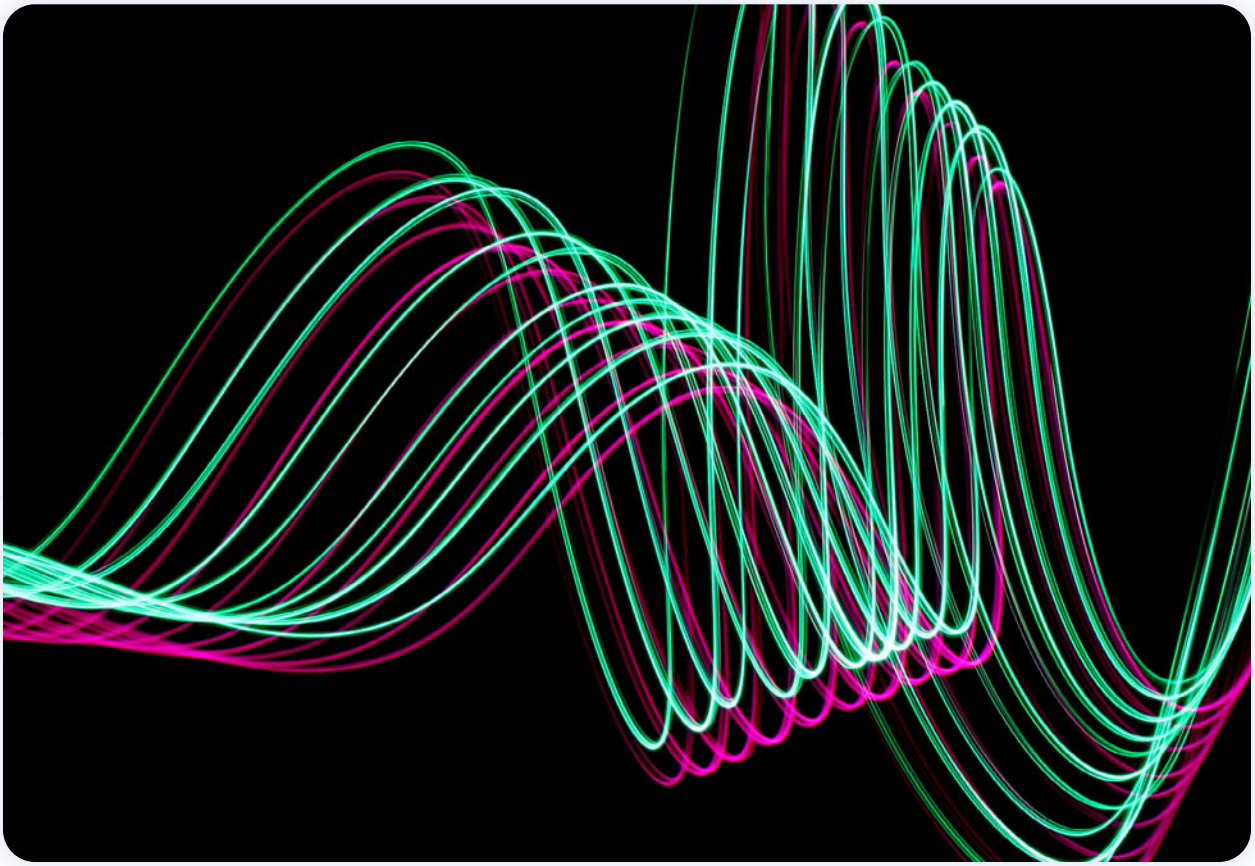


tokenex

De-identification Guide

How to protect personal data while preserving its business value

In today's IT environments, big data is big business. Modern organizations benefit greatly from the insights that can be derived through collecting and processing consumer data, and plenty of analytics-rich information has become available as the digital world continues to advance further into our everyday lives.

To gain access to such valuable business intelligence about consumer behavior, however, organizations first must gather and process customer data consensually and compliantly. Depending on how and where an organization operates, this can be a complex process that entails the understanding of and adherence to detailed requirements for collecting, storing, sharing, and other operations involving sensitive consumer data.

With these concerns in mind, many organizations find themselves in need of a solution that can protect the privacy of their customers as well as the utility of their data as they attempt to safely navigate the developing regulatory landscape. To accomplish this, we suggest several industry best practices as strategies to fulfill the litany of security and compliance obligations related to the management of sensitive personal data.

