Statice

The path to data anonymity

From de-identified to privacy-preserving synthetic data

Table of Contents

Summary

This presentation is a resource for organizations seeking to gain a deeper understanding of data protection techniques. It presents and compares common privacy preservation methods and explains the distinction between de-identified and anonymized data. Finally, it introduces privacy-preserving synthetic data as a compliant and granular response to the shortcomings of traditional methods.

At Statice, our mission is to enable organizations to unlock the potential of their data while safeguarding individuals' privacy. Privacy-preserving synthetic data represents an opportunity in that regard. This presentation also shares our experience in how customer-centric organizations benefit from using privacy-preserving synthetic data.

Introduction

The stakes of data privacy enforcement

In 2019, Forbes listed data privacy as the most important issue for the decade to come. At the same time, leveraging data is more than ever a predictor of success for organizations. Forrester estimated that "insights-driven" businesses will grow 8-10x faster than the global economy by 2021, driving 1.8 trillion in revenue along the way. But for many data-driven organizations, complying with regulations and tackling growing privacy and security concerns represents a challenge.

Today, businesses are operating in a volatile regulatory environment. Fast-evolving data protection laws are constantly reshaping the data landscape. Gartner predicted that "*by 2023, 65% of the world's population will have its personal information covered under modern privacy regulations, up from 10% today.*"

Companies must adapt to new requirements and restrictions. Failure to comply with such regulations exposes you to severe financial penalties from regulators, as well as risks of significant consumer backlash and the subsequent impact on company value. Since the introduction of the European General Data Protection Regulation (GDPR), European regulators have already issued over 1,5 billion euros worth of fines for non-compliance.

In parallel, data breaches continue to plague organizations globally, eroding consumers' trust. The Consumer Identity Breach Report predicts that 2022 will top last year's number of breaches, with already over 11 billion consumer records exposed over the past three years in the US only.

In the end, companies have sufficient motivations to ensure that their data protection measures are sufficient and reliable. But to achieve this, it is essential to

understand the ability of the technological options available to ensure efficient use of data while maintaining consumer privacy and regulatory compliance.

De-identification vs. anonymization

De-identification refers to the process of **separating** Personally Identifiable Information (PII), such as names or phone numbers, from sensitive data. On the other hand, anonymization refers to the process of **irreversibly transforming sensitive data** to prevent the identification of individuals.

Modern data privacy laws, such as the GDPR and the California Consumer Privacy Act (CCPA), clearly differentiate between de-identified and anonymized data. The GDPR defines anonymous data as "*information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable*". Thus, if data can lead to the re-identification of a person, it is legally not considered anonymous.

Today, many de-identification methods are incorrectly referred to as anonymization. Indeed, removing PIIs from sensitive data doesn't render the data anonymous. Researchers have repeatedly proven that solely removing or redacting PIIs in a dataset leaves a high risk of re-identification.

Netflix provided a famous illustration of this risk. In 2007, the company ran a contest to improve its movie recommendation algorithm. They released a supposedly anonymized database where some user information had been removed. This method of de-identification turned out to be inadequate. Researchers proved that such datasets could not only lead to the re-identification of users but also reveal more information about them, such as political affiliations in this case. As a result, Netflix ended up settling a privacy lawsuit for an undisclosed amount.

Legally, **this distinction between de-identified and anonymized data is crucial**. De-identified data is subject to personal data protection laws. Meaning that, regardless of the method used to protect data privacy, organizations can only process de-identified data under the conditions stipulated by data protection laws.

On the other hand, **anonymized data is not subject to personal data protection laws.** As stated by GDPR, *"the principles of data protection should therefore not apply to anonymous information*". For businesses working with sensitive data, this means that anonymizing their data alleviates the processing restrictions usually associated with the processing of personal data.

Anonymized data can be processed the same way as non-personal data. True data anonymization allows companies to work with sensitive data to comply with regulations and solve security and privacy concerns.

The methods presented in the next section belong to the category of de-identification. While they are a standard first layer of privacy and compliance frameworks, they also present privacy limitations that require additional protection efforts.

