aircloak

Data Anonymization in Digital Business Models

How to leverage sensitive data without putting privacy at risk

Introduction

A large variety of data-driven use cases rely on sensitive data. In order to use this data in a responsible way, a particularly high level of data protection and security measures is required.

Data anonymization can contribute a lot of value to a robust data protection concept. However, it is often not clear what kind of anonymization should be used or what opportunities and challenges are associated with it.

In this whitepaper, we want to address various questions that will help you evaluate data anonymization for your project. We will answer the following questions on the next pages:

- ➔ Why you should anonymize data
- ➔ What is the difference between pseudonymization and anonymization
- ➔ Which use cases are enabled by anonymization
- ➔ What kinds of anonymization approaches do currently exist
- ➔ Benefits of using Aircloak for privacy-preserving analytics

Why should I anonymize data?

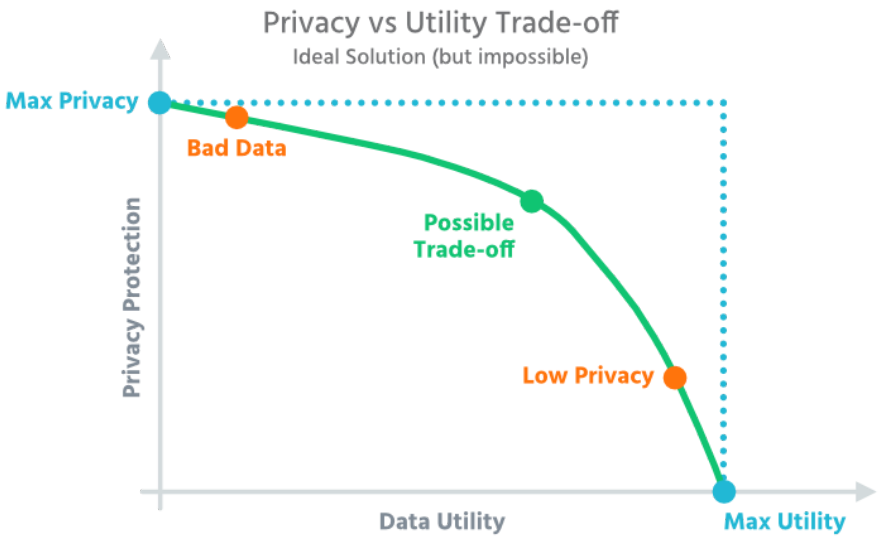
Since GDPR only relates to personal data, any data that is not personal does not underlie the regulation. This means that if you are able to modify the data in such a way that you can not reidentify an individual in the data set, the data is no longer subject of the regulation. This is where anonymization comes in. Anonymized data is free to use and enables a broad variety of use cases like **data sharing, data monetization or advanced analytics.**

Getting anonymization right is difficult

Data anonymization is difficult to achieve and it is characterized by a permanent conflict:

The tension between data protection and data quality.

A common problem is that analysts tend to opt for high data usability instead of high privacy in manual anonymization. As a result, data is merely pseudonymized.



Privacy vs. Utility Trade-off: The better the data utility has to be, the worse gets the privacy and vice versa.

Pseudonymization vs. Anonymization

Pseudonymization is the process of directly replacing personal identifiers with some form of other identifier. For instance, you might replace every occurrence of a name with fictional one or a random sequences of letters and replace a phone number with a random sequence of digits. The trouble is that although this may give the illusion of protecting personal data, it does not give strong protection against inference attacks.

Inference attacks work by combining data points in order to re-identify an individual with high probability and accuracy.

As an example take a dataset containing health records. You might pseudonymize the data by replacing all the names and healthcare IDs with unique random numbers.

An attacker may be able to use a combination of other data such as location, medical history, gender and age to make an accurate guess as to the real identity of the person.

Why Pseudonymization Isn't Enough

GDPR Recital 26 makes it clear that pseudonymization is not strong enough to be considered as non-personal data. not strong enough to protect personal data. Pseudonymization permits data controllers to handle their data more liberally, but it does not abolish all risks due to the possibility of re-identification.

