WIRTSCHAFTSUNIVERSITÄT WIEN
Vienna University of Economics and Business

Master Thesis

Title of Master Thesis:	A new approach for marketing analytics in an increasing environment of data regulation - synthetic data
Author (last name, first name):	Bogner, Johannes
Student ID number:	h1352800
Degree program:	Master in Marketing
Examiner (degree, first name, last name):	a.o. Univ.-Prof. Mag. Dr. Thomas, Reutterer

I hereby declare that:

1. I have written this Master thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.
2. This Master Thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.
3. This Master Thesis is identical with the thesis assessed by the examiner.
4. (only applicable if the thesis was written by more than one author): this Master thesis was written together with

   The individual contributions of each writer as well as the co-written passages have been indicated.

19.08.2020
Date

Signature

Master Thesis

at the Department of Marketing

Institute for Service Marketing and Tourism

Supervisor: a.o. Univ.-Prof. Mag. Dr. Thomas Reutterer

Co-supervisor: Stefan Vamosi MSc.

A new approach for marketing analytics in an increasing environment of data regulation – synthetic data

Johannes Bogner BSc.

Vienna, 19.08.2020

I would like to thank Stefan, Olha and Thomas for their great support.

I would also like to thank my father Walter and my mother Heike. Thank you for supporting me throughout my studies.

Table of contents

List of figures

List of tables

1 Abstract

The authorities have adopted rules to guarantee the privacy of each individual. This can be a challenge for users of personal data, such as marketing applications or market research. General Data Protection Regulation (GDPR) as effective of May 2018 has led to several ideas for solutions that increase privacy for customers while ensuring the utility of the data for companies. A new approach to marketing analytics is the synthesis of data. This study aims to determine whether companies of various industries see potential applications of synthetic data for market(ing) research activities. Moreover, the study provides deeper insights into the requirements for synthetic data for marketing procedures in the form of a use case analysis. Semi-structured interviews with companies are conducted as a qualitative research method. Using a qualitative research approach, empirical findings contribute to state-of-the-art knowledge of marketing managers and their perception of synthetic data. Regarding managerial value, the research outcomes are relevant for companies that use marketing analytics and data-driven innovation models. On top of that, it can also be of interest for policymakers and experts in the field of synthetic data. Moreover, future research topics in this domain are presented in the discussion.

2 Keywords

- GDPR
- Marketing analytics
- Comparative study
- Synthetic data

3 Introduction

With the enforcement of the GDPR in 2018, a landmark privacy law that illustrates the implications of corporate responsibility for the confidential handling of data (Goldberg et al., 2019), data privacy issues have globally gained importance. As a result, the state of California for instance, has passed a new privacy act that went into effect in 2020. Companies had sharpened their perspective on privacy concerns about customer data and the role and value of data analytics in privacy concerned environments are under scrutiny. Several studies have been conducted which illustrate the effect of the GDPR on business activities and public perception (Larsson and Teigland, 2020; Menon, 2019; Schweigert and Greyer-Schulz, 2019).

Moreover, the increasing environment of data regulations is a challenging obstacle for companies' marketing activities. These tasks result in higher cost for personal digital marketing channels and a change in customer behavior (Goldberg et al., 2019; Holtrop et al., 2017). Regulations also affect procedures that are vulnerable to breaches of the GDPR. Digital marketing activities such as churn management programs and user profiling are particularly susceptible to GDPR violations (Holtrop et al., 2017; Larsson and Teigland, 2020; Wedel and Kannan, 2016).

As a consequence of growing data protection, several implications have been developed to resolve vulnerable activities that violate the GDPR. Opt-ins as a way to allow third parties to disseminate data have become ubiquitous following the enforcement of the GDPR (Wieringa et al., 2019). Consumers have to opt-in to any program that collects personal data such as demographic information or purchase and clickstream histories. By opting in, they need to give their specific consent to a set of policies that regulate how that data can be used, traded or sold (Johnson et al., 2001). Wieringa et al. (2019) illustrate different methods to deal with big data while protecting privacy concerns, thus benefitting from the available information. Other developed approaches are anonymization techniques, injection of random noise, and aggregation (Drechsler and Reiter, 2010; Reiter, 2004; Schneider et al., 2018, 2017).

Wieringa et al. (2019) illustrate a clear picture of the personal data responsibilities amongst the three implementation levels – customer, intermediary, and firm. Their work distinguishes personal data responsibilities into 1) data collection, 2) data verification, 3) data storage and control, 4) deriving insights, and 5) disseminating insights. On the basis of this structure, they compare the levels of responsibility for the implementation of personal data protection legislation and present future research proposals. One of these propositions is a further analysis of the importance of synthetic data. Especially, Wieringa et al. (2019) propose to develop better approaches for generating synthetic data that are close to real-world data to use it for marketing research purposes. They ask for advanced models or machine learning approaches that make the data more valuable for marketing activities. Companies like Mostly AI[1] developed a possible approach for generating synthetic data. To answer the question how companies perceive synthetic data for market(ing) research activities the identification of the specific data requirements that must be met to be of value to companies' marketing activities is crucial – accuracy, legal and privacy, technical, frequency, latency, constraint, and any other including marketing specific requirements.

Marketing and statistical literature from journals such as the Journal of Marketing, the Journal of the American Statistical Association and Marketing Science present solutions to solve privacy issues. Schneider et al. (2017) demonstrate that there are several methods to protect customer privacy in an increasing environment of data regulation and data security. Considering these methods and investigating a new approach to synthesize data for marketing purposes, this thesis provides practical insights by companies that share their expertise in semi-structured interviews. In the course of the research project (ANITA: ANonymous bIg daTA[2]), funded by the Austrian Research Promotion Agency (FFG[3]), we are investigating together with the Viennese tech start-up Mostly AI, whether companies can process synthetically generated data for marketing purposes and the requirements that must apply. To evaluate the relevance of synthetic data and to specify these requirements, a use case template was developed by research associates and the management of Mostly AI. This template was used as a semi-structured interview guide provided to participants to gain insight and can be found in Appendix 10.2.

