largest fines (median=73,000€). Public data controllers are less likely to have human mistakes resulting in higher fines (median=69,000€) compared to private ones (median=29,000€). However, the pattern is reversed if one looks at the highest imposed fine - the highest fine for a public data controller was 900,000€ whereas for a private data controller it was 27,800,000€.

6.4 Danger points

We identify a set of actions or states that appear in or coincide with the infringement of data security under Art. 32, according to DPA decisions. We refer to them as *danger points.* Common danger points of primarily technical nature include:

i. System update, reset, restore or restart (appears in 3 cases);
ii. Migrations between platforms or versions (3);
iii. Moving data between physical locations (2);
iv. Code or system reuse (2);
v. Outsourcing to third company or sharing custody of systems (5);

In case of (i.) and (ii.) the problem arises when settings of a system are changed unintentionally, exposing (parts of) a system or removing protections. Point (iii.) has a high chance of leading to misplaced or lost data. With (iv.) when parts of a system or workflow are reused –especially outside of their original intention– there might be unwanted residual behaviour or propagation of (yet unrecognised) issues. Regarding (v.), there is a higher chance of miscommunication between people or unspecified cooperation of systems when outsourcing is deployed.

It is important to note that some of the actions recognised above as danger points are beneficial or even necessary and should not be avoided indiscriminately. Instead, special attention should be paid to them –if possible they should be minimised to a reasonable extent and treated with care when minimisation is not possible. Given the technical nature of mistakes, such danger points are especially relevant for the data controllers of Group 3 (general breach) and Group 4 (targeted attack).

Common danger points of primarily organisational/legal nature include:

vi. Ineffective accountability framework or lack of clarity in contractual obligations between controller and processor (4);
vii. Inconsistent (re)assessment of the effectiveness of organisational measures (6);
viii. Insufficient assessment of the imminent risk or harm for the rights and freedoms of data subjects (4).

Point (vi.) stems from the complex legal requirement of establishing a thorough accountability framework with specified contractual obligations, laying out the responsibility of every party participating in processing activities. Such an organisational malfunction is often apparent when the controller has outsourced certain processing activities or is connected for any other operational purpose to one or more processors. Furthermore, all security measures should be consistently re-assessed (vii.), in order to be able to address the risk level of processing activities and in accordance with state-of-the-art industry standards and practices. Such danger points and related actions are especially relevant for the data controllers in Group 1 (insufficient organisation measures) and -perhaps less formally- Group 2 (non-technical mistakes).

The GDPR requires –and largely relies on– data controllers to carry out an *ex ante* risk assessment when determining the appropriate measures for data security [11]. This is indispensable for evaluating all possible risk factors involved as well as for identifying the actual or potential harms inflicted or impending in each case (viii.). However, as discussed in the literature, this risk-based approach to determining the risks and harms to the rights and freedoms of data subjects may not comprehensively depict the number and magnitude of all human rights possibly infringed by a personal data breach [10] or the extent to which such infringements can be sustainably mitigated via a fine and *ad hoc post-factum* corrective actions. This final point leads us directly to the challenges we faced trying to critically analyse the decisions based on previous research, pointing us to exciting future work.

7 CONCLUSION

7.1 Discussion

In this paper, we uncover the patterns and commonly occurring pitfalls related to the security of processing under Art. 32. We do so by compiling a set of representative cases, coding a wide range of technical and legal characteristics of the cases, and using the coded variables to determine distinct groups of cases. In turn, this classification allows us to propose a number of high level fundamental actions that data controllers or processors should take to minimise the chances of a data incident and breach.

We note that our study is correlational and as such is unable to pin down the direction of causality. It is equally plausible that different group characteristics can cause a different type of breaches, or that different types of breaches are more likely to be investigated and consequently fined because of different characteristics. For instance, we find a lack of targeted technical attacks among the public controllers processing sensitive data of vulnerable subjects. Arguably, this is likely due to the increased scrutiny -as evidenced by the fine being issued without an incident- that the public controllers face and are therefore less vulnerable to external malicious attacks rather than due to a lack of interest from attackers.