Recital 26 states that

"The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymization, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person."

Crucially, this means pseudonymous data is still subject to privacy regulations under GDPR.

Recital 26 then explicitly distinguishes between pseudonymized and anonymized data. It explains that pseudonymized data

"…should be considered to be information on an identifiable natural person."

It then states:

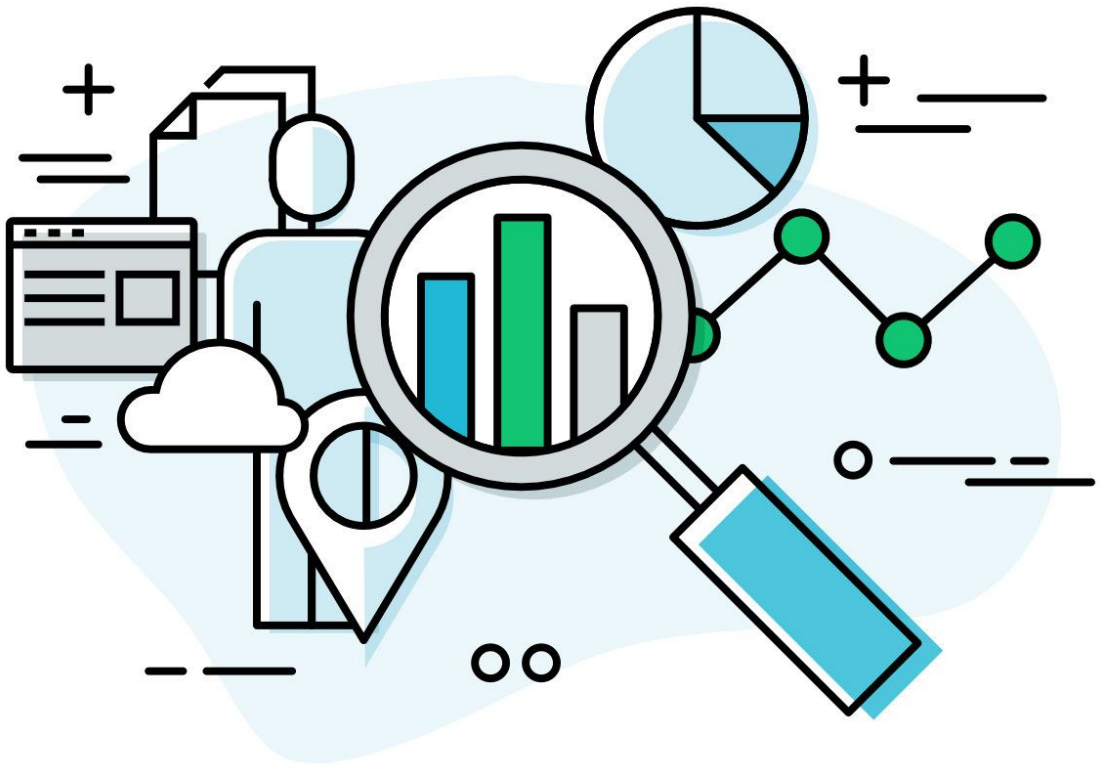
> "*The principles of data protection should […] not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."*

And it makes it clear:

> *"This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes."*

In short, if your data is anonymized it is no longer subject to the GDPR, while if it is simply pseudonymized GDPR still applies.

Even if it now sounds tempting and one prefers to simply anonymize all data, it is important to note that not every data record can be reliably anonymized. In some use cases, data must remain traceable to an individual person. The data should then be regarded as pseudonymous and personal, and protected accordingly.



Tasks for Data Anonymization

Anonymization enables and simplifies a variety of use cases with sensitive data that would otherwise not be possible due to compliance and risk reasons.

Companies in highly regulated markets such as finance or telecommunications, or in industries with very sensitive data such as patient data in healthcare or insurance can now create and offer new services to customers while maintaining the highest standards in data privacy.

Partnering
Data Partnerships are an ideal way of breaking the current data monopolies of Google, Apple, Facebook & Co. Since data is streaming in from many different sources, it is almost impossible for one company to have access to the many contact points a customer goes through the course of his customer journey. Anonymized data can be shared across a number of business units or companies, without risking privacy violations.

