Maria Catarina Caridade Batista

Synthetic Data and GDPR Compliance: Navigating the Legal and Ethical Landscape

Dissertation to obtain the Master's Degree in Law, with speciality in Business Law and Technology

Supervisor:
Dr Graça Canto Moniz, Professor of NOVA School of Law

December 2023

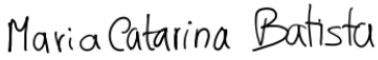
"*Aqueles que passam por nós, não vão sós, não nos deixam sós. Deixam um pouco de si, levam um pouco de nós.*"

Antoine de Saint-Exupéry

ANTI-PLAGIARISM STATEMENT

I hereby declare that the work I present, is my own work and that all my citations are correctly acknowledged. I am aware that the use of unacknowledged extraneous materials and sources, constitutes a serious ethical and disciplinary offence.

Maria Catarina Batista



Lisbon, 14 December 2023

ACKNOWLEDMENTS

First and foremost, I extend my deepest gratitude to my supervisor, Dr Graça Canto Moniz, for her invaluable guidance, patience, and expertise throughout my academic journey.

Special thanks are also due to Dr Gonçalo Ribeiro, whose guidance and encouragement have been instrumental in my growth and learning. The support provided by Dr. Ribeiro and the entire YData team has created an enriching environment that was crucial for the development of this study.

To my parents, António and Helena, and to my siblings, Beatriz and Afonso, I owe immense gratitude. Your unwavering love, belief, and support have been my anchor in this challenging year, offering comfort and strength throughout this time. Your presence has been a constant source of encouragement and inspiration.

I am equally thankful to my colleagues and friends, specially José Belo, Kate Francis and Maria Grácio. Your stimulating discussions, constructive feedback, and continuous encouragement have greatly enhanced my path of growth and development.

In closing, as Helen Keller aptly said, "Alone, we can do so little; together, we can do so much." This quote deeply resonates with me, as it underscores a vital truth: my journey to this point would have been impossible without the support and collaboration of all those acknowledged here.

Thank you all!

QUOTING AND OTHER CONVENTIONS

I. In this dissertation, quoting other academic works will be done in footnotes as well as in the References section. When applicable, a hyperlink to the document will be provided.

II. The quoting style that is used throughout the work is based on OSCOLA (Oxford University Standard for the Citation of Legal Authorities).

III. Due to the rapid evolution of this theme, this dissertation is updated with materials until 26 October 2023.

IV. The body of this dissertation has 156.862 characters, including spaces and footnotes.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AIML	Artificial Intelligence and Machine Learning
CJEU	Court of Justice of the European Union
DPA	Data Processing Agreement
E.g.	Example given
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
ENISA	European Union Agency for Cybersecurity
Etc.	Etcetera
GAN	Generative Adversarial Networks
GDPR	General Data Protection Regulation
ICO	Information Commissioner's Office
ML	Machine Learning
ONS	Office for National Statistics
PET	Privacy Enhancing Technology
SRB	Single Resolution Board
TFEU	Treaty on the Functioning of the European Union

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This thesis presents a detailed investigation of the evolving landscape of Personal Data in the digital era with a specific focus on the intricate relationship between Synthetic Data and their implications for Data Protection, especially concerning the General Data Protection Regulation ("GDPR"). This study begins with a deep analysis of the concept of Personal Data in Chapter 1. It then focuses on distinguishing between Personal and Non-Personal Data (according to the definitions in the GDPR). Chapter 2 extends the discussion to Data Synthesis, highlighting its potential and difficulties in balancing data utility and the Data Subjects' rights. Chapter 3 explores the transformative role of Synthetic Data in reinforcing Data Protection and ethical standards, while Chapter 4 critically examines ethical and practical challenges associated with Synthetic Data use. Lastly, Chapter 5 summarizes these insights, offering legal recommendations for using Synthetic Data within the GDPR framework and emphasizing the need for continuous adaptation and ethical mindfulness in this rapidly evolving field. This research contributes significant insights into the complexities of relying on Synthetic Data as a Privacy-Enhancing Technology within Personal Data processing activities.

RESUMO

Esta tese apresenta uma investigação detalhada da evolução do panorama dos Dados Pessoais na era digital, com um foco específico na complexa relação entre os Dados Sintéticos e suas implicações para a proteção de dados, especialmente no que respeita ao Regulamento Geral sobre a Proteção de Dados ("RGPD"). O estudo inicia-se com uma análise profunda acerca do conceito de Dados Pessoais (Capítulo 1). Incide, em seguida, sobre o desafio que a distinção entre Dados Pessoais e não Pessoais representa (de acordo com as definições que constam do RGPD). O Capítulo 2 estende a discussão para a Sintetização, com especial destaque para o obstáculo em estabelecer um equilíbrio entre a utilidade dos dados e os direitos dos Titulares dos Dados. O Capítulo 3 explora o papel transformador dos Dados Sintéticos no reforço da proteção de dados e padrões éticos, enquanto Capítulo 4 examina criticamente questões éticas práticas associadas ao uso de Dados Sintéticos. Por último, o Capítulo 5 sumariza as análises anteriormente concretizadas, oferecendo recomendações legais para o uso de dados sintéticos dentro do quadro do RGPD e enfatizando a necessidade de adaptação contínua e consciencialização ética neste campo, que se encontra em célere desenvolvimento. Em suma, esta pesquisa contribui com opiniões significativas sobre a complexidade inerente à utilização da Sintetização enquanto tecnologia de aprimoramento da privacidade nas atividades de tratamento de dados pessoais.

I. Introduction

With the exponential wave in data volume and the evolving landscape of data processing, the implications for privacy and Data Protection become increasingly complex. This era is marked not just by the quantity of data generated but also by the extensive nature of our digital footprints. As Personal Data permeates through various entities and processing activities, understanding how it is collected, processed, and protected becomes imperative. This thesis aims to dissect these dynamics, especially in the light of Synthetic Data's development, which further complicates the traditional notions of Personal Data and Data Protection in our society.

Before starting this research, I understood that the complexity of Data Protection issues within the current technological era demands an interdisciplinary approach. From my perspective, the intersection of law, technology and ethics is crucial for effectively addressing the contemporaneous encounters regarding Synthetic Data and Data Protection. Therefore, legal professionals, technologists and ethicists must work together to ensure that regulations like the GDPR are both technologically informed and ethically sound. Accordingly, this thesis embodies an interdisciplinary ethos, seeking to bridge the gap between legal theory and technological practice.

Data Synthesis, as a powerful Privacy Enhancing Technology ("PET"),[1] represents an important figure within the legal landscape, specifically in the realm of Data Protection. Its impact and relevance extend far beyond mere discussion, making it a transformative force in safeguarding Personal Data and upholding Data Subjects' rights. This research represents a comprehensive examination of the existing

[1] "PETs are technologies that embody fundamental Data Protection principles by minimizing Personal Data use, maximizing data security, and/or empowering individuals. Data Protection law does not define PETs. The concept covers many different technologies and techniques." Also, the European Union Agency for Cybersecurity ("ENISA") refers to PETs as: "'software and hardware solutions, ie systems encompassing technical processes, methods or knowledge to achieve specific privacy or Data Protection functionality or to protect against risks of privacy of an individual or a group of natural persons." See Information Commissioner's Office, 'Guide to Accountability and Governance' <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/accountability-and-governance/data-protection-by-design-and-default/> accessed 9 September 2023.

literature and engages with diverse professionals in the technology field to acquire a complete knowledge of Synthetic Data, the underlying Machine Learning models and their implications in the context of Data Protection.

The initial stage of this research was grounded in an extensive literature review, encompassing books and scientific articles, providing a theoretical foundation in Synthetic Data. Another key aspect of my research methodology is the incorporation of insights from Dr Khaled El Emam, a prominent figure in the field of Synthetic Data. Dr El Emam's knowledge as a leading authority on Synthetic Data and Anonymization in Canada provides significant depth to this study, especially in understanding the practical applications and ethical implications of Synthetic Data.

This theoretical base was significantly enhanced through direct collaboration with YData, a leading startup in Synthetic Data generation. The collaboration with YData offered me a unique hands-on perspective on the complexities of the Synthetic Data lifecycle, from algorithm development to dataset production, underscoring the interdisciplinary nature of this study. Engaging closely with YData, this thesis benefits from a blend of legal insights and technical expertise. This early-stage collaboration instigated detailed discussions and analysis of Datasets, thus providing a foundational technical understanding crucial for framing the legal analysis that forms the core of this study.

This investigation, at its core, seeks to unravel and explore the multifaceted nature of Synthetic Data in the context of Data Protection Laws, particularly the GDPR. The primary objective is to determine whether Synthetic Data falls within the purview of GDPR and, if so, at which stage in its Data Synthesis process. This inquiry involves a critical examination of the concept of Personal Data as defined by the GDPR, assessing the implications of Synthetic Data from a legal perspective.

The primary thrust of the research was to discern whether Synthetic Data could fall into the categories of Anonymized or Pseudonymized Data. Such examination led me through a labyrinth of legal frameworks, scholarly articles, and pertinent case studies to understand the differences among these types of data. However, the

further I progressed in my research, the clearer it became that Synthetic Data defies the conventional concept of Anonymized Data. Key insights emerged from dialogues with industry professionals, leading to a realization that unlike data where personal identifiers from a Data Subject are removed (Anonymized Data), Synthetic Data, even with a small dataset as input, can generate thousands hypothetical data corresponding to the same amount of artificial Data Subjects. This distinction was further discussed with Data Protection consultants and experts, revealing prevalent misconceptions about Synthetic Data being simply Anonymized Data.

Therefore, this research aims to dissect and clarify this confusion of concepts by analyzing three interrelated PETs: De-identification, Anonymization, and Data Synthesis. At a later moment of the study, it is examined whether Synthetic Data falls under the GDPR's definition of Personal Data and at which point in the Data Synthesis process GDPR considerations become pertinent. The research incorporates practical cases, including the accuracy of health data, synthetic population databases, and software testing scenarios, to demonstrate the real-world applications and benefits of Synthetic Data. It also probes into the ethical implications of Synthetic Data, particularly its growing significance in societal contexts.

Emphasizing the necessity for a dynamic and evolving concept of Personal Data to accommodate technological advancements, this thesis positions Synthetic Data as a key exemplar for such need. By addressing these objectives, the study endeavors to contribute to the ongoing discourse by offering insights and recommendations for policymakers, organizations, and researchers navigating the complex topic of Synthetic Data, promoting GDPR compliance and responsible data sharing and innovation.

Each section of this thesis weaves together a comprehensive narrative that intricately combines legal, technical, and ethical aspects of Synthetic Data, culminating in detailed legal recommendations tailored for each of these communities. Concluding, the cumulative effect of these chapters aims to provide

these communities with a thorough understanding and guidance on navigating the complex landscape of Synthetic Data and GDPR compliance.

1. Blurring Lines: The Concept of Personal Data

1.1. Distinguishing Personal from Non-Personal Data

In the realm of Data Protection, the dichotomy between Personal and Non-Personal Data has ignited vigorous debates among scholars and technologists.[2] In this chapter, we embark on a brief examination of this distinction, navigating through pertinent legal provisions and exploring the evolving nature of data. The result of such analysis sets the stage for the subsequent exploration of whether Synthetic Data meet the criteria for classification as Personal Data within the existing legal framework or if they require enhancements to the conceptual framework.

Article 4(1) of the General Data Protection Regulation (henceforth, "GDPR")[3] offers a defining framework for Personal Data. According to the GDPR, Personal Data include:

> *any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person*.

In essence, Personal Data constitute information directly or indirectly linked to an identified or identifiable individual – henceforth referred to as "Data Subject".[4]

[2] In exploring the legal dimensions of Synthetic Data, the insights of several esteemed professionals are particularly noteworthy. Michal S. Gal, a Professor and the Director of the Center of Law and Technology at the University of Haifa Faculty of Law, also serves as the President of the International Association of Competition Law Scholars (ASCOLA) and is a Visiting Professor at NYU School of Law. Orla Lynskey, an Associate Professor at the London School of Economics and Political Science and a Visiting Professor at the College of Europe in Bruges, offers another valuable perspective. Additionally, the work of Georgi Ganev, a PhD Researcher at University College London, contributes to the discourse. For technological insights, the extensive research and publications of Dr Khaled El Emam on Synthetic Data are indispensable.

[3] General Data Protection Regulation (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC).

[4] Article 4(1), GDPR.

Additionally, Regulation (EU) 2018/1807,[5] known as the Free Flow of Non-Personal Data Framework within the European Union (hereinafter referred to as the "Regulation of the Free Flow of Non-Personal Data"), establishes the rules on secure and easy exchanges of Non-Personal Data across the Union. It represents a conscientious effort by the European Union to establish a standardized approach to processing Non-Personal Data, thereby promoting innovation, competition, and data-driven growth while upholding the Principles of Security and Data Protection.[6]

Within this framework, detailed provisions are established to address crucial aspects of data management. These encompass guidelines concerning data localization requirements, which dictate where data can be stored or processed, and the obligations surrounding data accessibility for competent authorities. For example, Article 4 of the Regulation of the Free Flow of Non-Personal Data addresses data localization requirements by stipulating the conditions and constraints under which Non-Personal Data can be stored or processed within the European Union. This article sets forth the legal framework that prohibits Member States from enacting laws that unjustifiably restrict the storage or processing of Non-Personal Data across national borders within the European Union. Furthermore, the article under discussion effectively endorses the principle of Free Flow of Non-Personal Data, seeking to eliminate data localization barriers and thus fostering a digital single market. Nevertheless, this approach has been criticized for being too narrow and not providing a solid foundation for a new regime that would only apply when Non-Personal Data is being processed.[7] Indeed, the Regulation of the Free Flow of Non-Personal Data, in Article 3(1), explicitly delineates the concept of (Non-Personal) Data. It defines such data as information falling outside the scope of Personal Data, as explained in paragraph 1 of the preceding Article 4 of the GDPR.

[5] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of Non-Personal Data in the European Union (Text with EEA relevance.) 2018.
[6] Graef and others 'Towards a Holistic Regulatory Approach for the European Data Economy: Why the Illusive Notion of Non-Personal Data Is Counterproductive to Data Innovation' [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3256189> accessed 10 September 2023.
[7] Ibid.

On another hand, the recent Regulation (EU) 2022/868 (henceforth, "Data Governance Act") defines data as "any digital representation of acts, facts or information and any compilation of such acts, facts or information, including in the form of sound, visual or audiovisual recording". [8] Therefore, the Data Governance Act provides a broader definition of data, encompassing any digital representation of acts, facts, or information, regardless of its nature (Personal or Non-Personal).

We contend that (Non-Personal) Data has two distinct categories. The first category of (Non-Personal) Data pertain to any representation of acts, facts or information that lacks any connection to an identified or identifiable individual. Secondly, there exists a subset of (Non-Personal) Data that previously held the classification of Personal Data but has since undergone a meticulous Anonymization process. This process involves the thorough removal of all identifiers and associations with Data Subjects. As a result, these data shed their previous status as Personal Data and assume the form of (Non-Personal) Data, devoid of any discernible links to individuals – second category of data.[9]

Contrary to the conventional perspective held by many privacy professionals and scholars,[10] we argue that the distinction between Personal and Non-Personal Data is full of complexities and challenges. The fluidity of data classification complicates this differentiation. Data initially categorized as Non-Personal can transition into Personal Data through integration with other datasets. Please note that this transformation depends on whether the Data Controller possesses additional information that facilitates such a merger. Consequently, the ability categorize data as Non-Personal is not only challenging but also dependent on the broader context and circumstances of each entity participating in the processing activity. For instance, data initially categorized as Non-Personal – second category – when

[8] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) (Text with EEA relevance) 2022.
[9] Recital 26 of the GDPR. Recital 26 of the European Parliament and Council, Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of Personal Data and on the free movement of such data. OJ L281/31.
[10] This perspective has been shaped and refined through engaging discussions with colleagues who specialize in Data Protection consultancy and by participating in discussions with professors during classes of my master's degree in business law and technology at NOVA School of Law.

combined with the Real Dataset – not Anonymized – or under the influence of emerging technologies – such as aggregation of data – may transition into the realm of Personal Data.[11] This makes it challenging to create a regulatory framework that applies only to Non-Personal Data, as it is difficult to determine what data falls into this category.[12]

The GDPR's scope is confined to Personal Data, which notably encompasses Pseudonymized Data but expressly excludes Anonymized Data.[13] This distinction is critical because once Non-Personal Data are deemed personal due to their transformation or re-contextualization, they fall under the purview of GDPR. This shift not only brings about compliance requirements for Data Controllers but also activates the Data Protection rights of the individuals concerned, thereby significantly impacting their privacy and control over their data. Moreover, re-contextualization poses significant risks for the Data Subjects, such as the risk of re-identification, where (Non-Personal) Data initially Anonymized within a dataset might be traced back to the correspondent Data Subjects. There is also a pronounced lack of transparency and control, as Data Subjects may be unaware of and unable to manage the data that they thought it was Anonymized. Likewise, this transformation also escalates security risks, as the newly classified Personal Data becomes a more attractive target for breaches, increasing the vulnerability of the Data Subjects' right to Data Protection.

This nuanced delineation of data reveals a fundamental truth: that the concept of data is not static, but dynamic and it constantly transforms and adapts to new technological frontiers, as it will be shown in the next chapters of this thesis. Central to this ever-changing landscape is the intricate relationship between PETs, specifically Data Synthesis, and the notion of Personal Data. The forthcoming section will focus on pertinent legal frameworks to elucidate the evolving character of the concept of (Non-Personal) Data.

[11] Aggregation of data refers to the process of combining multiple pieces of data, often from various sources, into a single, comprehensive dataset. Raman A, 'What Is Data Aggregation: A Comprehensive Guide 101' (*Hevo*, 26 September 2023) <https://hevodata.com/learn/data-aggregation/> accessed 18 November 2023.
[12] Graef and others (n 6).
[13] Recital 26, GDPR.

1.2. Relevant Legal Provisions

1.2.1. Recital 26 of the GDPR

The dynamic nature of the (Non-Personal) Data concept gains particular significance when examining Recital 26 of the GDPR.[14] This Recital holds a pivotal position within the GDPR framework as it explicitly states that the core principles of Data Protection do not extend to Anonymized Data.

Anonymized Data, as discussed in the preceding section, comprises information stripped of any identifiers connecting it to a Data Subject. Furthermore, Anonymization is commonly described as

> *the process of creating 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.*[15]

From this provision, it is reasonable to infer that Recital 26 of the GDPR categorizes Anonymized Data as Non-Personal Data. Consequently, such data fall outside the scope of the GDPR's stringent restrictions since the processing of Non-Personal Data does not pose a significant risk to Data Subjects' rights. This distinction is instrumental in differentiating between data that necessitate rigorous Data Protection measures and data that can be handled with a lighter regulatory touch, aligning with the GDPR's overarching objective of balancing privacy with data utility.

This juxtaposition emphasizes the evolving and complex landscape of data governance. The careful balance that regulators (e.g., Data Protection Authorities and the European Data Protection Board) must strike, as discussed supra, becomes

[14] Recitals of the GDPR serve as a valuable interpretative guide to understanding and applying the regulation's articles. While these recitals are not legally binding in themselves, they provide essential context, clarification, and intentions behind the GDPR's provisions. European courts and Data Protection authorities often refer to these recitals for deeper insight into the Regulation's purpose, aiding in the interpretation and application of the GDPR's legally binding articles. See R. S. I. Security, 'What Are GDPR Recitals?' (*RSI Security*, 20 June 2018) <https://blog.rsisecurity.com/what-are-gdpr-recitals/> accessed 5 November 2023.

[15] European Commission, 'Anonymization' (*Collaboration in Research and Methodology for Official Statistics*, 28 April 2019) <https://cros-legacy.ec.europa.eu/content/Anonymization_en> accessed 19 November 2023.

all the more critical in light of these provisions. Therefore, De-identification and Anonymization techniques must continuously adapt to remain effective, and regulators must ensure that even Non-Personal Data maintain a high standard of protection, especially in the face of advancing technologies that could potentially re-identify seemingly Anonymous information – the following sections will delve deeper into these techniques.

In their report on Pseudonymization techniques and best practices, the European Union Agency for Cybersecurity ("ENISA") emphasizes the need for advancements in this field, particularly in the context of the big data era.[16] ENISA suggests that the research community should focus on developing more sophisticated Pseudonymization solutions to effectively address the unique challenges presented by the vast scale and complexity of big data.[17] Furthermore, ENISA notes that although Pseudonymization techniques "can motivate the relaxation, to a certain degree, of Data Controllers' legal obligations if properly applied", they acknowledge that these techniques may not be appropriate for more complex scenarios, which are more and more common.[18] In such instances, ENISA points out that "the use of more advanced (and robust) techniques, such as those arising from the area of Anonymization, will become increasingly needed."[19]

The prompt advancements in Artificial Intelligence ("AI") and Machine Learning ("ML") technologies are posing a distinctive challenge to the applicability of the European concept of (Non-Personal) Data.[20] These rapidly evolving technologies possess the capacity to process vast volumes of data and generate unexpected insights.[21] While promising great benefits, the implementation of AI and ML

[16] European Union Agency for Cybersecurity (ENISA), 'Pseudonymisation Techniques and Best Practices' (2019) Report/Study <https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices> accessed 19 November 2023.
[17] ENISA (n 16), 6.
[18] ENISA (n 16), 7.
[19] ENISA (n 16), 43.
[20] Batura O and Peeters R, 'European Union Data Challenge' (Policy Department for Economic, Scientific and Quality of Life Policies 2021) PE 662.939 <https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/662939/IPOL_BRI(2021)662939_EN.pdf> accessed 4 December 2023.
[21] Ibid.

technologies comes with many challenges.[22] AI might spot patterns in Anonymized Data used to train algorithms that are not immediately visible to the human eye, consequently increasing the re-identification risks of Anonymization.[23] [24] Furthermore, the concern regarding the inferences drawn by AI on Data Subjects, which could be biased, discriminatory, or otherwise inaccurate, is unquestionably a critical and urgent issue.[25] These issues have the potential to impact not only a single decision or prediction made about a Data Subject but also to propagate poor inferences through the AI system, thus influencing future decisions or predictions.[26] [27] This chain of influence, known as linking, can lead to a course of data-driven decisions that are potentially flawed.[28] Such inferences risk instigating unfair treatment, privacy violations, or discrimination against individuals based on erroneous or biased conclusions.[29] [30]

On the contrary, as explained in the following section, when these technologies accurately derive inferences from (Non-Personal) Data, they challenge the boundaries established by Recital 26 of the GDPR. Due to this process of linking, (Non-Personal) data might transition into the realm of Personal Data, notwithstanding the Anonymized or Non-Personal nature of the input Dataset. In other words, through the application of AI analysis, data transitions from a state where it was not linked to specific Data Subjects (Non-Personal Data) to a state where it can be associated with particular Data Subjects (Personal Data). As a

[22] Ibid.
[23] Ibid.
[24] Centre for Information Policy Leadership (CIPL), 'Artificial Intelligence and Data Protection How the GDPR Regulates AI' (2020) <https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf> accessed 9 September 2023.
[25] Mittelstadt BD and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 Big Data & Society. 15 <http://journals.sagepub.com/doi/10.1177/2053951716679679> accessed 10 September 2023.
[26] Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) <https://papers.ssrn.com/abstract=2477899> accessed 10 September 2023.
[27] Raso FA and others, 'Artificial Intelligence & Human Rights: Opportunities & Risks' (25 September 2018). 7 <https://papers.ssrn.com/abstract=3259344> accessed 10 September 2023.
[28] Data linking refers to the process of matching and combining data from multiple databases. See Information Technology Laboratory, 'Data Linking - Glossary' (*Computer Security Resource Center*) <https://csrc.nist.gov/glossary/term/data_linking> accessed 19 November 2023.
[29] CIPL (n 24).
[30] Barocas and Selbst (n 25).

result, data that was formerly categorized as Non-Personal or Anonymized are now facing the prospect of re-identification or the inference of personal traits, thereby blurring the line between the concepts of Personal and Non-Personal Data.[31] Moreover, Recital 26 of the GDPR explicitly states that

> *The principles of Data Protection should therefore 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*.