On the other hand, we notice that malicious actors mainly use technical means to gain unlawful access to the data. The majority of data controllers attacked by malicious actors are private controllers dealing with non-vulnerable subjects and non-sensitive data. This may reflect under-investment in technical measures by private data controllers.

Given the fact that increased scrutiny commonly finds insufficient organisational measures and potential internal threat actors in public controllers (Group 1), it is reasonable to assume that other -less scrutinised- groups suffer from similar problems. However, they are not uncovered until an external attacker targets the controller or a human mistake is made.

We find that human mistakes are the least costly for the non-compliant data controller, followed by technical and finally organisational mistakes. To an extent, this also correlates with the controller status and types of data being processed. For example, a small privately owned shop that only processes non-sensitive data of non-vulnerable subjects is more likely to suffer from a human mistake than a public hospital with extensive data handling procedures. Consequently, we also find that public companies face higher fines than private companies, which get higher maximum fines.

Finally, we notice that both human and technical mistakes commonly co-occur with organisational mistakes, while the latter sometimes appear on their own. We speculate that lack of organisational measures is often the precondition for human and technical mistakes and should therefore be the starting point for all attempts at GDPR compliance. However, more research is needed to confirm such statements.

All these observations emphasise that indeed appropriateness of measures and proportionality to risks involved gets differentially taken into account by DPAs. Our paper provides one empirical piece of evidence towards a better understanding of such differences and we expect to see more research in the topic in the near future.

7.2 Limitations and future work

Below are some limitations and potential future directions. Given the rapidly increasing number of fines related to Art. 32, a sample size of 50 is relatively small. We distinguish three main limiting factors: the lack of complete, officially-maintained datasets compiling the cases, the need for manual annotation, and the diversity of the cases (in terms of availability of full text, style of reporting and details given). To facilitate future work, there is great need for official statistics, datasets or repositories that collect fines across all EU Member States in a standardised manner. Community-maintained repositories, such as GDPRHub, certainly help bridge the gap, but also introduce a level of uncertainty. Where possible, our team has appended and amended decisions, contributing to the GDPRHub.

An additional level of uncertainty is added through human annotation of the fines, which is time consuming and inherently subjective, even with clearly defined tags. A way to mitigate this is through automatic annotation using machine learning. Using automatic annotation, the sample size can be scaled virtually infinitely. The potential challenges of this approach include providing sufficient and unbiased (due to subjective human annotation) training data and the training itself -the cases often quote full paragraphs from the GDPR-, which would likely skew the results. In addition to that, some human intuition is necessary when interpreting the cases -we often found similar phrasings used by the DPAs for very different contexts and decisions.

It is worth noting that a related challenge for our research has been the diversity of languages in the original sources and the need for English translations of the decisions. Such translations can deviate from the message the original text wanted to convey, which complicates some aspects of the technical and legal analysis as well as the attainment of definite conclusions. As before, human intuition and reading of context was helpful, however wider availability of cases in English would greatly benefit future research. This would also allow us to make further per-member-state supervisory authority comparisons which are outside of the scope of this paper due to the sample size and high-level of analysis. Such comparisons would be fruitful since the question of whether resources are similar across member states is reported as an important issue [31].

An additional avenue worth pursuing is structured interviews with legal practitioners, DPAs and technical personnel from various types of data controllers. Such information would provide a new dimension to explore in the form of good practices that the DPAs encounter, common problems or challenges that data controllers face when trying to ensure compliance, as well as insights from legal practitioners.

As mentioned in the previous section, prospective work could also incorporate a more detailed analysis of the GDPR-instructed risk assessment approach when addressing the possible risks to data subjects under the meaning of Art. 32. Such an analysis is necessary in order to distinguish whether the applied risk-based approach is sufficient and capable of capturing all possible harms for the rights and freedoms of data subjects, or a rights-based approach would be more effective for that purpose [9]. Such further research could also engage in determining whether this ostensibly purely risk-based approach inherently incorporates an assessment of the human rights affected and the harms caused as well, which seems to be the case at least in several of the analysed cases. An analysis of Art. 32 fines from a human rights perspective would not be exhaustive, if we would not further explore the rights-harms pairs in literature and try to associate them with specific organisational or technical security mistakes. However, such an investigation would be relying on inferences, due to the conceptual lacuna between the scholarly approach on the subject and the practical focus of DPA decisions.

