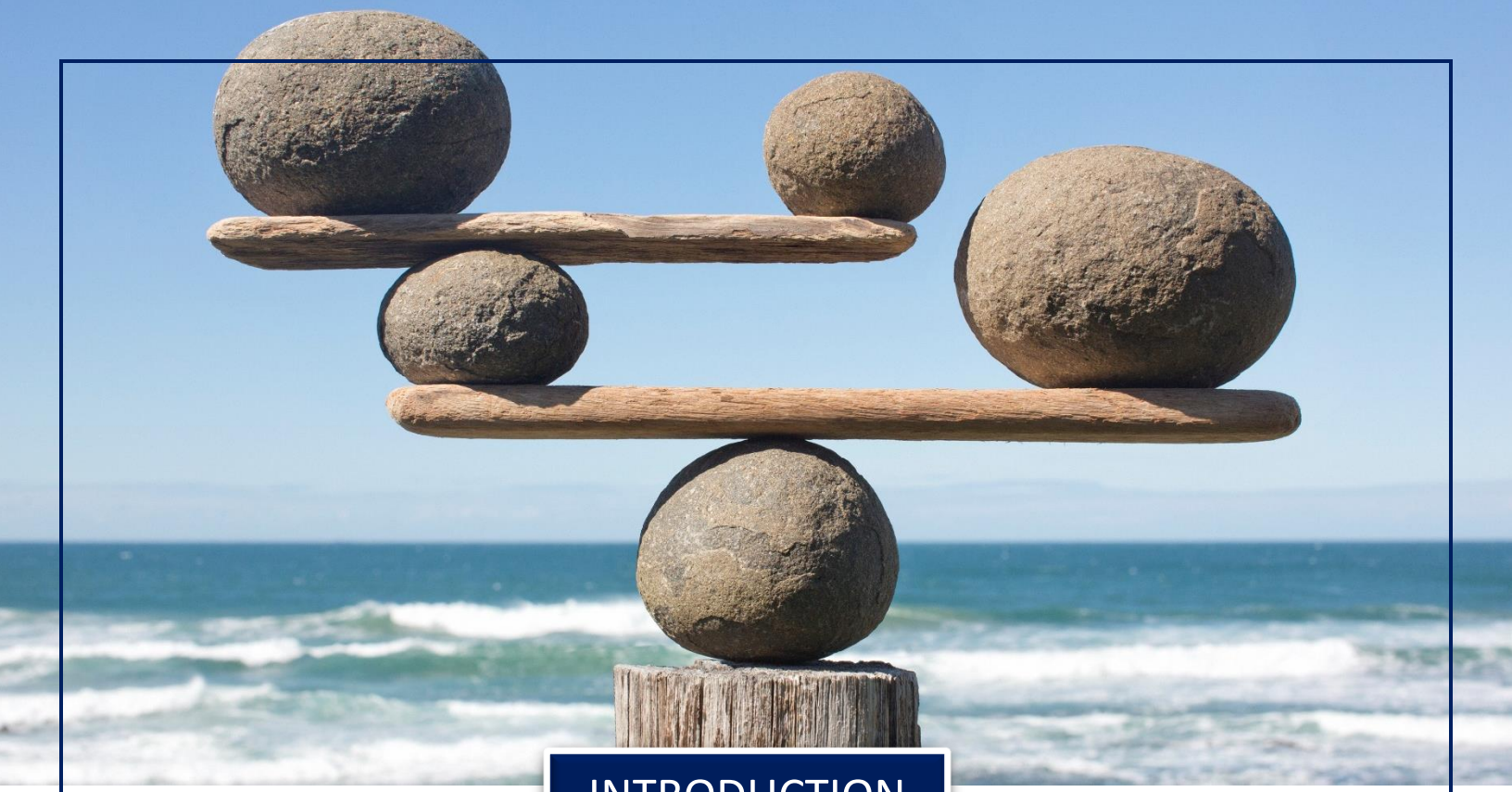
CHALLENGES IN DE-IDENTIFICATION

FINDING THE RIGHT BALANCE BETWEEN PRIVACY AND DATA UTILITY



INTRODUCTION

Imagine unlocking the full potential of your data without compromising on privacy or compliance.