Data Monetization
Once fully anonymized, a dataset can be analysed or sold to third parties without needing additional consent from the data subject. The anonymized data can be also used internally, for example to analyze detailed customer information that are then used by product development and marketing to improve products, services and campaigns.

Reporting to 3rd Parties
Reporting from anonymized sensitive datasets may be relevant for compliance reasons. The demands for such anonymization should be evaluated on a case-by-case basis. Relevant factors are, among others, what information the third party already holds, how quickly the dataset is changing and how interactive the report should be.

Privacy-preserving Machine Learning
Because machine learning is data-hungry, it can have a negative impact on data privacy. Companies avoid compliance risks when training machine learning algorithms on anonymized data instead of raw data. But since anonymization removes some of the informational value of the data, it can distort or completely destroy important correlations. However, introducing noise in the data can also be useful to avoid overfitting the models later on.

Open Data / Open Government Data
Companies are not alone in possessing valuable data. Universities, research institutes, and public authorities also have access to very interesting information that could be made available to the general public and be used for the greater good. These include for example mobility or transportation data, healthcare data or housing data.

Reporting Dashboards
Business intelligence and data analysis is extremely important at all levels in an organization. In the past, decisions used to be made intuitively or based on assumptions, today management has access to a vast amount of business and market data. With the help of anonymization, reporting dashboards can be improved by including anonymized sensitive data and can provide even more value.

Data Streaming
Anonymization of single events, often coming from a single client, is a much trickier problem to solve than anonymization of aggregate data. Yet exactly such approaches are vital for certain use cases such as in IoT projects or connected cars.

Data Retention
Data minimization is one of the core principles of the GDPR (§5) which includes strict rules for how long data can be kept. Once personal data has been anonymized, no restriction on retention periods apply anymore.

Where Anonymization Isn’t Possible

Choosing the correct anonymization solution for your project can be a daunting task. With the current lack of standardization or certificates, careful review of approaches before implementation is key. Sometimes, this also includes the insight that it might simply not be possible to anonymize data for your use case.

For example, to comply with the opinion on anonymization techniques that the European Data Protection Board (formerly known as Article 29 Working Party) laid out, one can argue that audiovisual files and free text can not truly be anonymized. Of course, that doesn’t mean that they can’t be adequately protected.

Anonymization Approaches

There are a number of technologies with which data can be made anonymous. What kind of technology you need depends on different factors:

- ➔ **Purpose**
  What are you planning to do with the data? What are your data quality requirements?
- ➔ **Type of data**
  What type of data do you want to anonymize? Is it structured, semi-structured, or unstructured data? Where is the data stored?
- ➔ **Processes / Occurrence**
  How often does the data need to be anonymized? Is it a one-time-project or do you need it on a regular basis?
- ➔ **Expertise**
  Are your employees skilled enough to achieve anonymization that is strong enough to free you from the GDPR while maintaining high data quality?

On the technological side, the currently most popular anonymization methods can be classified in three different categories:

Static anonymization, sometimes referred to as release-and-forget anonymization. With static anonymization the publisher anonymizes the database in its entirety. Third parties can then access the anonymized data. Static anonymization can be achieved manually or using one of a range of available tools.

With static anonymization you might achieve a good level of anonymization, but every time the data changes you will have to go through the process again, which, unless you are careful, increases the risk of reidentification. Even when using a tool, static anonymization requires a skilled analyst to be able to assess whether the resulting data is usable and anonymous.

Supported use cases: Data monetization, reporting to 3rd parties, open data / open government data, data retention

With **dynamic anonymization**, the anonymization is applied dynamically as you query the data. The analyst queries the data as usual, but instead of returning the raw results, the system will apply suitable noise and other masking techniques first. Probably the best-known interactive techniques are Differential Privacy and Aircloak Insights.