Overall, analysing these fines we have observed a diverse set of cases with problems ranging from small human oversights all the way to coordinated malicious attacks. Given the variance, there is no silver bullet to solve all potential problems. Instead, when looking for the appropriate security measures, the data controller should consider numerous risk factors, *i.e.*, the kind of data subject affected, the kind of controller, the scope of processing, the size of the dataset (*e.g.,* we cannot expect a small beauty salon to fully comply with ISO-family of standards), the nature of data in question (*e.g.,* sensitive health data should be handled more carefully compared to e-mail addresses), in order to identify the most fitting group. This in turn would likely allow the data controller to use the relevant findings of our research as custom compliance guidelines for data security.

ACKNOWLEDGMENTS

This research was conducted with the support of the Network Institute of Vrije Universiteit Amsterdam.

REFERENCES

[1] Rogers Alunge. 2021. Breach of security vs personal data breach: effect on EU data subject notification requirements. *International Data Privacy Law* 11, 2 (2021), 163–181.

[2] Article 29 Working Party. 2014. *Statement on the role of a risk-based approach in data protection legal frameworks.* Opinions, Working Documents, Recommendations 218. European Commission, Brussels. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf

[3] Catherine Barrett. 2020. Emerging Trends from the First Year of EU GDPR Enforcement. *Scitech Lawyer* 16, 3 (2020), 22–35.

[4] Cesare Bartolini, Said Daoudagh, Gabriele Lenzini, and Eda Marchetti. 2019. GDPR-based user stories in the access control perspective. In *International Conference on the Quality of Information and Communications Technology.* Springer, Springer International Publishing, Ciudad Real, Spain, 3–17.

[5] Daniel Bastos, Fabio Giubilo, Mark Shackleton, and Fadi El-Moussa. 2018. GDPR privacy implications for the Internet of Things. In *4th Annual IoT Security Foundation Conference*, Vol. 4. IEEE, London, United Kingdom, 1–8.

[6] Danielle Keats Citron and Daniel J Solove. 2022. Privacy harms. *Boston University Law Review* 102 (2022), 793–863.

[7] Nelis J. de Vos. 2015–2021. kmodes categorical clustering library. https://github.com/nicodv/kmodes.

[8] Vasiliki Diamantopoulou, Aggeliki Tsohou, and Maria Karyda. 2020. From ISO/IEC27001: 2013 and ISO/IEC27002: 2013 to GDPR compliance controls. *Information & Computer Security* 28, 4 (2020), 645–662.

[9] Raphael Gellert. 2016. We have always managed risks in data protection law: understanding the similarities and differences between the rights-based and the risk-based approaches to data protection. *European Data Protection Law Review* 2, 4 (2016), 481–492.

[10] Raphaël Gellert. 2020. *The Risk-Based Approach to Data Protection.* Oxford University Press, Oxford.

[11] Maria Eduarda Gonçalves. 2020. The risk-based approach under the new EU data protection regulation: a critical perspective. *Journal of Risk Research* 23, 2 (2020), 139–152.

[12] Richard W Hamming. 1950. Error detecting and error correcting codes. *The Bell system technical journal* 29, 2 (1950), 147–160.

[13] Majid Hatamian. 2020. Engineering privacy in smartphone apps: A technical guideline catalog for app developers. *IEEE Access* 8 (2020), 35429–35445.

[14] Kalle Hjerppe, Jukka Ruohonen, and Ville Leppänen. 2019. The general data protection regulation: requirements, architectures, and constraints. In *27th International Requirements Engineering Conference.* IEEE, Jeju Island, South Korea, 265–275.

[15] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law* 28, 1 (2019), 65–98.

[16] Zhexue Huang. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery* 2, 3 (1998), 283–304.

[17] ISO Central Secretary. 2013. *Information technology – Security techniques – Information security management systems – Requirements.* Standard ISO/IEC TR 29110-1:2013. International Organization for Standardization, Geneva, CH. https://www.iso.org/standard/54534.html

[18] Costas Lambrinoudakis. 2018. The general data protection regulation (GDPR) era: ten steps for compliance of data processors and data controllers. In *International Conference on Trust and Privacy in Digital Business.* Springer, Regensburg, Germany, 3–8.

