The Concept of Identifiability in ML Models

Stephanie von Maltzan
Karlsruhe Institute of Technology, Centre for Applied Legal Studies, Vincenz-Prießnitz-Str. 3, 76137 Karlsruhe, Germany

Keywords: Anonymisation, Pseudonymisation, ML Model, Adversarial Attacks, Privacy, Utility.

Abstract: Recent research indicates that the machine learning process can be reversed by adversarial attacks. These attacks can be used to derive personal information from the training. The supposedly anonymising machine learning process represents a process of pseudonymisation and is, therefore, subject to technical and organisational measures. Consequently, the unexamined belief in anonymisation as a guarantor for privacy cannot be easily upheld. It is, therefore, crucial to measure privacy through the lens of adversarial attacks and precisely distinguish what is meant by personal data and non-personal data and above all determine whether ML models represent pseudonyms from the training data.

1 INTRODUCTION

The debate about Privacy Preserving and in particular anonymisation techniques has intensified as a result of an increasing demand for stronger and more comprehensive protection of personal data. In recent years many companies have considered anonymisation to be the answer to all data protection and privacy issues. Companies exploiting anonymisation assume that it cannot breach privacy. This premise poses, nevertheless, some challenges.

Two crucial assumptions should, therefore, be considered. The first point to note is that personal data is only considered anonymous, if it is not possible to identify an individual. The General Data Protection Regulation (GDPR) assumes that there is some leeway in considering data to be anonymous. Determining anonymity is, therefore, fraught with difficulties and depends on criteria that change according to technical progress or even specific analysis, as the GDPR takes into account technical developments to determine anonymous data. This leads to constant uncertainty in the anonymisation process. Government standards are, therefore, indispensable. Secondly, a large body of research on the volatility and vulnerability of machine learning (ML) models points to the problem that training data used for ML models have higher probability of re-identification as a result of adversarial attacks (Fredrikson et al., 2014; Shokri et al., 2017). Based on the premise that anonymisation is not a risk-free mechanism, as increasingly acknowledged (Brasher, 2018; Mehmood et al., 2016; Piras et al., 2019; Pomares-Quimbaya et al., 532019), adversarial attacks against ML models became the focus of research. Overall, the review showed that ML models memorise sensitive information of the data used for training, indicating serious privacy risks (Carlini et al.; Fredrikson et al., 2015; Hayes et al., 2019; Hilprecht et al., 2019; Jayaraman & Evans, 2019; Pyrgelis et al.; Shokri et al., 2017; Song et al.; Yeom et al., 2017). This emerge to the problem that ML models might be personal data and fall, therefore, under the scope of the GDPR. Anonymised data is exempt from the GDPR. If the data is not personal data, as previously assumed for ML models (and the output), the GDPR does not apply.

Machine learning algorithms are regularly trained and evaluated on disjoint data sets. Hence, research and industry have been under the erroneous belief that it is not possible to retrospectively draw conclusions from the ML model about the data used for training. However, some ML techniques - as the above mentioned research has shown - can *remember* the training data of the model antiparallel to the predefined learning process. Despite its "artificial" nature the ML process contains some characteristics of the properties, patterns and correlations from the data used for training and, thus, does not protect from linkage and attribute inference. As indicated above, this raises the question of whether the ML models represent pseudonyms from the training data and could, therefore, fall under the definition of personal data. Classifying models as personal data raises

von Maltzan, S.
The Concept of Identifiability in ML Models.
DOI: 10.5220/0011081600003194
In *Proceedings of the 7th International Conference on Internet of Things, Big Data and Security (IoTBDS 2022)*, pages 215-222
ISBN: 978-989-758-564-7; ISSN: 2184-4976

further far-reaching problems concerning the widely used ML as a Service (MLaaS). In the event that malicious users were able to re-identify data used to train these models, the resulting information leakage and privacy breach would cause serious issues. Therefore, the unexamined belief in anonymisation as a guarantor for privacy cannot be easily upheld.

As a result, the surprising ease of identifying individuals or information about individuals in supposedly anonymous – even synthetic (Stadler et al., 2020b) – datasets creates a great deal of uncertainty about which technical measures are adequate to both legal standards and practical expectations. The well-known tension between utility and privacy is thus amplified. It is crucial to measure privacy through the lens of adversarial attacks and precisely distinguish what is meant by personal data and non-personal data and above all determine whether ML models represent pseudonyms from the training data. It is, therefore, preferable to adopt an approach that incorporates the above mentioned vulnerability and volatility of the models into the training while retaining utility. A GDPR-compliant use of ML models requires technical measures specified by government standards (yet to be developed).