However, differential privacy suffers from the curse of the “privacy budget”. Basically, the more you learn about the data (for example by running more than one query) the easier it becomes to account for the noise that was added. Hence each information you learn also takes from your privacy budget. When the budget is empty the system can no longer guarantee that the results will be anonymous. Aircloak Insights has no privacy budget restrictions.

Supported use cases: Data monetization, partnering, open data / open government data, reporting dashboards

Synthetic Data is artificially generated data that has approximately the same properties (i.e. values) as the raw data, but that does not allow conclusions to be drawn about the individuals in the original dataset. Synthetic data can be created to look sensible to humans, to serve as test data, or for machine learning purposes. Creating data for testing purposes requires significant levels of domain knowledge and manual effort.

Supported use cases: Privacy-preserving machine learning, creation of test data sets

Privacy-Preserving Analytics with Dynamic Anonymization

Aircloak Insights uses a patented and proven dynamic anonymization approach that provides GDPR-compliant insights into sensitive data without risk. It offers real-time database anonymization that allows analysts to query anonymized data exactly as if it were the original raw data.

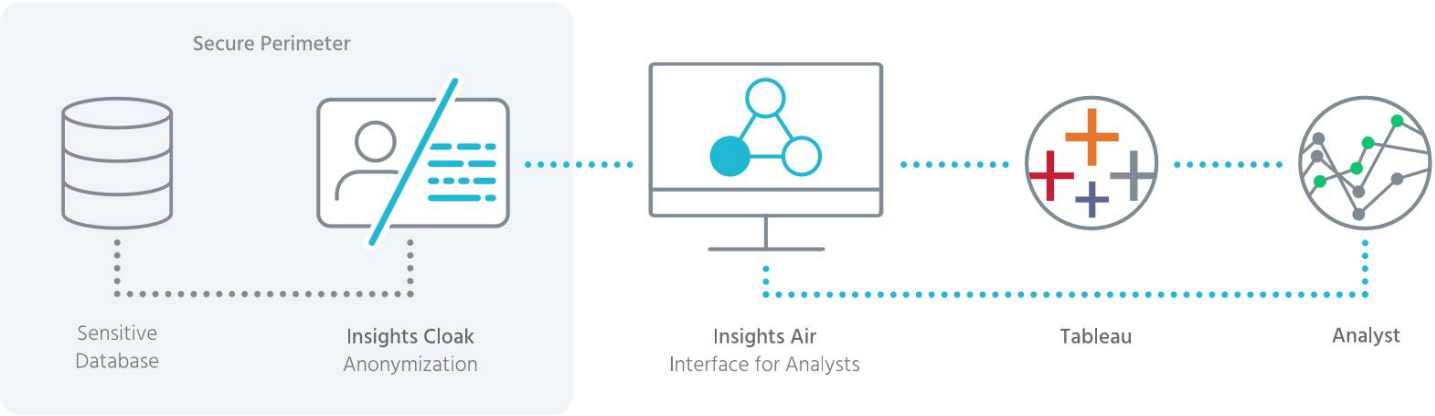
Of the earlier mentioned use cases, Aircloak's sweet spots are as follows:

- ➔ Data Monetization
- ➔ Data Sharing
- ➔ Partnering
- ➔ Customer Intelligence
- ➔ Open Data Initiatives and Platforms

In this section we will explain how Aircloak's technology works and show why it is the first GDPR-compliant tool for anonymized analytics.