[1] Mostly AI
[2] Project ANITA
[3] FFG

4 Research gap and research question

There is an existing gap in extant literature that this thesis aims to answer. The relatively new developments in an increasing environment of data regulations have brought forth exciting approaches within the last fifteen years. One approach has been used for the dissemination of data from statistical agencies in 2003 already (Reiter, 2005). Synthesis of data seems to be a promising approach that protects personal data and assures the usefulness of the data (Surendra and Mohan, 2017). However, synthetic data have not yet been investigated within the marketing domain. Therefore, this study aims to answer the following research questions:

How do companies perceive synthetic data for market(ing) research activities?

What are the requirements for synthetic data concerning accuracy and privacy?

On the one hand, the goal of empirical research is to present use cases of companies that see an application of synthetic data for their market(ing) research activities to answer the research questions and to identify future research topics. On the other hand, a structured overview of state-of-the-art literature is presented in the literature review.

5 Background

This section provides a brief overview of the current state of the literature. In the beginning, the most important terms are defined to ensure consistency of terminology, followed by the effects of the GDPR on companies' data privacy since its enforcement in 2018. Besides, the impact of these regulations on marketing activities is outlined in more detail. The effects include the marketing activities that tend to violate the GDPR, which are presented in Section 5.3. Also, approaches are outlined that help to address the privacy concerns associated with such activities. Finally, the data synthesis approach is presented.

5.1 Definitions

In this chapter, the terms marketing analytics, data protection, personal data, processing of data, and synthetic data are defined with regard to consistency throughout the thesis.

Marketing analytics (MA) is the method for measuring, analyzing, predicting, and managing marketing performance to maximize effectiveness and return on investment (ROI) (Wedel and Kannan, 2016).

Privacy has been defined in several environments. For this thesis, data protection is defined as controlling the dissemination and use of consumer information, which includes but is not limited to demographic information, search history, and personal profile information (Martin and Murphy, 2017).

Article 4[4] of the GDPR defines personal data as “any information relating to an identified or identifiable natural person,” and specifies further that “an identifiable natural person is a person 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.”(Wieringa et al., 2019, p. 2)

Moreover, article 4[5] of the GDPR also defines processing as "any operation or set of operations which are performed on personal data or sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration,

[4] Art. 4 GDPR
[5] Art. 4 GDPR

retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction."

“The fundamental idea of data synthesization involves sampling data from a pre-trained statistical model, then release the sample data in place of the original data. Synthetic data can be used in preserving privacy and confidentiality of the original data.” (Li et al., 2014, p. 4) Moreover, “the data are randomly generated with constraints to hide sensitive private information and retain certain statistical information or relationships between attributes in the original data.” (Surendra and Mohan, 2017, p. 95)

5.2 The GDPR and its effects on companies’ data privacy

Generated data have increased tremendously within the last decade (Menon, 2019). This development requires better approaches and methods to understand and process data. Moreover, the growth in the generation of personal data goes hand in hand with increasing data protection. (Menon, 2019). Therefore, companies in Europe and to some extent in the US have changed their privacy regulations. Emails are sent out by companies that inform about updates to privacy policies and a "change of terms". Moreover, pop-up links that ask consumers to opt-in for consent are some examples of data privacy policies forced by the GDPR. Companies are now required to get customers' consent to process their data, and a survey by Deloitte discovered that 86 percent of consumers believe they should be able to opt-out of the sale of their data (Sides & Rob, 2019). The increasing awareness of customers about data privacy results mainly from privacy breaches (Hart, 2018). One of the most famous examples of a privacy breach is the Cambridge Analytica case of Facebook in 2013. Even though Facebook ceased 200 apps in 2014 after its policy change, recent revelations show that the company was still sharing customer data with companies such as the Royal Bank of Canada and Nissan Motor Co., Ltd. (Menon, 2019). These revelations raise more questions about the credibility of companies, such as Facebook and their actual implementation of stricter privacy regulations.

Things might be different within the European Union after the enforcement of the GDPR as of May 2018. The GDPR regulates the legal foundation for the processing of personal data in article 28[6]. Companies may process data to fulfill contracts or legal obligations and safeguard the public or vital interests of individuals. In any other case, customers have to affirmatively and freely give their consent. This consent has to be granular to its purpose of processing data

[6] Art. 28 GDPR

and must state all third parties that process their customer data (Goldberg et al., 2019). Moreover, the GDPR is designed to synchronize the regulation for processing personal data by corporations and public authorities throughout the EU and it involves all individuals inside the EU and the European Economic Area (EEA). Additionally, it determines the export of personal data to countries outside the EU and EEA (Schweigert and Greyer-Schulz, 2019).

The seven key principles of the GDPR are the following: 1) lawfulness, fairness, and transparency; 2) purpose limitation; 3) data minimization; 4) accuracy; 5) storage limitation; 6) integrity and confidentiality, and; 7) accountability. These principles are stated in articles five to eleven of the GDPR (European Parliament, 2016). Moreover, it provides the following rights for individuals: 1) the right to be informed; 2) the right of access; 3) the right to rectification; 4) the right of erasure; 5) the right to restrict processing; 6) the right to data portability; 7) the right to object and 8) rights concerning automated decision making and profiling. These rights are presented in articles 12 to 23 in the GDPR (European Parliament, 2016). Thus, the GDPR has a wide range of implications for all industries that need to collect personal data. Even though "big tech" companies, such as Google and Amazon, are obviously more affected, it is likely that Small- and Medium-Sized Enterprises (SME), such as developers of online games, will be hit more strongly. Moreover, companies in the healthcare sector and companies from the bank and finance sectors are affected (Larsson and Teigland, 2020). Therefore, the GDPR has many ramifications from a legal point of view, but also with regard to the marketing activities of companies.