De-identification Methods

In this section, we take a closer look at the characteristics, benefits, and limitations of the most common de-identification methods:

- Pseudonymization
- K-anonymization
- l-diversity & t-closeness

Pseudonymization

Pseudonymization is one of the first methods developed to protect individuals' privacy. You remove PIIs, such as names and telephone numbers, from the data. The other "less sensitive" values, zip codes or ages for example, are left untouched. These types of attributes are referred to as quasi-identifiers.

phone	sex	zip code	income
202-555-0104	male	75080	$60,000
202-555-0172	female	75085	$80,000
202-555-0140	female	75090	$120,000
202-555-0182	male	30040	$0
202-555-0174	male	30035	$70,000
202-555-0173	female	30080	$50,000
202-555-0183	male	30023	$55,000

Original dataset

phone	sex	zip code	income
xxx	male	75080	$60,000
xxx	female	75085	$80,000
xxx	female	75090	$120,000
xxx	male	30040	$0
xxx	male	30035	$70,000
xxx	female	30080	$50,000
xxx	male	30023	$55,000

Pseudonymized dataset

Figure 1: The two datasets portray how an original dataset becomes a simple pseudonymized dataset, where phone numbers have been removed.

There are several technical approaches to pseudonymization (masking, tokenization, encryption, etc.) But whether PIIs are replaced with fictional values or removed from the data, these processes ultimately produce data that does not contain unique identifiers. As shown in Figure 1, the phone numbers of the original dataset are replaced with placeholders, "xxx", in the pseudonymized dataset.

Pseudonymization allows organizations to partially reduce the privacy risks related to the processing of personal data. But it also presents privacy limitations. In fact, solely removing PII from a dataset doesn't guarantee that data cannot be re-identified.

Researchers have continuously disproven the privacy-preserving aspect of these methods by performing attacks and exposing vulnerabilities, as presented below.

One common vulnerability is the fact that companies often keep a secondary dataset with the original data, allowing for the exact matching of identifiers to the pseudonymous data. The release of pseudonymized data sets also carries a high risk of re-identification because it can easily be linked to secondary data sources and, ultimately, expose individuals. This is known as a linkage attack.

A famous example of such re-identification dates back from the late '90s when Dr. Latanya Sweeney applied a linkage attack technique on a health record data set. It consisted of using auxiliary information sources to re-identify individuals from de-identified data. Dr. Sweeney successfully exposed the identity of the Governor of Massachusetts by linking hospital records with public electoral records.

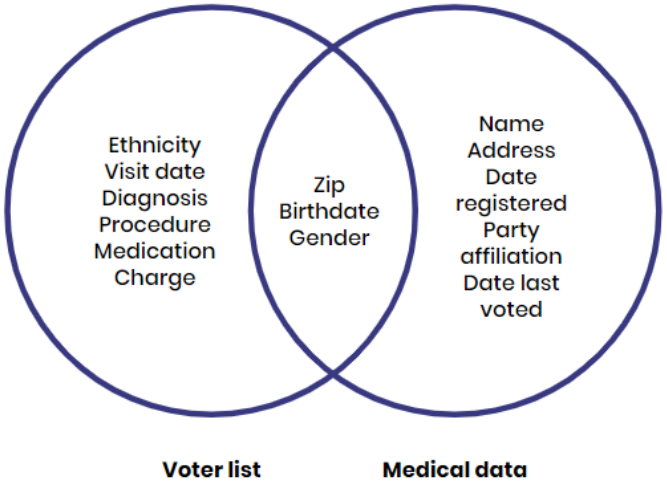


Figure 2: Illustration of the method used by Dr. Sweeney to re-identify the governor from pseudonymized medical data, using publicly available electoral records.

This privacy limitation explains why the GDPR requires companies to treat pseudonymized data as personal data. Indeed, from the legal perspective, *"personal data which have undergone pseudonymization, […] **should be considered to be information on an identifiable natural person** (Recital 26)* […]"

Its limitations regarding privacy guarantee make it an insufficient privacy measure on its own. According to the GDPR, pseudonymization is "*not intended to preclude any other measures of data protection*".

The bottom line is that privacy is not preserved by simply removing PIIs from the data. It still contains sensitive information, such as the quasi-identifiers, that can be traced back to real individuals if corroborated with auxiliary data.

k-anonymization

K-anonymization is a method developed by Dr. Sweeney and Dr. Samarati to counteract the re-identification described above. K-anonymization generalizes and/or removes attributes in datasets that would uniquely identify an individual. It is currently one of the most widely adopted privacy methods in the world.

