gretel

Understanding Synthetic Data, Privacy Regulations, and Risk:

The Definitive Guide to Navigating the GDPR and CCPA

Table of contents

Summary

This guide examines two European Union (EU) privacy regulations — Recital 26 of the GDPR and an Article 29 Data Protection Working Party opinion — to explore how risks associated with sensitive data can be mitigated by using synthetic data generation (SDG). Synthetic data is artificially annotated information that is generated by computer algorithms or simulations that is commonly used as an alternative to real-world data. Gretel's services provide privacy-preserving technologies to reduce privacy risks significantly by implementing best practices such as de-identification, overfitting prevention, differential privacy, similarity filters, outlier filters, and continuous monitoring. Gretel's services and best practices can provide safeguards against known adversarial attacks and provide protection from the applicability of GDPR, CCPA, and similar privacy regulations, to a dataset.

Overview

The European Union's General Data Protection Regulation (GDPR) has shaped how global companies design and build products, software, and systems since its inception. As data has become the foundation of most products and services, companies are increasingly collecting and analyzing customer data, and prioritizing respect for the privacy rights of customers while complying with various privacy regulations, such as the GDPR.

This guide provides a summary of European Union (EU) privacy regulations and how risks associated with sensitive data can be mitigated by using *synthetic data generation* (SDG). Synthetic data is artificially annotated information that is generated by computer algorithms or simulations that is commonly used as an alternative to real-world data. While artificial, synthetic data statistically reflects the insights of real-world data. Recent advancements in generative AI have made it possible to create synthetic data with accuracy that can be as effective as real-world data for training AI models, powering statistical insights, and fostering collaboration with sensitive datasets while offering strong privacy guarantees.

> At Gretel, we believe that synthetic data generated using our services could be considered anonymized information and thus not be subject to the GDPR or similar applicable privacy laws, including the California Consumer Privacy Act (CCPA). While the EU data protection authorities haven't directly opined on the nature of synthetic data, existing guidance in the EU regulations examined below supports the view that it would constitute anonymized data.

But any effort to implement strategies to mitigate privacy risk requires:

- A thorough understanding of the current, applicable regulations.
- An understanding of what synthetic data and related privacy-enhancing techniques are and how they can be used to significantly reduce or possibly eliminate risk.
- Best practices for implementing such a solution.

This guide covers these topics to provide readers with actionable knowledge on mitigating privacy risks with synthetic data.

Key Definitions and Regulatory Concepts

Before examining some of the key regulations, let's establish some shared vocabulary that we'll use throughout this guide. We'll cover the definition of synthetic data that Gretel uses and source definitions for pseudonymization and anonymization from relevant EU documents.

Definition of Synthetic Data

> Fundamentally, synthetic data is annotated information that computer simulations or algorithms generate as an alternative to real-world data.

Synthetic data is generated using a class of machine learning models that learn the underlying relationships in a real-world dataset and can generate new data instances. High-quality synthetic data will contain the same statistical relationship between variables (e.g., if real-world data reveals that weight and height are positively correlated, then this observation will hold true in the synthetic data). However, records of a synthetic dataset will not have a clear link to records of the original, real-world data that it is based on. Synthetic data is one step removed from real-world data. This means that synthetic data generated with the right level of privacy protection can provide safeguards against known adversarial attacks, something traditional anonymization techniques like masking or tokenization cannot promise.

Regulatory Concepts

- **Pseudonymization** is a "*data management and de-identification procedure by which personally identifiable information fields within a data record are replaced by one or more artificial identifiers, or pseudonyms.*"[1]
- **Anonymization** is the "*process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party.*"[2] In contrast to pseudonymization, anonymization is intended to prevent the re-identification of individuals within the dataset.

[1] "General Data Protection Regulation". 4(5)

[2] ISO 25237:2017 Health informatics -- Pseudonymization. ISO 2017 p. 7

Relevant EU regulations and opinions

With this conceptual foundation, we can examine the most relevant EU regulations. While EU privacy regulations and other related writings are voluminous, there is a subset of these documents that we believe forms a core set of requirements. The most relevant documents are:

- **Recital 26** of the GDPR
- **Article 29** Data Protection Working Party Opinion

Here are some of the key features of both documents for understanding their privacy impact.