While the GDPR deliberately excludes Anonymized Data from its scope, assuming lower risks to Data Subjects compared to Personal Data, the advancement in AI-driven inferences challenges these boundaries. The precision of these inferences, particularly in identifying Data Subjects from Anonymized Datasets, brings into question the adequacy of the current definitions and exclusions under GDPR. Given the evolving capabilities of AI and ML, there arises a controversial question: Should Anonymized Datasets, especially those employed in training AIML algorithms, be reconsidered under GDPR principles?

This shift in the classification of a dataset from Non-Personal to personal places it within the realm of the GDPR, subjecting it to the associated legal obligations. This underlying change underscores the importance of reevaluating the European Data Protection framework in light of ongoing technological advancements and the evolving standard of identifiability, a topic that will be elaborated upon in the following subsection.

1.2.2. The Standard of Identifiability

At the core of the concept of Personal Data lies the standard of identifiability, a criterion that is inherently tied to the timing of the data assessment, especially in the context of constant technological advancements, as discussed earlier. This section

[31] CIPL (n 24).

is dedicated to a deeper examination of this standard, and to the very definition of what constitutes Personal Data.

Recital 26 of the GDPR establishes a legal test for distinguishing between Personal and Non-Personal Data, based on the standard of identifiability of the Data Subject. This test requires the assessment of whether a natural person can be identified, considering all potential means that may reasonably be employed for direct or indirect identification. In this determination, objective factors play a pivotal role, encompassing considerations such as the associated costs, the time required for identification, and the available technology at the time of the processing and technological developments.

Hence, even though Article 4(1) of the GDPR provides a comprehensive definition of Personal Data, Recital 26 introduces a standard of "means likely reasonably to be used" to identify the Data Subject through the data. Therefore, it could be inferred that the definition of Personal Data has limitations and that the standard for identifiability should only be considered met if the means of identification are reasonably achievable to identify the Data Subject. In certain cases, even if identification cannot be completely ruled out, the level of risk may be negligible enough to consider data as Non-Personal.[32]

Furthermore, as outlined in Recital 26 of the GDPR, the assessment of whether means for a natural person's identification are reasonably likely to be employed involves a consideration of various objective factors. Recital 26 indeed acknowledges that the assessment of whether data can be considered personal is dynamic and should consider current technological capabilities. This dynamic nature of identifiability underlines that data initially deemed Anonymized could later be reclassified personal due to technological progress, without any changes to the data itself.[33] Consequently, it raises pertinent questions about whether the definition

[32] Finck M and Pallas F, 'They Who Must Not Be Identified - Distinguishing Personal from Non-Personal Data Under the GDPR' (1 October 2019) <https://papers.ssrn.com/abstract=3462948> accessed 17 June 2023.
[33] Purtova N, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40. 3. <https://www.tandfonline.com/doi/full/10.1080/17579961.2018.1452176> accessed 21 May 2023.

of Personal Data should also be dynamic and adaptable in response to evolving reality.

In essence, this perspective highlights the importance of assessing the likelihood and practicality of identifying an individual when classifying data as either. Recital 26 of the GDPR indeed acknowledges that the assessment of whether data can be considered as personal is dynamic. It emphasizes that the means likely to be used for identifying a natural person, including the costs, time, and available technology at the time of processing, are crucial factors. Subsequent chapters of this research will utilize the concepts explained above to delve into the central question of this study: "Do Synthetic Data meet the criteria to be classified as Personal Data?". Throughout this research, we develop and refine our investigation, ultimately culminating in a comprehensive answer to this query.

1.3. Exploring De-identification, Anonymization and Re-Identification Risks

In the pursuit of data protection, technologists have long relied on various technologies, including Privacy Enhancing Technologies (PETs), to manage Personal Data securely.[34] These technologies are designed to ensure the non-identifiability of Data Subjects, a critical aspect of Data Protection. However, the challenge of safeguarding Personal Data extends beyond just technological solutions.[35]

In this context, the GDPR plays a vital role. While it recognizes the importance of technological safeguards, the GDPR also introduces a comprehensive legal framework that enforces robust principles for Data Protection. These principles, as outlined in Article 5 of the GDPR, include lawfulness,[36] fairness,[37] transparency,[38]

[34] Ibid.
[35] Ibid.
[36] Article 5(1)(a), GDPR.
[37] Article 5(1)(a), GDPR.
[38] Article 5(1)(a), GDPR.

purpose limitation,[39] data minimization, [40] accuracy, [41] storage limitation, [42] integrity and confidentiality, [43] and accountability. [44] The regulation also addresses the handling of special categories of personal data, providing exceptions as per Article 9, GDPR. Crucially, the GDPR brings to the forefront the concepts of privacy by design, [45] and privacy by default, [46] thus integrating Data Protection into the developmental process of products and services. It also emphasizes information security as a key component of data handling.[47] Within this framework, the GDPR advocates for the use of Pseudonymization and Anonymization as key measures in enhancing data security.[48] These measures align with the broader objectives of the GDPR, striking a balance between protecting individual privacy and promoting data security. Accordingly, while the technological approaches to De-identification and Anonymization are foundational, their effectiveness and compliance are substantially reinforced by the GDPR's principles. This regulation not only guides but also mandates the application of these technologies in a manner that upholds the fundamental rights and freedoms of individuals, particularly regarding their Personal Data.

Although diverse De-identification approaches have instigated different notions of "De-identification", all these perspectives share common underlying objectives and principles. Hence, for the purpose of this study, we understand the concept in general terms, thus, we define De-identification techniques as any process of removing the association between a set of identifying data and the Data Subject.[49] Accordingly, De-identified data are records that have a re-identification code and have enough identifiable information removed or masked so that the remaining

[39] Article 5(1)(b), GDPR.
[40] Article 5(1)(c), GDPR.
[41] Article 5(1)(d), GDPR.
[42] Article 5(1)(e), GDPR.
[43] Article 5(1)(f), GDPR.
[44] Article 5(2), GDPR.
[45] Article 25(1), GDPR.
[46] Article 25(1), GDPR.
[47] Article 32(1), GDPR.
[48] See Recital 26, GDPR, Recital 28,GDPR, Article 25(1), GDPR, and Article 32(1)(a) GDPR.
[49] Information Technology Laboratory, 'De-identification - Glossary' (*Computer Security Resource Center*) <https://csrc.nist.gov/glossary/term/de_identification> accessed 25 November 2023.

information does not allow the identification of the Data Subject.[50] [51] Nevertheless, the re-identification code may allow the recipient to match information received from the same source.[52]

Pseudonymization is commonly described as a De-identification technique that replaces an identifier for a data principal with a pseudonym to hide the identity of that data principal.[53] The GDPR's definition of Pseudonymization, aligning with technical descriptions, sets a stringent implementation framework. It dictates that Personal Data should not be attributable to a specific individual without extra information.[54] As per GDPR Article 4(1), Pseudonymized Data must not allow direct or indirect identification of individuals without additional details. [55] This extends beyond shielding real world identity to include protecting indirect identifiers, like unique online identifiers.[56] Reversing Pseudonymization should be challenging for parties lacking this extra information. Nevertheless, according to the Opinion 05/2014 on Anonymization Techniques of the WP29,[57] pseudonymity is likely to allow for identifiability, and, therefore, must stay inside the scope of the legal regime of Data Protection.[58] Hence, the GDPR applies to Pseudonymized Data.

[50] Seastrom M, 'Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records' (23 November 2010). 6. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011601> accessed 19 September 2023

[51] Finch K, 'A Visual Guide to Practical Data De-identification' (*https://fpf.org/*, 25 April 2016) <https://fpf.org/blog/a-visual-guide-to-practical-data-De-identification/> accessed 19 September 2023.

[52] Ibid.

[53] Information Technology Laboratory, 'Pseudonymization' (*Computer Security Resource Center*) <https://csrc.nist.gov/glossary/term/Pseudonymization> accessed 25 November 2023.

[54] European Union Agency For Network and Information Security, 'Recommendations on Shaping Technology According to GDPR Provisions - An Overview on Data Pseudonymisation' (2018) 9-13 Report/Study <https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions> accessed 25 November 2023.

[55] Ibid.

[56] Ibid.

[57] The Article 29 Working Party was a European Union organization that worked as an independent advisory body on Data Protection and privacy. It consisted of the collected Data Protection authorities of the member states. The Article 29 Working Party was replaced by the similarly constituted European Data Protection Board (EDPB) on May 25, 2018, when the GDPR went into effect. See International Association of Privacy Professionals, 'Article 29 Working Party' (*ResourceCenter*) <https://iapp.org/resources/article/article-29-working-party/> accessed 6 December 2023.

[58] Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymisation Techniques' (2014) 0829/14/EN WP216 <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf> accessed 17 September 2023.

Furthermore, certain authors and publications tend to use the terms "Anonymous" and "Anonymization" interchangeably. However, for the clarity of this research, we make a clear distinction between the adjective "Anonymous" and Anonymization techniques. When we use the term "Anonymous Data", we are specifically referring to “information which does not relate to an identified or identifiable natural person”.[59] This implies that Anonymous Data are entirely disconnected from any specific Data Subject and may not have had any previous association with individuals. For instance, stock market data are Anonymous Data. In contrast, when we refer to 'Anonymized Data,' we are talking about Personal Data that have undergone a transformation process, rendering them Non-Personal or Anonymous. For example, government agencies release of Anonymized census data. Hence, Anonymized Data falls under the category of Anonymous Data, but not all Anonymous Data are necessarily Anonymized, as illustrated in Figure 1.

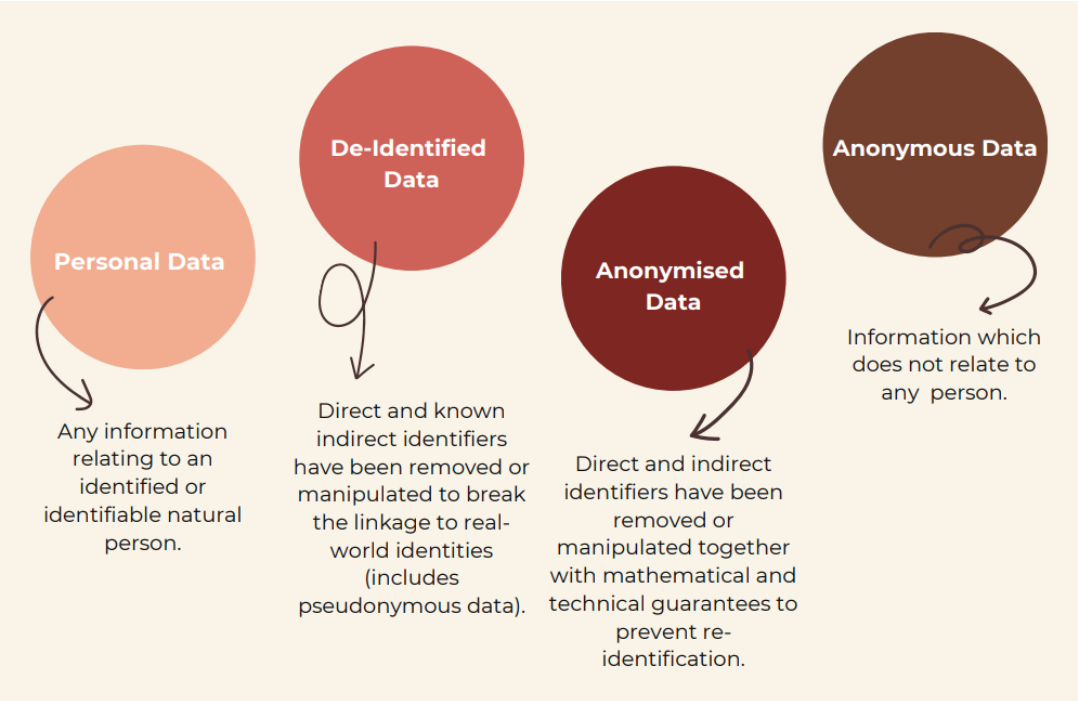


Figure 1 - Types of Data

While Anonymized Data no longer qualify as Personal Data under the GDPR, there is still a possibility of re-identification or risks stemming from residual information,

[59] Recital 26 of the GDPR.

as previously discussed. Consequently, if the Data Subject's re-identification takes place, data are reclassified as Personal Data and once again become subject to the GDPR.

1.3.1. The Evolution of the European Concept of Anonymized Data

During the GDPR's legislative procedure, the draft report proposed a definition of Anonymous Data, specifically excluding such data from the GDPR's scope.[60] This report defined them as

> *any data that cannot be related, directly or indirectly, alone or in combination with associated data, to a natural person or where establishing such a relation would require a disproportionate amount of time, expense, and effort, taking into account the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed*. [61]

Nevertheless, this definition was not included in the final redaction of the GDPR. A careful examination of Recital 26 of the GDPR offers a conceptual definition of Anonymization, thus, it asserts that for data to be considered "Anonymized", data must undergo a process that removes sufficient elements, rendering the data impossible to allow the identification of the Data Subject.[62] More precisely, this processing must ensure that no feasible means, reasonably employable by either the Controller or a third party, can be used to identify a natural person and this processing must be irreversible.[63] However, a significant change occurred in the adopted version of Recital 26. The revised version of the Recital under analysis marked a stringent position by removing the criterion of being resource-intensive to link an individual to the data. Instead, it now emphasizes that Anonymization must render it impossible to re-identify individuals from the data. Such a shift suggests a

[60] Amendment 14 proposed the following additional text to Recital 23 (now Recital 26). See European Parliament, 'REPORT on the Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) | A7-0402/2013 | European Parliament' (2013) A7-0402/2013 <https://www.europarl.europa.eu/doceo/document/A-7-2013-0402_EN.html> accessed 20 September 2023.
[61] Ibid.
[62] Article 29 Data Protection Working Party (n 58).
[63] Article 29 Data Protection Working Party (n 58).

more inflexible approach, requiring Anonymization techniques to achieve a level of protection where any possibility of re-identification is eliminated.

Furthermore, Recital 23 of the GDPR underscores the importance of considering "all the means likely to be reasonably used, either by the Controller or any other party, to identify the individual" when determining identifiability. However, the absence of criteria related to a disproportionate amount of time, expense, and effort required to identify a Data Subject has prompted scholars to question the precise standard for evaluating identifiability in the context of Anonymized Data.[64]

Proponents of a more radical position argue that any potential for re-identification, irrespective of its feasibility or likelihood, challenges the concept of valid Anonymization and presents privacy risks for individuals. [65] These advocates emphasize the need to develop rigorous measures to eliminate even the slightest possibility of re-identifying Data Subjects from Anonymized Data, as any residual risk undermines the fundamental right to the protection of Personal Data.[66] In opposition, critics of the stringent position argue that achieving the absolute impossibility of re-identification is often impractical or economically unviable.[67] They defend a pragmatic approach, thus focusing on reducing the risk of re-identification to a sufficiently low level, considering the contextual factors, available safeguards and legal requirements.[68] This approach aligns with the original text of Recital 26, therefore, by aiming to achieve a reasonable and attainable level of Anonymization while acknowledging the limitations and practical constraints, proponents of the pragmatic perspective seek a balanced approach to Data Protection, that respects individuals' rights without imposing undue burdens on Controllers.

[64] Ohm P, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' (13 August 2009) <https://papers.ssrn.com/abstract=1450006 > accessed 25 May 2023.
[65] Finck and Pallas (n 32), 5, 10 and 11.
[66] Finck and Pallas (n 32), 5, 10 and 11.
[67] Finck and Pallas (n 32), 5, 10 and 11.
[68] Finck and Pallas (n 32), 5, 10 and 11.

In April 2023, the Court of Justice of the European Union ruled that the term "re-identification" is subjective in nature.[69] This significant ruling, which will be further explored in the following sub-section, brought to light the complex nature of re-identification and emphasized the critical need to adopt a multifaceted approach when assessing the identifiability of the Data Subject.[70] In this judgment, it is suggested that the evaluation of the potential for re-identification of a Data Subject should take into account diverse perspectives, tailored to the unique circumstances of each case.[71] Therefore, according to distinct positions held by the parties considering a single processing activity, for instance, as Controllers or Processors, the ability to identify the Data Subjects will be different. Consequently, this ruling places a critical onus on supervisory authorities to conduct individualized assessments to ascertain whether the party involved possesses the means to access supplementary information that could lead to the re-identification of the Data Subjects.[72]

Moreover, Controllers inherently possess the capacity to re-identify Personal Data, unless they take deliberate steps to erase such information or impose access limitations on databases that could allow re-identification. The most recurrent problem in practice occurs when a Controller or Processor, having received Anonymized Data, refrain from re-identification but could potentially achieve re-identification through other information – or even databases – in their domain.[73] Hence, after an efficient Anonymization, data are no longer considered Personal Data, and thus Controllers/Processors are no longer subject to the associated stringent legal protections when processing such data. It is easy to imagine a scenario where a company receives a dataset that has been Anonymized to such an extent that it falls outside the scope of the GDPR. Now, envisage this company collaborating with another organization to process this Anonymized Dataset. Some critical questions emerge: With no clear Controller – as per GDPR terminology –

[69] *Single Resolution Board v European Data Protection Supervisor* [2023] General Court Case T-557/20, 80.
[70] Ibid.
[71] Ibid.
[72] Ibid.
[73] This observation is drawn from my professional experience as a Data Protection consultant, informed by practical insights and extensive discussions with colleagues in the field.

how can individuals' data be protected from re-identification by the new collaborating company? Are individuals whose data was Anonymized and now is within the Anonymized Dataset considered "Data Subjects" under GDPR?

Firstly, to answer to the former question, please note that the GDPR defines a "Data Subject" as an identified or identifiable natural person.[74] If the data has been Anonymized effectively, meaning individuals cannot be re-identified directly or indirectly, the GDPR does not consider it Personal Data anymore. Therefore, in such a case, the individuals represented in the dataset would not be considered "Data Subjects" under GDPR. However, if a poor anonymization procedure is applied, there is a reasonable chance of re-identification, especially with additional datasets or technical abilities, thus, these individuals might still fall under the GDPR's definition of Data Subjects.

Furthermore, to answer the first question, it is important to note that within this complex landscape remains ongoing scrutiny to define a concrete GDPR-based threshold for re-identification. Even Data Processing Agreements (“DPAs”), mandated by Article 28(3) of the GDPR and designed to detail the responsibilities and obligations of Controllers and Processors in managing Personal Data, cannot fully cover this particular scenario. In practice, DPAs can outline responsibilities and obligations for handling Personal Data, including provisions for re-identification, in several ways.[75] The DPA should clearly define data handling and security measures,[76] which, as a good practice, may include detailed procedures for processing De-identified or Anonymized Data to reduce re-identification risks. Additionally, for example, when two entities have a service agreement and share between them Anonymized Data, this relationship will not be regulated under the GDPR. However, it is a good practice to regulate this sharing and data processing in the master service agreement, thus, it could be added a clause which explicitly prohibits the re-identification of De-identified or Anonymized Data, unless

[74] Article 4 (1), GDPR.
[75] European Data Protection Supervisor, ‘Checklist 3: What Is Required in a Processing Agreement?’ (*Checklists and flowcharts on Data Protection*, 27 September 2019) <https://edps.europa.eu/data-protection/our-work/publications/factsheets/checklists-and-flowcharts-data-protection> accessed 26 November 2023.
[76] Ibid.

necessary and lawful under GDPR.[77] Nevertheless, this is not a universally adopted solution, and it is still challenging to establish a clear framework for safeguarding individuals' rights in a collaborative data processing environment.

Therefore, the question of how to guarantee the continued protection of Anonymized and De-identified data as it enters new processing contexts remains an open discussion in the Data Protection field. In the upcoming sub-subsection, we will explore a legal case that addresses the risk associated with de-identification.

1.3.1.1. Practical Case: SRB vs EDPS

In this section we will delve deeper into to case Single Resolution Board vs. European Data Protection Supervisor (Case T-557/20) which was mentioned previously.[78] This significant ruling instigated discussions around the risk of re-identification and the critical need to adopt a multifaceted approach when assessing the identifiability of the Data Subjects, further elucidating the practical and theoretical challenges it presents in the context of GDPR compliance.