5.3 Impact of the GDPR on marketing activities

As mentioned before, there are many implications, especially for marketers, who now have to explain to their customers what they are collecting the data for, how data are processed, and to what extent the data are passed on. As firms have to minimize personal data processing, they are required to get their customers' consent, which increases the costs of gathering web analytics data. A survey of consumers recently conducted by Deloitte revealed that 91 percent of them give their consent to the terms and conditions and privacy policy without reading them (Cakebread, 2017). Moreover, clickstream data are crucial for personalized digital marketing channels, which results in higher costs for those channels (Goldberg et al., 2019). Goldberg et al. focus on the effects of the GDPR on European web traffic and e-commerce outcomes. They describe the GDPR as a groundbreaking data protection law that protects individuals' privacy but thereby harms companies that rely on marketing analytics for decision-making and tailored marketing. The authors assume that the enforcement hurts online companies directly, as it confines online advertising and indirectly, as it moderates web analytics data. As one of the pioneers to examine the GDPR, the authors leverage the timing of enforcement as an event study using Adobe Analytics data. They both apply difference-in-differences and synthetic control models to determine the effects of the GDPR. The study covers 1,500 online companies, including 128 top global websites offering a wide range of content, e-commerce, and corporate sites. Their results show quantifiable impacts of the GDPR on critical economic outcomes such as page views, visits, orders, and revenues. The page views per week decrease by about four percent and the income per week by eight percent. These figures are high from an economic point of view since an eight percent drop in income per week corresponds to an average $8,000 drop in the weekly income of their sample. These results represent the difficulty and high costs of privacy regulation for companies, but do not include quantifying benefits to users of these privacy laws. Goldberg et al. postpone the understanding of the tradeoffs to future research.

In their study of 2019, Schweigert and Greyer-Schulz state that the new regulation affects all types of marketing activities. The authors are comparing the negative and positive impacts of the GDPR. Marketers need permission to contact customers, and they may only collect, process and store the necessary data. Another study from Holtrop et al. in 2017 illustrates that firms that do not pay attention to the increased awareness of data protection issues are confronted with adverse consequences such as customers' loss of confidence, customer behavior changes, or unfavorable stock market ratings. Although legal restrictions and public pressure to reduce the storing of excessive amounts of personal data are challenging, there are also positive aspects for marketers. Since the individual must give his or her consent, this self-selection enhances the performance of subsequent marketing activities and avoids annoyance to consumers. Moreover, customers get a feeling of security and trust, which is fostering the relationship to the firm (Schweigert and Greyer-Schulz, 2019).

Menon (2019) indicates that in 2018 a mere 36 percent of marketers have heard of the GDPR and approximately 15 percent have taken little preparatory action, thus, creating the risk of non-compliance. According to a different study, four weeks before the deadline, merely 28 percent of companies considered themselves fully compliant, whereas roughly 47 percent were confident that they would meet the deadline (“Marketing Post-GDPR: Two Tribes of Marketing,” 2018). Sirur et al. (2018) state that in a study carried out across the EU, UK, and the US, only one out of five organizations considered themselves fully compliant in 2018. A more recently published article from the International Conference on Text, Speech, and

Dialogue by Müller et al. in 2019 states that over 76 percent of the real privacy policies that were analyzed in his work did not contain all the requirements and therefore may not be fully compliant with the GDPR. The authors evaluated their data set based on five GDPR requirements, which were categorized into data protection officer, purpose, acquired data, data sharing, and rights. Their data set consisted of 250 privacy policies with a total number of 18.300 sentences. As in those above stated, the effects on marketing activities are tremendous. Even though some studies already try to set the focus on possible positive effects of the GDPR on such activities, there remain marketing activities that are prone to violate the regulations.

5.4 Marketing procedures that are prone to violating the GDPR

The big data hype for some companies turned out to be less useful as they might have thought (Wedel and Kannan, 2016). A possible reason for this minor usefulness is higher investments in data storage and capture and not into analytics. Wedel and Kannan (2016) investigated several marketing analytics methods and their potential to support marketing decisions. Considering data as "the oil" of the digital economy, their key domains for application of analytics are 1) customer relationship management (CRM), 2) marketing mix, 3) personalization of this very same, and 4) privacy and security. Based on these domains, the findings for a successful elaboration and application of marketing analytics in companies are two-fold. First, the organizations need to adapt structures and cultures that promote data-driven decision making, and, second, in order to make the most out of it, analytics professionals need to be educated and trained. The authors state that firms collect data from various sources and combine them to understand their customers better. This combination of data sets reveals information on consumers that should be private. For example, online reviews can be supported by recommendation algorithms and help companies refine their offers and increase their value proposition. Moreover, mobile retail analytics might support providing better recommendations or personalized promotions, thereby increasing spending. Nevertheless, these recommendation systems might process data that have been received inconspicuously from customers (Wedel and Kannan, 2016).

In their book of 2020, Larsson and Teigland illustrate various risks that marketing activities face about the GDPR. First of all, predictive analytics are suspect of determination decisions on people's creditworthiness or even job applications, which is part of discrimination. Moreover, data breaches may lead to the disclosure of personal information of customers and clients. Additionally, by combining several subsets of data it is possible to re-identify a person if only a few data sets are removed. The authors also mention the point of data brokerage, namely the sale of possibly unprotected or faulty data subjects. Another point that is mentioned is the storage of data on cloud services. The digitized world shows a shift to personal data stored on cloud services with an approximate number of ten trillion gigabytes in 2019 (Darra Hofman et al., 2017). This development requires secure servers and confidence in data controllers' capacity to ensure the correct management of the data to prevent the information from falling into the wrong hands. Thus, data storage can be seen as an activity that is prone to violate the GDPR.