Recital 26

Recital 26 of the GDPR makes several important statements:

- It makes a distinction between anonymization and pseudonymization, with the latter NOT providing protection against privacy regulations.
- It makes strong statements around anonymized information, saying, “The principles of data protection should therefore not apply to anonymous information” and “This Regulation does not therefore concern the processing of such anonymous information.”[3]

While the statements are strong, there are several caveats about Recital 26 that readers should note:

- Recital 26 is not binding
- The use of the term “should” indicates that there is some ambiguity to the Recital
- It does say that “account should be taken of all the means reasonably likely to be used” to identify a natural person

While aspects of the Recital seem to leave some ambiguity, other EU writings are very helpful in clarifying the criteria for when anonymization can successfully make a dataset not subject to privacy regulations.

[3]In the context of this document, data and information are interchangeable.

Article 29 Data Protection Working Party Opinion 05/2014 (A29)

This lengthy document is very helpful in clarifying criteria for anonymizing information, such as those noted in Recital 26. In this course of this clarification, it:

- States that “Anonymized data do fall out of the scope of data protection Legislation” (p. 3) but repeats some of the caveats from Recital 26.
- Adds the caveat that “their application [must be] engineered appropriately.”
- Reiterates that the measure of the success of an anonymization technique is the ability to “identify a natural person.”
- Lays out three, detailed tests to determine if a technique meets the standard for identifying a natural person.
- Provides a survey of anonymization techniques and their efficacy relative to the three tests.
- Provides a list of best practices to be observed when engineering an anonymization strategy.

The three tests are:

- *Singling Out* – the ability to isolate some or all records which identify an individual from a dataset
- *Linkability* – the ability to link at least two records concerning the same data subject or group of data subjects
- *Inference* – the possibility to deduce the value of an attribute from the values of a set of other attributes

A29 then discusses these tests against several anonymization technologies, and Gretel supports these technologies mentioned in **A29**:

- Substitution
- Hashing/Tokenization
- Noise addition
- Differential Privacy

A29 indicates — and Gretel agrees — that Differential Privacy is one of, if not the strongest, standards, with other techniques that rely on noise addition being very effective. Gretel's view is that using these techniques *in combination* can provide very strong protection from the applicability of GDPR privacy regulations to a dataset, *provided* that best practices are followed. We'll cover a set of best practices in a later section.

> Caveat: Note that even **A29** states that there is "no prescriptive standard in EU legislation" for which anonymization technique is to be used (A29, p.6). Properly applying techniques and best practices is critical, and Gretel can assist customers in their selection and application.

De-identification, anonymization, and the CCPA

While the CCPA is an American rather than EU regulation, many companies dealing with EU privacy regulations also aim to adhere to the CCPA, which was a first-of-its-kind privacy law in the United States.

Under the CCPA, the relevant concept is "de-identification" rather than anonymization. De-identification is a lower standard than the anonymization required by GDPR and could be satisfied with Gretel's Transform services by simply removing, replacing, or encrypting personal identifiers in data. The additional step of applying the use of Gretel's anonymization techniques via synthetic data generation is not necessary. However, Gretel recommends that customers seeking to ensure their data is outside the scope of CCPA use Gretel's synthetic data.

De-identified information is exempt from the CCPA. "De-identified information" under the CCPA means information that cannot reasonably identify, relate to, describe, be capable of being associated with or be linked, directly or indirectly, to a particular consumer. It requires that the business that uses such de-identified information implement technical safeguards and business processes that prohibit re-identification and processes to prevent inadvertent release of the de-identified information. This definition effectively exempts data de-identified in accordance with HIPAA standards. When customers use the Gretel "best practices" described below to maximize data anonymization of their datasets, they are significantly more likely to make re-identification difficult.

Gretel's solution for mitigating privacy risks

As a synthetic data platform that uses generative AI, Gretel provides multiple services and privacy-preserving technologies that significantly reduce privacy risks.

When using AI-based generative models, there is inherent stochasticity (randomness or noise) in generating synthetic data that offers inherent privacy benefits. Commonly used methods, like batch-based optimization (e.g., stochastic gradient descent) and regularization (e.g., dropout layers in deep neural networks), are designed to ensure that the models generating synthetic data do not memorize their inputs. They are intended to improve the generalization of models, which aids anonymization.

Gretel's synthetic data models can be augmented by applying "best practices," privacy-preserving techniques to achieve a higher level of anonymization. Such best practices are described further below.