[19] Isabel Maria Lopes, Teresa Guarda, and Pedro Oliveira. 2019. How ISO 27001 can help achieve GDPR compliance. In *14th Iberian Conference on Information Systems and Technologies (CISTI).* IEEE, Coimbra, Portugal, 1–6.

[20] Gianclaudio Malgieri and Jędrzej Niklas. 2020. Vulnerable data subjects. *Computer Law & Security Review* 37 (2020), 105415.

[21] Yod-Samuel Martin and Antonio Kung. 2018. Methods and tools for GDPR compliance through privacy and data protection engineering. In *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW).* IEEE, London, United Kingdom, 108–111.

[22] Jayashree Mohan, Melissa Wasserman, and Vijay Chidambaram. 2019. Analyzing GDPR compliance through the lens of privacy policy. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare.* Springer, Los Angeles, United States, 82–95.

[23] Xin Pei, Xuefeng Li, Xiaochuan Wu, Liang Sun, and Yixin Cao. 2020. UDPP: Blockchain based Open Platform as a Privacy Enabler. In *10th Annual Computing and Communication Workshop and Conference (CCWC).* IEEE, Las Vegas, United States, 500–505.

[24] Sandra Domenique Ringmann, Hanno Langweg, and Marcel Waldvogel. 2018. Requirements for legally compliant software based on the GDPR. In *OTM Confederated International Conferences On the Move to Meaningful Internet Systems.* Springer, Valletta, Malta, 258–276.

[25] Jukka Ruohonen and Kalle Hjerppe. 2022. The GDPR enforcement fines at glance. *Information Systems* 106 (2022), 101876.

[26] Sean Sirur, Jason RC Nurse, and Helena Webb. 2018. Are we there yet? Understanding the challenges faced in complying with the General Data Protection Regulation (GDPR). In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security.* ACM, Seoul, 88–95.

[27] Nickolay Smirnov. 1948. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics* 19, 2 (1948), 279–281.

[28] Daniel J Solove. 2005. A taxonomy of privacy. *University of Pennsylvania Law Review* 154, 3 (2005), 477–564.

[29] Damian A Tamburri. 2020. Design principles for the General Data Protection Regulation (GDPR): A formal concept analysis and its evaluation. *Information Systems* 91 (2020), 101469.

[30] Nguyen Binh Truong, Kai Sun, Gyu Myoung Lee, and Yike Guo. 2019. GDPR-compliant personal data management: A blockchain-based solution. *IEEE Transactions on Information Forensics and Security* 15 (2019), 1746–1761.

[31] Siddharth Venkataramakrishnan. 2020. GDPR accused of being toothless because of lack of resources. https://www.ft.com/content/a915ae62-034e-4b13-b787-4b0ac2aaff7e. Financial Times, Accessed: 14-12-2022.

[32] Josephine Wolff and Nicole Atallah. 2021. Early GDPR Penalties: Analysis of Implementation and Fines Through May 2020. *Journal of Information Policy* 11 (2021), 63–103.

A APPENDIX

This appendix provides detailed information about all the variables we initially coded as part of our full ontology, but were not used for clustering. Table A1 presents all those variables. Table A2 includes all the cases arranged by their corresponding cluster group; a subset of those cases was presented in Table 3.

Table A1: Additional variables in our ontology

Category	**Possible values**
General	
Decision number	
Other GDPR Articles	
Technical	
Stage of processing	collection, storage, handling, disposal
Data type	digital, physical
Requirements broken	access control, confidentiality, integrity, availability, testing and auditing
Legal - Infringement	
Type of breach	unlawful processing, accidental destruction, unauthorised access, unauthorised disclosure
Source of breach	cyber attack, weak authorisation, lack of diligence
Consequences of breach	data exposure, loss of control over data
Legal - Risk Factors	
Scope of processing	large data quantity, extensive data processing, large number of data subjects
Data specification	health, financial, education, employment, personal identifiers, online identifiers
Subject specification	patients, students, children, employees, customers, digital subscribers
Legal - Harm	
Likelihood & Severity	high, medium, low
Type of harm	material, moral
Harm specification	identity theft, online fraud, financial loss, emotional distress, chilling effect
Legal - Technical & Organisations measures	
Operational readiness	staff training, security standards or certifications, adequate security measures, diligence
Post-factum remedies	risk assessment, swift notification, cooperation with DPAs, adherence to internal policies