France and Ghose (2019) illustrate several business areas that benefit from marketing analytics. Their study is based on the topics of visualization, segmentation, and class prediction. Churn prediction is a marketing activity prone to violate the GDPR as such programs simulate customer retention, thus predicting customer lifetime value (CLV). Holtrop et al. state in their study of 2017 that increasing amounts of stored customer data for churn management programs lead to increased public awareness, which raises the question of how the need for marketing analysis can be reconciled with the protection of customer privacy. Companies either violate the regulations or include measures to protect privacy at the expense of analytical operations.

In summary, many marketing activities, especially digital marketing activities, such as the use of recommendation algorithms, a combination of data sets, data brokerage, or data storage, are prone to violate the GDPR. Therefore, to overcome these potential violations of privacy, new approaches to data processing are on the rise.

5.5 Approaches to the prevention of GDPR violation

Due to the enforcement of the GDPR and other data protection regulations, research has been conducted to deal with these evolving regulatory requirements. Drechsler and Reiter (2010) illustrate several approaches to deal with stricter data protection regulations. First, random noise injection describes a method where sensitive or identifying values can be disguised by deviation of the true values by some noise, e.g. sampled noise from a normal distribution with zero-mean value. By injecting random noise to data sets, potential data that has been received inconspicuously might be dissolved. To ensure confidentiality companies may have to resort to a widely dispersed distribution which results in decreased utility of the data (Drechsler and

Reiter, 2010). Second, data swapping allows agencies to exchange data values for targeted records, such as swapping age, race, and gender values for sensitive records with those for non-sensitive records. To prevent users from re-matching, it can be based on inaccurate data. Data swapping is an approach that helps to deal with data breaches and re-identification, as it assures confidentiality. Joining a statistical database might be a concern for individuals, as the curator gathers sensitive information. Differential privacy implies that the calculations must be insensitive to changes in the dataset of a particular person, which limits data leakage by the results (Dwork and Roth, 2013). Nevertheless, high-level swapping disrupts relationships that affect the variables that are exchanged and those that are not (Drechsler and Reiter, 2010). Even modest swaps can be complicated and based on data from the Survey of Youth in Custody, Mitra and Reiter (2006) observed that a 5 percent random swap of two identifying variables leads to low confidence interval coverage rates for these variables for regression coefficients. This means a loss of data utility, as the distributions and statistical properties are distorted severely (Winkler, 2005). Third, top coding, where e.g. financial variables and age are displayed with top codes and occasionally also with bottom codes. By definition, top or bottom coding censors values above a pre-selected 'top code' or 'bottom code'. For instance, in surveys that measure income, very high income ratings are perceived as the most sensitive and have the potential to disclose the respondents' identity. Therefore, sensitive income ratings that are higher than the threshold are recoded (An and Little, 2007). Top coding might solve the problem of discrimination within predictive analytics. However, Kennickel and Lane (2006) illustrate that frequently used top codes bias assumptions about the Gini coefficient, a crucial measure of disparity in income. Fourth, aggregation mitigates disclosure risks by transforming atypical records - which are generally the most vulnerable - into typical records. For example, a county may have only one person with a particular combination of demographic characteristics, but a country may have many people with these characteristics (Drechsler and Reiter, 2010). Aggregation is one solution for the issue with data brokerage.

Where missing data are involved, agencies usually use multiple imputations to deal with these defects while protecting confidentiality. Sharing data between different stakeholders is not only beneficial for the parties themselves but also for the general public. For instance, within the healthcare industry, sharing data can fasten the development of medication for diseases. Aggregation makes it difficult and often impossible to analyze at a more detailed level and leads to problems drawing economical conclusions. Another example is the sharing of market research data from companies such as AC Nielsen towards retailers in an aggregated form.

Although data aggregation is protecting privacy, there is a tradeoff between privacy and utility of the processing of this data (Schneider et al., 2018).

Back in 2004, Reiter already suggested an idea to use a regular partial synthesis setting. In his study, Reiter notes that statistical offices used or considered multiple imputations to reduce the risks of disclosing the identities of respondents. As these agencies release microdata in public use files, they aim to disseminate them safe from attacks but keep them still informative for statistical analyses and easy to handle for standard statistical methods. A major drawback of survey data in terms of its practical implications is that most surveys lack data from units that do not respond to some or all of the items. Reiter illustrates a multiple imputation model that is dealing both with missing data and disclosure limitations. However, the findings state many challenges about partially synthetic data models. Reiter mentions the tradeoff between disclosure risk and data utility. In further studies, Reiter and other authors investigate the possibilities of fully synthetic data.

5.6 Overview of methods for data synthesis

Holtrop et al. 's study of 2017 presents a model to predict customer churn rates while keeping customer privacy. Thus, their method, a generalized mixture of Kalman filter (GMOK), illustrates that privacy protection does not go hand in hand with a loss of analytical procedures. In comparison to Ascarza & Hardie (2013), the model does not require storing of protection-relevant panel data at an individual level on customer behavior in the past but achieves data anonymization by aggregating information from previous periods into the model parameters. In fact, to update the model, it only needs to be filled with new cross-sectional information from the present period. They compare various methods such as the Hidden Markov Model, logistic regression, classification trees, and restricted versions of their proposed GMOK model, a dynamics only, and a heterogeneity only model. Nevertheless, their model rather implies a method to fulfill data minimization and data anonymization for preserving privacy. This is one approach to deal with the increasing environment of data security but mainly focuses on churn prediction.