Privacy strategies

Strategies for meeting the obligations of privacy regulations and effectively protecting personal data can be divided into two areas: data-oriented methods and process-oriented methods. Data-oriented methods involve protecting the data itself via techniques such as minimization, separation, abstraction, and concealment.

Two basic strategies

Data-oriented

Process-oriented

Other strategies, known as process-oriented methods, deal with the management and handling of data to try to prevent its misuse. This is done by keeping consumers informed about how their data is being used, giving them control over that use, demonstrating a commitment to privacy principles, and enforcing the data's proper use in accordance with those principles.

Data-oriented

Minimize	Separate	Abstract	Hide
Limit the amount of data in your systems.	Distribute or isolate the personal data to prevent correlation.	Limit the detail in which personal data is processed.	Prevent the exposure of personal data.

Process-oriented

Inform	Control	Demonstrate	Enforce
Inform data subjects about the processing of their personal data.	Provide data subjects control over the processing of their personal data.	Demonstrate your prioritization of protecting privacy when processing personal data.	Commit and enforce processing personal data in a privacy-friendly way.

In this resource, we'll be discussing in detail the data-oriented privacy strategy of de-identification, specifically de-identification via tokenization. Tokenization combines principles of separation and abstraction to desensitize data in an organization's environment, creating instead a database consisting of pseudonymized placeholder data.

Personally identifiable information

Per the Privacy Act of 1974, PII is defined as "information that can be used to distinguish or trace an individual's identity, either alone or when combined with other personal or identifying information that is linked or linkable to a specific individual." For the purposes of this guide, we'll be using PII broadly to refer to all types of sensitive data that are associated with individuals protected by privacy regulations.

** Source for Graphic: The Hitchhiker's Guide to Privacy by Design*

Depending on which regulatory compliance obligations your organization is required to follow, there can be differences in terms of what types of information are protected. Here are definitions for "PII" according to common privacy regulations.

Individually identifiable health information

Information that is a subset of health information, including demographic information collected from an individual.

Personal data (GDPR)

Any information relating to an identified or identifiable natural person.

Personal information (CCPA)

Information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household.

Despite these differences, the term "personal information" is commonly confused for and used interchangeably with "personal data." Distinguishing between the two might seem pedantic, but the distinction is crucial for fully understanding the compliance requirements for their respective regulations. Not all data related to a person has the capacity to identify an individual, so only data from which a person's identity can be derived falls under the umbrella of personal information.

De-identification

De-identification is the process of removing certain identifying elements from sets of sensitive data so that they no longer identify the individual from whom it was collected. According to HITRUST's "De-identification Framework," which is based on the HIPAA Privacy Rule's De-identification Standard, sensitive data can be grouped into three states of identification.

Identifiable

Information that can be used to identify an individual directly or indirectly.

De-identified (pseudonymized)

Information with sufficient data elements removed so that the identification of the original individual is very unlikely.

Non-identifiable (anonymized)

Any information relating to an identified or identifiable natural person.

However, de-identification can be complicated by the conditional nature of what can be considered identifying data. In other words, what might identify an individual in one context might not do so in another or in the absence of certain additional information.

Identifiers

This brings us to the concept of identifiers: data sets that contain identifying elements. As mentioned above, in order for data to be de-identified, it must be altered in a way that prohibits the data from being used—independently or in combination with additional information—to identify the individual it represents.

Identifiers can be categorized as either direct or indirect ("quasi-identifiers"). Direct identifiers are pieces of data that can directly identify an individual without the use of additional information. Typical examples of direct identifiers include name, address, Social Security number, driver's license number, phone numbers, passport numbers, IP addresses, and other data points that are commonly used to directly identify an individual.

Indirect identifiers are pieces of data that can be used to identify an individual only when combined with additional information or appearing within a certain context. Examples of indirect identifiers include information such as a state of residence, descriptions of physical features/characteristics, demographic data, and even consumer behavior.

Individual identifiers specifically referenced in HIPAA

- **Geographic locations**
- **Certificate or license numbers**
- **Birth dates, admission dates, discharge dates, dates of birth**
- **Vehicle identifiers, serial numbers & license plates**
- **Telephone & fax numbers**
- **Device identifiers**
- **Email addresses**
- **URLs**
- **Social Security numbers**
- **IP addresses**
- **Medical record numbers**
- **Biometric identifiers**
- **Health plan beneficiary numbers**
- **Photographic images**
- **Account numbers**
- **Any other unique identifying number, characteristic or code**

The above collection of identifying information is especially important in regard to HIPAA's "Safe Harbor" rule for de-identifying data. The "Safe Harbor" method states that if all of the information above is removed from a data set, then that data set is considered de-identified and therefore is compliant with HIPAA's privacy rules.