In this whitepaper you will learn more about the main challenges and best practices regarding de-identification.

In today's data-driven world, organizations must comply with privacy regulations like the General Data Protection Regulation (GDPR) in the European Union and the Health Insurance Portability and Accountability Act (HIPAA) in the United States. Most organizations are caught in a conflict between the need for insights and the imperative of privacy protection. The cost of getting it wrong? Millions in fines and irreparable reputation damage.

This conflict led to a possible solution in finding the correct balance between Privacy and Data Utility: **De-identification**. De-identification is the process of removing, masking or aggregating personal identifiers in such a way that data subjects in a complete dataset, tool or algorithm are **less likely to be identified** or even no longer identifiable. It can include both pseudonymization and anonymization techniques, such as masking, suppression or obfuscation.

De-identification is a crucial method for protecting individual privacy while still enabling the use of data for research and analysis. However, achieving an effective de-identification process faces many challenges, primarily due to the delicate balance required between maintaining data utility and ensuring privacy.

WHAT ARE THE MAIN CHALLENGES?

Lack of knowledge

One of the primary challenges in de-identification is the lack of comprehensive knowledge among organizations. Many are not fully aware of the technical possibilities and limitations of current de-identification methods. For instance, while removing direct identifiers such as names and social security numbers is relatively straightforward, understanding the nuances of re-identification risk caused by indirect identifiers, like combinations of age, gender, and location, **requires a deep statistical and technical knowledge**.

A common misunderstanding is the belief that simply removing columns of identifiable information guarantees anonymity. However, as illustrated by Jessilyn Dunn[1], an assistant professor of biomedical engineering at Duke University, re-identification risk can remain. Dunn's research[2] revealed that de-identified health data could be re-identified by matching it with other datasets between 86% and 100% of the time. This highlights the need for organizations to invest in specialized training and stay updated on state-of-the-art de-identification techniques to effectively mitigate these risks.

Risk of re-identification

The risk of re-identification remains a significant challenge, particularly in the era of big data where numerous datasets can be cross-referenced. Re-identification occurs when anonymized data is matched with external data, revealing the identity of the data subject. For example, Jessilyn Dunn's research at Duke University found that wearable health data, even when de-identified, could be re-identified by matching it with other datasets. This research showed that a few seconds of sensor data could be enough to re-identify individuals by correlating it with known data.

This case underscores the importance of **continuous risk assessment** and the implementation of robust de-identification techniques. Besides using state-of-the-art techniques, organizations must regularly evaluate their effectiveness as new data and technologies emerge. **Setting an acceptable risk threshold** and ensuring ongoing monitoring is crucial to maintaining privacy.

[1] Link to the bio of Jessilyn Dunn

[2] Link to Dunn's research

Failing to properly remove identifiable information

Personal identifiable information (PII) consists of **direct and indirect identifiers**. Effective de-identification requires addressing both direct and indirect identifiers to prevent re-identification. Direct identifiers, like names and social security numbers, can be masked, hashed, or suppressed. However, indirect identifiers, such as gender, age, and event dates, pose a more complex challenge. The indirect identifiers can be combined in ways that uniquely identify individuals.