This case revolves around a dispute between the European Single Resolution Board ("SRB") and the European Data Protection Supervisor ("EDPS") concerning the processing of Personal Data during the SRB's decision-making process related to a Spanish bank the SRB had under resolution.[79] In short, SRB engaged Deloitte and, within the context of this contractual arrangement, transferred datasets containing comments to questions the SRB posed to shareholders and creditors.[80] However, the transferred dataset did not reveal their identities, instead, it enclosed an alphanumeric code.[81] The EDPS claimed that the datasets contained Personal Data and found that SRB had infringed the GDPR by not informing shareholders and creditors that their Personal Data (which was Pseudonymized) would be shared

[77] This insight is based on my experience in the field of Data Protection consulting.
[78] *Single Resolution Board v European Data Protection Supervisor* (n 69).
[79] *Baumgartner U, 'New Options for Anonymization Ahead?' (IAPP The Privacy Advisor, 18 May 2023) <https://iapp.org/news/a/new-options-for-Anonymization-ahead/> accessed 11 December 2023.*
[80] Ibid.
[81] Ibid.

with Deloitte.[82] SRB claimed that the datasets shared with Deloitte did not constitute Personal Data and solicited the Court of Justice of the European Union ("General Court") to annul EDPS's decision.[83]

The General Court decided by a relative approach to identifiability, determining that it is possible that a piece of information can be qualified as Personal Data for someone who is able to identify the Data Subject, whereas the same piece of information can be Anonymous for someone without such ability.[84] Furthermore, the General Court stated that the burden of proof that Personal Data is being processed, and thus that the GDPR applies, lies with the supervisory authorities.[85] Due to the fact that the EDPS failed to examine whether the authors of the comments were reidentifiable for Deloitte and whether such reidentification was reasonably possible, the General Court's ruling annulled EDPS's decision.[86] Ultimately, unsatisfied with the outcome of the case, the EDPS has filed an appeal against the decision of the General Court rendered on April 26, 2023, under Case C-413/23 P.[87] Nevertheless, the General Court did not elaborate on the threshold for identifiability, and when it can be regarded as practically impossible, leading this threshold to be determined in each individual case.[88]

To sum up, the case under analysis marks a significant turning point in the interpretation and application of the GDPR, particularly concerning the concept of Personal Data and the identifiability standard.[89] This case arises from a complex debate over data processing, Anonymization and Pseudonymization techniques, revealing the urgent need for a flexible and context-aware understanding of what constitutes Personal Data under GDPR, especially considering the rapid advancement of technology and the distinct roles of Data Controllers and Processors.[90] Consequently, in the forthcoming subsection, we delve into the

[82] Ibid.
[83] *Single Resolution Board v European Data Protection Supervisor* (n 69), 81.
[84] *Baumgartner* (n 79).
[85] *Baumgartner* (n 79).
[86] *Baumgartner* (n 79).
[87] *Baumgartner* (n 79).
[88] *Baumgartner* (n 79).
[89] *Single Resolution Board v European Data Protection Supervisor* (n 69), 81.
[90] *Single Resolution Board v European Data Protection Supervisor* (n 69), 81.

differentiation between Anonymized and Synthetic Data, as well as the examination of various theories of re-identification. This exploration aims to provide a deeper understanding of these ambiguous and challenging areas of Data Protection.

1.3.2. The Distinction Between Anonymized and Synthetic Data: Re-Identification Theories

In the realm of Data Protection, there is a significant debate surrounding the approaches to Anonymization, characterized by stringent and flexible perspectives, as previously explained in subsection 1.3.1. These different views fundamentally disagree on how to handle the risk of re-identification in Anonymized Data. In the present subsection, we provide a comprehensive distinction between Data Synthesis and Anonymization, specifically through the lens of re-identification risk theories.

Firstly, Data Synthesis is a process crucially distinct from Anonymization, – as explained in the previous sections – since it involves generating new data based on patterns and characteristics of existing data, without directly doing a one-to-one correlation with the original input.[91] Data Synthesis, while oriented towards generating Anonymous Data, – data disconnected from any Data Subject – differs notably from Anonymization.[92]

Anonymized Data, which are Personal Data rendered Anonymous, have sparked contrasting views,[93] that inform our understanding of Synthetic Data. On the one hand, proponents of a more pragmatic approach acknowledge data as Anonymized, thus Anonymous, if the risk of re-identification of the Data Subjects is sufficiently low.[94] By applying the line of reasoning of the pragmatic theory to Synthetic Data, we argue that when Data Synthesis reduces the risk of re-identifying individuals to a very low threshold, then, by analogy, Synthetic Data could be regarded as Anonymous. This approach aligns with a more flexible understanding of Data

[91] El Emam and others, *Practical Synthetic Data Generation* (O'Reilly Media, Inc 2020). P9.
[92] Ibid.
[93] Finck and Pallas (n 32), 5, 10 and 11.
[94] Finck and Pallas (n 32), 5, 10 and 11.

Protection, placing Synthetic Data outside the scope of the GDPR, if they maintain a low probability of compromising individuals' rights through re-identification. Furthermore, this perspective is aligned with the subjective approach to the identifiability standard of Personal Data defended by the General Court in Case T-557/20, as discussed in the previous sub-section of this research.

On the other hand, advocates of the stringent approach to Anonymization argue for the absolute elimination of re-identification risk, without compromise.[95] When this stringent standard is applied to Synthetic Data, it implies that Synthetic Data cannot be considered Anonymous if there is any level of re-identification risk, no matter how low.[96] Hence, one could argue that any detectable potential for re-identifying individuals, regardless of the entity that has such control, would disqualify Synthetic Data from being classified as Anonymous. In such a case, Synthetic Data would be considered Personal Data and fall under the purview of the GDPR.

Indeed, the emergence of Data Synthesis introduces a significant new aspect to the ongoing debate concerning the re-identification risks of Anonymized Data. This development challenges the established classifications of Personal and Non-Personal Data, prompting a reevaluation of how we categorize data as Anonymous. It brings to the forefront critical questions about the suitability and efficiency of these classifications in Data Protection, demonstrating the need for adaptive and nuanced approaches in data categorization. Chapter 1 sets the stage for an in-depth exploration of Data Synthesis in Chapter 2, where we examine the Data Synthesis procedure, characteristics, and applications of Synthetic Data in detail, revealing its emerging role in our society.

[95] Finck and Pallas (n 32), 5, 10 and 11.
[96] Finck and Pallas (n 32), 5, 10 and 11.

2. Data Synthesis

2.1. Technological background

Data are commonly collected from diverse sources in the physical world, encompassing a wide range of information. For this study, data obtained from real-world sources are referred to as "Real Data". When this data pertains to an individual and enables their identification, Real Data are categorized as Personal Data, as explained in Chapter 1.

The increasing demand for data in today's digital society has sparked a widespread interest in PETs, as a bridge that facilitates responsible data sharing among Data Controllers and Processors.[97] According to the Information Commissioner's Office (henceforth, "ICO"), "PETs are technologies that embody fundamental Data Protection principles by maximizing Personal Data use, maximizing data security, and empowering individuals."[98] These technologies have become a focal point for government institutions, businesses, and the general public, as they seek to explore the potential benefits and limitations they offer. [99]

Among the PETs gaining significant attention is Data Synthesis, which has emerged as a promising solution to address Data Protection concerns while enabling valuable insights to be extracted from datasets.[100] Within this framework, Data Synthesis unfolds as a transformative and innovative process. It enables the generation of Synthetic Data that accurately reflect the characteristics of Real Data without necessitating the direct manipulation or processing of personal information. This technique serves as a pivotal approach, allowing for an intricate balance between data utility and privacy, ensuring the preservation of essential attributes and patterns inherent to the Real Data while safeguarding individuals' Data Protection rights.[101]

[97] Office of the Privacy Commissioner of Canada, 'Privacy Tech-Know Blog: When What Is Old Is New Again – The Reality of Synthetic Data' (12 October 2022) <https://www.priv.gc.ca/en/blog/20221012/?id=7777-6-493564> accessed 11 June 2023.
[98] Information Commissioner's Office, 'Chapter 5: Privacy-Enhancing Technologies (PETs)' (2022) 3 <https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf> accessed 1 November 2023.
[99] Ibid.
[100] Office of the Privacy Commissioner of Canada (n 96).
[101] Information Commissioner's Office (n 97), 35.

Data Synthesis, originally an established De-identification technique dating back to the 1980s, has undergone a remarkable transformation in its functionality and application.[102] This significant evolution can be attributed to advancements in AIML projects,[103] which have enhanced the data processing and analytics capabilities of Synthetic Data.[104] These cutting-edge advancements have led Synthetic Data into a revolutionary phase, expanding their potential and establishing them as invaluable assets across numerous fields. This innovative approach to data is transforming how organizations operate, offering unprecedented opportunities for analysis without compromising individuals' rights.

In the past five years, Data Synthesis has evolved into a sophisticated tool that addresses privacy and accuracy challenges in data-intensive environments.[105] Recognizing its potential, in 2020, Gartner advised organizations to incorporate Synthetic Data into their overall data strategy and to explore its applications, as Synthetic Datasets are mostly perceived to be both scalable and privacy-compliant.[106]

Both commercial enterprises and governmental institutions are harnessing the transformative potential of Synthetic Data, employing it to advance research, optimize services, and enhance decision-making processes. Such influence of Synthetic Data is evident across diverse fields, from healthcare to finance, offering a myriad of applications while adhering to Data Protection standards. Two notable examples illustrating this are Synthea and Capital One. Synthea, an open-source project, generates synthetic patient data, allowing extensive research in healthcare while preserving patient confidentiality.[107] It is a noteworthy exemplar of how Synthetic Data can emulate realistic patient information for diverse studies and

[102] Liew CK and others, 'A Data Distortion by Probability Distribution' (1985) 10 ACM Transactions on Database Systems 395 <https://dl.acm.org/doi/10.1145/3979.4017> accessed 20 September 2023.
[103] We adopt the definition used by computer scientists. We define an AIML project quite broadly to include projects run in various industries, for example, the development of software applications that have AIML components. See El Emam and others, (n 91).
[104] El Emam, *Accelerating AI with Synthetic Data* (O'Reilly Media, Inc 2020), 9.
[105] Office of the Privacy Commissioner of Canada (n 96).
[106] Judah and others, 'Predicts 2021: Data and Analytics Strategies to Govern, Scale and Transform Digital Business' (*Gartner Research*, 2 December 2020) <https://www.gartner.com/en/documents/3993855> accessed 2 December 2023.
[107] SyntheticMass, 'Synthea' <https://synthea.mitre.org/about> accessed 22 September 2023.

analyses, enabling advancements in medical research and healthcare services.[108] Similarly, a finance titan, Capital One, has harnessed Synthetic Data to craft and validate new products and services.[109] Capital One is exploiting Synthetic Data to refine its financial models and ensure the security of customer data, achieving a balance between innovation and compliance with Data Protection.[110] This deployment of Synthetic Data underscores its crucial role in sharpening financial strategies and safeguarding customer interactions in the finance sector.[111]

As we venture further into this study, we will unravel the extensive ways various sectors are leveraging the strengths of Data Synthesis, showcasing its diverse applications and its impact on European legal norms and procedures. In the next section of this chapter, our objective is to systematically explore the concept of Synthetic Data and delve into its ramifications within the GDPR and its associated principles. We place particular emphasis on three fundamental pillars of the European Data Protection regime: data access, privacy, and data accuracy.

2.2. Data Synthesis: Concept and Synthetic Data Generation

In this section, we delve further into the process of Data Synthesis and explore the concept of Synthetic Data in more detail. According to Dr El Emam, at a conceptual level, Synthetic Data can be defined as "*data that has been generated from Real Data and that has the same statistical properties as the Real Data.*"[112] This definition recognizes the artificial nature of Synthetic Data while retaining the statistical characteristics of the Real Data.

[108] Gonzales and others, 'Synthetic Data in Health Care: A Narrative Review' (2023) 2 PLOS Digital Health <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082> accessed 4 November 2023.
[109] Harris S, 'The Rising Role of Synthetic Data in the Automotive Industry' (*Automotive Testing Technology International*, 5 June 2023) <https://www.automotivetestingtechnologyinternational.com/industry-opinion/the-rising-role-of-synthetic-data-in-the-automotive-industry.html> accessed 22 September 2023.
[110] Harris (n 109).
[111] Harris (n 109).
[112] El Emam and others (n 91), 9.

First and foremost, it is important to define Synthetic Data as artificially generated data that incorporates the correlations and insights from the Real Dataset, while avoiding the direct replication of the Data Subjects' information.[113] Synthetic Data can be generated either based on an existing Real Dataset or through deductions made by the coder.[114] [115] Such inferences could be derived through AI or ML algorithms, or via human analysis, contingent on the variables present within the dataset.[116] In his work "*Practical Synthetic Data Generation*", Dr El Emam presents various methods of generating Synthetic Data.[117] His approach distinguishes in Data Synthesis three distinct types of Synthetic Data, each defined by the degree of their reliance on Real Data.[118]

Our decision to focus on this classification by Dr El Emam, thus distinguishing three types of Synthetic Data, is driven by two main factors. The first is that this classification provides a comprehensive view of the Synthetic Data generation spectrum, from real-data-based to entirely theoretical models. By examining each type of Synthetic Data, we aim to offer a holistic understanding of the capabilities and limitations of Synthetic Data as a PET. Secondly, we have selected Dr El Emam's classification due to its widespread recognition and adoption in data science. Although alternative classifications exist, this framework is recognized for its practical applicability and clarity, making it a valuable reference for both academic and industrial applications.

To conclude, Dr El Emam's classification due provides a structured foundation to explore how each type of Synthetic Data interacts with legal considerations under GDPR, ensuring a thorough and coherent analysis aligned with the thesis's overarching objectives. In the following sections, we discuss each type of Synthetic Data in detail. This exploration illuminates the unique characteristics and

[113] El Emam and others (n 91), 9.
[114] Rubin D, 'Statistical Disclosure Limitation', vol 9 (2nd edn, J OFF STAT 1993), 461-462 <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf> accessed 1 November 2023.
[115] Gal M and Lynskey O, 'Synthetic Data: Legal Implications of the Data-Generation Revolution' (10 April 2023), 7 <https://papers.ssrn.com/abstract=4414385> accessed 21 September 2023
[116] Ibid.
[117] El Emam and others (n 91), 9.
[118] El Emam and others (n 91), 9.

methodologies of Data Synthesis and its practical applications and potential impacts on Data Protection.

2.2.1. Data Synthesis from Real Data

When Synthetic Data is created using Real Data as their source material, it means that the initial dataset used for the Data Synthesis process is obtained from real-world sources.[119] Please note that this source data can encompass various types of data, including Personal Data about individuals or Non-Personal Data.

The synthesis procedure begins by using a dataset, and constructing and training a model, in order to capture the unique patterns and structural attributes intrinsic to the source dataset. [120] Various techniques can be used to train the model/generator.[121] After fitting the model, it can be applied to generate the Synthetic Dataset.[122] Therefore, the trained model encapsulates and reproduces the essential statistical properties of the Real Data to a new dataset while retaining none of the personal identifiers – see Figure 2.

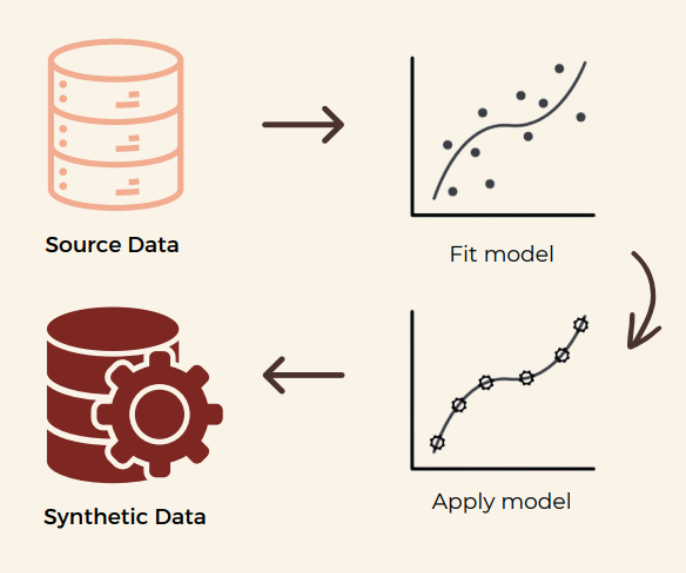


Figure 2- Data Synthesis

[119] El Emam and others (n 91), 9-10.
[120] El Emam and others (n 91), 9-10.
[121] El Emam and others (n 104), 15.
[122] El Emam and others (n 91), 9-10.

Advanced Synthetic Data generation techniques have the potential to address privacy concerns, mitigate biases, and rectify imbalances within datasets, ultimately enabling the creation of more precise scenarios and enhancing the quality of simulations and analyses.[123] Determining the most suitable approach for creating Synthetic Data is difficult, since there is not a one-size-fits-all method.[124] [125] The selection of the most suitable Data Synthesis method depends on various factors, including the intended use case (e.g., data-sharing, privacy preservation, or ML development), data characteristics (e.g., data volume, structure, type - structured, unstructured, tabular, time-series), the planned application (e.g., supervised or unsupervised learning), and any domain-specific constraints (e.g., telecommunications, healthcare, finance, retail).[126]

There are various models capable of generating Synthetic Data, nonetheless, Generative Adversarial Networks ("GANs") are among the most popular.[127] In the upcoming subsection, we provide a more detailed explanation of the GAN model and delve into the procedure employed to train this model. In the following section, we break down and examine each key step involved in the data synthesis process.

[123] Riemann R, 'Synthetic Data' (*European Data Protection Supervisor*, 20 October 2023) <https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data> accessed 22 October 2023.
[124] YData, 'Synthetic Data: The Future Standard for Data Science Development' (2 April 2020) <https://ydata.ai/resources/synthetic-data-the-future-standard-for-data-science-development> accessed 22 October 2023.
[125] YData, '10 Most Frequently Asked Questions about Synthetic Data' (21 March 2023) <https://ydata.ai/resources/10-most-frequently-asked-questions-about-synthetic-data> accessed 22 October 2023.
[126] Ibid.
[127] **"**A generative adversarial network, or GAN, is a deep neural network framework which is able to learn from a set of training data and generate new data with the same characteristics as the training data. For example, a generative adversarial network trained on photographs of human faces can generate realistic-looking faces which are entirely fictitious. Generative adversarial networks consist of two neural networks, the generator and the discriminator, which compete against each other. The generator is trained to produce fake data, and the discriminator is trained to distinguish the generator's fake data from real examples. If the generator produces fake data that the discriminator can easily recognize as implausible, such as an image that is clearly not a face, the generator is penalized. Over time, the generator learns to generate more plausible examples." Wood T, 'Generative Adversarial Network' (DeepAI, 22 July 2020) <https://deepai.org/machine-learning-glossary-and-terms/generative-adversarial-network> accessed 22 October 2023.

2.2.1.1. Key Steps in Data Synthesis

Briefly put, the process of generating Synthetic Data through GANs involves 5 main steps – see Figure 2.[128] This process begins with data preparation and computing the metrics of the Real Dataset (step 1).[129] [130] This phase entails calculating a range of statistical metrics for the real dataset, which offers a condensed illustration of the dataset's principal attributes.[131] The data preparation involves cleansing the Real Data to eliminate errors, ensuring uniformity in coding schemes on the dataset, and harmonizing data from multiple sources into a common typology.[132]

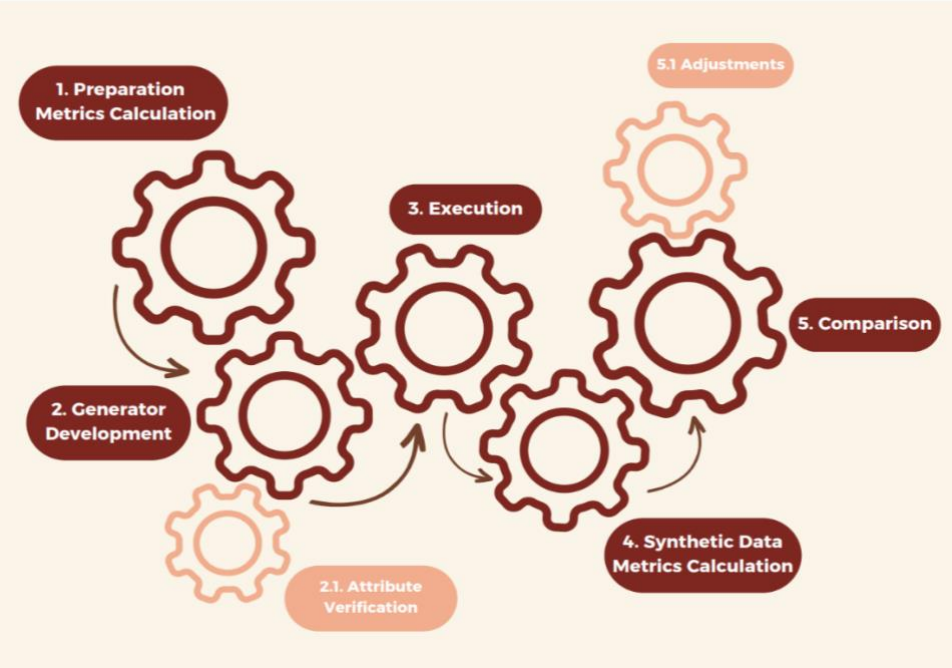


Figure 3 - Data Synthesis Procedure from a Real Dataset

The subsequent phase involves developing a data generator (step 2) to produce Synthetic Data based on manipulations of the Real Dataset, thus, the generator's algorithm is going to evaluate the metrics of the Real Data, followed by setting parameters that will guide the generation process.[133] However, to maintain logical

[128] J.P. Morgan, 'Synthetic Data' (*J.P. Morgan AI Research*) <https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data> accessed 21 September 2023.
[129] Ibid.
[130] El Emam and others (n 104), 12-26.
[131] J.P. Morgan (n 128).
[132] J.P. Morgan (n 128).
[133] El Emam and others (n 104), 12-26.

consistency, it may be necessary to verify certain attributes of the Real Dataset. For example, if we are working with a dataset about voting habits, it would be necessary to ensure there are no instances where individuals under the legal voting age are marked as voters (step 2.1 – this step is optional).[134] Following this, the next stage involves running the previously developed data generator to produce the Synthetic Dataset (step 3).[135] Once the Synthetic Dataset has been created, the subsequent step is to calculate its metrics (step 4), followed by comparing the metrics of the Real Data and the Synthetic Data using an entity referred to as a discriminator (step 5).[136] This component assesses the utility of the Synthetic Dataset by determining if its statistical properties resemble those of the Real Dataset closely.[137]

By computing metrics for both the Real and Synthetic Data, it is possible to statistically compare the two datasets.[138] For example, if the Real Dataset represents an average age of the Data Subjects of 35.5 years, and the Synthetic Dataset reflects an average age of 36 years, this would be a strong indicator that the Synthetic Data is accurately replicating the age distribution in the Real Data. However, if the average age in the Synthetic Data were 50 years, this could suggest that the Synthetic Data is not accurately mirroring the age distribution in the Real Data – see Annex II for practical examples of comparison between synthetic data and real data variables. This depends on the similarity factor that the developer requires for the study.[139]

In opposition, if the comparison reveals that the substance of the Synthetic Dataset greatly diverges from the Real Dataset, adjustments should be made to the generation parameters and a new dataset of Synthetic Data must be produced.[140] This iterative process should continue until Synthetic Data satisfactorily mirrors the

[134] J.P. Morgan (n 128).
[135] J.P. Morgan (n 128).
[136] J.P. Morgan (n 128).
[137] J.P. Morgan (n 128).
[138] El Emam and others (n 104), 17-26.
[139] El Emam and others (n 104), 17-26.
[140] El Emam and others (n 104), 17-26.