Hence, this work will first address the concept of identifiability and the scope of privacy criteria that lead to effective anonymisation (or pseudonymisation) and transfer these findings to ML models. Building on the principles arising from the GDPR, the Guidelines from the Art. 29 Data Protection Working Group, nowadays known as the European Data Protection Board (EDPB) and the European Union Agency for Cybersecurity (ENISA) as well as the CJEU's Breyer judgement were used to provide the underlying rationale. Further guidance – albeit with a contrary view – was above all also provided by Mourby (Mourby et al., 2018), Stalla-Bourdillon (Hu et al., 2017; Stalla-Bourdillon & Knight, 2017) and Groos (Groos & van Veen, 2020).

Based on this research, the aim of this paper is not only to outline the legally relevant scope of identifiability – which has not been discussed so far in the context of ML models – but also to combine the respective concepts in an interdisciplinary manner. This can only be achieved by drawing on existing research and established legal definitions and concepts. To the best of my knowledge, there is no conceptual and cross-cutting work that is in line with the recent research concerning the legal outcome of adversarial attacks and anonymising effects of ML models and above all the identification of ML models as pseudonyms of the data used for training.

2 THE CONCEPT OF IDENTIFIABILITY

Approaches to define identifiability can be found in the GDPR and are closely linked to the concept of personal data. Personal data under Art. 4 (1) GDPR means any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly. The terms *identified* and *identifiable* are, therefore, of crucial importance to distinguish the different types of data and to determine (Stalla-Bourdillon & Knight, 2017) whether data should be considered personal data.

These vague criteria allow different interpretations, which leads to a dynamic and thus different understanding and perception of the definition. The notion of personal data is, therefore, quite difficult to define. However, the objective depends on the question of what personal data is.
Recital 26 expands the notion of identifiability and makes a distinction between personal data and anonymised data, excluding anonymised data from the scope of the GDPR. In view of the aforementioned risk of re-identification and in order to avoid misunderstandings and conceptual ambiguities, the distinction from anonymisation is of crucial importance. This is primarily due to the fact that uncertainties exist regarding the classification of pseudonymised data as personal data or the classification of technical and organisational measures that are considered pseudonymisation measures (Mourby et al., 2018) and not anonymisation measures. In light of the existing uncertainties, a differentiation of the anonymising (or pseudonymising) effect is necessary.

2.1 Anonymising Effect of ML Models

Anonymisation effectively serves as a privacy protection technique and as a way to remove the personal character of the data. However, as is increasingly acknowledged, anonymisation is not a risk-free mechanism (Brasher, 2018; Mehmood et al., 2016; Piras et al., 2019; Pomares-Quimbaya et al., 532019), especially with regard to ML models; and as demonstrated by Stadler (Stadler et al., 2020a), this also applies to synthetic data.

Anonymisation is regarded as a process whereby a data subject can no longer be identified, directly or indirectly, either by the controller or by a third party on the basis of irreversibly altered personal data. The key factor is that a person is not or no longer identifiable after the data has been anonymised. With

the criterion of *all means reasonably likely to be used* Recital 26 provides guidance in addressing this issue. The decisive factor is whether identifying the data subjects is possible with the data and the additional knowledge. However, the extent to which additional knowledge and means of third parties should be included is a matter of dispute. The previous attempts to define the concept of identifiability and anonymisation from literature and judiciary do not draw a consistent picture.

Following the European Court of Justice's (CJEU) Breyer ruling, the additional knowledge of third parties has to be attributed to the controller if the additional knowledge "constitutes a means which may reasonably be used to identify the data subject" (*Case C-582/14, Patrick Breyer V Bundesrepublik Deutschland*, 2016). This criterion is not met, according to the CJEU, if the identification of the data subject is prohibited by law or practically impossible because it requires a disproportionate effort in terms of time, costs and manpower, so that the risk of identification appears to be insignificant. The legally permissible means are therefore the decisive criterion to be discussed.