Stage of processing refers to the stage of processing in which a breach occurred or may occur. We distinguish between collection, storage, handling (retrieval, transmission and disclosure) and disposal. Similar to the mistake type, a (potential) incident may involve any number or combination of stages. *Requirement broken* refers to a specific requirement that was broken or is at risk of being broken. We distinguish confidentiality, integrity, availability, access control, and testing and evaluation. Again, any combination of multiple requirements may be broken. *Data type* describes the format of the personal data in question. We distinguish between digital and physical data. Note that the two can coexist –*e.g.,* a USB stick (physical) containing patient data in text format (digital) is stolen.

Characteristics concerning the GDPR infringement include attributes of a decision from a legal and/or organisational perspective, and usually help determine the severity of the infringement and the amount of the DPA fine. Particularly, *type of data breach* refers to the type of breach as determined by the DPA, in accordance with Art 4., taking into account the implications of the breach to the data subject(s). Common types include unlawful processing, accidental destruction, unauthorised access to or (public) disclosure of personal data. *Source of breach* refers to the incident (*e.g.,* cyber attack) or the insufficiency or lack of technical (*e.g.,* weak authorisation mechanism) and/or organisational (*e.g.,* lack of diligence or over-authorisation for access into a database) measures. *Consequence of breach* refers to the result of the security breach and/or to the adverse effect on personal data, such as public exposure of data or online data leak, unlawful dissemination of personal data, and accidental loss of control over data.

Data specification refers to the specific type of data processed, *i.e.* data on health, employment, education, the financial situation or payment details, personal identifiers (*e.g.,* name, address, ID or passport number) or online identifiers (*e.g.,* username, password, IP address). *Data subject specification* refers to all possible types of data subjects we came across while analysing our subset of decisions, for example patients, children, students, customers, employees, insured persons or digital subscribers.

After reviewing seminal literature on privacy risks and harms [6, 28], we tried to connect these established theories with the specific Art. 32 violations presented in our subset of decisions. Usually, DPAs establish such violations by extensively analysing the facts of the case, generally referring to possible risks to the rights and freedoms of data subjects and, finally, merely stating the consequence of the infringement or data breach. Our research moves a step further, as we try to conceptualise which risk theories and specific harms correspond to each type of non-compliance behaviour in our cases. Therefore, we codify the following categories of characteristics. *Likelihood & severity* refers to data processing properties which may (significantly) enhance the risk and/or likelihood of a breach (and even increase the fine itself), such as large data quantity, extensive data processing (numerous kinds of personal data processed) or the large number of data subjects. *Type of harm* refers to the distinction between material and moral (non-material or non-pecuniary harm), *i.e.* the consequences of the breach (when risk materialises) or the potential consequences of the likely risk(s). *Harm categorisation* refers to all the possible types of harm a breach can inflict upon the data subject, whether material or moral, for example online fraud, financial loss or emotional distress and angst from the exposure of sensitive data.

The final category consists of the security measures that data controllers or processors have in place at the moment of the data incident or breach, or the remedial measures they take once the security of processing requirement has been violated and/or the competent DPA has been notified and has taken relevant action. Specifically, *operational readiness* refers to the technical and organisational measures the data controller has in place when the breach happens and a complaint is filed to the DPA. *Post-factum remedies* include all the possible technical and organisational measures the data controller deploys after the breach, whether before or after the respective DPA suggestion for mitigation.