The latest developments in technologies lead to firms' and organizations' feasible approaches to gathering, storing, and processing vast amounts of microdata. As the business world is interconnected and companies collaborate, personal data have a higher value if shared amongst various parties to foster research and innovation. To circumvent possible privacy and disclosure risks and to guarantee customers' safety, but still share raw microdata, synthetic data generation is a new approach. In recent years, research on synthetic data has been conducted both focused on privacy concerns due to publishing and the support of validation of algorithms and applications as they can be trained with such data (Surendra and Mohan, 2017). Table 1. illustrates various methods of data synthesization which are mentioned in this study. A more elaborated overview of methods for data synthesis is presented by Surendra and Mohan (2017). In their article, they illustrate the different types of synthetic data, including their limitations and techniques. One type of published synthetic data are partially synthetic data. In the process of generating these, only those values that carry a high risk of disclosure are replaced with synthetic values to hinder re-identification. On top of that, these methods allow imputing missing values in the original dataset. Partially synthetic data, that impute missing values, however poses a risk of disclosure, as the data set includes original data and imputed synthetic data. Therefore, fully synthetic data generation is a better approach concerning disclosure and privacy risk. Multiple imputations and bootstrap techniques are used to generate a completely synthetic data set that does not contain any original data. Although this process goes hand in hand with strong privacy protection, the data's truthfulness is influenced. Last but not least, hybrid synthetic data generation is a middle road between the former two methods. Therefore, the author state that hybrid synthetic data delivers adequate privacy protection with a high utility of the generated data. On top of that, they conducted a comparative study of synthetic data generation methods from 2006 to 2016, illustrating the type of generation, the proposed method, and the study's limitations. In conclusion, Surendra and Mohan define the need for the development of synthetic data generation techniques that ensure individuals' privacy and have a high utility for further application of the data.

In his paper of 2005, Reiter presents an empirical study on the disclosure of fully synthetic microdata. The author uses models based on data from the US Current Population Survey:

(a) to assess the potential validity of conclusions based on fully synthetic data for several descriptive and analytical estimators,

(b) to evaluate the level of confidentiality protection provided by the synthetic data as a whole; and

(c) to illustrate the requirement to specify synthetic data implementation models.

By publishing fully synthetic data, the risk of re-identification is practically non-existent. Nearly none of the published synthetic entities are included in the original sample because they were randomly drawn from the sampling design. Their data collection values are simulated so that no sensitive values are released for these entities. Besides, synthetic data records cannot be meaningfully compared with other data records. For instance, administrative data, as the values of the variables released for the survey, are more likely to be simulated than actual values and are thus not the same as those in administrative datasets (Reiter, 2005). Reiter demonstrates the relevance of the specification of exact imputation models when creating fully synthetic data.

Table 1.

Author	Year	Method
Haoran Li, Li Ziong, Lifan Zhang and Xiaoqian Jiang	2014	• Differentially private data synthesizer "DPSynthesizer" • The data is modelled by creating histograms for each characteristic of the original data set • A dependency matrix is created with a Gaussian copula function based on the original data
Patki, N., Wedge, R., Veeramachaneni, K.	2016	• Synthetic data vault (SDV) • A state-of-the-art multivariate modeling approach is used • All possible relations are iterated that create a model for the entire database • This method allows to sample from any part of the database, thereby synthesizing data using a Gaussian Copula process
Karras, T., Aila, T., Laine, S., & Lehtinen, J.	2018	• Generative adversarial networks (GANs) • These networks start with low resolutions and add new layers for finer details as training progresses, which allows to create images of unprecedented quality
Schneider, M.J., Jagpal, S., Gupta, S., Li, S., Yu, Y.	2018	• Bayesian probability model • The model uses a decision criterion, which allows to protect variables that have the most power to predict store IDs

Table 1: *Methods for data synthesization*

5.7 Synthetic data for marketing activities

Data synthesis is relevant concerning the marketing domain in various applications. In their study of 2019, Nabbosa and Iftikhar state that synthetic data can be processed to predict customer shopping behavior and provide customized services without risking the potential identification of individual data subjects.

Starting in 2014 already, Jarmin et al. illustrated the positive impacts of synthetic data for the US CENSUS BUREAU. In their study, the authors discuss the benefits and challenges of increasing the range of synthetic data products in official statistics. The authors illustrate that public-use microdata can be released if protected by synthesization of the data. Nevertheless, various users have different requirements and, therefore, no single method is available to satisfy the needs of these users while protecting confidentiality.

In their paper of 2016, Patki et al. propose a system that they call Synthetic Data Vault that allows them to generate as much synthetic data as is required, thereby keeping the same structure and format as the original data. They created three versions of data for the comparison with different noisy conditions. The first condition was synthesized data without noise. The second condition was synthesized data with table noise, which effectively reduces the strength of covariance. The third condition was synthesized data with key noise, which randomly samples a primary key. The comparison of the control group and the different conditions demonstrate utility for predictive models using this synthetic data. The method applied allows the synthesis of data in two categories. The first one is model-based and allows the company to synthesize a complete database of customer information. The second category is knowledge-based and allows the user to synthesize information, for instance, particular types of customers (male, 35, married). Their study shows that there is no statistically significant difference in the accuracy values between control data and synthesized data. The results indicate that the work to be done with synthesized data can be as productive as with control data (Patki et al., 2016).

Schneider et al., in their article of 2018, propose a method where the data provider can offset the tradeoff between information loss arising from data protection and the risk of disclosure to intruders. The authors state that data providers are not only motivated to protect data because of legal reasons but also because protection is an important pillar for data providers' brand positioning. Outcomes of their study illustrate that the approach of synthetic data outperforms seven benchmark data protection methods – namely, "true" or unprotected store-level data, random noise, rounding, top coding, 20 percent swapping, 50 percent swapping and market-level. Their model allows the data provider to choose the data protection preference upfront, which influences the information loss. Their model incorporates three parties: the data provider, the data processer, and a potential data intruder (Schneider et al., 2018).