When it is critical to include sensitive and potentially identifying fields like names in a dataset, a user will first replace real names with fabricated ones using Gretel Transform and then pass the dataset to a synthetic data-generating model. The resulting synthetic data will not have any such identifiers from the real-world dataset.

Additional privacy mechanisms, such as differential privacy and privacy filters, can be used with Gretel models. Differential privacy is a mathematical standard discussed in Article 29 that users can configure Gretel models to utilize. Gretel also introduces privacy-enhancing technologies, including privacy filters that aim to thwart adversarial attacks by ensuring that no synthetic record is overly similar to any record in the real-world data or is an outlier. These additional technologies provide strong protection against data linkage, membership inference, and re-identification attacks with only a minimal loss in data accuracy or utility.

Using synthetic data models can mitigate the likelihood of singling out an individual by linking records relating to an individual or inferring information about an individual. However, if a dataset contains information primarily about one individual, then the likelihood of exposure for that individual increases. This caveat is not unique to synthetic data and applies to every anonymization method.

Some anonymization techniques show inherent limitations. These limitations must be considered seriously before data controllers craft an anonymization process using a given technique. They must have regard for the purposes to be achieved through anonymization — such as protecting individuals' privacy when publishing a dataset or allowing a piece of information to be retrieved from a dataset.

With Gretel's services, anonymization efforts can adopt several levels of protection, both to offer increasing guarantees for the protection of personal data and also to help find the right balance between privacy and data utility appropriate for the use case. While all three of the levels outlined below provide protection from the GDPR, implementations that adopt the "Better" or "Best" approaches defined below most optimally minimize risk.

- **Good**: De-identification and synthetic data generation (without any additional protections)
- **Better**: De-identification and synthetic data generation with similarity filters and outlier filters
- **Best**: De-identification and synthetic data generation with differential privacy, overfitting prevention, and continuous monitoring

Best practices for anonymization

The EU regulations we've covered, particularly **A29** (pp. 24 - 25), specify best practices for applying anonymization techniques.

A29 suggests that:

- "The optimal solution should be decided on a case-by-case basis."
- "A thorough evaluation of the identification risk should be performed."
- The anonymization technique used should be disclosed.
- "Rare attributes / quasi-identifiers should be removed from the dataset." (Gretel's Transform services can be used to implement this best practice.)

It also suggest that companies:

- Identify new risks, re-evaluate, and adjust regularly to monitor and control for risks.
- "Take into account the identification potential of the non-anonymized portion of a dataset (if any) [to determine the risk of] possible correlations between attributes."
- Consider all the relevant contextual elements such as sample size, availability of public information sources, and whether the information will be released to third parties.
- Consider the appeal of the data to possible attackers based on the "sensitivity of the information and nature of the data"

Where possible, Gretel's platform follows these best practices automatically, or the Gretel team can assist with ensuring that all of these are followed.

Interestingly, one of the best practices mentioned in A29 is dealt with automatically by Gretel's technical approach. A29 states that "When relying on differential privacy (in randomization), account should be taken of the need to keep track of queries so as to detect privacy-intrusive queries." This best practice is relevant for systems that place a differential privacy-based filter between the raw data and the entity that is querying that data. With Gretel's approach, where the synthetic data is generated and separated entirely from the original data, this "multi-query risk" is eliminated.

Additionally, Gretel recommends this best practice:

The fewer the records per individual, the better: the more records in a dataset about a single individual, the higher the likelihood of exposure for that individual. While this is not always possible, pre-processing the data to limit the number of records containing information about any one individual as a fraction of all individuals in the data increases the effectiveness of every anonymization method, including synthetic data.

Privacy best practices implemented by Gretel

Many privacy-focused systems or platforms require developers or users to implement best practices manually. Gretel, however, implements many automatically.

These fall into three categories:

1 **Best practices to implement before model training:**

- De-identification: this can be done by pre-processing real-world data to detect and replace personal data with fake data using Gretel Transform before training synthetic data models. Effective de-identification includes named-entity recognition (NER) to also detect and replace personally identifiable information (PII) present as part of both structured and unstructured text fields. While de-identification on its own is not sufficient to meet GDPR anonymization standards, it provides an effective safeguard against synthetic data models inadvertently memorizing real PII in their training data.

2 **Best practices to implement during model training:**

- Overfitting prevention: this includes methods such as batch-based optimization, regularization, and early stopping, which are designed to ensure that the models generating synthetic data do not memorize their inputs.
- Differential privacy: a provable mathematical guarantee of privacy obtained by inserting specific types of noise in the process of training synthetic data models. It is effective against both known and unknown attacks aimed at inferring information about the original data.