This method prevents linking to external data records from indirect identifiers. *k*-anonymization maintains privacy by ensuring that in every record there is a number "k" of indistinguishable copies. This ensures that no row in the table is unique because of at least "k" others.

phone	sex	zip code	income
202-555-0104	male	75080	$60,000
202-555-0172	female	75085	$80,000
202-555-0140	female	75090	$120,000
202-555-0182	male	30040	$0
202-555-0174	male	30035	$70,000
202-555-0173	female	30080	$50,000
202-555-0183	male	30023	$55,000

Original dataset

phone	sex	zip code	income
xxx	male	750*	$60,000
xxx	female	750*	$80,000
xxx	female	750*	$120,000
xxx	male	300*	$0
xxx	male	300*	$70,000
xxx	female	300*	$50,000
xxx	male	300*	$55,000

K-anonymized dataset

Figure 3: The two datasets portray one original dataset, which has been k-anonymized by suppression and generalization in the second dataset.

K-anonymization typically uses two methods of de-identification: suppression and generalization. The first method replaces critical identifying information in the original dataset with a placeholder, such as “xxx” in the figure above. Generalization adds uncertainty by aggregating pieces of data in a row, shown in Figure 3 under zip code.

K-anonymization offers a greater privacy guarantee than pseudonymization. But from the usability point of view, it compromises the usefulness of the data. Because the values are highly aggregated or wholly removed, the resulting data loses granularity.

Besides, there is a considerable risk of violating data privacy if the sensitive attributes are homogeneous or have skewed distributions. An attacker may discover sensitive information from a k-anonymized dataset with external knowledge.

Researchers demonstrated an inevitable trade-off in k-anonymization, with a loss in utility to preserve privacy and vice versa.

l-diversity and t-closeness

l-diversity or t-closeness are other methodologies in this area with similar shortcomings. l-diversity is an extension of k-anonymity and is effective in protecting categorical attributes (data points that can be classified as countable numbers of distinct groups or categories). Yet, it is still vulnerable to linkage attacks.

The t-closeness method extends on l-diversity by treating the values of an attribute distinctly. While t-closeness protects against attribute disclosure, it does not protect against identity disclosure.

A common characteristic of these de-identification methods is that they aggregate and distort the original data to such an extent that the data can no longer be used in complex analysis.

Researchers have also continuously demonstrated that methods "*that treat privacy as a property of the output, such as k-anonymization and other traditional statistical disclosure limitation techniques, will fail to protect privacy*".

Overview and limitations of de-identification methods

	Pseudonymization	k-anonymization	l-diversity and t-closeness
Personal information is irreversibly deleted.	❌	✔	✔
No individual can be identified by combining several data points.	❌	❌	❌
The statistical significance of the data remains largely unchanged.	✔	❌	❌

The data produced falls outside the GDPR protection regulations related to “personal data”	❌	❌	❌

For applications such as business intelligence, data-driven product development, internal or external collaboration, and AI/ML development, **these methods do not offer enough privacy and utility guarantees.**

In most cases, there is a one-to-one correspondence between the original dataset and the de-identified one, which is the leading cause of re-identification attacks. The processing of de-identified data is also strongly limited by modern data privacy legislation, as **the risk of re-identification requires companies to process it as personal data.**

The GDPR definition of anonymous data demands more from businesses than removing unique identifiers. It increases the standards for modern data privacy solutions and services.

Customers are demanding more personalized services while, at the same time, valuing their data privacy rights. To meet these demands, **balancing the privacy/utility trade-off in favor of both parties is crucial**. This has pushed the community to develop improved methods to generate anonymous yet statistically representative data.

Overcoming the privacy and utility limitations of de-identification methods

For decades, data scientists have tried to overcome the trade-off between data privacy and data utility presented by traditional de-identification methods. This led to new research on the anonymization and synthetic generation methods that we present in this section.

Differential Privacy: the path to anonymization

What constitutes data anonymization methods is often debated within the scientific community. Regulators also are still to provide further technical definitions of anonymized data within legal frameworks. Before the GDPR definition, Article 29 from the European Working Party defined three criteria to assert the validity of anonymization methods.

They should produce data from which:

- It is no longer possible to single out an individual.
- It is no longer possible to link records relating to an individual.
- Information can not be inferred concerning an individual.

In recent years, one method that has gained popularity within the legal and scientific communities is Differential Privacy. Differential Privacy is the current “gold standard” of privacy-preserving techniques. Its mathematical framework guarantees that data participants will not be affected by allowing their data to be used.

At the core of Differential Privacy is the idea of adding “just enough” noise to mask the presence of any individual in the analysis (see the figure below).

Differential Privacy introduces a privacy parameter that measures the strength of the privacy protection. Offering a quantitative measure of privacy, it makes it possible to strive for the optimal privacy/utility trade-off: maximizing the accuracy of the analysis while limiting and minimizing the potential of exposing sensitive information about individuals.