Another article by Schneider et al. (2017) focuses on protecting customer privacy when marketing with second-party data. To get the most out of customer data, companies try to boost their value by augmenting their information with customer-level data from a second company. This strategic approach of data sharing is a widespread marketing initiative throughout several industries. However, certain events, such as the Cambridge Analytica scandal, can be very harmful to data providers. Therefore, the authors came up with a decision-theoretic approach that protects customer data before entering data-sharing agreements. Using synthetic segment memberships for each client and not true segment memberships, the customer list is protected against a breach of an intruder. This approach provides managers with the opportunity to decide on the tradeoff between profitability and privacy, based on data protection level, expected profits, misclassification costs, and estimated data breach costs.

To sum it up, there already exist several applications for synthetic data in marketing, but there are still many other potential ways to explore. As Wieringa et al. in their study of (2019) state, one area of great potential is the development of better models for the generation of synthetic, high-dimensional data at the individual level that simulate real entities. Considering the background of state-of-the-art literature, the empirical part of this thesis investigates the requirements for synthetic data in the following.

6 Methodology

The methodology that was applied to conduct this master thesis is a qualitative research design. In qualitative research, scholars must understand the subjective and socially construed meanings conveyed by the investigated phenomenon. This thesis is based on both primary and secondary data research. The qualitative approach used for gathering primary data is mainly through semi-structured interviews, whereas secondary data were gathered in a literature review through articles, books, journals, and databases. This chapter portrays the method of a literature review to gather secondary data and semi-structured interviews to gather primary data.

6.1 Literature review

Secondary data were collected in a six-step approach to define and classify relevant literature. (1) Scoping of literature, (2) application of search tools, (3) selection of appropriate journals, (4) implementation of forward and backward search, (5) title and abstract screening, (6) classification of articles into categories (Booth et al., 2016).

First of all, the literature was searched in broad outlines to get familiar with the topic and volume of existing literature. Thus, the scoping process consisted of the search in the database of Google Scholar with terms "synthetic data", "marketing analytics", and "marketing activities" which resulted in 221 papers. In addition, the quantity was narrowed down by using the EBSCO Host's business source premier databases with the same terms, thereby also adding similar phrases such as "fully synthetic data" and "market research activities". Throughout this process, several search terms particularly important for answering the research question were identified, what is introduced in the next paragraph. Applying this keyword-based approach helped to identify relevant literature for this thesis.

The terms synthetic data or partially/fully/hybrid synthetic data are substantial to cover the main aspect of data synthesis. The terms marketing, marketing analytics, marketing activities, and market research activities incorporate the field of interest based on the business and, especially, the marketing literature. The terms data regulation, GDPR, and data security implement the legal perspective. Even though every keyword search has its limits, the preferred selection covers the most relevant literature for this thesis.

Second, a Boolean phrase was applied to the Google Scholar search tool and EBSCO Host's business source premier databases. The period from the last 15 years from January 2005 to the latest issues in 2020 was considered to research the most recent literature on synthetic data. However, some studies older than 2005 which contribute to this topic are included to create a transition from the past to the latest research. The Boolean phrase "synthetic data" AND "marketing" OR "marketing analytics" AND ("GDPR" OR "data security") used with EBSCO over the past 15 years identified 39 relevant journal articles.

Third, subject areas were identified to set the scope for relevant journals of this thesis. The main topic areas of this research include marketing, data science, and law. The Academic Journal Guide published in 2018 by the Chartered Association of Business Schools (CABS) consists of several subject areas that offer structural support in the classification of literature into specific research areas. Subject areas for this thesis are "Marketing Management", "General Management, Ethics, Gender, and Social Responsibility" and areas from the technology and law domain. The incorporation of research from these various subcategories of journals was crucial to ensure the complete coverage of high-quality research with a focus on marketing analytics. To ensure the quality of this work, refereed journals of the CABS with a ranking of at least three were primarily used, including journals such as the International Journal of Research in Marketing, Journal of Business Research and Marketing Science. However, some studies published in other journals, which were identified by more in-depth analysis and are highly relevant to the research topic, were supplemented to the sample.

Fourth, forward and backward search was added to work out further articles that are highly relevant to the research topic. Forward search examines articles that cite the initial article, whereas backward search examines articles that are listed in the references of the initial article. Key databases offer bibliographic mining and citation searches as a helpful tool for identifying appropriate literature. Examining the 39 potential articles with this approach was crucial to detect several studies that were not discovered by the keyword-based method but relevant to the research topic. These were added to the initial potential sample.

Fifth, through title and abstract review, the articles selected in the first four steps were investigated, which assisted the final selection of relevant or irrelevant articles. Based on their abstracts, the literature had to indicate interconnections to the three aspects listed in the search terms. As a result of this examination, 28 articles were considered relevant to the research topic and included in the literature review.

Lastly, in order to structure the review, a classification into categories was used. In the course of the whole process of in-depth analysis of the relevant articles, six categories with thematic similarity emerged. (1) The GDPR and its effects on companies data privacy, (2) impact of the GDPR on marketing activities (3) marketing procedures that are prone to violating the GDPR, (4) approaches to the prevention of GDPR violation, (5) overview of methods for data synthesis, and (6) synthetic data for marketing activities. This data in a structured line is highlighted in the literature review of this paper and builds the foundation for the further examination of primary data and its collection. Moreover, it is a factual basis for a critical discussion of the literature and the outcomes and results of the interview data.

6.2 Semi-structured interviews

After building the foundation with a critical discussion on the state-of-the-art literature, the methodology is illustrated for gathering the primary data. In order to cover the empirical part of this thesis, mainly semi-structured telephone interviews were conducted to answer the research question. In the following, the approach for the empirical part of data collection and analysis is outlined.

Saunders et al. (2012) propose three types of research designs which are exploratory, descriptive, and explanatory. Exploratory research designs are used to gain new insights and examine the development of new approaches. Descriptive studies are used to establish an accurate image of incidents, people or situations. Explanatory research investigates causal relationships between variables (Saunders et al., 2012). For this study, semi-structured interviews are conducted, and mainly open questions are asked to discover what is happening and gain insights into synthetic data as a subject of marketing research. An exploratory approach, therefore, seemed the most appropriate.