Favouring a broad interpretation of personal data the European Data Protection Board (EDPB, formerly Article 29 Working Party) still refers to its Opinion 5/2014 on anonymisation techniques (Working Paper 216), in which it proposes a high threshold for achieving successful anonymisation and refers to a technique comparable to permanent erasure, i.e. "it must not be possible to further process the personal data" (Article 29 Working Party, 2014a). According to this opinion, inference is considered one of the key risks, which is why a very broad definition has been chosen (Article 29 Working Party, 2014a). As far as anonymisation is concerned, WP 216 states that the data set can be considered anonymous if the controller *aggregates* the data at a level where individual events are no longer identifiable.

This Opinion has not become obsolete after the Breyer decision. In determining whether a natural person is identifiable, all the means reasonably available, either to the controller or to any other person, to identify the natural person directly or indirectly should be considered. This includes all objective factors such as the cost and time required for identification, the technology available at the time of processing and technological developments.

The potential additional knowledge of the controller or a third party is, therefore, relevant. The determining factor is whether there are (unlawful) means that can reasonably be used to link the data held by the controller with the additional information from the third party to enable re-identification.

Some researchers assume that only the capabilities of the data controller should be taken into account, thus excluding the capabilities of third parties or at least seeing such capabilities as insignificant (Groos & van Veen, 2020) in regard to time, cost and manpower. Recital 26 states *means reasonably likely to be used (...) by the controller or another person (...)*, which emphasis not only the controller's but also the third person's capabilities. Notwithstanding the fact that non-mandatory law is not binding under Article 288 of the Treaty on the Functioning of the European Union, recitals add a layer of understanding and define what the rules mean in the context of a particular case. The recitals must, therefore, be respected.

Contrary to the court's opinion and statements from the literature, I argue that even unlawful means should be included in the discussion. Attacks by third parties should not be ignored here, as there are often legally impermissible but technically easy to implement measures for re-identifying individuals. For the evaluation of the re-identification potential, it is, therefore, important to consider not only the legally permissible measures, but also the technically possible ones. The evaluation of the likelihood must apply an objective criterion, i.e. it must not depend on the motivation or the intention to obtain the means or to actually use it in a particular case. Taking into account the high risk of re-identification mentioned above it seems questionable whether the CJEU took this into account in its deliberations. The risks of re-identification, therefore, also address the way attackers can identify data subjects in data sets. This means that if prohibited means allow re-identification the data is not considered anonymised. With the attack methods mentioned above it becomes evident that despite disjoint data sets in the ML process inferences cannot be completely prevented. The generated additional knowledge can be accomplished with reasonable effort. With Privacy Preserving ML – as will be shown – there are nevertheless measures that mitigate the risk of re-identification (Article 29 Working Party, 2007) and can be used to a reasonable extent. By implementing technical and organisational measures as part of the training process, and especially by considering the inherent risk of adversarial attacks, all means that could reasonably be used to identify the data subject are taken into account in determining the risks.

The ML models can consequently be interpreted as personal data by re-identifying the data contained in the training data through an adversarial attack. All

information that relates to a person - no matter how trivial or banal it may seem - is considered personal data. If one follows this line of reasoning, information obtained through an adversarial attack is also personal data.

2.2 Pseudonymising Effect of ML Models

As already indicated the concept of additional information is closely linked to that of pseudonymisation. A pseudonym is considered a piece of information that – depending on the pseudonymisation function – are associated to an identifier of a data subject (with different degrees of linkability) and, therefore, carries the risk of being subject to a re-identification attack, as those described above. Pseudonymisation's decisive feature is that, according to Art. 4 No. 5 GDPR, pseudonyms can no longer be associated with a specific person without the use of additional information (ENISA European Union Agency for Network and Information Security, 2019).

Both background knowledge and personal knowledge must be included in the criterion of additional knowledge. The latter is the information that could be kept separately from the dataset by technical and organisational measures whereas background knowledge corresponds to the knowledge that is publicly accessible to an average, reasonably competent individual which cannot be physically separated from the dataset and can have a high impact on re-identification risk. Personal knowledge, on the other hand, can vary from one person to another and represents information that is not publicly accessible to an average, reasonably competent individual, but to some qualified individuals. In combination with anonymised data, this personal knowledge in conjunction with the derived attribute(s) can lead to re-identification or at least disclosure of (potentially sensitive) information about an individual. Therefore, the use of additional information is central to the re-identification risk and on the same time strength of the pseudonymisation. This process can be more or less complex depending on the pseudonymisation function.

It remains to be ascertained if ML models represent pseudonyms.