Real Data (step 5.1. - optional).[141] In the next session, we explore the synthesis procedure without the use of Real Data.

2.2.2. Synthesis Without the Use of Real Data

The second category of Synthetic Data pertains to a Synthetic Dataset not directly derived from a Real Dataset. Instead, it is produced by a simulation model, drawing upon pre-existing models or on the analyst's background knowledge – see Figure 4.[142]

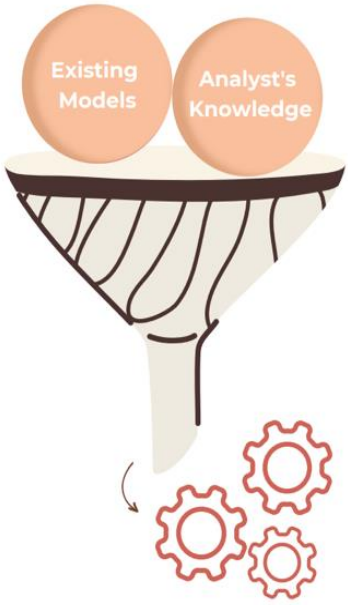


Figure 4 - Data Synthesis Procedure Without Real Data

On the one hand, existing models can be statistical models of an activity developed through surveys or other data collection mechanisms, or they can be simulations, such as simulation systems that produce consumer data characterized by specific attributes, like age and gender.[143] On the other hand, background knowledge might encompass understanding derived from textbook explanations of financial market

[141] El Emam and others (n 104), 12.
[142] El Emam and others (n 91), 11.
[143] El Emam and others (n 91), 11.

behaviors or from observing historical stock price fluctuations under diverse conditions.[144]

Therefore, Synthetic Data generated in this manner is rooted in the coder's assumptions concerning the statistical characteristics of pertinent data attributes, nonetheless, following a predetermined set of rules that define the relationships between significant features.[145] Hence, when the analyst's comprehension of the process is complete and precise, Synthetic Data will align closely with the Real Data.[146]

Furthermore, there are circumstances where Real Data may not exist. This could occur when the analyst is aiming to model an entirely novel concept, or when generating or collecting a fresh dataset may be financially unviable or logistically challenging.[147] Note that this method of Synthetic Data generation is largely influenced by the complexity of the intended dataset usage.[148] Increasingly, these simulators are becoming the primary tools for training and testing Machine Learning algorithms due to their adaptability and efficiency.[149]

Consider, for example, a case where a simulator operates independently, not requiring the usage of any Real Dataset. The simulator develops a dataset to train an ML model for testing an algorithm's capacity to predict prime numbers within a certain range. A Synthetic Data generator, which programmatically creates numbers based on the mathematical rules governing primes, could serve this purpose effectively.[150] Despite not requiring any Real Data, such a simulation has the potential to provide a robust platform for testing and validating a prime predicting model.[151] Otherwise, even in scenarios where simulations do not rely on pre-

[144] El Emam and others (n 91), 11.
[145] Gal M and Lynskey (n 115).
[146] El Emam and others (n 91), 11.
[147] El Emam and others (n 104), 8.
[148] Gal M and Lynskey (n 115), 11.
[149] Subramanyam J and Ramos L, 'Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI' (*Gartner Research*, 24 June 2021) <https://www.gartner.com/en/documents/4002912> accessed 21 September 2023.
[150] Gal M and Lynskey (n 115), 11.
[151] Gal M and Lynskey (n 115), 11.

existing Real Datasets, the simulator exhibits the remarkable capacity to generate novel information based on the distributions and correlations of the variables intrinsic to the process.[152]

2.2.2.1. Practical case: AlphaGo

While the Data Synthesis procedure without real Personal Data may not wholly mirror the properties of Real Data, it retains utility for certain purposes. Such potential becomes palpable when observing ML implementations in domains such as board games. For instance, AlphaGo, an AI system designed to master the game Go, was set with the game's foundational rules and subsequently honed its skills through cycles of self-play simulations.[153]

The algorithm's proficiency was significantly strengthened through iterative cycles of self-play simulations, with reinforcement learning methodologies deployed to enhance its performance incrementally.[154] As part of this process, each iteration of the algorithm effectively created its own Synthetic Dataset of game moves, thereby serving as a learning resource for its subsequent iterations.[155]

Therefore, AlphaGo through its self-learning process, created new strategies for playing Go, thereby generating novel data on the board game.[156] Consequently, not only did AlphaGo excel in the game, but it also devised innovative strategies, proving that simulations can indeed generate Real Data.[157] The following section delves into the significance of the utility parameter in this context.

[152] Gal M and Lynskey (n 115), 11.
[153] Silver D and Hassabis D, 'AlphaGo: Mastering the Ancient Game of Go with Machine Learning' (27 January 2016) <https://blog.research.google/2016/01/alphago-mastering-ancient-game-of-go.html> accessed 21 September 2023.
[154] Ibid.
[155] Rogers A, 'Council Post: What Deep Blue And AlphaGo Can Teach Us About Explainable AI' (*Forbes*) <https://www.forbes.com/sites/forbestechcouncil/2019/05/09/what-deep-blue-and-alphago-can-teach-us-about-explainable-ai/> accessed 21 September 2023.
[156] Hern A, 'AlphaGo: Its Creator on the Computer That Learns by Thinking' *The Guardian* (15 March 2016) <https://www.theguardian.com/technology/2016/mar/15/alphago-what-does-google-advanced-software-go-next> accessed 21 September 2023.
[157] Ibid.

2.3. The Relevance of the Utility Parameter

As we study deeper Synthetic Data, it becomes evident that utility stands as the most important parameter of their relevance and application.[158] This characteristic is not just fundamental but is also thoroughly evaluated during the Synthetization Procedure.[159] While certain scenarios demand data with extremely high utility, there are instances where lower utility may be acceptable.[160]

When processing Real Data, it is essential to understand its implications on both privacy and utility. While the imperatives of privacy dictate that Synthetic Data may not encapsulate all the statistical intricacies inherent in the Real Dataset, this manipulation of data is not without its compromises.[161] It necessitates a balance, forging an inherent trade-off between the safeguarding of data and the practical utility of such data.[162] Therefore, such trade-off is typically quantified by measuring the accuracy of the Synthetic Data in relation to the Real Data. Hence, the higher the degree of privacy preservation incorporated, the more likely the Synthetic Data is to diverge from the statistical relationships present in the Real Dataset. [163]

This balancing test is crucial in scenarios where the preservation of underlying patterns and relationships is paramount for analytical accuracy and reliability. It emphasizes the need for meticulous approaches in Data Synthesis to reconcile the demands of privacy with the imperatives of data utility.[164] For instance, when the goal is to construct AIML models to forecast consumer actions and develop marketing strategies, the necessity for high utility is superior.[165] Conversely, if the aim is to assess the software's capability to manage an extensive volume of transactions, the expectations surrounding data utility would be significantly diminished.[166]

[158] El Emam and others (n 91), 12.
[159] El Emam and others (n 91), 12.
[160] El Emam and others (n 91), 12.
[161] El Emam and others (n 91), 12.
[162] El Emam and others (n 91), 12-13.
[163] El Emam and others (n 91), 12-13.
[164] El Emam and others (n 91), 12-13.
[165] El Emam and others (n 91), 12-13.
[166] El Emam and others (n 91), 12-13.

In Table 1, we offer a comprehensive breakdown by type of Synthetic Data. Each type is classified based on the nature of the Real Dataset from which it is derived, following the qualification system proposed by Dr El Emam.[167] Alongside this, the expected utility of each type is stated. By understanding these distinctions, it is possible to make informed decisions on which kind of Synthetic Data are most suitable for a given purpose.

Type of Synthetic Data	Expected Utility
Generated from real nonpublic datasets	Can be quite high
Generated from real public data	Can be high, although there are limitations because public data tends to be de-identified or aggregated
Generated from an existing model of a process, which can also be represented in a simulation engine	Will depend on the fidelity of the existing generating model
Based on analyst knowledge	Will depend on how well the analyst knows the domain and the complexity of the phenomenon
Generated from generic assumptions not specific to the phenomenon	Will likely be low

Table 1 - Different types of Data Synthesis with their utility implications.

The information delineated in Table 1 lays a basis for the forthcoming categorization of Synthetic Datasets. This first categorization is fundamentally anchored in the variations of the input data and the consequent Synthetic Data that are generated.[168] This typology resonates with the understanding that distinct data types correlate to varied stages in the data processing sequence.[169] Therefore, it is shown a causal nexus, where datasets, even if temporary, reciprocally affect each other. [170] For instance, data initially collected can serve as a partial foundation for fabricating Synthetic Data and, subsequently, this Synthetic Data can be used in

[167] El Emam and others (n 91), 12-13.
[168] Benthall S, 'Situated Information Flow Theory', Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security (Association for Computing Machinery 2019) <https://doi.org/10.1145/3314058.3314066> accessed 25 September 2023.
[169] Ibid.
[170] Gal M and Lynskey (n 115), 11.

simulations, thereby generating a new dataset, with a distinct variant of Synthetic Data – see Figure 5. [171]

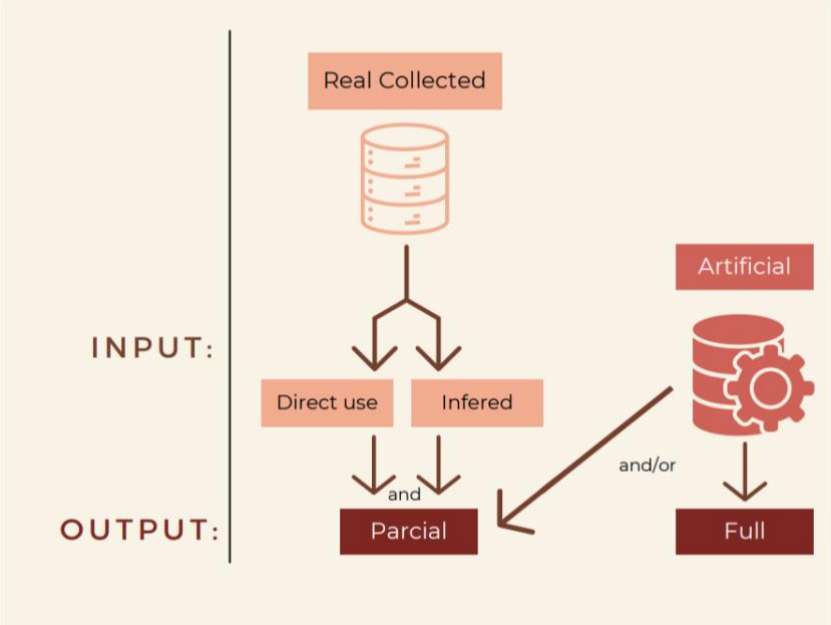


Figure 5 - Types of Synthetic Datasets[172]

A prime exemplification of this process is evident in the training regimes of autonomous vehicles.[173] [174] In such instances, a concise set of acquired images is deployed either in the Generator or the Discriminator, culminating in the generation of an extensive array of synthetic images. [175] These images then become instrumental in simulating diverse road conditions, essential for the ML training modules. [176]

Therefore, these simulations exemplify the transformative capacities of Data Synthesis, illustrating its consequential impact in evolving technological domains, such as autonomous vehicle development, by offering nuanced insights and enhancing the adaptive learning capabilities of the systems involved. [177] In the

[171] Gal M and Lynskey (n 115), 11.
[172] This Figure was elaborated according to the typology adopted in Gal M and Lynskey (n 115), 12.
[173] El Emam and others (n 91), 32-33.
[174] Andrews G, 'What Is Synthetic Data?' (*NVIDIA Blog*, 8 June 2021) <https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/> accessed 25 September 2023.
[175] Toews R, 'Synthetic Data Is About To Transform Artificial Intelligence' (*Forbes*) <https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/> accessed 1 December 2023.
[176] Andrews G (n 174).
[177] Andrews G (n 174).

upcoming section, we focus more closely on the intricacies of synthetic data and its relationship with the GDPR.

2.4. Data Synthesis and the GDPR

As we conclude this chapter on Data Synthesis, it seems evident from our detailed exploration that the process of generating Synthetic Data is very complex and multifaceted. To facilitate the study of this procedure, we have highlighted five essential steps into this process, as presented in Figure 3. We cannot overstate the importance of meticulously completing each of these steps. It is through this comprehensive process that the Synthetic Data obtains their key attributes, particularly their Anonymous nature.

The generation of Synthetic Data goes beyond mere data transformation; it fundamentally erases any connections to the Real Dataset. This aspect is crucial, especially considering the GDPR's definition of Personal Data. We maintain that when Synthetic Data is generated using Real Datasets or incorporates a hybrid approach, the entire Data Synthesis procedure falls within the jurisdiction of the GDPR until its completion. However, the final outputs of this process, the Synthetic Data themselves, are not governed by the GDPR. This distinction arises since the GDPR defines Personal Data as information that can be linked to an identifiable or identified individual and the fact that the Data Synthesis process ensures the elimination of any traceable connections to the original input data. Furthermore, the principle of purpose limitation, stated in Article 5(1)(b), GDPR, is especially pertinent when Synthetic Data is derived from Real Data containing personal information. The Real Dataset must be collected for clear, explicit, and legitimate purposes, and any further processing for Data Synthesis should align with these initial purposes.

In summary, when the input of the procedure is Real Data, even when it is a hybrid input, the generated Synthetic Data do not fit within the conventional boundaries of Personal Data. As a result, processing such data is exempt from the requirements and stipulations of the GDPR. In the next chapter we assess the legal positioning

of Synthetic Data, examining its significant role in reshaping some European Data Protection and ethical considerations.

3. Synthetic Data's Role in Reshaping Data Protection and Ethics

Having delved into the intricacies of Personal Data and gained a comprehensive understanding of Synthetic Data's nature, this chapter embarks on a journey to explore the complex intersection between this cutting-edge technology and the GDPR. Our aim is to unravel, through practical cases, how Synthetic Data aligns with, or diverges from, the principles of the GDPR.

Firstly, Article 8(1) of the Charter of Fundamental Rights of the European Union ("Charter")[178] and Article 16(1) of the Treaty on the Functioning of the European Union ("TFEU")[179] specifically declare the protection of natural persons in relation to the processing of their Personal Data as a fundamental right. This right is further acknowledged in Recital 1 of the GDPR. These pieces of European legislation unequivocally reflect the essence of the European Union' commitment to enhance the protection of Data Subjects' rights when processing Personal Data.[180] Hence, such provisions establish a solid legal foundation for enforcing Data Protection measures and emphasize the inviolability of the Data Subjects' rights.[181]

In this section, we will scrutinize the intricate equilibrium between the instrumental value of data – commonly referred to as "utility", as explored in section 2.3. – and the preservation of the fundamental right to Data Protection, particularly in the context of Data Synthesis. More precisely, we will explore in-depth three Data Protection principles and assess the ramifications of processing Personal Data within the procedure of Data Synthesis.

3.1. The Promise of Synthetic Data to Reduce the Re-identification Risk

Indeed, Data Synthesis emerges as a powerful PET that finds seamless integration within the technical and organizational strategies for processing activities, aligning

[178] Charter of Fundamental Rights of the European Union (2012) OJ C 326/02.
[179] Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union (2012) OJ C 326/01.
[180] Kuner C and others, The EU General Data Protection Regulation (GDPR): A Commentary (Oxford University Press 2020) 48-49.
[181] Ibid.

harmoniously with the provisions of Article 32 of the GDPR. This article focuses on the security of processing activities, requiring that Controllers and Processors implement appropriate technical and organizational measures to ensure a level of security appropriate to the risk of the data processing activities. Since Data Synthesis involves creating artificial datasets that simulate the statistical properties of Real Data without containing Real Data - meaning that Personal Data is not directly used or exposed in data processing activities – processing Synthetic Data significantly reduces the risk of data breaches and unauthorized access. This is crucial in scenarios where data needs to be shared or used for analysis, research, or development purposes.

When Synthetic Data retain substantial utility and accurately reflects the attributes of Real Data, Data Synthesis serves effectively as a proxy for Real Dataset.[182] In practical terms, Synthetic Data works as a tool for addressing real-world challenges while concurrently mitigating privacy risks associated with processing Real Data.[183] Hence, re-identification of the Data Subjects would be possible if Real Data appeared within the Synthetic Dataset. Since Synthetic Data do not directly correspond to Real Data, the risk of re-identification – linking data back to the individual it pertains to – is significantly lowered.[184]

Nevertheless, some argue that the risks of re-identification in De-identified Datasets – discussed in section 1.3.2. – are similar to those for Synthetic Datasets.[185] For instance, if the algorithm replicates the statistical properties of the Real Data too closely or with high accuracy, – meaning it "overfits" the Synthetic Data – the dataset would essentially duplicate Real Data.[186] This can make re-identification more straightforward. However, without overfitting, there is a smaller yet present risk of

[182] El Emam and others (n 91), 15.
[183] El Emam and others (n 91), 15.
[184] El Emam and others, 'Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation' (2020) 22 Journal of Medical Internet Research e23139 <https://www.jmir.org/2020/11/e23139> accessed 5 October 2023.
[185] Griffith I, 'International: Is Synthetic Data the Future of Privacy?' (*DataGuidance*, 27 February 2023) <https://www.dataguidance.com/opinion/international-synthetic-data-future-privacy> accessed 5 December 2023.
[186] Ibid.

inadvertently replicating the real data purely by chance.[187] Nevertheless, we do not agree with the stringent perspective and perceive Synthetic Data as a powerful security measure, as we will explain later.

Recent research on the security of AI models has revealed a new category of re-identification or reconstruction attacks, such as the "membership inference" and "attribute disclosure" attacks.[188] First, this type of attack involves the attacker scrutinizing the characteristics of Synthetic Data to ascertain if information about a specific individual exists in the original Real Dataset.[189] In certain scenarios, an individual, just for being part of a dataset, can reveal their Personal Data.[190] For instance, if a dataset is exclusive to individuals with a certain medical condition, identifying a Data Subject within it automatically exposes their health information.[191] Therefore, Data Synthesis, while reducing some privacy risks, it does not offer full immunity against such attacks, thereby still presenting a challenge in the context of GDPR's stringent privacy requirements and re-identification theories – examined in section 1.3.2.

On another hand, there is the privacy risk of attribution inference.[192] In attribution inference, an attacker deduces confidential attributes of an individual without necessarily identifying the exact Data Subject or their data.[193] This attack often involves correlating an individual with a group possessing a shared characteristic, either through deterministic or probabilistic means.

Synthetic Data, while presenting a relatively lower risk to privacy, does not offer complete protection against attribute disclosure or the associated danger of Personal Data extraction from published datasets.[194] Despite Data Synthesis' susceptibility to such attacks, it is important to acknowledge that even traditional

[187] Ibid.

[188] Ibid.
[189] Ibid.
[190] Ibid.
[191] Ibid.
[192] El Emam and others (n 184).
[193] Griffith (n 185).
[194] Griffith (n 185).

De-identification or Anonymization techniques fall short in providing an absolute defense against attribute disclosure.[195] Such risk presents a significant encounter in ensuring that relying only on Synthetic Data as a PET ensures a comprehensive Data Protection and security measures. In the next subsection we will delve deeper into this matter and assess Dr El Emam's research on risk disclosure.

3.1.1. Practical case: Identity Disclosure Risk in Synthetic Data

To further illustrate the effectiveness of Synthetic Data as a PET, it is crucial to delve into the recent research by Dr El Emam, which illuminates the capabilities and risks of Synthetic Data as a PET through empirical analysis.[196] This research employed a careful methodology for evaluating the identity disclosure risks of fully Synthetic Data, scrutinizing both identity disclosure and the adversary's potential to extract new information on the Data Subjects when Synthetic Data matched a Real Dataset.[197] Hence, the methodology conducted in this study assessed not only the identity disclosure risks of fully Synthetic Data but also the potential for an adversary to assemble new information when Synthetic Data matched a real person.[198]

Two primary datasets were at the focus of Dr El Emam's research. [199] The first being the Washington State Inpatient Database (SID) for 2007 – a comprehensive dataset encompassing population hospital discharges for the entire year, with a rich set of 206 variables and 644,902 observations. [200] The second dataset entailed the

[195] Griffith (n 185).
[196] El Emam and others (n 184).
[197] We refer to the term “adversary” as a person, group, organization, or government that conducts or has the intent to conduct detrimental activities, as defined by the NIST glossary. Information Technology Laboratory, ‘Adversary - Glossary’ (*COMPUTER SECURITY RESOURCE CENTER*) <https://csrc.nist.gov/glossary/term/adversary> accessed 5 October 2023.
[198] El Emam and others (n 184).
[199] El Emam and others (n 184).
[200] Sweeney L, ‘Matching Known Patients to Health Records in Washington State Data’ (5 June 2013) <https://papers.ssrn.com/abstract=2289850> accessed 10 December 2023.