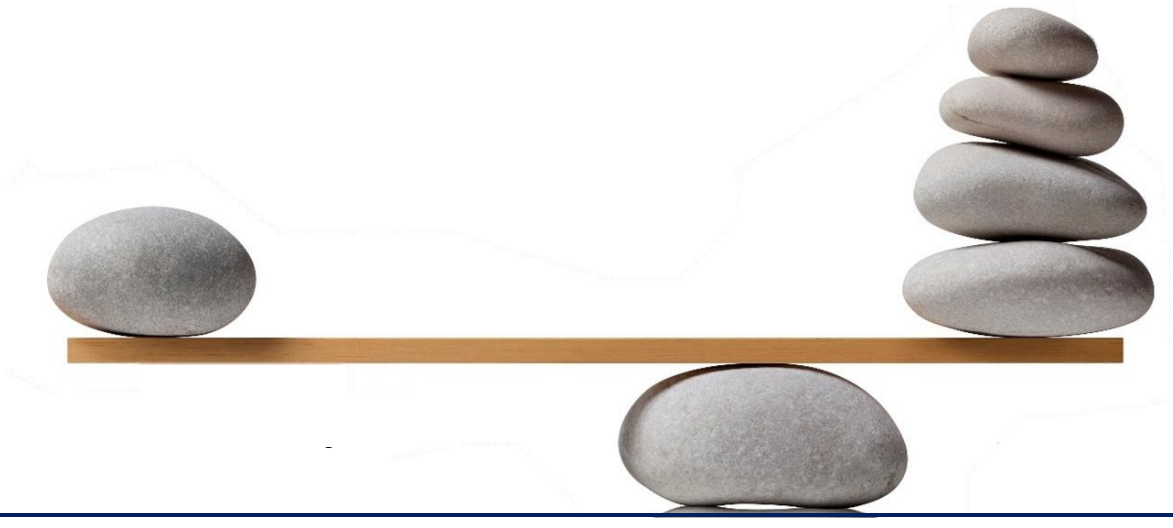
Consider the case of a hospital sharing patient data with a research institute. Even after removing direct identifiers, the combination of indirect identifiers (e.g., diagnosis date and treatment type) allowed researchers to re-identify patients by cross-referencing the data with other publicly available information.

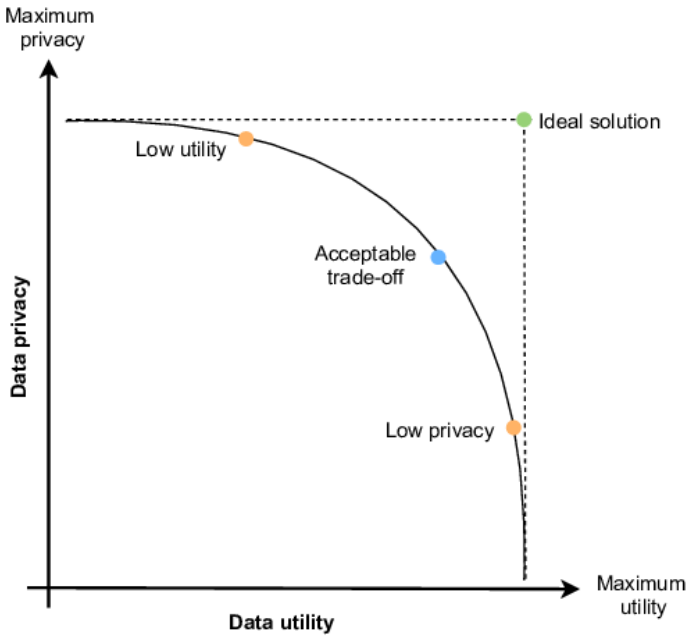
Therefore, these indirect identifiers would be subject to a risk assessment to calculate the probability of re-identification before any mitigations are done. This is what makes it complex, since you must **understand the business needs** to be able to recommend the best mitigations while not losing the most important variables.

Not involving other specialists

De-identification is not solely a technical challenge. It requires a multidisciplinary approach, involving **security experts, business users and privacy legal experts**. Each specialist brings a unique perspective that is crucial for developing legally compliant and technically sound de-identification strategies. For example, privacy legal experts ensure that de-identification practices adhere to regulations such as GDPR and HIPAA, while business users provide insights into the data's practical applications and necessary utility.

Consider you are working in a pharmaceutical company and your project involves clinical trial data. Involving biostatisticians, IT security professionals, and legal experts will then be crucial to ensure that the de-identification process meets regulatory standards and maintains the utility of the data for further research. Coordinating these efforts can be challenging but **aligning is essential for achieving effective de-identification**.





Finding the right balance between privacy and utility (source: ResearchGate)

Losing too much data utility

Preserving the utility of de-identified data while ensuring privacy is a delicate balance. Data must **retain enough detail to be useful** for research and analysis without compromising individual privacy. For example, in research, data must often include specific details about the participants' conditions, treatments and steps taken. However, these details can also make it easier to re-identify individuals.

One approach is to involve end users in the de-identification process to **understand the specific needs and purposes** of the data. For instance, when de-identifying health data, collaborating with medical researchers can help determine which data elements are essential for analysis and which can be safely modified or removed to protect privacy.