It is, therefore, sufficient if the data subject can be identified and statements can be made about his or her factual and personal circumstances (Article 29 Working Party, 2007). In the training process certain properties of the training datasets are stored in the model as feature vectors - regardless of whether they are labelled or stored, which basically depends on the application or learning technique. Support Vector Machines or k-nearest neighbour classification methods store the feature vectors whereas neural networks, for example, do not, but can remember them unintentionally (Carlini et al.). In the latter, a model inversion attack, for instance, generates feature vectors similar to those used to train the model by using the outputs obtained from the model. Such training data sets, which consist of a set of features and an associated output, may contain sensitive information - like medical records or images - and thus have quasi-identifiers or values of other features that can be used to identify individuals. According to the GDPR, this information is considered personal data. The ML process is reversible, insofar as an external assignment rule remains and thus a general possibility of re-identification exists.

Based on this, the following constellation was developed to illustrate the effects of Model Inversion Attacks:

> It is assumed that there is a generated data set with personal data A and an ML model B trained with personal data B. Access is given either via the model directly (white box) or via an interface (black box).

In such a constellation model B represents the pseudonymised version of the training data set B while data set A represents the key or the assignment rule which can be used to (partially) re-identify this data. If an attacker has access to A and model B and the model inversion attack is successful, it seems possible to consider not only A but also model B (and its output) as personal data. If model B has been published and A as well as model B are kept by different persons, model B is also considered personal data.

This constellation can also be applied to Membership Inference Attacks and Model Manipulation Attacks. In contrast to Model Inversion Attacks, in a Membership Inference Attack, the shadow models comparatively represent the key or assignment rule, whereas in a (black box) Model Manipulation Attack, the key or assignment rule is considered to be the enriched randomly generated but unique data that can be used to retrieve the information stored in the labels.

This information cannot be obfuscated in ML models that are vulnerable to adversarial attacks to the extent that re-identification is no longer possible. Reversibly pseudonymised data is considered indirectly identifiable information about individuals - even if the disclosure is not made consciously

possible - with the exception of the model manipulation attack - under previously specified conditions. Conceptually, the above is thus close to the notion of pseudonymisation in the GDPR.

Consequently, a ML model attacked by adversarial attacks can, therefore, no longer be considered anonymous. Re-identification of the training data seems to be within the realm of possibility. Such an ML model should, therefore, be subject to similar principles and regulations that apply to identifiable data.

2.3 Technical and Organisational Measures

These results are extremely problematic from the perspective of a researcher or company that wants to use ML models and attribute an anonymising effect to them.

As described above, the personal identifiability of data depends on the context. Therefore, it is necessary to regularly assess whether data can (still) be considered anonymous. As a consequence, personal identifiability has to be determined dynamically and risk-dependently. A change in the situation can also lead to a change in the risk of (re-)identification. This is affected, for example, by the knowledge of third parties or by new developments in (de-)anonymisation techniques (Article 29 Working Party, 2014b) A non-recurrent risk analysis is therefore insufficient. Anonymisation in general is, therefore, subject to a number of uncertainties, especially with regard to the fact that the relevant technical and social circumstances can change rapidly over time. Despite anonymisation, a residual risk may remain for the data subject (Article 29 Working Party, 2014b). These risks also apply to ML models. Inferences cannot be completely prevented, as seen above. It is possible to retrospectively draw conclusions from the ML model to the data used for training. Adversarial attacks can, therefore, lead to a different classification of the (output of) ML models that were previously considered anonymous. Consequently, if vulnerable ML models do not represent anonymous data, methods must be found that can guarantee both utility and data protection. This also applies to the training process of ML models.

The question is how one can reliably defend the above mentioned attacks on pseudonymisation. Firstly, the whole training dataset and all data values should be considered. Secondly, any knowledge and inferences should be eliminated if possible. Therefore, an effective privacy protection technique should be applied. However, the challenge is to ensure privacy protection without reducing utility. The trade-off between protection and utility is apparent.