Canadian COVID-19 case data, compiled with meticulous detail by Esri Canada, featuring 7 variables and encompassing 100,220 records.[201]

Data set	Synthetic Data	Real Data
Washington Hospital	0.0197	0.098
Canadian Covid Data	0.0086	0.034

Table 2 - Privacy risks[202]

Therefore, table 2 underscores the potential of Synthetic Data, quantitatively highlighting the meaningful identity disclosure risks tied to the synthesized Washington State Hospital discharge database (0.0198) and the Canadian COVID-19 cases database (0.0086).[203] By setting them alongside, the risks from their respective Real Datasets, the table not only illustrates the substantial reduction in meaningful identity disclosure risks achieved through Synthetic Data techniques, but also demonstrates the pragmatic application of Synthetic Data as a robust shield in safeguarding privacy whilst retaining data utility for profound analysis and research.[204]

Furthermore, as explained before in Chapter 1 of this dissertation, Anonymization poses a significant challenge by often requiring the removal or alteration of identifiable information, ultimately reducing the analytical value of the data.[205] In contrast, Data Synthesis maintains higher utility of the dataset for meaningful analysis and modeling.[206] Therefore, Synthetic Datasets preserve the analytical richness without compromising on privacy. [207]

Data Synthesis also offers an alternative to Pseudonymization, transcending its limitations. While Pseudonymization focuses on masking real Personal Data with pseudonyms,[208] Data Synthesis pioneers an active approach by generating entirely

[201] Canada E, 'Covid-19 Resources' (17 May 2023) <https://resources-covid19canada.hub.arcgis.com/> accessed 5 October 2023.
[202] El Emam and others (n 184).
[203] El Emam and others (n 184).
[204] El Emam and others (n 184).
[205] Finck and Pallas (n 32), 5, 10 and 11.
[206] El Emam and others (n 91), 12.
[207] El Emam and others (n 91), 15.
[208] Seastrom M (n 50)

new data, echoing the statistical properties and relationships intrinsic to real Personal Data.[209]

Data Synthesis serves as a versatile technical and organizational instrument that finds applications across a spectrum of use cases, ranging from ML model training to comprehensive data analysis. Its adaptability and multifunctionality make it a valuable asset in various domains. To leverage Data Synthesis effectively, it is crucial to recognize its potential as a forward-looking strategy for mitigating risks associated with data processing activities.

Therefore, by harnessing its capabilities, organizations can proactively address GDPR's requirements and cultivate an environment that aligns with the legal and ethical principles underpinning this regulation. In the next section we discuss how Data Synthesis emerges as a pivotal tool for the responsible and innovative use of data, effectively balancing the European principles of Data Protection and the imperatives of privacy in our current data-driven world.

3.2. Enhancement of Data Protection Principles

Taking into consideration the Data Protection principles outlined in the GDPR, in the present part of this research we demonstrate how Synthetic Data not only aligns with these principles but also enhances some of their major features. This discussion immediately directs our attention to three key principles of the GDPR: data security, data minimization and data accuracy.

3.2.1. Data accuracy

The principle of data accuracy, enshrined in Article 5(1)(d) of the GDPR, stands beyond being a regulatory requirement, it is a pivotal element that embodies the trustworthiness and reliability of the Data Subjects in the processing of Personal

[209] Finck and Pallas (n 32), 5, 10 and 11.

Data. This principle mandates Controllers and Processors to maintain the precision of processed data and to immediately rectify inaccuracies when they arise.[210]

However, when Synthetic Data enters the equation, data accuracy takes on a nuanced complexion. Synthetic Data is a simulated, artificial representation that no longer directly corresponds to a Data Subject. This pleads the question: How does one ensure data accuracy in Synthetic Data that is not directly tied to the Data Subjects? We maintain that the answer to this question lies in the commitment to conscientious generation and utilization of Synthetic Data. On the one hand, the developers of Synthetic Data have the vital responsibility to meticulously ensure that the Synthetic Dataset faithfully mirrors the statistical properties, distributions, and relationships present in the input (Real Dataset). [211] Accordingly, this commitment to data accuracy is seamlessly integrated into the synthesis procedure, as detailed in Section 2.3.1. Therefore, once a Synthetic Dataset has been crafted, the subsequent phases of the procedure entail calculating their metrics (step 4) and conducting a comparative analysis with the metrics derived from Real Data (step 5), thereby ensuring their accuracy.[212] [213]

On the other hand, we argue that Synthetic Data provide effects of increased data quality, which encompasses several interrelated yet distinct dimensions, with completeness and accuracy standing out as characteristics. [214] Therefore, completeness guarantees that specific data attributes are not missing or omitted within a dataset, while accuracy ensures that these attributes are faithfully represented without distortion or misrepresentation in the Synthetic Dataset.[215] As a result, Synthetic Data proves to be highly useful in scenarios where Real Data is unavailable or impractical to gather. [216] The following subsection will focus on one practical application of Synthetic Data, specifically in the context of health data.

[210] Kuner C and others (n 180).
[211] Gal M and Lynskey (n 115).
[212] J.P. Morgan (n 128).
[213] El Emam and others (n 104), 12-26.
[214] Gal M and Lynskey (n 115), 46-57.
[215] Gal M and Lynskey (n 115), 46-57.
[216] El Emam and others (n 91), 14.

3.2.1.1. Practical Case: Health Data Accuracy

In this subsection, we explore the critical role of data accuracy in healthcare, particularly through the lens of a practical case study: a clinical trial assessing the efficiency of a new medication. A notable challenge in such trials is often the inaccurate gender distribution of participants. For instance, consider a scenario where most participants of a trial are male. This imbalance poses a significant barrier to understanding the medication's effects on females, leading to a gap in comprehensive healthcare.

To address this disparity, Synthetic Data – constructed to mirror the health profiles and characteristics typical of female participants – can be incorporated into the dataset in analysis. Even though Synthetic Data are not directly sourced from real female participants, they work as a valuable resource to facilitate a more inclusive and balanced study.[217] Therefore, we argue that by including Synthetic Data within the research database, it can be achieved an unbiased and inclusive dataset. Hence, as shown in this practical case, Synthetic Data can be used to develop medical research and ensure that treatments are effective and safe for a broader demographic.[218] The use of Synthetic Data in this context not only enhances the accuracy of decision-making in medical research but also embodies a commitment to equitable healthcare.[219]

To conclude, Data Synthesis emerges as a pragmatic solution that facilitates cost-effective research. Its capacity to generate data, which may be challenging, unfeasible, or ethically questionable to obtain in the real world, underscores its transformative potential. By addressing these barriers, Synthetic Data play a pivotal role in enhancing the principle of data accuracy, reducing bias and fostering more

[217] Goncalves A and others, 'Generation and Evaluation of Synthetic Patient Data' (2020) 20 BMC Medical Research Methodology 108 <https://doi.org/10.1186/s12874-020-00977-1> accessed 4 November 2023.
[218] Gonzales and others (n 108).
[219] Benaim AR and others, 'Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies' (2020) 8 JMIR Medical Informatics <https://medinform.jmir.org/2020/2/e16492> accessed 4 November 2023.

robust research and decision-making. Moving forward, we shift our focus to how Synthetic Data plays a transformative role in the realm of Data Minimization.

3.2.2. Data Minimization

In this section, we explore how Synthetic Data not only complies with but also enhances the Principle of Data Minimization, stated in Article 5(1)(c) of the GDPR. This principle emphasizes the need to collect and process only the minimum amount of Personal Data adequate, relevant, and limited to what is necessary for the specific purposes for which they are processed.[220]

The processing of Synthetic Data by organizations naturally reduces the necessity of retaining datasets containing Personal Data for future processing activities.[221] This technological shift alters the organizations' perspective, moving their focus away from accumulating vast real-world datasets and toward the strategic application of Synthetic Data for well-defined purposes.[222] Moreover, within the context of AIML, the importance of labeled data in model development cannot be overlooked.[223] Real Data may exist, but it frequently exhibits poor quality and low utility due to its unrefined and raw nature.[224] Also, the process of labeling Real Data is often fraught with challenges, demanding considerable time and effort and is prone to errors, as it typically relies on extensive manual work.[225]

In opposition, Data Synthesis effectively bypasses the challenges inherent in labeling Real Data. The generation of Synthetic Data is a controlled and customizable process, allowing for the creation of datasets that are precisely tailored to specific requirements and purposes.[226] This results in Synthetic Data that is inherently clean and labeled from its generation, making them an invaluable asset for training and refining AIML models. This approach eliminates the typical

[220] Kuner C and others (n 180), 317.
[221] El Emam and others (n 91), 14.
[222] El Emam and others (n 91), 14.
[223] El Emam and others (n 91), 14.
[224] El Emam and others (n 91), 14.
[225] El Emam and others (n 91), 14.
[226] El Emam and others (n 91), 14.

drawbacks associated with the collection and labeling of Real Data, offering a more efficient and reliable alternative for model development.

Furthermore, Synthetic Data stands out from De-identification and Anonymization methods. Unlike the latter, which require Personal Data at the start to transform them into a non-identifiable format, Synthetic Data processes may not require Personal Data at all – depending on the context and purposes of the Data Synthesis.

In essence, Synthetic Data can be used as a strategic asset in two key ways. First, by minimizing the necessity for extensive Real Data collection, thereby reducing potential privacy risks associated with it.[227] Second, by fostering efficiency in model development, all while upholding the fundamental GDPR tenet of data minimization.[228] In the following subsection, we delve into a practical example illustrating how Synthetic Data can effectively advance the principle of data minimization.

3.2.2.1. Practical Case: Synthetic Population Database

The stringent privacy laws in the United States and Europe often pose challenges for healthcare research. To illustrate a solution to these challenges, in this subsection, we evaluate a practical case concerning health synthetic data.

In the United States, health data is regulated under the Health Insurance Portability and Accountability Act ("HIPAA").[229] In Europe, health-related information is categorized as a special category of Personal Data, as per Article 9(1) of the GDPR. This classification of health data demands the application of more rigorous protective measures to the processing activities, reflecting the heightened sensitivity and privacy concerns associated with health data.[230] This regulatory environment, while essential for protecting personal information, inadvertently acts

[227] El Emam and others (n 91), 15.
[228] El Emam and others (n 91), 15.
[229] Health Insurance Portability and Accountability Act of 1996, Pub L No 104-191, 110 Stat 1936.
[230] El Emam and others (n 104), 36-42.

as a barrier to the advancement of research in healthcare. This tension between Data Protection and research progress underscores the need for innovative solutions that respect privacy laws while facilitating essential healthcare studies.

Furthermore, the financial burden associated with the collection of health data complicates the process.[231] For example, the collection of data from several locations in clinical trials is costly.[232] Picture a team of public health researchers aiming to assess the impact of a new public health intervention across different demographics and locations within the United States. Traditionally, this would involve collecting detailed household and individual-level data from surveys or health records, which poses significant privacy concerns and logistical challenges. However, with the US Synthetic Household Population database at their disposal, researchers can avoid these challenges.[233] This database provides them with Synthetic Data that represent the sociodemographic attributes of the U.S. population, both at the individual and household levels.[234] Since the database statistically mirrors the actual population and includes accurate geographical information, it serves as a robust foundation for various simulations.[235]

Therefore, researchers can use this Synthetic Database to conduct microsimulations to understand how different public health strategies might play out across various population segments.[236] Also, this database can be used to evaluate public health interventions by analyzing their likely effectiveness across different demographic groups without the need for actual personal health data.[237] Since Synthetic Data statistically matches the real population without corresponding to any real individuals' personal data, researchers can perform comprehensive analyses with minimal privacy risks.[238]

[231] El Emam and others (n 104), 36-42.
[232] El Emam and others (n 104), 36-42.
[233] Wheaton W and Rineer (n 233)
[234] 'RTI U.S. Synthetic Household Population™' (*RTI International*) <https://www.rti.org/impact/rti-us-synthetic-household-population%E2%84%A2> accessed 10 December 2023.
[235] Ibid.
[236] Gonzales and others (n 108).
[237] Gonzales and others (n 108).
[238] El Emam and others (n 104), 36-42.

Therefore, it can be concluded that by using a Synthetic Dataset, researchers can bypass the need to collect sensitive Personal Data, thus adhering closely to the Principle of Minimization. Furthermore, this approach not only respects individuals' privacy but also allows for the broad application of the research findings, enhancing their utility for policymakers and public health officials.[239] In the upcoming section, we turn our focus to how Data Synthesis significantly strengthens the principle of data security.

3.2.3. Data Security

Building on the discussions in section 3.1, Data Synthesis, in its role as a PET, effectively meets the stipulations of Article 32 of the GDPR. This key article underscores the European obligation of Data Security, requiring Controllers to implement technical and organizational measures proportional to the risks involved in their processing activities.[240]

Data Synthesis, by generating datasets that resemble Real Data, while erasing real Personal Data, substantially decreases the likelihood of Personal Data exposure.[241] This reduction of risk is crucial, particularly in safeguarding against unauthorized access, potential data breaches and re-identification of Data Subjects.[242] Furthermore, Synthetic Data allows for the sharing of valuable data insights with high utility without the risk associated with transferring Real Data.[243] For example, by employing Data Synthesis on AIML models instead of Real Data, – even if De-identified or Anonymized – it can significantly elevate the reliability and performance of the models.[244] Additionally, in scenarios where data sharing is essential to achieve the purpose of the processing, such as in collaborative research or cross-organizational projects, Synthetic Data provides a more secure alternative.[245]

[239] Gonzales and others (n 108).
[240] Kuner C and others (n 180), 630.
[241] El Emam and others (n 104), 36-42.
[242] El Emam and others (n 104), 36-42.
[243] El Emam and others (n 104), 36-42.
[244] El Emam and others (n 104), 36-42.
[245] El Emam and others (n 104), 36-42.

To conclude, by employing Data Synthesis, Controllers can generate representative datasets that enable them to better develop technologies and research, due to the high utility of Synthetic Datasets, while safeguarding Data Subjects' rights.[246] In the upcoming subsection, we delve into a real-world application of Synthetic Data, particularly in enhancing data security within the realm of software testing.

3.2.3.1. Practical Case: Software Testing

In this section, we focus on a practical example of Synthetic Data usage in software testing and development, specifically in the financial services sector. Synthetic Data serves as a safe, efficient, and versatile alternative to Real Data for software developers.[247]

For instance, consider using Synthetic Data to test a fraud detection software within a financial institution.[248] This data, while not authentic, is designed to mirror Real Data closely, allowing the software to be tested under countless scenarios without processing Real Data.[249] In this context, the primary advantage of using Synthetic Data is the security it provides, as mentioned before.[250] By employing Synthetic Data, during the early stages of the model's development, Real Data remains protected, to be used later once the software is proven secure.[251] Using Synthetic Data during the development phase also speeds up the software refinement process and lessens computational demands, due to the high quality of the labeled Synthetic Data, as explained before for the principle of data minimization.[252] However, to respect the accuracy principle and to effectively evaluate the software's performance in the financial services sector, the Synthetic Dataset should encompass edge cases or unusual scenarios, like spikes in trading volumes due to

[246] El Emam and others (n 104), 36-42.
[247] El Emam and others (n 91), 15.
[248] El Emam and others (n 91), 15.
[249] El Emam and others (n 91), 15.
[250] El Emam and others (n 91), 15.
[251] El Emam and others (n 91), 15.
[252] El Emam and others (n 91), 15.

external events.[253] These outlines ensure the model's robustness in unlikely circumstances.

In conclusion, processing Synthetic Data significantly bolsters Data Protection and diminishes the risk of data breaches, thereby lowering the chances of compromising the confidentiality, integrity, or availability of Personal Data. This method underscores the pivotal role of Synthetic Data allowing research advancements, while simultaneously enhancing the principle of data security. In this chapter, we have thoroughly examined the significant Data Protection risks associated with Data Synthesis in the context of the GDPR. Moving forward, the next chapter will shift our perspective to these practical cases, analyzing them through an ethical lens, providing a deeper understanding of the moral implications and responsibilities entailed in the use of Synthetic Data.

[253] El Emam and others (n 91), 15.

4. Unveiling the Challenges and Risks in Research using Synthetic Data

Synthetic Data, as an innovative approach, offers a way to tackle the challenges posed by limited data availability and privacy concerns. However, it also introduces some ethical dilemmas. Expanding on the legal insights presented in the previous chapter, this section aims to shed light on the ethical considerations surrounding Synthetic Data. In this segment of the research, we revisit the case studies discussed in the previous chapter, nevertheless, through an ethical lens. In this part of the discussion, we emphasize the importance of transparent and ethical decision-making processes when processing Personal Data. This includes ensuring clear communication about how Synthetic Data is handled and processed, as well as maintaining ethical standards to preserve public trust and uphold the principles of truth and integrity.

4.1. Bias and Loss of Public Trust

This segment explores significant challenges associated with the use of Synthetic Data in social contexts. It highlights their beneficial side in enhancing reproducibility and diversity in research and AIML system development, while also acknowledging their potential to reduce biases in data. Ethical challenges are explored, particularly in the context of synthetic media and deepfakes, underscoring the risks of misinformation and societal distrust.

Data Synthesis facilitates early-stage product development and collaboration in research, ensuring privacy, reproducibility and diversity within the input and, consequently, the output of the study.[254] By focusing on practical data interaction, structural biases can be addressed, and move towards responsible, equitable inputs which instigate a better development of AIML systems.[255] Therefore, Synthetic Data

[254] Rodriguez L and Howe B, 'In Defense of Synthetic Data' (arXiv, 3 May 2019) <http://arxiv.org/abs/1905.01351> accessed 1 December 2023.
[255] Ibid.

has the potential to protect disadvantaged groups from harmful bias present in input datasets when the Real Data reflect discrimination.[256]

During the Data Synthesis procedure, in particular, during the models' development phase, developers should engage with domain experts, who are deeply familiar with the relevant issues related to the purpose of the Data Synthesis, and stakeholders representing the populations in the data.[257] Experts can provide in-depth insights into the data's source, the methods used in data collection and identify any potential biases or shortcomings within the input dataset.[258] This comprehensive understanding is crucial for validating the authenticity and applicability of the Synthetic Data.[259] Therefore, expert determination plays a key role in assuring the accuracy and trustworthiness of Synthetic Data.[260] We argue that to ensure transparency regarding the decision-making during Data Synthesis, especially concerning bias mitigation measures applied, it is required clear communication to the final handlers of the Synthetic Data concerning the procedures applied to the data, ensuring users are fully aware of the dataset's nature and limitations.

The intensification of the use of Synthetic Data has brought forth a compelling challenge: the compilation of valid and invalid information within the same dataset.[261] The loss of public trust in the Synthetic Datasets, is mentioned by Dr Rune Klingenberg from the Danish National Center for Ethics, as a pivotal concern among the usage of Synthetic Data.[262] While Synthetic Data opens up new avenues, as discussed before, they can also enable the spread of misinformation, casting uncertainty on the trustworthiness of digital content.[263] Take synthetic media as a prime example, which is a subset of Synthetic Data, focusing specifically

[256] Ibid.
[257] Ibid.
[258] Ibid
[259] Ibid.
[260] J. Kleczyk E, 'Importance of Patient Privacy in Healthcare Analytics Research', *Ethics - Scientific Research, Ethical Issues, Artificial Intelligence and Education [Working Title]* (IntechOpen 2023) <https://www.intechopen.com/online-first/1120485> accessed 3 December 2023.
[261] Hansen R, 'AI Image Generator: This Is Someone Thinking About Data Ethics · Dataetisk Tænkehandletank' (*Dataetisk Tænkehandletank*, 15 August 2022) <https://dataethics.eu/ai-image-generator-this-is-someone-thinking-about-data-ethics/> accessed 7 October 2023.
[262] Ibid.
[263] Ibid.

on media content created using AI techniques.[264] Its primary function is to replicate real-world content, like images, videos, or audio.[265] Deepfakes are a notable instance, where this media alters images and videos to falsely show individuals saying or doing things they have not actually said or done.[266] Therefore, we argue that a paramount ethical issue regarding synthetic media, following violations of privacy and consent, is its capacity for misrepresentation.[267] [268] This can result in misleading perceptions and harm individuals' reputations.[269] [270] In short, the existence of deepfakes can lead to a broader societal distrust in media, which can erode social cohesion and amplify skepticism towards real information.[271]

To ensure societal welfare, it is imperative to approach Synthetic Data generation and usage with ethical diligence, ensuring that technological advancements do not unduly compromise public faith and that they uphold a principled adherence to truth, accuracy, and ethical integrity.[272] When Synthetic Data mirrors real individuals too closely, they might inadvertently compromise privacy, or enable malicious entities to manipulate public perception and behavior.[273] Therefore, Synthetic Data might inadvertently lead to cases of mistaken identity.[274] When creating a synthetic persona, there is also the possibility that the synthetic individual could be mistaken for the individual within the Real Dataset used as input for the Data Synthesis.[275] In the following section, we re-examine the case studies from the previous chapter, but this time with a focus on ethical considerations.

---

[264] Lancaster University, 'Lancashire Cyber Foundry an Introduction to Deepfakes' (*Digital Resources*) <https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/cyber-foundry/lcf-articles/LCFArticle-Josh-Deepfakes_WEB.pdf> accessed 3 December 2023.
[265] Ibid.
[266] Ibid.
[267] Sensity Team, 'How to Detect AI Generated Images with Sensity in 2023' (10 May 2023) <https://sensity.ai/blog/deepfake-detection/how-to-detect-ai-generated-im/> accessed 3 December 2023.
[268] Ibid.
[269] Lancaster University (n 264)..
[270] Dolhansky B and others, 'The Deepfake Detection Challenge (DFDC) Preview Dataset' (arXiv, 23 October 2019) <http://arxiv.org/abs/1910.08854> accessed 3 December 2023.
[271] Lancaster University (n 264).
[272] Lancaster University (n 264).
[273] Hansen (n 261).
[274] Hansen (n 261).
[275] Hansen (n 261).