However, not all of this information is capable of independently identifying an individual. As we established earlier, the identifiability of a given data set varies depending on the context in which the data is presented—in other words, what other data elements are available and whether those additional elements can be combined to create a data set capable of reliably identifying an individual.

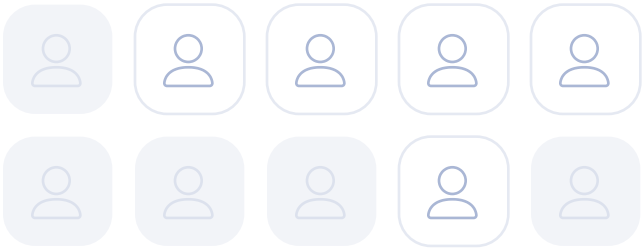
** Source for Graphic: Guidance on De-identification of Protected Health Information*

For this example, let's say we've collected personal data from 10 people. By looking at individual elements of that data, we'll show how the identifiability of indirect identifiers is dependent on the other data included in the data set.

Indirect identifiers scale

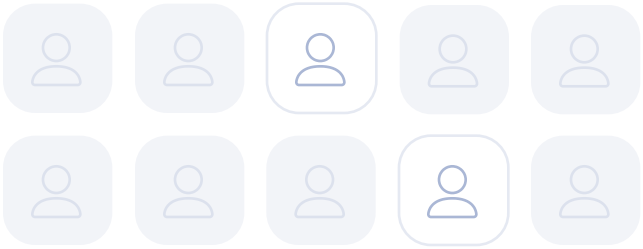
1. Name: John

Here, we have a first name. Five of the 10 people included in this data set are named John, so in this context, a name alone would not be considered identifying.



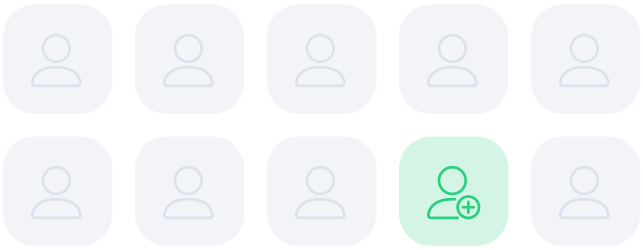
2. Name: John State: California

Next, we have a state of residence along with a name. Only two of the 10 people included here share both, but that's still not enough information to identify one of them. Neither would be considered identifying.



3. Name: John State: California DOB: 11/23/88

Finally, we add a date of birth, which narrows it down to one person. Although none of those elements would be considered individually (directly) identifying, when grouped together, they make it possible to identify someone. So in this context, they are identifying.



Methods for de-identifying PII

So, now that we've established what de-identification is and covered the differences between direct and indirect identifiers, let's look at various methods for de-identification. The visuals and descriptions below are designed to help explain the degrees to which a set of data can be de-identified and how that level of obfuscation impacts its utility and identifiability. In general, the greater a data's utility, the greater its risk of re-identification, so it's critical to find the appropriate balance for your organization's needs surrounding data protection and business operations.

De-identification scale

Name
Jim Smith

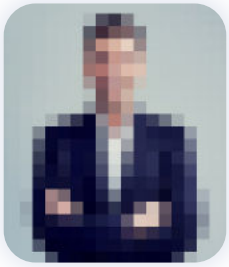
Email
jsmith@example.com

State
California

Zip
94121

SSN
123456789

DOB
11/23/88



Name
8dj3hbs9r

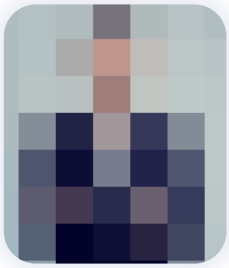
Email
h4k9fekema0l37y3h

State
California

Zip
94121

SSN
837495728

DOB
11/23/88



Name
xxxxxxxxx

Email
xxxxxxxxxxxxxxxxxx

State
xxxxxxxxxx

Zip
xxxxx

SSN
xxxxxxxxx

DOB
xxxxxxxx

Pseudonymization

Pseudonymization is the process of replacing identifying or sensitive data with a pseudonym and implementing safeguards to prevent the reversal or re-identification of that data. It is defined in Article 4(5) of the GDPR as:

> **The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.**

As illustrated in the graphic on the previous page, pseudonymized data lies in the middle of the de-identification spectrum—with fully identifiable, explicit personal data such as full names and Social Security numbers at one end and anonymized data with no identifiable personal information at the other. In the interest of retaining the utility of a data subject's sensitive information, pseudonymized data can be reidentified, or associated, with that individual by replacing the pseudonyms with the actual data. But by observing appropriate security measures, organizations can minimize the likelihood of an unauthorized party reversing the de-identification process and gaining access to the original data.

Anonymization

Anonymized data is data that's altered in such a way that it cannot be linked back to the individual(s) with which it corresponds. Anonymized data is not subject to the data protection obligations instituted by the GDPR, HIPAA, or CCPA, so it's value for compliance is clear.

However, anonymizing all of the sensitive data within an organization's systems often isn't a realistic solution. First, fully anonymizing a data set is a difficult task that requires significant

obfuscation of the original data and the development of stringent security controls to prevent the reversal of that obfuscation. Second, by definition, anonymized data can't be linked back to identifiable individuals, which renders it useless for almost anything but very high-level data aggregation and analysis. So although anonymized data can be considered "safer," it loses its value for business intelligence.

Pseudonymization represents the ideal de-identification method for balancing your organization's needs for data utility with its concerns regarding risk and compliance.

Focusing on direct identifiers

Of the two processes explained above, we recommend pseudonymization. When executed correctly, pseudonymization represents the ideal de-identification method for balancing your organization's needs for data utility with its concerns regarding risk and compliance. What little risk it introduces for possible re-identification can be minimized via sufficient security practices and is far outweighed by the value of the data it preserves.

The key to retaining that value is to know which elements of the data to de-identify and which to keep intact—which brings us back to direct and indirect identifiers. Let's say, for example, your organization collects demographic information about an individual's appearance: height, weight, eye color, etc. Pieces of data such as height or hair color probably wouldn't be enough to independently identify someone if you're looking at a data set for an entire state. However, if your data set was only for a single department within an organization or if an individual's hair color was included in a data set that also contained the person's name or address, the likelihood that it could be used to identify that individual would be much greater.

Because of this, the process of de-identification can be less intensive than some might think. Rather than needing to completely remove all identifying elements of a data set—and likely rendering it unusable for analytics and other business intelligence purposes—organizations need only to remove enough direct identifiers to reduce the possibility of the data being identified.

Preserving analytics

Further, for analytics purposes, indirect identifiers are frequently collected to create consumer profiles and other aggregations that can provide valuable statistical insights. However, due to the requirements of certain privacy regulations, the types of indirect identifiers an organization can compliantly gather and the sets of data they can then compile can be limited inasmuch as those groups of data could potentially be cross referenced to identify the individuals from whom they were collected in the first place.

Often, though, indirect identifiers can be segmented in a way that significantly reduces the likelihood of re-identification. Adhering to industry best practices for data compilation and aggregation and focusing on removing direct identifiers from data sets can usually enable organizations to satisfy their regulatory compliance obligations while still preserving much of the business utility of the personal data in their possession.

Because direct identifiers are closely linked with individuals and possess limited data utility, de-identifying them can provide greater protection from re-identification without adversely affecting the value of sensitive data sets. This provides obvious benefits for both data protection and preservation.

Because direct identifiers are closely linked with individuals and possess limited data utility, de-identifying them can provide greater protection from re-identification without adversely affecting the value of sensitive data sets.

Pseudonymization via tokenization

As we mentioned earlier, our recommended method for pseudonymizing data is tokenization. Tokenization is the process of securing and desensitizing data by removing it from an organization's internal systems and exchanging it with nonsensitive placeholders called tokens. These tokens remain in an organization's network for business use while the original, sensitive data is safely stored outside of that environment.

This pseudonymization technique combines principles of separation and abstraction to minimize the likelihood of exposure, theft, re-identification, and other threats to privacy. This form of pseudonymization is a particularly effective form of data protection because it can prevent the compromise of tokenized data in the event of a breach. Even if a tokenized environment is infiltrated, the only data accessible as a result would be tokens—the sensitive data they represent remains securely stored offsite by the tokenization provider.