Based on the risks highlighted above, it is essential to be cautious when using ML models to process sensitive information. One has to consider not only what kind of model is used and how it is provided, but also how the data should be prepared before the training process. Many algorithms commonly used are based on the assumption that they need raw data. However, with Privacy Preserving ML, there are methods (Al-Rubaie & Chang, 2019; Gambs et al., 2021; Jia et al., 2019; Milad Nasr et al., 2018; Mukherjee et al., 2021; Nasr et al., 2019) to reduce the effectiveness of the above attacks while preserving utility. There are several approaches depending on the application and model. Gambs (Gambs et al., 2021) demonstrate, for example, an approach for synthesised data based on Differential Privacy that reduces the risk of adversarial attacks while preserving utility whereas Mukherjee (Mukherjee et al., 2021) optimise current approaches to mitigate Membership Inference Attacks on GAN models that previously resulted in poorer generated sample quality. The authors stated not only that their method provide protection against Membership Inference Attacks “while leading to negligible loss in downstream performances” (Mukherjee et al., 2021) but also that their algorithm prevent memorisation of the training data set. In order to prevent Membership Inference Attacks it is also proposed to limit the number of classes that a model can predict to the most commonly used classes. Avoiding overfitting a model can also be beneficial (Yang et al., 2020). The use of regularisation techniques like dropout (Srivastava et al., 2014) may contribute to prevent overfitting and also to strengthen privacy (Jain et al., 2015) in neural networks. However, no guarantee exists that a model is completely invulnerable to attack. In some cases, it has been shown that such attacks are successful even without overfitting the model (Yeom et al., 2017). However, overfitting is not the only reason that causes Membership Inference Attacks. Even if ML models are overfitted they could *leak* different amounts of membership information. Specifically, due to their different structures, they might remember information about the data used for training.

Nonetheless, if no raw data is used for training, the risk assessment of whether personal data is affected could be completely different. However, it should clearly be stated that previous approaches to defend against Membership Inference Attacks have limited effect on Model Inversion Attacks. To my

knowledge, there is no known method that adequately defends both attacks.

3 CONCLUSION

As outlined above, the ambiguous terms of the GDPR as well as the classification of ML models as pseudonyms cause a number of problems which need to be addressed. It is, therefore, important that adjusted and comprehensive guidelines and best practices are elaborated to eliminate the uncertainty as to which technical offerings meet legal standards as well as adequately match practical expectations. Measuring privacy through the lens of adversarial attacks is therefore crucial. As stated earlier, one should not rely solely on the belief in a data anonymising ML model but also that the model be assessed with respect to a privacy test based on adversarial attacks to quantify the privacy protection provided by the processing method. It is, therefore, of crucial importance that the question of whether a data set is considered to be personal, pseudonymised or anonymised can be answered without ambiguity. The GDPR could apply depending on the outcome.

Concerning ML, anonymous data itself does not guarantee privacy without the support of other techniques. Which does not mean that anonymisation is a useless tool, but it must be applied with the support of other Privacy Preserving mechanisms. In order to properly assess and mitigate privacy threats, a risk-based approach to be evaluated regularly should be adopted, taking into account the purpose and overall context of the processing of personal data, as well as the degree of utility and scalability. Choosing technical and organisational measures depends on various parameters, e.g. the level of data protection and the utility of the pseudonymised data, which may lead to different approaches or even variations of approaches. The trade-off between utility and data protection should carefully be analysed. On one side, utility need to be optimised for the intended purposes while keeping a strong data protection. This field of privacy preserving ML is gradually becoming a highly debated topic and is a challenging one, with a high dependency on matters of context, involved entities, data types and additional knowledge. There is consequently no single approach that fit for all possible scenarios. A one-size-fits-all solution is not sufficient. Applying robust pseudonymisation to reduce the risk of adversarial attacks and maintain the utility of the pseudonymised data requires a high level of competence.

It is therefore necessary to develop a holistic and legally binding concept consisting of governmental and technical measures. The following criteria can provide preliminary orientation.

The Data Protection Authorities as well as the EDPB should, therefore, provide practical guidance with regard to the assessment of the risk and best practices in the field of pseudonymisation and anonymisation. The definition and explanation of the state of the art is of crucial importance. Furthermore, the notion of identifiability and anonymisation needs to be readdressed. The adversarial attacks are evolving which leads to more and more challenging anonymisation and pseudonymisation process. The authorities should, therefore, extend the current techniques to more advanced solutions addressing the special challenges appearing with ML models. The relevant EU institutions should provide support and disseminate these efforts. To achieve more legal certainty, manageable standards for anonymisation associated with presumption rules in the event of compliance should be established at EU level. Considering the advancing technical development, it would also be advisable to provide the standards with a temporal validity. Furthermore, standardised test procedures should be developed to check supposedly anonymous data for personal identifiability. Guidance on appropriate or inappropriate techniques is therefore indispensable. The Article 29 Data Protection Working Party as well as ENISA provided guidance about the use of various privacy technologies – including Differential Privacy. The anonymisation and pseudonymisation techniques should be revisited concerning the aforementioned highly probably adversarial attacks and the inherent flexible determinability of the degree of anonymity. Orientation could be formulated at EU level by means of a guideline for determining suitable anonymisation procedures, as well as addressing criteria for determining appropriate procedural parameters. The regularly updated technical guideline of the Bundesamt für Sicherheit in der Informationstechnik (Federal Office for Information Security in Germany) „Kryptographische Verfahren: Empfehlungen und Schlüssellängen“ (Cryptographic Procedures: Recommendations and Key Lengths), which gives recommendations for cryptographic procedures could be the role model in this respect.