4.2. Evaluating Ethical Dimensions in Synthetic Data Use Cases

4.2.1. Practical Case: Health Data Accuracy

In the realm of clinical trials, the accuracy and integrity of decision-making processes are of utmost importance. Increasingly, AIML models, which require substantial and precise training data, are being utilized in healthcare for various applications such as medical imaging, patient data analytics, and drug discovery.[276] In these contexts, the integration of Synthetic Data alongside real patient data is becoming more and more common. [277] [278] [279] It is mostly seen as a strategic feature, that is improving the performance and reliability of AIML models, leading to more informed and accurate results.[280] [281] [282] However, unlike Real Datasets where the authenticity of every entry can be assured, Synthetic Datasets require a more cautious approach.[283] Errors in Synthetic Data can diminish trust in the Real Data source. [284] Therefore, analysts and researchers must proceed with caution, acknowledging that not every pattern or correlation observed might be grounded in real-world truth.

Consider the scenario presented in section 3.2.1.1., where a clinical trial assessing the effectiveness of a new medication predominantly involves male participants. In such a case, incorporating Synthetic Data that accurately reflects the health profiles and characteristics of female participants can significantly enhance the dataset's inclusivity and accuracy. This approach addresses the gender imbalance in the trial, ensuring that the study's findings are more representative of the broader population.

[276] Pappalardo F and others, 'In Silico Clinical Trials: Concepts and Early Adoptions' (2019) 20 Briefings in Bioinformatics 1699 <https://academic.oup.com/bib/article/20/5/1699/5032454> accessed 29 November 2023.
[277] Dilmegani C, 'Synthetic Data for Healthcare: Benefits & Case Studies in 2023' (*AI Multiple*, 22 December 2022) <https://research.aimultiple.com/synthetic-data-healthcare/> accessed 29 November 2023.
[278] Bamford S, 'Synthetic Data and Privacy - Experiences Implementing Data Synthesis in a Global Life Sciences Company' <https://edps.europa.eu/system/files/2021-06/01_stephen_bamford_en_0.pdf> accessed 3 December 2023.
[279] Ibid.
[280] Ibid.
[281] Pappalardo (n 276).
[282] Dilmegani (n 277).
[283] Hansen (n 261).
[284] Hansen (n 261).

However, the inclusion of Synthetic Data in the trial may also raise certain ethical concerns. While it supports balancing the dataset, relying on Synthetic Data brings into question the authenticity and integrity of the trial's findings. Using artificial data to represent a demographic not adequately included in the real trial data could also lead to potential inaccuracies in understanding how the medication affects that specific group. Hence, balancing these ethical considerations is crucial to maintaining the accuracy and trustworthiness of the clinical research. On the one hand, Synthetic Data can be used to address the underrepresentation of certain groups in clinical trials, solving a major gap that can lead to biased or incomplete medical observations.[285] [286] Hence, Data Synthesis can be used to generate edge cases, rapidly producing high volumes of perfectly labeled synthetic health data. On the other hand, Synthetic Data must be accurate enough to ensure that decisions based on them are reliable and reflective of real-world scenarios. Accordingly, if Synthetic Data inaccurately represent the Real Data, they could lead to erroneous conclusions, with dangerous consequences for real patients.[287] [288]

A good example of the limitations of Synthetic Data is the partial synthesis of survey data collected by the Cancer Care Outcomes Research and Surveillance (CanCORS) project.[289] In this case, upon evaluating the Synthetic Data, which was developed using the project's model, researchers concluded that the dataset was only useful for preliminary data analysis purposes, due to issues regarding data correlations.[290]

Another essential aspect to consider in the use of Synthetic Data in research is their potential limitation in capturing Real Data's outliers. Synthetic Data, primarily designed to mimic real-world datasets, often focuses on reproducing general patterns and trends. This approach can sometimes overlook outliers and the

[285] Dilmegani (n 277).
[286] Pappalardo (n 276).
[287] Dilmegani (n 277).
[288] Pappalardo (n 276).
[289] Loong B and others, 'Disclosure Control Using Partially Synthetic Data for Large-scale Health Surveys, with Applications to CanCORS' (2013) 32 Statistics in Medicine 4139 <https://onlinelibrary.wiley.com/doi/10.1002/sim.5841> accessed 29 November 2023.
[290] Ibid.

significance of these "black swans" cannot be underestimated.[291] Instead of disregarding outliers, we should pay more attention to them, as they can reveal valuable insights about the potential for extreme and unexpected variations within a dataset. In the realm of healthcare and clinical trials, outliers represent rare but pivotal cases that could significantly influence the overall findings and conclusions of the trial. For instance, in a dataset containing patient reactions to a new medication, the typical reaction patterns are certainly important. However, the rare, extreme cases of "black swans" might hold vital insights into unexpected side effects that could have profound implications for patient safety. Therefore, Synthetic Data must be accurate and complete, to avoid inaccurate decision-making.

The previously discussed shortcomings highlight the necessity of rigorously validating Synthetic Data. The validation process should involve experts in the specific field of study to ensure the credibility and applicability of the methods used to generate such data. In short, we stress the critical role of validation in the Data Synthesis process. This essential step is what ensures the Synthetic Data's fidelity to real-world scenarios and confirms its suitability for the intended applications. As of the writing of this thesis, there are no standard method for validating Synthetic Data. We contend that there is an urgent need further explore this issue to establish a universal validation framework for Synthetic Data. This is standard is required to accurately evaluate whether a specific Synthetic Dataset is valid and reliable.

At the moment, the onus of conducting and reviewing Synthetic Data through validation inspections rests on researchers and data users. We contend that to foster trust in Synthetic Data, it is essential to quantifiably assess key qualities within the Data Synthesis Procedure, including utility, fidelity, diversity, accuracy, among

[291] In his book *The Black Swan,* Nassim Nicholas Taleb explores the impact of highly improbable events, which he terms "black swans". These events are rare and unpredictable; however they have significant and widespread consequences. Taleb demonstrates the importance of outliers as data points or occurrences that deviate greatly from the norm. Outliers do not fit the regular patterns or expectations. In the context of complex systems, whether in finance, science, or social dynamics, these outliers are commonly overlooked since they do not happen frequently and are hard to predict. However, Taleb argues that these outlier events can have an intense impact on our understanding and functioning of complex systems. See Taleb, *The Black Swan: The Impact of the Highly Improbable* (Random House 2016).

others.[292] However, a significant difficulty persists in determining how to establish effective validation procedures and thresholds. This accuracy issue arises not only from the difficulty of measuring these qualities but also from the inherent issue of defining them precisely.[293] In the next subsection, we assess the use of Synthetic Population Databases for policy making.

4.2.2. Practical Case: Synthetic Population Database

In this section, we delve into the intricate role of Data Synthesis in shaping modern policy-making and research. We assess the cases studied in section 3.2.2.1., and the UK and EU perspectives on synthetic population data as an invaluable asset for policy analysts and researchers. Considering that Synthetic Datasets raise critical implications for transparency and communication to the Data Subjects,[294] when presenting findings based on Synthetic Data, it becomes crucial to ensure that audiences are made highly aware of this combination of data.[295] The onus of transparency also requires compliance with ethical standards to reduce potential reservations tied to Synthetic Data.[296] Accordingly, in the next subsection, we delve into the UK Statistics Authority and Office of National Statistics considerations on research relying on Synthetic Data.

4.2.2.1. UK Statistics Authority: Ethical Considerations Relating to the Creation and Use of Synthetic Data

The ethical implications surrounding Synthetic Data are not mere academic discussions. Recognizing the gravity of these concerns, bodies like the UK Statistics Authority and the Office for National Statistics ("ONS") have taken proactive steps,

[292] Arthur L and others, 'On the Challenges of Deploying Privacy-Preserving Synthetic Data in the Enterprise' (arXiv, 9 July 2023) <http://arxiv.org/abs/2307.04208> accessed 1 December 2023.
[293] Breugel B and others, 'Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data' <https://arxiv.org/abs/2305.09235> accessed 1 December 2023.
[294] Wheaton W and Rineer (n 233).
[295] Hansen (n 261).
[296] Hansen (n 261).

formulating comprehensive guidelines on Synthetic Data.[297] Such guidelines, including the ONS policy on Synthetic Data, address key legal and ethical issues, such as confidentiality and data disclosure risks, offering an essential framework for responsible data usage in statistical research.[298] These steps are crucial in shaping the ethical handling of Synthetic Data in research and analysis, ensuring compliance with legal standards and reducing potential liabilities.[299] The selection of the UK Statistics Authority's guidelines for Synthetic Data to study within this research is justified by their alignment with the UK GDPR, which closely mirrors the European GDPR, ensuring relevance and applicability in a broader European context.[300] Notably, the EU lacks specific guidelines on Synthetic Data, making the UK's framework particularly pertinent.

Therefore, the UK Statistics Authority provides comprehensive guidance on the ethical considerations related to the creation and use of Synthetic Data.[301] This includes the importance of ethical practices in the realm of Data Synthesis, an overview of ethical considerations and mitigation strategies and an ethics checklist.[302] [303] [304] Additionally, the ONS policy on Synthetic Data outlines key considerations for their use in statistical research, emphasizing confidentiality and reduced disclosure risks.[305] These principles are crucial in the handling of Synthetic

[297] The UK Statistics Authority, as a foremost authority in statistics and data ethics, alongside the Office for National Statistics as the UK's principal provider of statistical data, significantly enhances the importance of these guidelines. The expertise in statistical analysis and data processing activities of these entities ensure these guidelines were develop through comprehensive knowledge and practical experience on the field, thus, being an important element of this research.
[298] Office for National Statistics, 'Synthetic Data Policy - Office for National Statistics' <https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/syntheticdatapolicy/> accessed 10 December 2023.
[299] UK Statistics Authority, 'Ethical Considerations Relating to the Creation and Use of Synthetic Data' (*UK Statistics Authority - Statistics for the Public Good*) <https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/> accessed 27 September 2023.
[300] Data Protection Act 2018, c 12 (incorporating the UK General Data Protection Regulation).
[301] UK Statistics Authority (n 299).
[302] Ibid.
[303] UK Statistics Authority, 'Ethics Self-Assessment Tool' (*UK Statistics Authority - Statistics for the Public Good*) <https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/> accessed 27 September 2023
[304] UK Statistics Authority, 'Ethical Principles' (*UK Statistics Authority - Statistics for the Public Good*) <https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethical-principles/> accessed 27 September 2023.
[305] Office for National Statistics (n 298).

Data, providing an essential ethical framework for researchers and analysts across all jurisdictions.[306] Furthermore, the UK Statistics Authority has developed ethical principles and an ethics self-assessment tool to guide researchers and statisticians in addressing ethical issues in various projects, including those involving Synthetic Data.[307] Such principles emphasize the public good, data confidentiality, risk assessment, legal compliance, public perception and transparency in data collection and usage.[308] Therefore, by consistently incorporating a thoughtful ethical framework into each project, it is possible to address these concerns, ensuring both the integrity of the research and the continued trust of the Data Subjects in the synthesis procedure.[309][310]

The UK Statistics Authority stresses the importance of Synthetic Data in research.[311] The Authority underscores the crucial need to balance utility, which is the data's practical usefulness, and fidelity, their authenticity.[312] This balance is a significant element in assessing the efficiency of Synthetic Data, ensuring they serve their intended purpose while accurately representing the Real Data.[313] In other words, utility represents the efficiency with which Synthetic Data aligns with and satisfies specific research or analytical purposes.[314] This could range from training ML algorithms to in-depth statistical analyses, as elaborated in Section 2.3. of this study. Moreover, Synthetic Data retaining substantial fidelity means that they accurately reflect the attributes of Real Data, consequently successfully serving as an alternative for the Real Dataset.[315]

Picture a mirror attempting to reflect a complex scene; the clearer the reflection, the higher its fidelity.[316] On the one hand, if Synthetic Data bends too far from the Real Dataset, their utility for research could be compromised due to a lack of

[306] UK Statistics Authority (n 298).
[307] UK Statistics Authority (n 303).
[308] UK Statistics Authority (n 304).
[309] UK Statistics Authority (n 303).
[310] UK Statistics Authority (n 304).
[311] UK Statistics Authority (n 299).
[312] UK Statistics Authority (n 299).
[313] UK Statistics Authority (n 299).
[314] UK Statistics Authority (n 299).
[315] El Emam and others (n 91), 15.
[316] UK Statistics Authority (n 299).

authenticity.[317] On the other hand, if it adheres too closely to the Real Dataset, it might accidentally reveal Personal Data, for instance, through inference, thus interfering with Data Protection norms and ethical considerations.[318] Therefore, while high fidelity datasets are suitable for complex applications like hypothesis generation and AI model testing, on the contrary, low fidelity datasets offer lower disclosure risk and primarily serve the purpose of understanding and scoping challenges regarding the research.[319] Accordingly, the US Synthetic Household Population Database, by having high utility, aligns closely with the intended research objectives and by having high fidelity addresses the extent to which the dataset replicates the characteristics of the Real Dataset from which it derives.[320] Consequently, this Synthetic Database signifies a groundbreaking development in both policy-making and research fields. It facilitates thorough behavioral profiling and in-depth analysis of populations, all the while adhering to Data Protection principles, owing to its absence of Personal Data.

Furthermore, the Office for National Statistics' Policy on Synthetic Data offers essential guidelines for Synthetic Data use in statistical research.[321] This policy delves into the importance of preserving data confidentiality and minimizing the risks of data disclosure.[322] It also stresses major implications when producing, using, and sharing Synthetic Data, predominantly focusing on how Synthetic Data can be used responsibly to protect individuals' privacy while maintaining utility and offering valuable insights.[323] The importance of this policy lies in its function as a standard for ideal data management practices in the UK's statistical research industry, in harmony with wider legal norms like the UK GDPR. Organizations and researchers who follow these guidelines can achieve adherence to legal mandates related to Data Protection, consequently diminishing legal risks linked to the utilization and management of Synthetic Data. In the following subsection, we explore the applications of a multipurpose synthetic population in 3 distinct use cases.

[317] UK Statistics Authority (n 299).
[318] UK Statistics Authority (n 299).
[319] UK Statistics Authority (n 299).
[320] UK Statistics Authority (n 299).
[321] Office for National Statistics (n 298).
[322] Office for National Statistics (n 298).
[323] Office for National Statistics (n 298).

4.2.2.2. Joint Research Centre: Multipurpose Synthetic Population for Policy Applications

The technical report by the Joint Research Centre ("JRC") – the European Commission's dedicated scientific and knowledge service – provides evidence-based scientific support to advance Synthetic Data within European policy formulation.[324] They note that this data mirrors the statistical characteristics of real populations, serving as a "digital twin."[325] In this capacity, the demographic attributes of synthetic populations are blended with Real Data, thereby enriching the quality and complexity of the dataset.[326] This report delves into the creation of a synthetic population for specific policy-related purposes: providing policy advice, simulating the effects of policy measures on the population, and generating population estimates.

Firstly, In use case 1 (henceforth, "French use case") the synthetic population is used to produce policy advice.[327] The population is described on a structured population graph that allows for the linking of behavioral profiles to a common database.[328] The French use case demonstrates that through the use of synthetic populations, it can be tested various policy alternatives, thereby providing policymakers with valuable theoretical guidance for their decisions.

Secondly, use case 2 (henceforth, "Italian use case") has a more pragmatic approach and it involves the use of a synthetic population to simulate the impact of policy interventions on the real population.[329] This allows policymakers to analyze the results of various policy options in a simulated environment before implementing them in real-world scenarios.[330] The Italian use case specifically encompasses the use of a synthetic population to simulate the impact of a policy intervention aimed

[324] European Commission, Joint Research Centre, *Multipurpose Synthetic Population for Policy Applications.* (Publications Office 2022) <https://data.europa.eu/doi/10.2760/50072> accessed 12 October 2023.
[325] European Commission (n 324).
[326] European Commission (n 324).
[327] European Commission (n 324),19-26.
[328] European Commission (n 324),19-26.
[329] European Commission (n 324), 27-42.
[330] European Commission (n 324), 27-42.

at reducing the consumption of sugary drinks.[331] In particular, this study encompasses different scenarios, such as the introduction of a tax on sugary drinks and the implementation of a public health campaign.[332] The results of such simulation are used to evaluate the effectiveness of different policy options and inform evidence-based policymaking.[333] Therefore, these two use cases demonstrate the potential of multipurpose synthetic populations to support policy decisions and evidence-based policymaking.[334]

Lastly, use case 3 (henceforth, “Dutch use case”) involves a synthetic population to improve the accuracy of population estimates, which are commonly used to inform a wide range of policy decisions.[335] The traditional methods of estimating the population can be inaccurate and outdated.[336] The synthetic population solves this issue since it is generated by using a statistical model that takes into account a combination of survey data, administrative data and a wide range of demographic and socioeconomic factors, thus generating more accurate and up-to-date population estimates.[337] The results of the Dutch use case demonstrate the potential of synthetic population data to improve the accuracy of population estimates and, consequently, of policy decision-making.[338]

In summary, the three use cases presented in the report by the JRC demonstrate the potential of synthetic population databases to support evidence-based policymaking and improve the effectiveness of policy interventions. Moving on to the next subsection, our focus shifts to the last practical case study of this thesis: exploring how Synthetic Data is utilized in the context of software testing, through an ethics perspective.

[331] European Commission (n 324), 27-42.
[332] European Commission (n 324), 27-42.
[333] European Commission (n 324), 27-42.
[334] European Commission (n 324), 44-58.
[335] European Commission (n 324), 44-58.
[336] European Commission (n 324), 44-58.
[337] European Commission (n 324), 44-58.
[338] European Commission (n 324), 44-58.

4.2.3. Practical Case: Software Testing

In addressing the ethical concerns of using Synthetic Data in software testing is crucial to understand both their advantages and potential drawbacks. In section 3.2.3.1., it was discussed the use of Synthetic Data for software testing in a financial scenario, thus, demonstrating their value as a PET. While Data Synthesis offer a privacy-compliant alternative for testing with Real Data, ensuring extra security, their effectiveness hinges on accurately replicating real-world scenarios, including outliers and edge cases. This is vital for detecting irregular patterns like those in financial fraud.

While Data Synthesis reduces risks of the processing activities by providing a secure testing environment, it is crucial to communicate transparently about its use, limitations, and how it is processed. This ensures ethical compliance and maintains the integrity of the research. The accuracy and quality of Synthetic Data are paramount; if these standards are not met, it could lead to ineffective tools and negative societal impacts.

As mentioned before when discussing health data accuracy - in section 4.2.1.1 - Synthetic Data's primary objective is to reproduce the general pattern or trend of the Real Data, not necessarily its rare exceptions. For instance, if one were relying on software to evaluate pay equity within the company, an exceptional salary might indicate potential equity issues, demanding further investigation. Picture that most employees of a company earn between €800 and €2,000 per month, but one individual earns €40,000, hence standing out as an outlier. Consequently, if one were to evaluate pay equity within this company through software tested with Synthetic Data, that exceptional €40,000 salary might indicate potential equity issues.

Therefore, if the Synthetic Dataset does not capture outliers present within a financial dataset, the software might miss unusual transaction patterns that could be, for example, indicative of money laundering.[339] Accordingly, when relying solely

[339] Mostly AI, 'Enhancing Fraud Detection Models with Synthetic Data' <https://mostly.ai/case-study/synthetic-training-data-for-machine-learning-fraud-detection> accessed 2 December 2023.

on Synthetic Data for analysis of the efficiency of a model, as in this case, there is a risk of missing crucial indicators, if the Synthetic Dataset is not properly developed. In light of this, some authors argue that this issue can be avoided by correctly balancing the model before advancing on the Data Synthesis procedure.[340]

Furthermore, addressing data accuracy involves confronting biases and outliers. Gartner forecasts that by 2022, 85% of AI algorithms may produce erroneous results due to bias.[341] Rectifying these imbalances is not merely a theoretical ethical issue; it is becoming a practical necessity to avoid incorrect inferences, and loss of trust in AI, as discussed before. Hence, the ability to use Synthetic Data for correcting these biases or statistical imbalances in datasets is becoming essential and will be a key focus of data engineering in the near future.[342] [343]

Concluding, Synthetic Data needs to be well-balanced, accurate, and inclusive of outliers to serve effectively for software testing. Employing Synthetic Data during software development shields Real Data from the risks of unpredictable software environments. In the next chapter, we bring this research to a close with an in-depth final analysis and present a series of carefully formulated legal recommendations tailored to our findings on the processing of Synthetic Data.

[340] Platzer M, 'Boost Your Machine Learning Accuracy with Synthetic Data' (*Mostly AI*, 7 August 2020) <https://mostly.ai/blog/boost-machine-learning-accuracy-with-synthetic-data> accessed 2 December 2023
[341] Judah and others (n 106).
[342] Mostly AI, 'GUIDE: How to Leverage AI-Powered Synthetic Data in Enterprises' <https://mostly.ai/ebook/synthetic-data-for-enterprises> accessed 1 December 2023.
[343] Platzer (n 340).

5. Final remarks

In this final chapter, our objective is to present the conclusions drawn from this research, alongside informed suggestions for entities either considering or currently engaged in processing Synthetic Data. To facilitate a clear understanding, the chapter is divided into two main sections: the first offers insightful suggestions based on the GDPR and academic analysis pertinent to Synthetic Data processing, and the second section is dedicated to the final conclusions of the study. It is important to note that the suggestions provided here are based on academic research and should not be construed as professional legal advice.

5.1. Legal recommendations

This section outlines the key compliance aspects to consider when processing Synthetic Data. To enhance understanding, the discussion is organized into two distinct parts: the compliance requirements before Data Synthesis and those that apply after the Data Synthesis.

A) Before the Data Synthesis

When handling Personal Data, Article 6 of the GDPR requires the Controller to have a legitimate reason for this processing. This specific legal basis will depend on the concrete situation. For example, it could be the consent of the individual whose data is being processed or the legitimate interests pursued by the Controller. Therefore, when a dataset containing Personal Data is intended to be used as input within Data Synthesis, the Controller must have a lawful basis for such processing activity, as per Article 6(1) of the GDPR – Principle of Lawfulness.

Furthermore, to comply with the Principles of Accountability (Article 5(2), GDPR), Purpose Limitation (Article 5(1)(b), GDPR), and Data Minimization (Article 5(1)(c), GDPR) it is imperative to meticulously define and document not only the specific and legitimate purposes for which the Personal Data will be collected but also the precise methods that will be carried out during the Data Synthesis (which is a processing activity). Only adequate and relevant data to achieve the purposes of

the processing activity can be lawfully processed. Also, the documentation will serve as evidence to support the lawful and transparent processing of Personal Data for the purpose of Data Synthesis.