To answer the research question, in-depth interviews were conducted. An in-depth interview is a qualitative technique in which one talks to individuals to determine their perspectives. This method is useful when the interviewer wants to obtain detailed information about the thoughts of a person (Booth et al., 2016). In a semi-structured approach, the questions are pre-formulated but not strictly followed. These questions are already coded beforehand and allow a better structuring of the data afterward. Furthermore, new questions may arise during the interview, which evaluate the topics in more detail. For each interview, the author sticks to the questions throughout the process, thereby assuring consistency. On the one hand, this might decrease the number of new insights. On the other hand, it combines unstructured and structured interviews, which decrease risks and save time (Saunders et al., 2012).

The approach of semi-structured interviews allows gathering information through relatively open questions. Moreover, the outcomes of semi-structured interviews are more beneficial as interviewees are not limited when answering the questions. Additionally, the interviewer can guide the interview in a particular direction, thereby not losing focus on what is crucial to be answered. Personal interviews are the best method as they allow more precise insights and more profound personal experience to the topic (Saunders et al., 2012), but, due to the current situation of COVID-19 as the interviews were conducted, the approach was changed to telephone and virtual interviews mainly. Gathering data through interviews provides meaningful insights that might differ from the information detected by the literature beforehand. Thus, the newly generated data present a different outcome and can be used for the discussion and conclusion of the thesis with implications for managers, policymakers, and experts.

6.3 Collection of data and analysis

As mentioned in the section before, semi-structured interviews are the method for data collection. To be able to conduct such interviews, possible interview partners have to be identified. In the next step, these partners have to be approached and acquired. Moreover, data have to be collected throughout the interviews. Lastly, the collected data must be analyzed with a specific approach, illustrated in section 6.3.3. This chapter presents the process of data collection and analysis.

6.3.1 Identification of possible interview partners

In order to collect data for the qualitative research design in this study, partners for interviews from various industries were to be defined. In the first step, relevant industries with specific criteria were selected based on the Statistical Classification of Economic Activities in the European Community, commonly referred to as NACE, classification rules[7]. For the scope of the project, industries had to use personal data on the one hand. On the other hand, data protection regulations such as the GDPR had to influence companies within these industries. In close collaboration with the project partners and after thorough analysis, seven industries seemed to be most relevant for this study. These industries can be found in Table 2. Based on

[7] NACE classification rules

these industries, the identification of five companies as key players in each of these industries led to a further specification of possibly relevant partners. Additionally, due to geographical reasons and the scope of the study, these key players are mainly located in the Austrian area with their headquarter in Vienna. For each of these companies, at least one person of contact with expertise in data analytics was identified. Possible key players with either one or more persons of contact lead to 93 different possible interview partners.

Table 2.

1	Wholesale and retail
2	Electricity, gas, steam and air conditioning supply (Energy) (A, C, I)
3	Telecommunications (B, K, L)
4	Financial service activities (Finance) (G)
5	Insurance (F, H, J)
6	Human health activities (Health) (E, M)
7	Transport and storage (Mobility) (D)

Table 2: *Industries and the respective interviewees in brackets.*

6.3.2 Approaching the interview partners and data collection

After identifying potential contacts for interviews, the crucial step of approaching them in the right way had to be set up. The attempt to contact the interview partners via LinkedIn was unsuccessful and did not lead to the expected outcome to gather interviews. Therefore, a short text in the form of an article was created to draw the attention of potential partners. On top of that, a short interview guide with eleven questions for phone calls was drafted to evaluate companies' attitudes towards the topic (Appendix 10.1). Thorough research through LinkedIn and the companies’ websites helped to identify the respective contact person’s details to get in touch. In the first wave, at least one contact person of each company was approached via email from the 7th of April to the 8th of April. The first wave was not successful as many companies were either strongly affected by the COVID-19 crisis or the contacts were already on the Easter holidays. In a second wave, the contacts that were not approached in the first wave were addressed after the Easter holidays from the 15th to the 17th of April. However, many potential partners still did not answer to this email approach. Therefore, another email was sent later as a reminder of the project. On the one hand, 80 potential interview partners could not participate in the study due to the current COVID-19 situation or a lack of interest. On the other hand, 13 emails led to the next step of the data collection process, which is a short interview with an interview guideline in Appendix 10.1. These initial interviews were meant to figure out whether the potential contacts are willing to share a use case in a longer follow-up interview. The use case template can be found in Appendix 10.2. From these 13 interviews, four agreed on a follow-up interview to fill out the use case template. The 13 telephone interviews and the four use case interviews represent the basis of the data analysis.

6.3.3 Data analysis

After the interviews were conducted, the transcription in Microsoft Word followed. To comply with the GDPR, the names of the respondents were anonymized and they gave their consent to use the transcript for research purposes and analysis. This was crucial to meet data protection and to address privacy concerns. After the transcription of the audio records, the different interviews were assigned to respective cases. The cases themselves needed to be analyzed with a specific approach, which can be divided into five steps. (1) identification of categories or codes to be able to understand the data; (2) introduction of additional data from other sources to the respective categories and codes to integrate the data; (3) deduction of analytical codes to detect patterns and relationships; (4) development of verifiable proposals; (5) drawing and verifying conclusions (Saunders et al., 2012).

In a first step, categories or codes have to be identified. This method helps to rearrange original data into the respective analytical codes. Moreover, the process allows us to provide an emergent structure relevant to answering the research questions and organizing and analyzing the data in the next steps. As the interviews were conducted following an interview guideline, categories such as relevance, kind of data, impact of the GDPR and current approach were already predefined and an inductive approach was used from the beginning. Nevertheless, throughout the analysis of the data, other categories evolved, and the existing categories were developed further. Specifically, as the categories were predefined, the process of analyzing the data follows the method of template analysis. The main categories that were either introduced from the beginning or arose throughout the data analysis are stated in the following chapter.