Strong Anonymization for Data Analytics

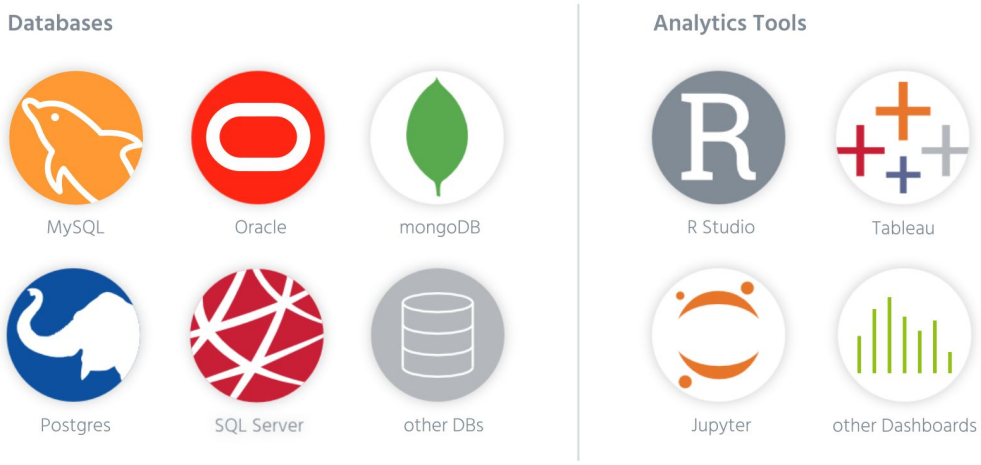
Aircloak Insights is a transparent proxy that sits between analysts and the data they need to work with. Because it is transparent, analysts are able to query the data as if they were doing it directly. They are able to construct queries in SQL or create dashboards using tools like Tableau. Aircloak Insights will intercept these queries and will convert them into a form suitable for the backend (which may be a structured SQL database, or a NoSQL data lake). Results are fully anonymized. Anonymization is achieved through outlier suppression, the addition of noise, and suppression of infrequent values.



Aircloak is installed in front of and connects to your existing database servers.It also connects to business intelligence tools like Tableau or other dashboards that know how to communicate using the Postgres Message Protocol

The benefit of Aircloak Insights is that it allows analysts to use the tools they are familiar with, without any need for expertise in anonymization.

It is also compatible with all the most common databases, both SQL and NoSQL. This means it can be directly integrated into your existing analytics stack without the need for any major modifications.



Aircloak supports different databases and analytics tools

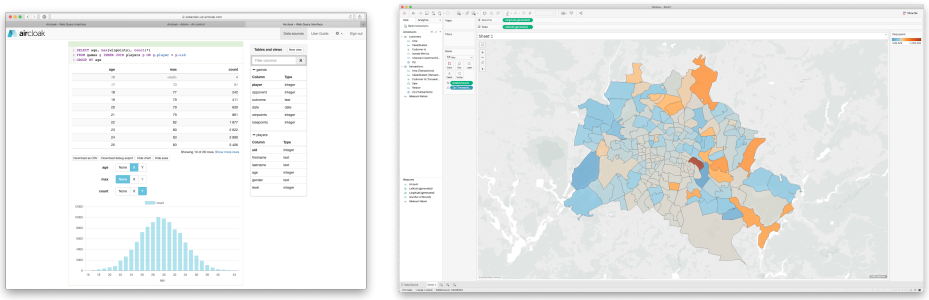
The Concept Behind Aircloak Insights

Instead of anonymizing or pseudonymizing the source data prior to analysis, Aircloak Insights still performs analytics on the unprocessed data. The analytics output is then instantly and automatically anonymized as queries are run.

This approach avoids the usual problems faced when doing data anonymization. If the data is anonymized before it is analysed, you may well lose much of the utility in the data. But if the data is anonymized after the analysis you don't have sufficient knowledge of the underlying data to know whether the anonymization is strong enough.

The key to this design is the way that the analytics and anonymization steps are performed simultaneously.

Additionally, Aircloak's approach has engineered away the need for a privacy budget by producing tailored pseudo-random noise values that do not average away. Repeated or semantically equivalent queries produce the same noise values, which in turn leads to the ability to ask as many queries of your dataset as you desire, in contrast to Differential Privacy where you depend on the privacy budget and only have a limited amount of queries available.



Screenshots of the Aircloak Insights User Interface and Tableau connection

Aircloak Insights consists of two discrete modules that work together. Both components are installed on-premise, deployed as standard Docker containers and can run equally well on bare metal servers or your own virtual server infrastructure:

Insights Air

Insights Air is a web-based control centre that ships as an integral part of Aircloak Insights. Insights Air offers you full granular control over who sees your data and provides all the necessary data security and audit functions needed to ensure you are compliant with any data protection legislation. Insights Air supports LDAP and Active Directory and provides full authentication and authorisation functionality to verify the identity of analysts and ensure they only have access to the datasets they are allowed to see. Administrators get full visibility of the system health and a detailed audit log tracking all queries and system actions.