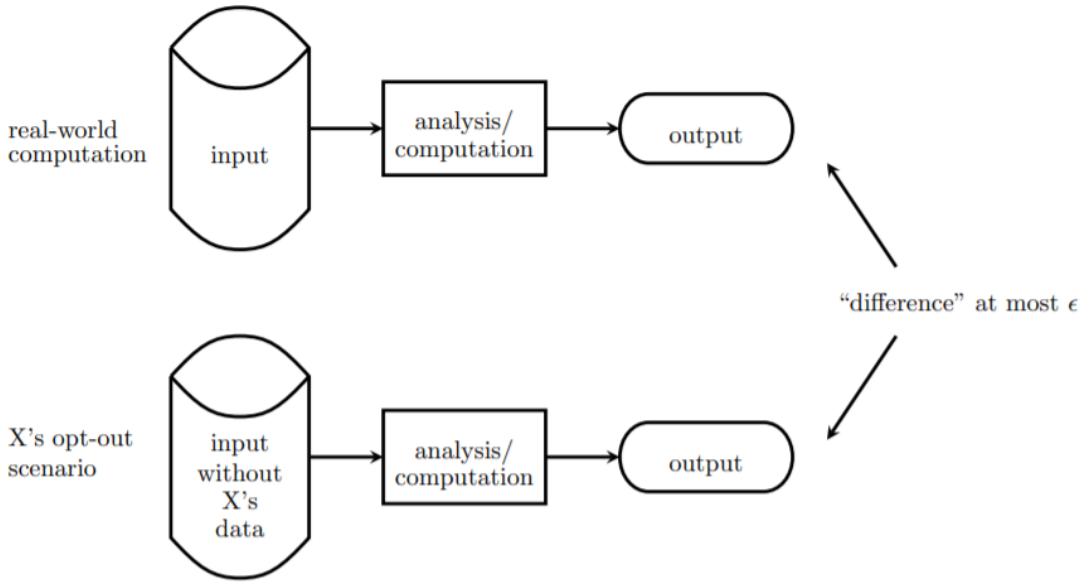


Figure 4: "Differential Privacy makes it arbitrarily hard to tell whether one's microdata has even been included in the analysis. The maximum deviation between the opt-out scenario (participant X negates their data) and real-world computation (participant X consent usage of their data) should hold simultaneously for each individual X whose information is included in the input." as explained in [Differential Privacy: A Primer for a Non-technical Audience]

For 15 years, the continuous work of researchers has advanced the efficiency and scalability of Differential Privacy. In a paper published in 2006, Cynthia Dwork provided methods to calculate the amount of noise required to protect the privacy of every individual in a database.

Although the technique of applying Differential Privacy is relatively new, it still provides a better way of assuring privacy than any previous method. According to

privacy researchers and academics, it is one of the most robust and secure privacy guarantees.

The concept has been adapted in various real-world applications thanks to research developments. Today, Differential Privacy is used in production environments by companies likeApple and Uber. It has been included in popular analytics and machine learning libraries, such as PyTorch and TensorFlow.

Maintaining dataset granularity with synthetic data

Synthetic data is artificially generated instead of collected from the real world. Synthetic data is often used when real-world data is costly to collect or hard to come by. It is also used in the context of privacy preservation to make information available for processing when regulations or other privacy concerns restrict access to the original data.

In this context, companies generate synthetic data by sampling from machine learning models trained to learn a sensitive target dataset's structural and statistical characteristics. Such data preserves the statistical properties of the original data to a very high degree. Machine learning algorithms learn the statistical characteristics of the original data and create new data points from them.

The synthetic data maintains the statistical properties of the original data, except where such statistical properties would compromise privacy. This property makes the data suitable for almost any use-case initially intended for the original data. For a large class of analyses, results are similar whether operations are performed on the synthetic or original data.