Table A2: Cluster groups of cases

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	Organisation
Norway	2021	4,750€	-	Internal	-	O	-	✓	✓	Public
Sweden	2021	35,000€	-	Internal	-	O	-	✓	-	Public
Norway	2020	73,000€	-	Internal	-	O,T	-	✓	✓	Public
Germany	2019	105,000€	-	Internal	-	O	✓	✓	✓	Public
Sweden	2020	394,000€	-	Internal	-	O	-	✓	✓	Public
Portugal	2018	400,000€	-	Internal	-	O	-	✓	✓	Public
Netherlands	2020	440,000€	-	Internal	-	O	-	✓	✓	Public

(a) Group 1: Insufficient organisational measures.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	Organisation
Slovenia	2021	0€	-	External	-	O	✓	-	-	Public
Romania	2021	500€	-	External	-	H,O	✓	-	✓	Private
Romania	2020	1,500€	-	Internal	-	H,O	✓	-	-	Private
Romania	2021	2,000€	-	Internal	-	H,O	✓	✓	-	Private
Romania	2020	2,000€	-	External	-	H,T	✓	-	-	Private
Spain	2021	3,000€	-	External	-	H,O	✓	-	✓	Private
Iceland	2020	8,600€	-	External	-	H,O	✓	✓	✓	Public
UK	2021	29,000€	-	External	-	H,O	✓	✓	✓	Private
Ireland	2020	65,000€	-	External	-	H,O	✓	✓	✓	Public
Ireland	2020	75,000€	-	External	-	H,O	✓	✓	-	Public
Spain	2021	150,000€	-	External	-	O	✓	-	-	Private
Italy	2021	4,500,000€	-	External	-	O	✓	-	-	Private

(b) Group 2: Non-technical mistake.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	Organisation
Spain	2020	0€	-	External	-	O,T	✓	✓	✓	Public
Denmark	2020	0€	-	External	-	T	✓	-	-	Private
Romania	2020	500€	-	External	-	H,O,T	✓	-	-	Private
France	2020	3,000€	-	External	-	H,T	-	✓	✓	Public
Spain	2020	3,000€	-	External	-	T	✓	-	-	Private
Hungary	2021	7,000€	-	External	-	O,T	✓	-	✓	Private
Cyprus	2020	15,000€	-	Internal	-	O	✓	✓	-	Private
Poland	2021	22,000€	-	External	-	O,T	✓	✓	-	Public
Italy	2019	30,000€	-	External	-	O,T	✓	✓	-	Public
Hungary	2020	56,000€	-	External	-	H, O, T	✓	-	-	Private
Italy	2020	80,000€	-	External	-	O,T	✓	-	✓	Public
Germany	2019	100,000€	-	External	-	O,T	✓	-	-	Private
Belgium	2021	100,000€	-	Internal	✓	O	✓	-	-	Private
Spain	2021	120,000€	-	External	-	O	-	-	✓	Private
France	2020	250,000€	-	External	-	T	-	-	-	Private
Netherlands	2019	900,000€	-	External	-	O,T	-	✓	✓	Public
France	2020	2,250,000€	-	External	-	O,T	-	-	-	Private
Italy	2020	27,800,000€	-	Internal	-	H,O,T	✓	-	-	Private

(c) Group 3: General breach.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	organisation
Spain	2020	0€	-	External	✓	T	✓	-	-	Private
Czechia	2020	1,000€	-	External	✓	H,T	✓	-	-	Private
Germany	2020	20,000€	-	External	✓	O,T	✓	-	-	Private
Norway	2018	120,000€	-	External	✓	O,T	✓	✓	✓	Public
Greece	2019	150,000€	-	External	✓	O,T	✓	✓	-	Private
Poland	2020	235,000€	-	External	✓	H,T	✓	-	✓	Private
Italy	2020	600,000€	-	External	✓	O,T	✓	-	-	Private
UK	2020	1,434,000€	-	External	✓	O,T	✓	-	-	Private

(d) Group 4: Targeted attack.

Country	Year	Fine	Compliant	Threat	Malicious	Mistake	Incident	Vulnerable	Sensitive	Organisation
Belgium	2020	0€	✓	-	-	-	✓	-	-	Private
Spain	2021	0€	✓	-	-	-	-	-	-	Private
Denmark	2019	0€	✓	-	-	-	-	-	✓	Private
Spain	2019	0€	✓	-	-	-	✓	-	-	Private
Denmark	2021	0€	✓	-	-	-	-	-	-	Public

(e) Group 5: GDPR compliant.

Notes: The cases are ordered by fine size within each sub-table. Mistake type: H=Human, O=Organisational, T=Technical