Complying with different regulatory requirements

In the current complex regulatory landscape, complying can be really challenging, especially for companies operating in several jurisdictions. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the USA and the General Data Protection Regulation (GDPR) in Europe have stringent requirements for data privacy, each with its nuances. For instance, while GDPR focuses on anonymization to remove data from its scope, HIPAA allows for data to be de-identified through expert determination or via the safe harbor method.

Organizations must ensure that their **de-identification processes are compliant with all relevant regulations**. This often involves creating jurisdiction-neutral processes that can adapt to different legal requirements. Regular audits and updates to de-identification strategies can help organizations remain compliant and protect individual privacy effectively.

BEST PRACTICES FOR DE-IDENTIFICATION

To overcome these challenges in de-identification, your organization can start by adopting the following best practices. These are meant to build a robust and effective de-identification process to be implemented.

Conduct thorough risk assessments

Before starting the de-identification process, it is essential to conduct a thorough risk assessment on the indirect identifiers left in the dataset. Direct identifiers should be removed or masked to prevent a 100% re-identification risk. Evaluate both data risk and context risk (the environment in which the dataset will be stored and used) to understand the probability of re-identification. Identifying high-risk elements and addressing them proactively can significantly reduce the probability of re-identification.

In our previous example the hospital wanted to share patient data for a research study without direct identifiers, only the combination of treatment dates and specific medical conditions could potentially re-identify patients when matched with external data. By recognizing these risks, additional transformations such as date shifts could be applied.

Involve a multidisciplinary team

Effective de-identification requires collaboration among security experts, business users, and privacy legal experts. This multidisciplinary approach ensures that de-identification strategies are legally compliant, technically sound, and aligned with business needs. Regular training and awareness programs can help build a robust de-identification process within your organization.

If a company does not involve experts, such as IT security professionals and business users, leaving only privacy legal counsels to have provided their perspective, they might not be able to maintain the utility of the data for further research. Collaboration is crucial to achieve the ultimate balance between utility and privacy.

Use appropriate techniques

Proper de-identification should ensure that the residual risk of re-identification is low enough to not be able to single out a specific individual. It is crucial to differentiate between pseudonymization and anonymization techniques. Pseudonymized data can still be linked back to individuals if additional information is available. Anonymization, on the other hand, should ensure that data cannot be associated with specific individuals, rendering it outside the scope of GDPR. However, fully anonymized data can be challenging to achieve and maintain.

A company might use pseudonymization techniques to key code the personal data of a participant. Besides this, that company could also consider generalizing ages into ranges and removing specific event dates or using a date-shift technique with random offset (both anonymization techniques) before they transfer the data to a third party for a research study.

Leverage State-of-the-Art Tools and Techniques

Stay updated on the best and latest de-identification techniques and tools. Encryption for example is not a foolproof anonymization technique, as this can be reversed. Therefore, ensure that the tools and techniques used are fit for purpose, considering security, privacy, and business requirements. Besides this, make sure you are properly trained in all these techniques and tools.

Educate and train colleagues

Ensure that not only the experts within your company understand why this is so important. Creating awareness within all departments of who they can contact if they have any questions and ensuring that you have several ambassadors within your organization will help in building a robust process for all datasets within your organization to follow.

A large device provider implemented a comprehensive training program to educate their staff across all departments on the importance of de-identification. This program included workshops and regular updates. Besides this, clear information on the point of contact, a legal privacy expert, was provided for the employees to know who to contact in case of questions.

Engage with external experts

Fortunately, you are not alone in this matter. Besides experts in the field of de-identification, there are many authoritative sources you can reach out to for guidance in these topics, such as IAPP - International Association of Privacy Professionals for all privacy professionals.