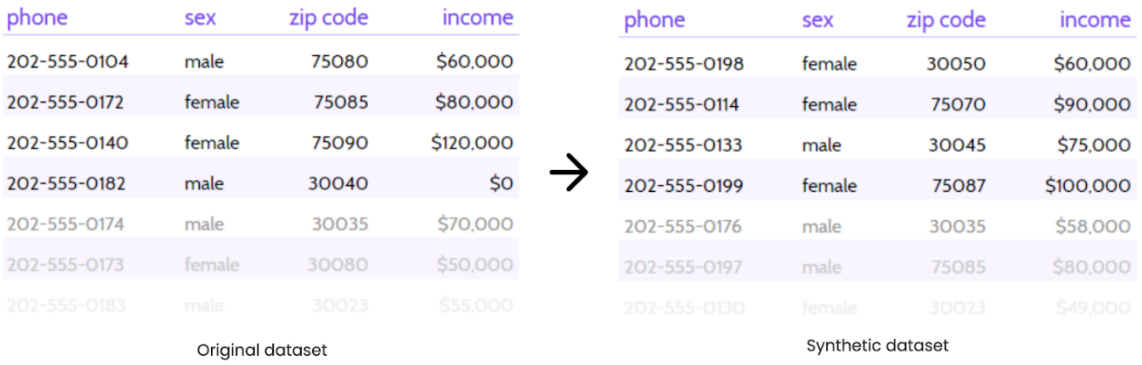


Figure 5: The figure shows how the Statice software generates privacy-preserving synthetic data consisting of new data points while reflecting the original data in its structure and statistical properties.

Figure 5 shows a synthetic dataset and compares it to the original data. Notice that all fields contain data similar to the original data and that no data is redacted.

Implementing privacy protection guarantees is also possible, so the algorithms generate "privacy-preserving" synthetic data. Various mechanisms protect the sensitive information of an individual present in the original dataset event after the release of the synthetic dataset. One way to protect privacy in particular is to generate differentially private synthetic data.

Differentially private synthetic data: maximizing utility and privacy

In legal terms, the Differential Privacy guarantee would ensure compliance with even the most inflexible privacy regulations and policies and thus allow for data sharing and use without any legal obstacles.

As pointed out by researchers: *"Differential Privacy may be seen as a technical solution for analyzing and sharing data while protecting the privacy of individuals in accordance with existing legal or policy requirements for de-identification or disclosure limitation."*

Differentially private synthetic data builds upon state-of-the-art machine learning algorithms and the latest privacy research. It allows organizations to maximize data utility and privacy and offer many benefits for customers-centric industries.

The robust mathematical guarantees of Differential Privacy "*calm not only the fear of data leakage but also the risks of adversarial machine learning*", according to researchers.

The Statice data anonymization engine builds on this. The core technology relies on deep learning algorithms to generate privacy-preserving synthetic data. The model is trained using algorithms that satisfy the definition of Differential Privacy, which guarantees that the synthetic data is robust against all sorts of privacy attacks. This approach preserves the statistical properties of the original data to a very high degree while offering strong privacy guarantees. Statice clients use this technology to generate privacy-preserving synthetic data sets for business intelligence, product development, and compliance use cases.

De-identification vs. privacy-preserving synthetic data: Comparison overview

	Pseudonymization	k-anonymization	l-diversity and t-closeness	Privacy-preserving synthetic data
Personal information is irreversibly deleted.	❌	✔	✔	✔
No individual can be identified by combining several data points.	❌	❌	❌	✔
The statistical significance of the data remains largely unchanged.	✔	❌	❌	✔
The data produced falls the GDPR protection regulations related to “personal data”	❌	❌	❌	✔

Benefits of privacy-preserving synthetic data for businesses

The capability to generate privacy-preserving synthetic data is an asset for any organization that collects or produces sensitive data. Because of its nature, it offers several advantages.

Future-proofing the compliance of data operations

The most obvious benefit is that differentially private synthetic data is privacy-preserving and compliant with the requirements of data anonymization of data protection laws.

Anonymous data doesn't fall under the scope of personal data processing regulations. Thus, contrary to personal data, the processing of anonymous data is not subject to regulation. This means that data can be used for secondary purposes, for example, product development, business intelligence. It can also be retained over time without infringing on individual privacy or clashing with regulations.