Depending on the types of data and purposes of the processing, we may suggest performing a Data Protection Impact Assessment ("DPIA") on the Data Synthesis. For instance, when processing health data, it is required to perform a DPIA, as per Article 9 (1) and Article 35 of the GDPR. The DPIA is an instrument that facilitates a comprehensive evaluation of processing operations, examining the necessity and proportionality of these activities in comparison to the risk to the Data Subjects' rights. It serves as a strategic framework for implementing measures to mitigate identified risks, ensuring that the Data Synthesis process is in stringent adherence to the obligations mandated by the GDPR. Therefore, this assessment should be comprehensive, encompassing a range of considerations, such as the methodology of the Data Synthesis, the characteristics of the real dataset, and the risk of re-identification. The assessment of this risk may include, for example, the possibility of external data sources used for re-identification of the Data Subjects.

It is also imperative to ensure the correct application of Data Synthesis techniques. Entities must ensure that these techniques are applied in a manner that inherently results in the generation of Anonymous Synthetic Data. To comply with the Principles of Data Protection by Design and by Default (Article 25 of the GDPR), it is essential to ensure that privacy and Data Protection standards are firmly embedded within the system's architecture and operational procedures, starting from the initial data collection phase and continuing through to the generation of Synthetic Data.

Additionally, we argue that it is crucial to regularly review and update the Data Synthesis methods to keep pace with evolving technologies and rising privacy concerns. This may include employing advanced algorithms that minimize the risk of re-identification while maintaining data and conducting periodic risk assessments to assess the effectiveness of current synthesis techniques. Entities should also be transparent about their Data Synthesis practices, facilitating stakeholders'

understanding of how their data is being used and protected, thus reinforcing trust and compliance with GDPR requirements.

B) After the Data Synthesis

After the Data Synthesis course, it is crucial for the responsible entity to keep adhering to good practices in line with the GDPR principles of Data Protection by Design and by Default. This ensures ongoing compliance and integrity in the handling of Synthetic Data.

Please note that when Synthetic Data is created using Real Datasets, the Data Synthesis procedure falls within the jurisdiction of the GDPR until its completion. However, the final output, the Synthetic Data, is not governed by the GDPR, as it does not fit within the boundaries of the GDPR's Personal Data concept. The recommendations delineated in this section, likewise, should be viewed as best practices aimed at reinforcing Data Protection and ensuring responsible data management, rather than strict legal requirements. Therefore, as a good practice, the Data Protection measures established during the design and execution of the Data Synthesis process, should continue to be effective and integral during the handling of the Synthetic Data. This means maintaining the privacy safeguards that were implemented during the Data Synthesis process, such as ensuring that the Synthetic Data remains non-identifiable and cannot be reverse engineered to reveal Personal Data.

In conclusion, we argue that entities should not only focus on the adherence to legal standards but also take into account evolving technological advancements and emerging legal interpretations. Managing periodic reviews of the Data Synthesis serve a dual purpose: they ensure that the procedure remains within the legal boundaries set by GDPR and they allow for timely adjustments in response to new technological challenges and opportunities, thereby reducing re-identification risks. This proactive approach is crucial in maintaining the integrity of the Data Synthesis process and safeguarding the data protection rights of individuals used in the input

datasets. In the following section, we arrive at the concluding part of this thesis, marking the culmination of our discussions and findings.

5.2. Conclusion

Our research was initiated in Chapter 1 with an exploration of the intricate and evolving nature of Personal Data. We delved into the increasingly ambiguous distinction between Personal and Non-Personal Data, shaped by ongoing legal interpretations and technological advancements, therefore revealing the dinamic nature of these concepts. A detailed analysis of GDPR's Recital 26 and the concept of identifiability highlighted the challenges in consistently applying these notions. The risks and methodologies associated with De-identification and Anonymization were also examined, showcasing their significant roles in Data Protection and security. Throughout this chapter, we introduced four pivotal questions, subsequently addressed and extensively discussed in the later phases of our study.

Concretely, we posed the question, "Should Anonymized Datasets, particularly those used in training AIML algorithms, be reevaluated under GDPR principles?" Our findings affirmatively concluded that Anonymized Data falls outside the scope of the GDPR. Additionally, we inquired, "Do Synthetic Data qualify as Personal Data under GDPR criteria?" Through rigorous analysis, it became clear that although Synthetic Data may originate from Personal Data, the extent of transformation rendering them Anonymous signifies that they do not identify any individual, thereby exempting them from being classified as Personal Data under the GDPR.

As we progressed deeper into the first chapter, two questions appeared concerning the sharing of Anonymized Data with third parties. Firstly, we considered, "In the absence of a clear Controller, as defined by GDPR, how can re-identification risks be mitigated for data handled by a new collaborating company?" Addressing this query led to an essential debate within the Data Protection domain. We firmly believe that the entity (which formerly acted as a Controller, as stipulated by the GDPR) retains primary responsibility for data the processing, even extending to data transformed into an Anonymized or synthetic state. In alignment with this view,

we advocate for stringent best practices, particularly during the transfer of Synthetic Data. The entity must diligently ensure that any recipient of Synthetic Data upholds security standards equivalent to those maintained during the Data Synthesis process. This responsible strategy for transferring data ensures continuous protection and aligns with the ethical and social responsibilities inherent to the processing. This proactive approach to transferring Synthetic Data demonstrates a profound understanding of the complexities in Data Protection, hence of the GDPR, highlighting the need for vigilant and proactive data management, especially in scenarios involving multiple transformations and crossing various jurisdictional and operational boundaries.

Furthermore, we pondered, "Are individuals whose data has been Anonymized and included in an Anonymized Dataset still considered 'Data Subjects' under GDPR?" We believe that individuals are no longer identifiable as Data Subjects once data has been effectively Anonymized. Therefore, within the realm of Data Synthesis, when Personal Data serves as the input and the Data Synthesis process is successfully completed, these individuals have no remaining association with the resulting Synthetic Data, thus, they are not Data Subjects in light of the GDPR.

In Chapter 2, we focused on the technological underpinnings and methodologies of Data Synthesis, emphasizing its ability to generate valuable Synthetic Data while prioritizing Data Protection. The relevance of Data Synthesis, derived from both Personal and Non-Personal Data, in relation to the GDPR, was critically assessed, balancing the trade-off between data utility and Data Protection. Instances like AlphaGo were used to exemplify these concepts.

Furthermore, Chapter 3 illuminated that Synthetic Data necessitates a multifaceted approach, emphasizing accuracy and reliability. Central to this approach is the application of robust Data Synthesis methodologies. We also explored the question, "How can data accuracy be assured in Synthetic Data that is indirectly linked to Data Subjects?" We conclude that this aspect of the process requires careful assessment by the developers of Synthetic Data, given the absence of an accuracy threshold or legal standard. We advocate for a careful and responsible approach to

generating and utilizing Synthetic Data, ensuring they accurately reflect the statistical properties and relationships inherent in the real data. Also, in this chapter, we discussed the transformative impact of Synthetic Data in enhancing Data Protection. We highlighted how Synthetic Data can mitigate re-identification risks and facilitate GDPR compliance. Our examination of data accuracy, minimization, and security in the Synthetic Data context provided practical insights for adhering to these principles in various scenarios, including health data and software testing.

Chapter 4 addressed the ethical and practical challenges of using Synthetic Data. We delved into issues of bias, loss of public trust, and the imperative to uphold ethical standards. By analyzing practical cases in health data, synthetic population databases, and software testing, we demonstrated the nuanced ethical considerations necessary for responsible Synthetic Data usage, closely linked to the considerations of responsible processing in light of the GDPR.

Finally, the present chapter synthesizes the insights from previous chapters, offering a summative perspective on legal recommendations for entities interested in processing Synthetic Data. This conclusion brings together the key themes of our research: the complexity of defining Personal Data, the transformative potential and challenges of Data Synthesis and the role of Synthetic Data in contemporary Data Protection and ethical considerations.

REFERENCES

Andrews G, 'What Is Synthetic Data?' (*NVIDIA Blog*, 8 June 2021) <https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data/> accessed 25 September 2023.

Arthur L and others, 'On the Challenges of Deploying Privacy-Preserving Synthetic Data in the Enterprise' (arXiv, 9 July 2023) <http://arxiv.org/abs/2307.04208> accessed 1 December 2023.

Article 29 Data Protection Working Party, 'Opinion 05/2014 on Anonymisation Techniques' (2014) 0829/14/EN WP216 <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf> accessed 17 September 2023.

Bamford S, 'Synthetic Data and Privacy - Experiences Implementing Data Synthesis in a Global Life Sciences Company' <https://edps.europa.eu/system/files/2021-06/01_stephen_bamford_en_0.pdf> accessed 3 December 2023.

Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) <https://papers.ssrn.com/abstract=2477899> accessed 10 September 2023.

Batura O and Peeters R, 'European Union Data Challenge' (Policy Department for Economic, Scientific and Quality of Life Policies 2021) PE 662.939 <https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/662939/IPOL_BRI(2021)662939_EN.pdf> accessed 4 December 2023.

Baumgartner U, 'New Options for Anonymization Ahead?' (*IAPP The Privacy Advisor*, 18 May 2023) <https://iapp.org/news/a/new-options-for-anonymization-ahead/> accessed 11 December 2023.

Benaim AR and others, 'Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five

Observational Studies' (2020) 8 JMIR Medical Informatics <https://medinform.jmir.org/2020/2/e16492> accessed 4 November 2023.

Benthall S, 'Situated Information Flow Theory', *Proceedings of the 6th Annual Symposium on Hot Topics in the Science of Security* (Association for Computing Machinery 2019) <https://doi.org/10.1145/3314058.3314066> accessed 25 September 2023.

Breugel B and others, 'Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data' <https://arxiv.org/abs/2305.09235> accessed 1 December 2023.

Brittain B and Brittain B, 'Google Hit with Class-Action Lawsuit over AI Data Scraping' *Reuters* (12 July 2023) <https://www.reuters.com/legal/litigation/google-hit-with-class-action-lawsuit-over-ai-data-scraping-2023-07-11/> accessed 25 July 2023.

Canada E, 'Covid-19 Resources' (17 May 2023) <https://resources-covid19canada.hub.arcgis.com/> accessed 5 October 2023.

Centre for Information Policy Leadership (CIPL), 'Artificial Intelligence and Data Protection How the GDPR Regulates AI' (2020) <https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl-hunton_andrews_kurth_legal_note_-_how_gdpr_regulates_ai__12_march_2020_.pdf> accessed 9 September 2023.

Chaudhuri A, 'Internet of Things Data Protection and Privacy in the Era of the General Data Protection Regulation' [2016] Journal of Data Protection & Privacy <https://hstalks.com/article/2661/internet-of-things-data-protection-and-privacy-in-/> accessed 25 July 2023.

Dilmegani C, 'Synthetic Data for Healthcare: Benefits & Case Studies in 2023' (*AI Multiple*, 22 December 2022) <https://research.aimultiple.com/synthetic-data-healthcare/> accessed 29 November 2023.

——, ‘What Is Synthetic Data? Use Cases & Benefits in 2023’ (26 October 2023) <https://research.aimultiple.com/synthetic-data/> accessed 29 November 2023

Dolhansky B and others, ‘The Deepfake Detection Challenge (DFDC) Preview Dataset’ (arXiv, 23 October 2019) <http://arxiv.org/abs/1910.08854> accessed 3 December 2023.

El Emam K, *Accelerating AI with Synthetic Data* (O’Reilly Media, Inc 2020).

El Emam K, Mosquera L and Hoptroff R, *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data* (First Edition, O’Reilly Media, Inc 2020).

El Emam K and others, ‘Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation’ (2020) 22 Journal of Medical Internet Research e23139 <https://www.jmir.org/2020/11/e23139> accessed 5 October 2023.

European Commission, ‘Anonymization’ (*Collaboration in Research and Methodology for Official Statistics*, 28 April 2019) <https://cros-legacy.ec.europa.eu/content/anonymization_en> accessed 19 November 2023.

European Commission, Joint Research Centre, *Multipurpose Synthetic Population for Policy Applications.* (Publications Office 2022) <https://data.europa.eu/doi/10.2760/50072> accessed 12 October 2023.

European Data Protection Supervisor, ‘Checklist 3: What Is Required in a Processing Agreement?’ (*Checklists and flowcharts on Data Protection*, 27 September 2019) <https://edps.europa.eu/data-protection/our-work/publications/factsheets/checklists-and-flowcharts-data-protection> accessed 26 November 2023.

European Data Protection Supervisor and Agencia Española de Protección de Datos, ‘10 Misunderstandings Related to Anonymisation’ (2021)

<https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en> accessed 9 December 2023.

European Parliament, 'REPORT on the Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) | A7-0402/2013 | European Parliament' (2013) A7-0402/2013 <https://www.europarl.europa.eu/doceo/document/A-7-2013-0402_EN.html> accessed 20 September 2023.

European Union Agency for Cybersecurity (ENISA), 'Pseudonymisation Techniques and Best Practices' (2019) Report/Study <https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices> accessed 19 November 2023.

European Union Agency For Network and Information Security, 'Recommendations on Shaping Technology According to GDPR Provisions - An Overview on Data Pseudonymisation' (2018) Report/Study <https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions> accessed 25 November 2023.

Finch K, 'A Visual Guide to Practical Data De-Identification' (*https://fpf.org/*, 25 April 2016) <https://fpf.org/blog/a-visual-guide-to-practical-data-de-identification/> accessed 19 September 2023.

Finck M and Pallas F, 'They Who Must Not Be Identified - Distinguishing Personal from Non-Personal Data Under the GDPR' (1 October 2019) <https://papers.ssrn.com/abstract=3462948> accessed 17 June 2023.

Gal M and Lynskey O, 'Synthetic Data: Legal Implications of the Data-Generation Revolution' (10 April 2023) <https://papers.ssrn.com/abstract=4414385> accessed 21 September 2023.

Ganev G, 'When Synthetic Data Met Regulation' (arXiv, 1 July 2023) <http://arxiv.org/abs/2307.00359> accessed 8 October 2023.

Goncalves A and others, 'Generation and Evaluation of Synthetic Patient Data' (2020) 20 BMC Medical Research Methodology 108 <https://doi.org/10.1186/s12874-020-00977-1> accessed 4 November 2023.

Gonzales A, Guruswamy G and Smith SR, 'Synthetic Data in Health Care: A Narrative Review' (2023) 2 PLOS Digital Health <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082> accessed 4 November 2023.

Graef and others, 'Towards a Holistic Regulatory Approach for the European Data Economy: Why the Illusive Notion of Non-Personal Data Is Counterproductive to Data Innovation' [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3256189> accessed 10 September 2023.

Griffith I, 'International: Is Synthetic Data the Future of Privacy?' (*DataGuidance*, 27 February 2023) <https://www.dataguidance.com/opinion/international-synthetic-data-future-privacy> accessed 5 December 2023.

Hansen RK, 'AI Image Generator: This Is Someone Thinking About Data Ethics · Dataetisk Tænkehandletank' (*Dataetisk Tænkehandletank*, 15 August 2022) <https://dataethics.eu/ai-image-generator-this-is-someone-thinking-about-data-ethics/> accessed 7 October 2023.

Harris S, 'The Rising Role of Synthetic Data in the Automotive Industry' (*Automotive Testing Technology International*, 5 June 2023) <https://www.automotivetestingtechnologyinternational.com/industry-opinion/the-rising-role-of-synthetic-data-in-the-automotive-industry.html> accessed 22 September 2023.

Hendrawirawan D, 'Synthetic Data for AI: Definition, Risks, and Strategies' (*Eckerson Group*, 11 October 2022) <https://www.eckerson.com/articles/synthetic-data-for-ai-definition-risks-and-strategies> accessed 29 November 2023.

Hern A, 'AlphaGo: Its Creator on the Computer That Learns by Thinking' *The Guardian* (15 March 2016) <https://www.theguardian.com/technology/2016/mar/15/alphago-what-does-google-advanced-software-go-next> accessed 21 September 2023.

IBM Corporation, 'Outlier Detection' (19 March 2021) <https://www.ibm.com/docs/en/guardium/10.6?topic=audit-outlier-detection> accessed 11 October 2023.

Information Commissioner's Office, 'Introduction to Anonymisation' (2021) <https://ico.org.uk/media/about-the-ico/consultations/2619862/anonymisation-intro-and-first-chapter.pdf>.

——, 'Chapter 5: Privacy-Enhancing Technologies (PETs)' (2022) <https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf> accessed 1 November 2023.

——, 'Guide to Accountability and Governance' <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/accountability-and-governance/guide-to-accountability-and-governance/accountability-and-governance/data-protection-by-design-and-default/> accessed 9 September 2023.

Information Technology Laboratory, 'Adversary - Glossary' (*COMPUTER SECURITY RESOURCE CENTER*) <https://csrc.nist.gov/glossary/term/adversary> accessed 5 October 2023.

——, 'Data Linking - Glossary' (*Computer Security Resource Center*) <https://csrc.nist.gov/glossary/term/data_linking> accessed 19 November 2023.

——, ‘De-Identification - Glossary’ (*Computer Security Resource Center*) <https://csrc.nist.gov/glossary/term/de_identification> accessed 25 November 2023.

——, ‘Pseudonymization’ (*Computer Security Resource Center*) <https://csrc.nist.gov/glossary/term/pseudonymization> accessed 25 November 2023.

International Association of Privacy Professionals, ‘Article 29 Working Party’ (*ResourceCenter*) <https://iapp.org/resources/article/article-29-working-party/> accessed 6 December 2023.

J. Kleczyk E, ‘Importance of Patient Privacy in Healthcare Analytics Research’, *Ethics - Scientific Research, Ethical Issues, Artificial Intelligence and Education [Working Title]* (IntechOpen 2023) <https://www.intechopen.com/online-first/1120485> accessed 3 December 2023.

Johnston A, ‘Keeping It Fake: The Legal and Ethical Implications of Synthetic Data’ (*Salinger Privacy*, 6 July 2023) <https://www.salingerprivacy.com.au/2023/07/06/synthetic-data/> accessed 7 October 2023.

J.P. Morgan, ‘Synthetic Data’ (*J.P. Morgan AI Research*) <https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data> accessed 21 September 2023.

Judah S, White A and Sicular S, ‘Predicts 2021: Data and Analytics Strategies to Govern, Scale and Transform Digital Business’ (*Gartner Research*, 2 December 2020) <https://www.gartner.com/en/documents/3993855> accessed 2 December 2023.

Khanna A and Kaur S, 'Internet of Things (IoT), Applications and Challenges: A Comprehensive Review' (2020) 114 Wireless Personal Communications 1687 <https://doi.org/10.1007/s11277-020-07446-4> accessed 25 July 2023.

Kuner C and others (eds), *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford University Press 2020) <https://academic.oup.com/book/41324> accessed 20 September 2023.

Lancaster University, 'Lancashire Cyber Foundry an Introduction to Deepfakes' (*Digital Resources*) <https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/cyber-foundry/lcf-articles/LCFArticle-Josh-Deepfakes_WEB.pdf> accessed 3 December 2023.

Liew CK and others, 'A Data Distortion by Probability Distribution' (1985) 10 ACM Transactions on Database Systems 395 <https://dl.acm.org/doi/10.1145/3979.4017> accessed 20 September 2023.

Loong B and others, 'Disclosure Control Using Partially Synthetic Data for Large-scale Health Surveys, with Applications to CanCORS' (2013) 32 Statistics in Medicine 4139 <https://onlinelibrary.wiley.com/doi/10.1002/sim.5841> accessed 29 November 2023.

López CAF and Elbi A, 'On the Legal Nature of Synthetic Data' (2022) <https://openreview.net/forum?id=M0KMbGL2yr&noteId=0mH-aK63WH> accessed 8 October 2023.

Mittelstadt BD and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 Big Data & Society. 15 <http://journals.sagepub.com/doi/10.1177/2053951716679679> accessed 10 September 2023.

Mostly AI, ‘Enhancing Fraud Detection Models with Synthetic Data’ <https://mostly.ai/case-study/synthetic-training-data-for-machine-learning-fraud-detection> accessed 2 December 2023.

——, ‘GUIDE: How to Leverage AI-Powered Synthetic Data in Enterprises’ <https://mostly.ai/ebook/synthetic-data-for-enterprises> accessed 1 December 2023.

Narayanan A and Shmatikov V, ‘Robust De-Anonymization of Large Sparse Datasets’, *2008 IEEE Symposium on Security and Privacy (sp 2008)* (2008).

Office for National Statistics, ‘Synthetic Data Policy - Office for National Statistics’ <https://www.ons.gov.uk/aboutus/transparencyandgovernance/datastrategy/datapolicies/syntheticdatapolicy/> accessed 10 December 2023.

Office of the Privacy Commissioner of Canada, ‘Privacy Tech-Know Blog: When What Is Old Is New Again – The Reality of Synthetic Data’ (12 October 2022) <https://www.priv.gc.ca/en/blog/20221012/?id=7777-6-493564> accessed 11 June 2023.

Ohm P, ‘Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization’ (13 August 2009) <https://papers.ssrn.com/abstract=1450006> accessed 25 May 2023.

Pappalardo F and others, ‘In Silico Clinical Trials: Concepts and Early Adoptions’ (2019) 20 Briefings in Bioinformatics 1699 <https://academic.oup.com/bib/article/20/5/1699/5032454> accessed 29 November 2023.

Platzer M, ‘Boost Your Machine Learning Accuracy with Synthetic Data’ (*Mostly AI*, 7 August 2020) <https://mostly.ai/blog/boost-machine-learning-accuracy-with-synthetic-data> accessed 2 December 2023.

Pratt MK, 'What Is Data Curation? - Definition from SearchBusinessAnalytics' (*Business Analytics*, January 2022) <https://www.techtarget.com/searchbusinessanalytics/definition/data-curation> accessed 21 September 2023.

Purtova N, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40 <https://www.tandfonline.com/doi/full/10.1080/17579961.2018.1452176> accessed 21 May 2023.

R. S. I. Security, 'What Are GDPR Recitals?' (*RSI Security*, 20 June 2018) <https://blog.rsisecurity.com/what-are-gdpr-recitals/> accessed 5 November 2023.