3 **Best practices to implement after model training:**

- Similarity filters: a post-processing checkpoint that actively removes any synthetic data record that is overly similar to a training record, ensuring that no such record slips through the cracks, even in the case of accidental overfitting.
- Outlier filters: a post-processing checkpoint that actively removes any synthetic record that is an outlier with respect to the training data. Outliers revealed in the synthetic dataset can be exploited by membership inference attacks, attribute inference, and a wide variety of other adversarial attacks.

- Monitoring: the regular monitoring and evaluation of synthetic data models and their generated data ensure that they continue to meet privacy requirements beyond their initial deployment.

For more detail on the synthetic data privacy and protection features of Gretel's services, please visit Gretel's Privacy Protection FAQ.

Conclusion

At Gretel, we believe that the proper implementation of a privacy solution using Gretel services can dramatically reduce — and possibly eliminate — the risk that a dataset is subject to privacy regulations. As discussed in this guide, it is vital that the right techniques are used, and that best practices are followed during implementation and thereafter. Gretel's team is happy to assist customers with the implementation of privacy solutions using synthetic data.

For advice or support with privacy issues using Gretel services, please contact support@gretel.ai.

Frequently Asked Questions

1 **Is synthetic data the same as pseudonymized data, which is considered personal data under the GDPR?**

No. Pseudonymized data is considered personal data and is thus subject to the GDPR. Pseudonymized data differs from anonymized data, which is not personal data under the GDPR.[4] As described above, since synthetic data is akin to anonymized data, synthetic data is not pseudonymized data. Pseudonymization of data refers to a procedure whereby personal identifiers in a set of information are replaced with artificial identifiers or pseudonyms. It merely reduces the linkability of a dataset with the original identity of a data subject and is a useful security measure. Pseudonymization is not a method of anonymization. In pseudonymization, individuals are not identifiable from the dataset itself but can be identified by referring to other information held separately. By contrast, anonymization means that individuals are not identifiable and cannot be re-identified by any means reasonably likely to be used.

2 **Can my data be re-identified once it has been synthesized by Gretel?**

No. Once data is de-identified and synthesized via Gretel, re-identification is significantly reduced or no longer possible, depending on which of Gretel's privacy filters are used before synthetic data generation. If properly generated using Gretel's data privacy and protection "best practices" described above, the possibility of synthetic data being attributed to a specific data subject is significantly mitigated. The Information Commissioner's Office, the UK's data protection regulator, in its draft guidance on anonymization and pseudonymization, has recognized synthetic data as a privacy-enhancing technology that aims to weaken or break the connection between an individual in the original personal data and the derived data.[5]

3 **Is the initial dataset from which the synthetic data is created regulated under the GDPR or other privacy laws?**

Yes. As discussed above, once the original raw data is synthesized, it is no longer tied to an identified or identifiable natural person. However, because

[4] Recital 26 of the GDPR, https://www.privacy-regulation.eu/en/recital-26-GDPR.htm

[5] https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf, page 6.

anonymization is a technique applied to personal data to achieve irreversible deidentification, the personal data in the raw dataset must have been collected and processed in compliance with the GDPR or other applicable privacy laws. Further, the Working Party (A29) points out that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of its dataset (e.g., after removal or masking of identifiable data), the resulting dataset is still personal data.[6]

4 Do I need to list Gretel as a GDPR data processor?

When using Gretel in a SaaS or Hybrid Cloud deployment, Gretel will be a data processor but not a data controller. If you are using the Gretel Cloud infrastructure service to train your synthetic models, the original data is only required during model training and is not required to be saved on the Gretel cloud instance. In this mode, Gretel will be a data processor of this initial raw dataset until it is deleted, and the customer will be the data controller of the personal data in the raw dataset. When deploying in Hybrid Cloud, sensitive data used to train synthetic data models never leaves your VPC or cloud environment.

As a data processor, Gretel agrees to provide assistance to the customer so that you can demonstrate your privacy obligations in your role as data controller. For purposes of the CCPA, Gretel would be considered a service provider of its customers who meet the definition of “business” under the CCPA and would need to represent to its customers who are subject to the CCPA that it does not sell personal data.

[6] Article 29 Working Party, Opinion 05/2014 on Anonymization Techniques, WP216, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, page 9.