This work has the limitation of being purely theoretical. Nonetheless, it provides not only a revisited evaluation of the notion of identifiability and the classification of ML models as pseudonymised data but also an insight into the inherent risk for ML models as well as sufficient technical and

organisational measures which has to be standardised. Overall, the work provided ample background information on relevant concepts concerning anonymisation and pseudonymisation and how to deal with the fact that the ML process does not have an anonymising effect. Clearly, more practical interdisciplinary work linking up adversarial attacks and Privacy Preserving techniques with regulation and data protection efforts needs to ramp up. Above all, it is important that the uncertainty associated with adversarial attacks is surmounted by governmental and technical standards, which will be developed in the future.

REFERENCES

Al-Rubaie, M., & Chang, J. M. (2019). Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security & Privacy Magazine*, *17*(2), 49–58. https://doi.org/10.1109/MSEC.2018.2888775

Article 29 Working Party. (2007). *WP 136: Opinion 4/2007 on the concept of personal data*.

Article 29 Working Party. (2014a). *Opinion 05/2014 on Anonymisation Techniques WP216*.

Article 29 Working Party. (2014b). *WP 217: Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC,*

Brasher, E. A. (2018). Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation. *Columbia Business Law Review*, *2018*, 209. https://heinonline.org/HOL/Page?handle=hein.journals/colb2018&id=215&div=&collection=

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium (SEC'19). USENIX Association*.

Case C-582/14, Patrick Breyer v Bundesrepublik Deutschland. (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62014CN0582

ENISA European Union Agency for Network and Information Security (2019). Guidelines-on-shaping-technology-according-to-GDPR-provisions. https://www.ledecodeur.ch/wp-content/uploads/2019/12/Guidelines-on-shaping-technology-according-to-GDPR-provisions.pdf

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In I. Ray, N. Li, & C. Kruegel (Eds.), *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15* (pp. 1322–1333). ACM Press. https://doi.org/10.1145/2810103.2813677

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., & Ristenpart, T. (2014). Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In K. Fu (Ed.), *23rd USENIX Security Symposium: August 20 - 22, 2014, San Diego, CA* (pp. 17–32). USENIX Association.

Gambs, S., Ladouceur, F., Laurent, A., & Roy-Gaumond, A. (2021). Growing synthetic data through differentially-private vine copulas. *Proceedings on Privacy Enhancing Technologies*, *2021*(3), 122–141. https://doi.org/10.2478/popets-2021-0040

Groos, D., & van Veen, E. (2020). Anonymised Data and the Rule of Law. *European Data Protection Law Review*, *6*(4), 498–508. https://doi.org/10.21552/edpl/2020/4/6

Hayes, J., Melis, L., Danezis, G., & Cristofaro, E. D. (2019). LOGAN: Membership Inference Attacks Against Generative Models. *Proceedings on Privacy Enhancing Technologies*, *2019*(1), 133–152. https://doi.org/10.2478/popets-2019-0008

Hilprecht, B., Härterich, M., & Bernau, D. (2019). Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. *Proceedings on Privacy Enhancing Technologies*, *2019*(4), 232–249. https://doi.org/10.2478/popets-2019-0067

Hu, R., Stalla-Bourdillon, S., Yang, M., Schiavo, V., & Sassone, V. (2017). *Bridging Policy, Regulation, and Practice? A Techno-Legal Analysis of Three Types of Data in the GDPR.*

Jain, P., Kulkarni, V., Thakurta, A., & Williams, O. (2015, March 6). *To Drop or Not to Drop: Robustness, Consistency and Differential Privacy Properties of Dropout*. https://arxiv.org/pdf/1503.02031