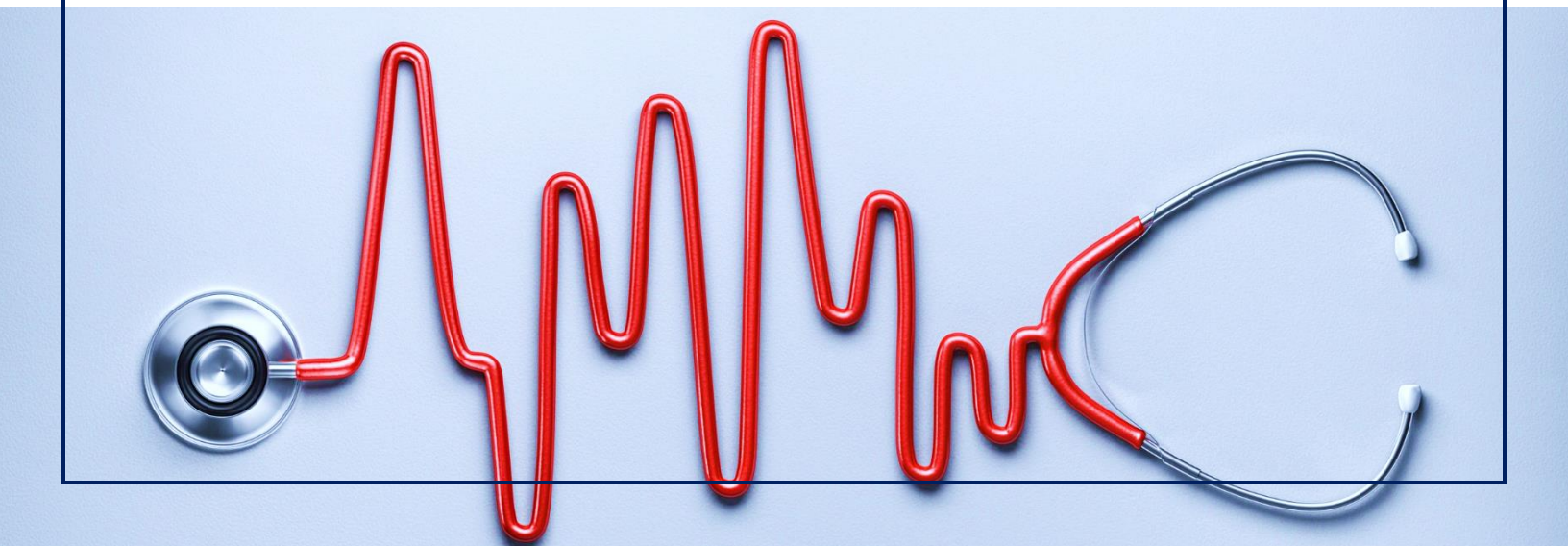
For companies in the healthcare business you have MedTech Europe supporting medical device companies and EFPIA - European Federation of Pharmaceutical Industries and Associations supporting pharmaceutical companies.

Collaboration can provide you with valuable insights and result in a more structured way of working.

Continuous Monitoring and Adaptation

De-identification is not a one-time process; it requires continuous monitoring and adaptation. Technological advances and new data releases can compromise previous de-identification efforts. Set a validity time for the risk assessments done, for example, two years considering the conditions are not changed. Ensure to regularly update your de-identification strategies to address new risks and maintain the balance between data utility and privacy.

A company regularly reviewed and updated its de-identification processes in response to new technological developments and potential data breaches. By setting a review cycle and continuously adapting their strategies, they were able to maintain data subject privacy while enabling valuable research



CONCLUSION

Walking the fine line between data privacy and utility is a continuous challenge. By embracing a multidisciplinary approach and staying updated with the latest techniques, organizations can achieve a balance that protects individual privacy while maximizing the utility of their data. This process cannot be fully automated, it needs a human expert oversight to continuously challenge what de-identification techniques would be best used, supported by the right tooling.

Ensuring the challenges are met and implementing the best practices provided could be a great first step to unlock the full potential of your data without compromising privacy or compliance.

Click here if you want to learn more about the differences between anonymization, pseudonymization and de-identification.

ABOUT B.A.I.M.E

B.A.I.M.E was founded by Renate van Kempen and stands for Business A.I. Made Easy.

B.A.I.M.E can support anonymization, de-identification and pseudonymization challenges!
B.A.I.M.E can also perform a quick-scan risk assessment on your datasets.

If you are interested in learning more about the possibilities, feel free to contact Renate van Kempen via renate@baime.nl or +31 6 481 737 35.
More information can also be found on the website www.baime.nl.

www.baime.nl