Raman A, 'What Is Data Aggregation: A Comprehensive Guide 101' (*Hevo*, 26 September 2023) <https://hevodata.com/learn/data-aggregation/> accessed 18 November 2023.

Raso FA and others, 'Artificial Intelligence & Human Rights: Opportunities & Risks' (25 September 2018) <https://papers.ssrn.com/abstract=3259344> accessed 10 September 2023.

Riemann R, 'Synthetic Data' (*European Data Protection Supervisor*, 20 October 2023) <https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data> accessed 22 October 2023.

Rodriguez L and Howe B, 'In Defense of Synthetic Data' (arXiv, 3 May 2019) <http://arxiv.org/abs/1905.01351> accessed 1 December 2023.

Rogers A, 'Council Post: What Deep Blue And AlphaGo Can Teach Us About Explainable AI' (*Forbes*) <https://www.forbes.com/sites/forbestechcouncil/2019/05/09/what-deep-blue-and-alphago-can-teach-us-about-explainable-ai/> accessed 21 September 2023.

'RTI U.S. Synthetic Household Population™' (*RTI International*) <https://www.rti.org/impact/rti-us-synthetic-household-population%E2%84%A2> accessed 10 December 2023.

Rubin D, 'Statistical Disclosure Limitation', vol 9 (2nd edn, J OFF STAT 1993) <https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf> accessed 1 November 2023.

Rubinstein I and Hartzog W, 'Anonymization and Risk' (17 August 2015) <https://papers.ssrn.com/abstract=2646185> accessed 17 June 2023.

Schwartz PM and Solove DJ, 'Reconciling Personal Information in the United States and European Union' (6 September 2013) <https://papers.ssrn.com/abstract=2271442> accessed 20 May 2023.

Seastrom M, 'Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records' (23 November 2010) <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011601> accessed 19 September 2023.

Sensity Team, 'How to Detect AI Generated Images with Sensity in 2023' (10 May 2023) <https://sensity.ai/blog/deepfake-detection/how-to-detect-ai-generated-im/> accessed 3 December 2023.

Silver D and Hassabis D, 'AlphaGo: Mastering the Ancient Game of Go with Machine Learning' (27 January 2016) <https://blog.research.google/2016/01/alphago-mastering-ancient-game-of-go.html> accessed 21 September 2023.

Stephen B, 'Synthetic Data and Privacy Experiences - Implementing Data Synthesis in a Global Life Sciences Company' (June 2021) <https://edps.europa.eu/system/files/2021-06/01_stephen_bamford_en_0.pdf> accessed 27 September 2023.

Subramanyam J and Ramos L, 'Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI' (*Gartner Research*, 24 June 2021) <https://www.gartner.com/en/documents/4002912> accessed 21 September 2023.

Sweeney L, 'Matching Known Patients to Health Records in Washington State Data' (5 June 2013) <https://papers.ssrn.com/abstract=2289850> accessed 10 December 2023.

SyntheticMass, 'Synthea' <https://synthea.mitre.org/about> accessed 22 September 2023.

Taleb NN, *The Black Swan: The Impact of the Highly Improbable* (Random House trade paperback edition, Random House 2016).

Toews R, 'Synthetic Data Is About To Transform Artificial Intelligence' (*Forbes*, 12 June 2022) <https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/> accessed 1 December 2023.

Torre I and others, 'A Framework for Personal Data Protection in the IoT', *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)* (IEEE 2016) <http://ieeexplore.ieee.org/document/7856735/> accessed 25 July 2023.

UK Statistics Authority, 'Ethical Considerations Relating to the Creation and Use of Synthetic Data' (*UK Statistics Authority - Statistics for the Public Good*) <https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/> accessed 27 September 2023.

——, 'Ethical Principles' (*UK Statistics Authority - Statistics for the Public Good*) <https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethical-principles/> accessed 27 September 2023.

——, ‘Ethics Self-Assessment Tool’ (*UK Statistics Authority - Statistics for the Public Good*) <https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/> accessed 27 September 2023.

Wheaton W and Rineer J, ‘RTI U.S. Synthetic Household Population™’ <https://www.rti.org/impact/rti-us-synthetic-household-population%E2%84%A2> accessed 4 November 2023.

Wood T, ‘Generative Adversarial Network’ (*DeepAI*, 22 July 2020) <https://deepai.org/machine-learning-glossary-and-terms/generative-adversarial-network> accessed 22 October 2023.

YData, ‘Synthetic Data: The Future Standard for Data Science Development’ (2 April 2020) <https://ydata.ai/resources/synthetic-data-the-future-standard-for-data-science-development> accessed 22 October 2023.

——, ‘10 Most Frequently Asked Questions about Synthetic Data’ (21 March 2023) <https://ydata.ai/resources/10-most-frequently-asked-questions-about-synthetic-data> accessed 22 October 2023.

Single Resolution Board v European Data Protection Supervisor [2023] General Court Case T-557/20.

Charter of Fundamental Rights of the European Union 2012 (2012/C).

Consolidated version of the Treaty on the Functioning of the European Union 2012 (OJ C).

Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of Personal Data and on the free movement of such data. OJ L281/31.

Health Insurance Portability and Accountability Act of 1996.

NAI Code of Conduct 2020 7;8.

Protecting Consumer Privacy in an Era of Rapid Change: Recommendations For Businesses and Policymakers 2012 20;21.

Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of Non-Personal Data in the European Union (Text with EEA relevance.) 2018 (OJ L).

ANNEX I - DEFINITIONS

Concept	Adopted Definition	Source
AIML Project	We define an AIML project quite broadly to include projects run in various industries, for example, the development of software applications that have AIML components.	El Emam K, Mosquera L and Hoptroff R, Practical Synthetic Data Generation (O'Reilly Media, Inc 2020).
Anonymization	Procedure of generating Anonymized Data.	El Emam K, Mosquera L and Hoptroff R, Practical Synthetic Data Generation (O'Reilly Media, Inc 2020).
Anonymized Data	Personal Data that have undergone a transformation process, rendering them into Non-Personal or Anonymous.	Recital 26, GDPR
Anonymous Data	Information that does not relate to any person.	Recital 26, GDPR
Data Controller/Controller	In general, the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the Processing of Personal Data.	Article 4(7), GDPR
Data Processor/Processor	In general, a natural or legal person, public authority, agency or other	Article 4(8), GDPR

	body which Processes Personal Data on behalf of the Data Controller.	
Data Protection Laws	The Regulation and complementary national Data Protection Laws of the European member-states, including any guidance and / or codes of practice issued by the relevant Supervisory Authorities within the EU.	
Data Subject	An identified or identifiable natural person.	Article 4(1), GDPR
Data Synthesis	Procedure of generating Synthetic Data.	
De-identification	Any process of removing the association between a set of identifying data and the Data Subject.	Information Technology Laboratory, 'De-identification - Glossary' (Computer Security Resource Center) <https://csrc.nist.gov/glossary/term/de_identification> accessed 25 November 2023.
De-identified Data	Records that have a re-identification code and have enough identifiable information removed or masked so that the remaining information does not allow the identification of the Data Subject.	Seastrom M, 'Basic Concepts and Definitions for Privacy and Confidentiality in Student Education

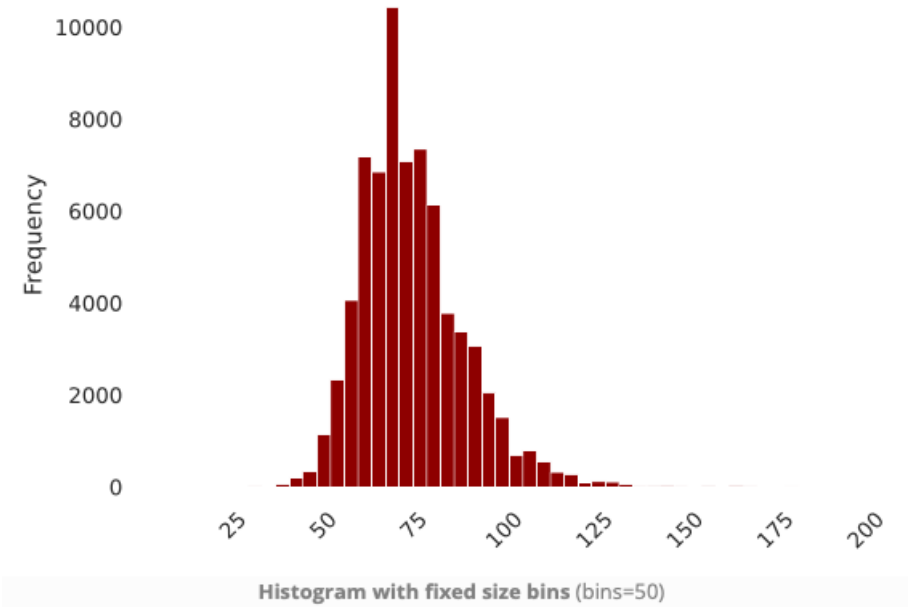
		Records' (23 November 2010). 6. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011601> accessed 19 September 2023.
Non-Personal Data	Information falling outside the scope of Personal Data, as explained in paragraph 1 of the preceding Article 4 of the GDPR.	Article 3(1), Regulation of the Free Flow of Non-Personal Data.
Personal Data	Any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.	Article 4(1), GDPR
Privacy Enhancing Technology (PET)	Privacy Enhancing Technologies (PETs) are technologies that embody fundamental Data Protection principles by minimising Personal Data use, maximising data	Information Commissioner's Office, 'Chapter 5: Privacy-Enhancing

	security, and empowering individuals.	Technologies (PETs)' (2022).
Processing	Any operation, or set of operations, which is performed on Personal Data, or on sets of Personal Data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.	Article 4(2), GDPR
Pseudonymization	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.	Article 4(5), GDPR
Real Data	Data obtained from real-world sources.	
Synthetic Dataset	Dataset generated through Data Synthesis.	El Emam K, Mosquera L and Hoptroff R, Practical Synthetic

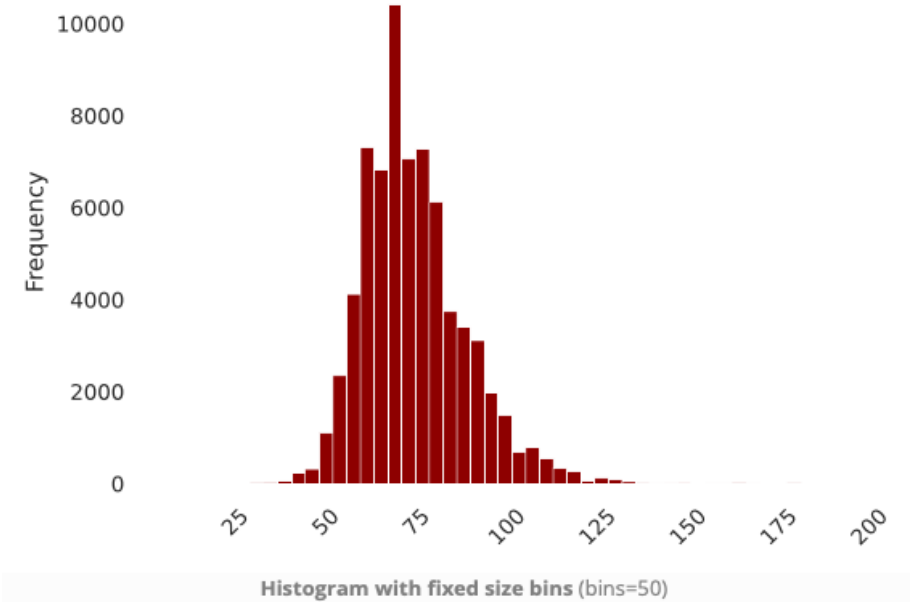
		Data Generation (O'Reilly Media, Inc 2020).

ANNEX II - COMPARISON BETWEEN SYNTHETIC DATA AND REAL DATA VARIABLES

Variable Weight Comparison by Histogram

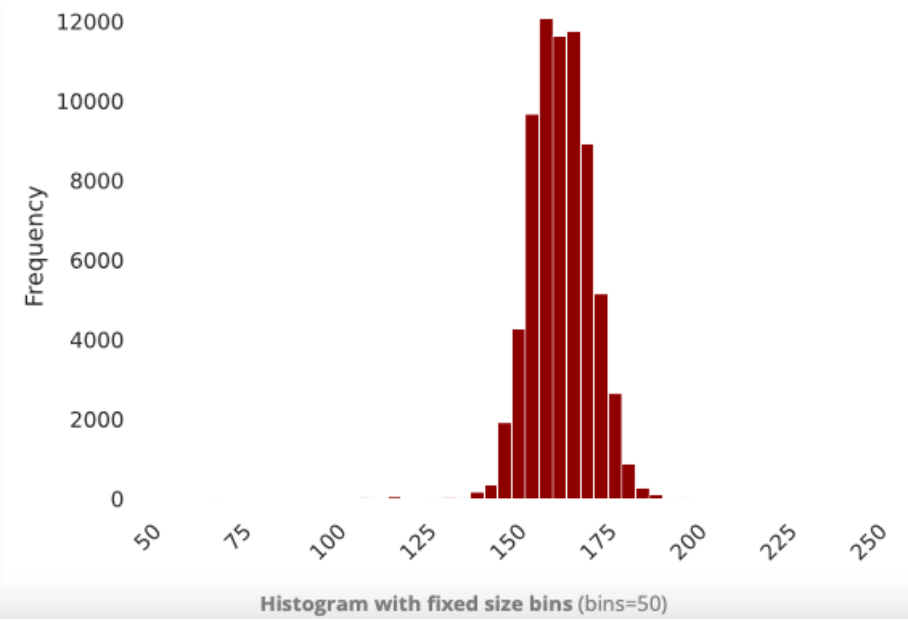


Variable Weight Synthetic Data

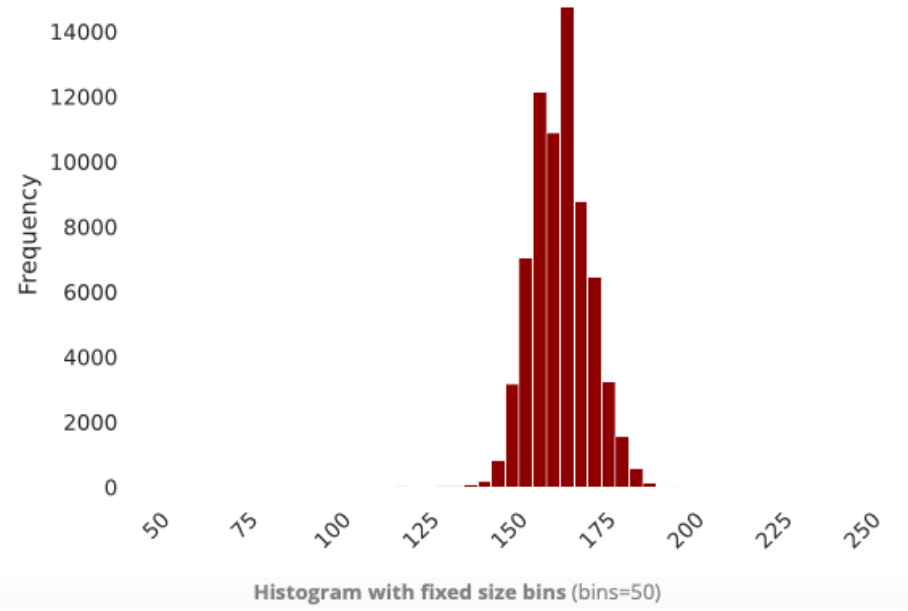


Variable Weight Real Data

Variable Height Comparison by Histogram

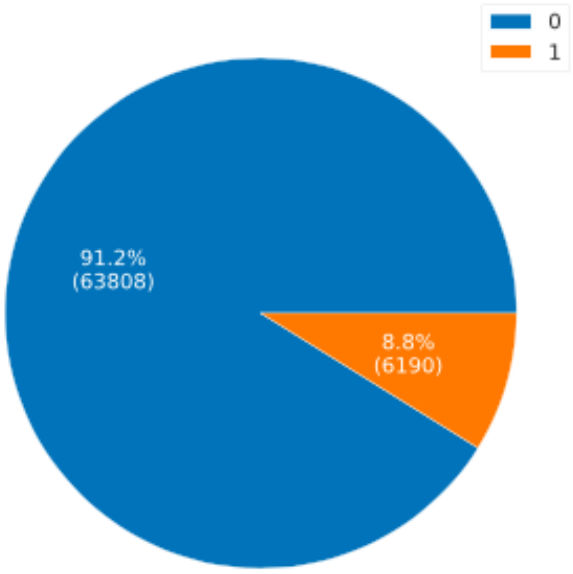


Variable Height Synthetic Data

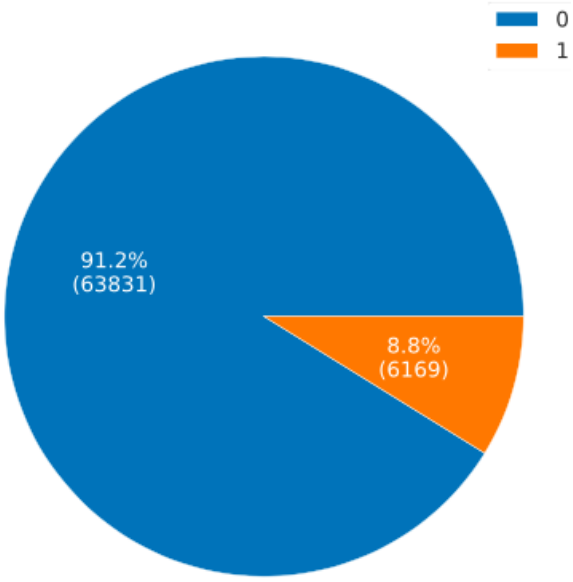


Variable Height Real Data

Variable Smoke Comparison by Chart

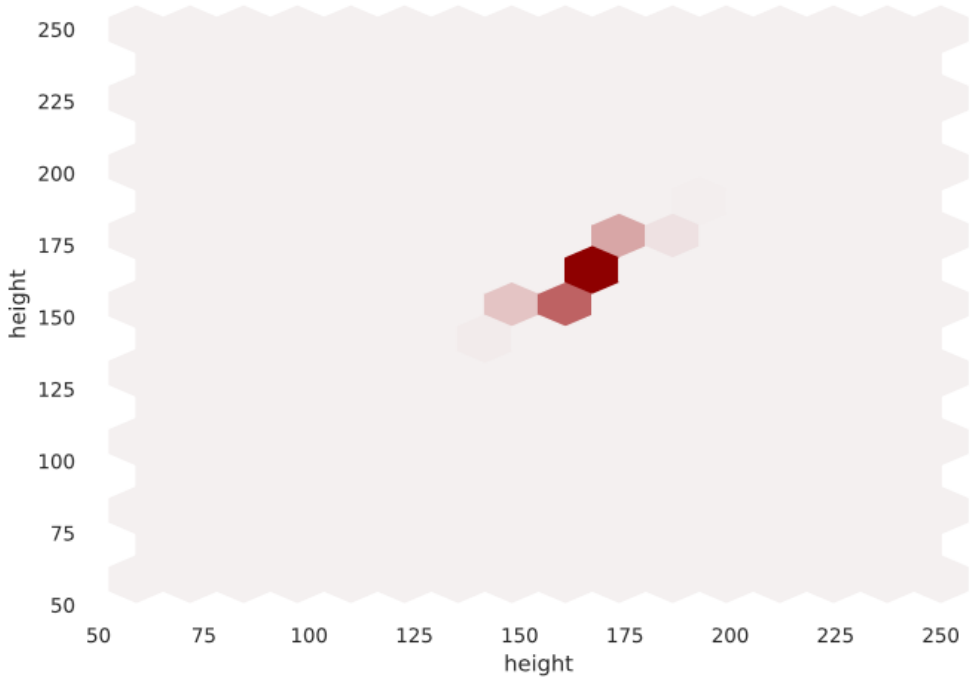


Variable Smoke Synthetic Data

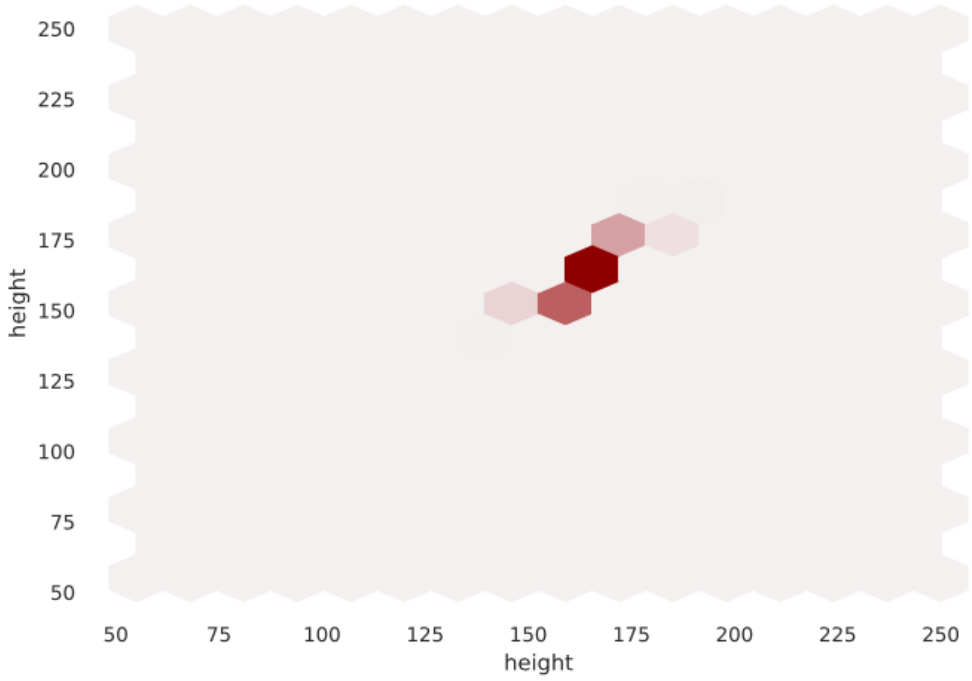


Variable Smoke Real Data

Variable Height Comparison by Heatmap



Variable Height Synthetic Data



Variable Height Real Data