Benefiting from greater agility

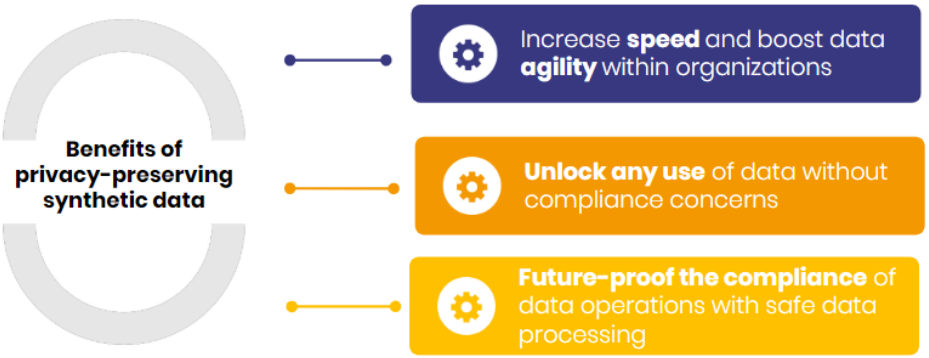
Privacy-preserving synthetic data no longer contains PII, which means that it generally has a much lower overhead for use in terms of compliance and security. This, in turn, allows for quicker and easier internal or external data sharing, making data teams more efficient and allowing better resource allocation. While sensitive data sharing agreements and compliance processes can take weeks to months, privacy-preserving synthetic data can be shared in a fraction of the time.

Additionally, privacy-preserving synthetic data is safe to use with public cloud providers, which means that data teams can leverage the latest tools from Amazon, Google, and co, without the long, tedious process normally associated

with getting sensitive data onto outside systems or forcing the provisioning of expensive computation resources on-premise.

Gaining the ability to monetize data

Data is a key resource for product development, whether leveraged to create new products or packaged and sold to third parties in direct data deals. Privacy-preserving synthetic data offers an opportunity to build on data streams that are otherwise too sensitive to use for such purposes under normal circumstances, which means that organizations can build new data-derived revenue streams at will, without risking individual privacy.



Statice

Conclusion

Data authorities and citizens worldwide require more privacy for sensitive customer data. Stringent regulations and risks associated with data breaches create challenges for businesses. To avoid missing out on valuable assets, data-driven organizations must find ways to innovate while safeguarding data privacy.

Privacy-preserving methods and tools are available to help companies protect their sensitive data. De-identification techniques such as pseudonymization offer the first layer for risk mitigation. But to comply with the latest legal framework and freely leverage sensitive data, organizations must go a step further.

Recent developments have shown the limitation of traditional de-identification methods. Methods such as pseudonymization display clear privacy risks, while others such as k-anonymization raise concerns regarding utility. Today, however, companies no longer have to compromise privacy in the name of utility or vice versa. The privacy research community has developed state-of-the-art techniques that protect data privacy while maintaining the quality and utility of data.

Differentially private synthetic data is the latest and safest method that enables the best of both worlds: leveraging data and safeguarding privacy. Organizations can use privacy-preserving synthetic data for almost any use-case possible on the original data, while remaining compliant and without the risk of exposing sensitive information.

We hope this presentation sheds light on protection methods for working with sensitive customer data. We hope it also provided insights into the benefits of privacy-preserving synthetic data as a solution for data-driven innovation.

References

- Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D.R., Steike, T., Vadhan, S., Wood, A. (2018). "Differential privacy: A non-technical primer"
- Bellovin. S.M., Preetam, K.M., Reitinger, N. (2019). "Privacy and Synthetic datasets". Stanford Technology Law Review. Vol. 22:1.
- California Consumer Privacy Act of 2018 [1798.100 - 1798.199]. (2019). California Legislative Information.
- Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006). "Calibrating Noise to Sensitivity in Private Data Analysis".
- GDPR. (2018). Recital 26 Not applicable to anonymous data.
- GDPR. (2018). Recital 28 Introduction of pseudonymisation.
- Gorde, K., Jajodia, S., Kim, Y., Mohammadi, M., Park, H., Park, N. (2018). "Data Synthesis based on Generative Adversarial Networks". Cornell University.
- Howe, B. & Rodriguez, L. (2019) "In Defense of Synthetic Data". Cornell University.
- Meehan, M. (2019). "Data Privacy Will Be The Most Important Issue In The Next Decade". Forbes.
- Narananyan, A., Shmatikov, V. (2008). "Robust De-anonymization of Large Sparse Datasets". The University of Texas, Austin.

About Statice

Statice develops state-of-the-art data privacy technology that helps companies double-down on data-driven innovation while safeguarding the privacy of individuals. Thanks to the privacy guarantees of the Statice data anonymization software, companies generate privacy-preserving synthetic data compliant for

any type of data integration, processing and dissemination. With Statice, enterprises from the financial, insurance, and healthcare industries can drive data agility and unlock the creation of value along their data lifecycle. Safely train machine learning models, finally process your data in the cloud or easily share it with partners with Statice.

Learn more about Statice on www.statice.ai or get in touch with us at info@statice.ai


www.statice.ai