Jayaraman, B., & Evans, D. (2019). *Proceedings of the 28th USENIX Security Symposium: August 14-16, 2019, Santa Clara, CA, USA*. USENIX Association. https://atc.usenix.org/system/files/sec19-jayaraman.pdf

Jia, J., Salem, A., Backes, M., Zhang, Y., & Gong, N. Z. (2019). Memguard. In L. Cavallaro (Ed.), *ACM Digital Library, Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (pp. 259–274). Association for Computing Machinery. https://doi.org/10.1145/3319535.3363201

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., & Guo, S. (2016). Protection of Big Data Privacy. *IEEE Access*, *4*, 1821–1834. https://doi.org/10.1109/ACCESS.2016.2558446

Milad Nasr, Reza Shokri, & Amir Houmansadr (2018). Machine Learning with Membership Privacy using Adversarial Regularization. In *ACM Conference on Computer and Communications Security (CCS)*. https://www.researchgate.net/publication/326782376_Machine_Learning_with_Membership_Privacy_using_Adversarial_Regularization

Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., Smith, H., Aidinlis, S., & Kaye, J. (2018). Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK. *Computer Law & Security Review*, *34*(2), 222–233. https://doi.org/10.1016/j.clsr.2018.01.002

Mukherjee, S., Xu, Y., Trivedi, A., Patowary, N., & Ferres, J. L. (2021). privGAN: Protecting GANs from membership inference attacks at low cost to utility. *Proceedings on Privacy Enhancing Technologies*, *2021*(3), 142–163. https://doi.org/10.2478/popets-2021-0041

Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy: Sp 2019 : San Francisco, California, USA, 19-23 May 2019 : Proceedings* (pp. 739–753). IEEE. https://doi.org/10.1109/SP.2019.00065

Piras, L., Al-Obeidallah, M. G., Praitano, A., Tsohou, A., Mouratidis, H., Gallego-Nicasio Crespo, B., Bernard, J. B., Fiorani, M., Magkos, E., Sanz, A. C., Pavlidis, M., D'Addario, R., & Zorzino, G. G. (2019). DEFeND Architecture: A Privacy by Design Platform for GDPR Compliance. In S. Gritzalis, E. R. Weippl, S. K. Katsikas, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Trust, Privacy and Security in Digital Business* (pp. 78–93). Springer International Publishing.

Pomares-Quimbaya, A., Sierra-Múnera, A., Mendoza-Mendoza, J., Malaver-Moreno, J., Carvajal, H., & Moncayo, V. (532019). Anonylitics: From a Small Data to a Big Data Anonymization System for Analytical Projects. In *Proceedings of the 21st International Conference on Enterprise Information Systems* (pp. 61–71). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0007685200610071

Pyrgelis, A., Troncoso, C., & Cristofaro, E. D. Knock Knock, Who's There? Membership Inference on Aggregate Location Data. http://arxiv.org/pdf/1708.06145v2

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May 22–26). Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 3–18). IEEE. https://doi.org/10.1109/SP.2017.41

Song, C., Ristenpart, T., Shmatikov, V., & 2017. Machine Learning Models that Remember Too Much. In *Thuraisingham, Evans et al. (Hg.) 2017 – Proceedings of the 2017 ACM* (pp. 587–601). https://doi.org/10.1145/3133956.3134077

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. https://dl.acm.org/doi/pdf/10.5555/2627435.2670313

Stadler, T., Oprisanu, B., & Troncoso, C. (2020a, November 13). *Synthetic Data -- Anonymisation Groundhog Day*. https://arxiv.org/pdf/2011.07018.pdf

Stadler, T., Oprisanu, B., & Troncoso, C. (2020b, November 13). *Synthetic Data -- Anonymisation Groundhog Day*. https://arxiv.org/pdf/2011.07018

Stalla-Bourdillon, S., & Knight, A. (2017). Anonymous Data v. Personal Data — A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data. *Wisconsin International Law Journal*, *34*(2), 285–322. http://hosted.law.wisc.edu/wordpress/wilj/files/2017/12/Stalla-Bourdillon_Final.pdf

Yang, Z., Shao, B., Xuan, B., Chang, E.□C., & Zhang, F. (2020, May 8). *Defending Model Inversion and Membership Inference Attacks via Prediction Purification*. https://arxiv.org/pdf/2005.03915v1.pdf

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2017, September 5). *Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting*. https://arxiv.org/pdf/1709.01604