Insights Cloak

Insights Cloak handles the actual queries, passing them to the data storage and anonymizing the results as they come back. It is able to do the anonymization properly because it understands the query semantics as well as having access to the actual raw un-anonymized data. This means it can apply the most suitable anonymization to the results before passing these back to the analyst.

What Other Benefits Does This Approach Bring

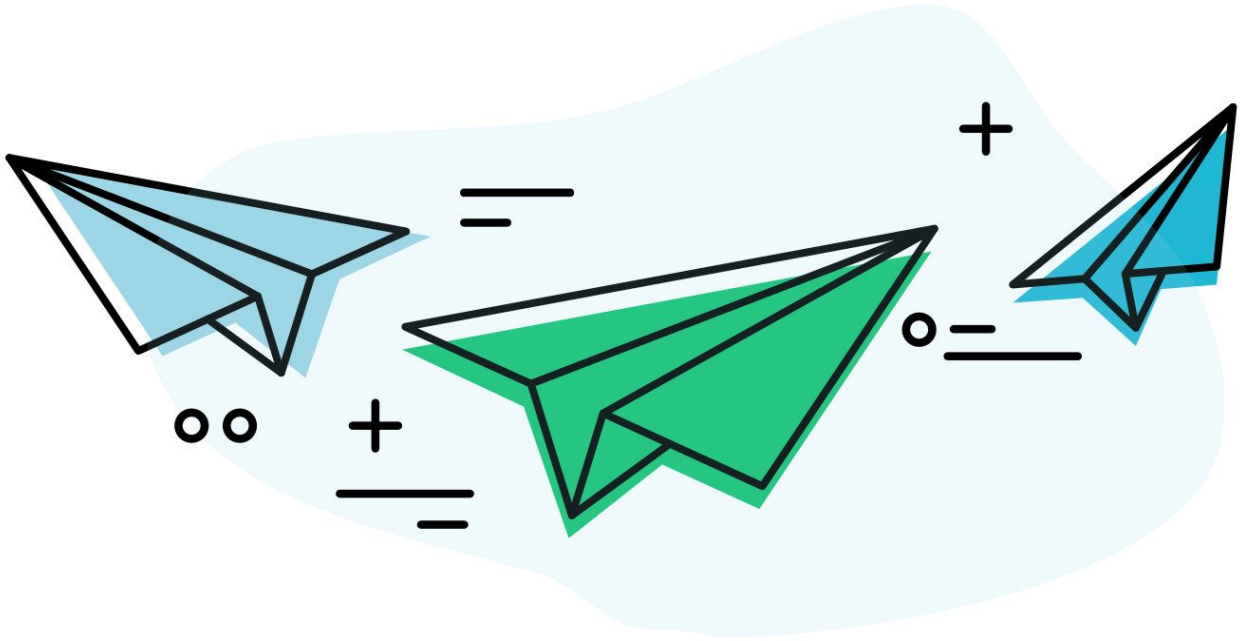
This new approach to anonymization helps you streamline the process or running analytics. Chief among these is the fact that any suitable team member can run analytics over the whole dataset without the need for specific per-use-case authorizations or anonymization.

This means all organizations who store valuable data, such as health, location, or financial data, can now instantly and safely monetize the data regardless of its format.

About Aircloak

Aircloak's privacy preserving analytics solution has been developed after years of research in close collaboration with the Max Planck Institute for Software Systems. It has been confirmed to be fully compliant with European Guidelines for anonymization by the French Data Protection Authority CNIL and is already in use at European healthcare, finance and telecommunication companies.

Do you want to learn more?
Write us at solutions@aircloak.com or schedule a demo!



Aircloak GmbH
Brunnenstraße 185
10119 Berlin
solutions@aircloak.com
www.aircloak.com

Aircloak GmbH, Kaiserslautern, Germany | Amtsgericht Kaiserslautern, HRB 31665 | Managing Director: Felix Bauer