Additionally, tokenization can retain elements of the original data via length- and format-preserving token schemes. These customizable schemes can obfuscate certain portions of the data while others remain intact, such as the last four digits of a credit card number. This further allows for the aforementioned preservation of business analytics for nearly any structured data set. See a list of common token schemes below.

Common token schemes

Token scheme	Example data	Example token
sixTOKENfour	4242424242424242	424242925164242
fourTOKENfour	4242424242424242	424276225864242
TOKENfour	4242424242424242	635276725864242
GUID	ThisIsATest	25892e17-80f6-415f-9c65-7395632f0223
SSN	123456789	958475126
nGUID	25947582	25892e17-80f6-415f-9c65-7395632f0223
sixANTOKENfour	4242424242424242	424242YAV5164242
fourATOKENfour	4242424242424242	4242ZYAV516342 42
ANTOKENfour	4242424242424242	9TY2ZYAV516342 42
ANTOKEN	4242424242424242	5FR962FGT0
TOKEN	ThisIsATest	DUH3JSLDTAYHUCO51MXY7IINZ8HLNDU90FMTTM
PCI	4242424242424242	424242APT0Oa4242
Ascii	ZvT4'{vdLd	KOC8ExkNin7J6q91JwDT

Choosing a provider that promotes positive business outcomes

Ultimately, the goal of a tokenization platform is to protect sensitive data by removing it from an organization's internal systems. However, security and compliance technologies that inhibit business operations aren't particularly useful. Rather, these platforms derive their true value from the business initiatives and positive outcomes they enable—whether that's reducing friction, minimizing cost, simplifying systems, or facilitating business processes.

A truly exceptional provider can offer additional benefits, such as the consolidation of multiple data types from diverse technologies, the unification of compliance within a single platform, a modern security architecture that can serve as the central data hub of business operations, and other capabilities that transcend basic security to facilitate organization-wide success.

With its flexible integrations and data-centric principles, TokenEx is a provider that delivers unmatched data protection and business enablement. Our nimble cloud platform functions as the fabric of a unified digital ecosystem for security, privacy, and compliance—aligning security with operations and empowering our clients to build for tomorrow as well as today.

Growing cost of compliance

2022

$8 Billion+

Total worldwide compliance spending.

2023

65%

Percentage of the population's personal information that will be protected by privacy regulations.

30%

Companies who prioritize privacy will generate 30 percent more ecommerce profits than their competitors.

2024

80%

Percentage of the world's organizations that will be regulated by privacy and data protection obligations.

** Source for Graphic: Gartner's State of Privacy and Personal Data Protection*

The projections on the previous page show the increasing cost and prevalence of compliance obligations, providing a forecast for a regulatory climate that is rapidly approaching. For some, this calm before the storm can introduce uncertainty. Attempting to pre-empt upcoming regulations by implementing costly or disruptive processes and controls might seem unnecessarily burdensome—and even then, compliance needs can change as regulations are revised and amended. Conversely, waiting too long to act increases the risk of facing penalties, fines, and other adverse economic effects due to noncompliance—or worse, breaches.

The bottom line is that global privacy regulations are already on their way and the frequency of data breaches continues to increase. Delaying the inevitable will only make an organization slow to react and compound the difficulty of altering existing operations to achieve compliance. By anticipating upcoming regulations and preparing for them now, organizations can better position themselves to adapt quickly and prevent the financial and reputational damage of not prioritizing consumer privacy. }

Bibliography

De-Identification Framework: A Consistent, Managed Methodology for the De-Identification of Personal Data and the Sharing of Compliance and Risk Information. E-book, HITRUST, 2015.

"Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." **HHS**, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#:~:text=(a)%20Standard%3A%20de%2D,not%20individually%20identifiable%20health%20information. Accessed 17 September 2020.

Henein, Nader, et al. The State of Privacy and Personal Data Protection, 2020-2022. E-book, Gartner, 2020.

Maldoff, Gabe. "Top 10 operational impacts of the GDPR: Part 8 – Pseudonymization." IAPP, https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-8-pseudonymization/. Accessed 17 September 2020.

Near, Joseph, et al. "Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series." NIST, https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-da ta-analysis-introduction-our. Accessed 17 September 2020.

Peruskovic, Barbara. *The Hitchhiker's Guide to Privacy by Design.* E-book, Protegrity, 2018.