7 Findings

Various findings have emerged from the analysis of the interviews. Company partners from different industries were asked questions concerning the topic of synthetic data as a new approach for marketing activities. In the first step, findings of the short telephone interviews are illustrated. In a second step, findings from the extended telephone interviews in the form of a use case discussion are evaluated.

7.1 General interest for synthetic data

Contact persons which agreed on a short telephone interview were asked several questions that should assure their general interest in the topic area that this study covers. Therefore, an interview guide (see Appendix 10.1) was used throughout the interviews. The guide covers the following topics: Relevance of personal data, type of data that help to improve service operations, the impact of the GDPR on companies, current approaches to deal with the GDPR, machine learning and AI as methods for data privacy, and, synthetic data as a way to foster marketing activities. In this section, the findings that were collected within these interviews are illustrated.

7.1.1 Relevance of personal data

In the first step, the interviewees were asked about the importance of personal data for their company. The participants were able to clarify what they refer to as personal data and ranked its relevance on a Likert scale from not at all (1) to very relevant (5) (Boone and Boone, 2012).

Interviewee C states that "in the context of digitalization and online business personal data becomes more and more important and the more data, the better it is. Nevertheless, the company sees that it is partly not able to process this data."

Interviewee A specifies that “[the company] can live without data and [they] try to work as much without it as possible.”

The outcome of the answers to this question is relatively similar throughout the different industries. A mean of 3.85 on the scale indicates that personal data are rather important for most industries. Only for the health industry, with one rating of (1) and one rating of (2) and one outlier in the energy industry, with a rating of (2), personal data seems relatively unimportant. Means for all industries can be found in Fig. 1. The total sample size is 13. The clarification of the relevance of personal data was essential to determine whether respective interview partners show interest in the research topic. Based on this general approach, the interview was continued.

Figure 1.

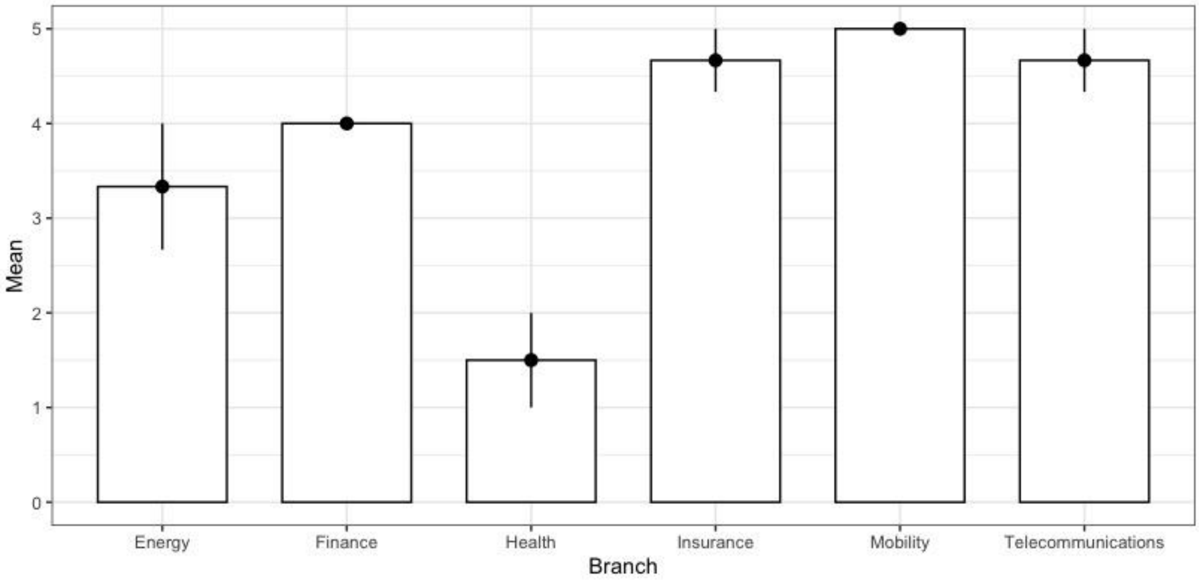

Figure 1: *Mean values of the relevance of personal data in the various industries from the sample of 13 respondents on a Likert scale from (1) not at all - (5) very relevant.*

Table 3.

Branch	Energy	Finance	Health	Insurance	Mobility	Telecommunications
	(n=3)	(n=1)	(n=2)	(n=3)	(n=1)	(n=3)
Mean	**3.33**	**4**	**1.5**	**4.67**	**5**	**4.67**
SD	**1.15**	**NA**	**0.71**	**0.58**	**NA**	**0.58**

Table 3: *Mean values and standard deviations (SD) from figure 1.*

7.1.2 Type of data which helps to improve service operations

In a second step, the interviewees were asked about the type of data which helps their respective companies to improve their service operations. More precisely, the type of data crucial to improve service operations, such as aggregated data, anonymized data, or pseudonymized data.

Throughout the interviews, it becomes obvious that for more than half of the companies (10) personal data are still playing an essential role in their service operations. Interviewee B illustrates that "with numerous customers [they] have agreements regarding marketing options that [they] also can contact them with individualized offers." This is consistent with interviewee F's statement that they need "personal data, which also allows conclusions to be drawn about personal preferences and customer orientation." Aggregated data are relevant for almost half of the companies (6). "Aggregated data are relevant for reporting purposes," as interviewee G declares. For five out of 13 interview partners, anonymized customer data are also crucial for the improvement of their services. As interviewee A states, "no more non-anonymized data are leaving the systems without consent." Moreover, for industries such as the mobility industry, booking data are also fundamental. Finally, third-party customer data are still very often processed in the fintech sector, but also in anonymized form. An overview of the types of data can be found in Figure 2.

Figure 2.

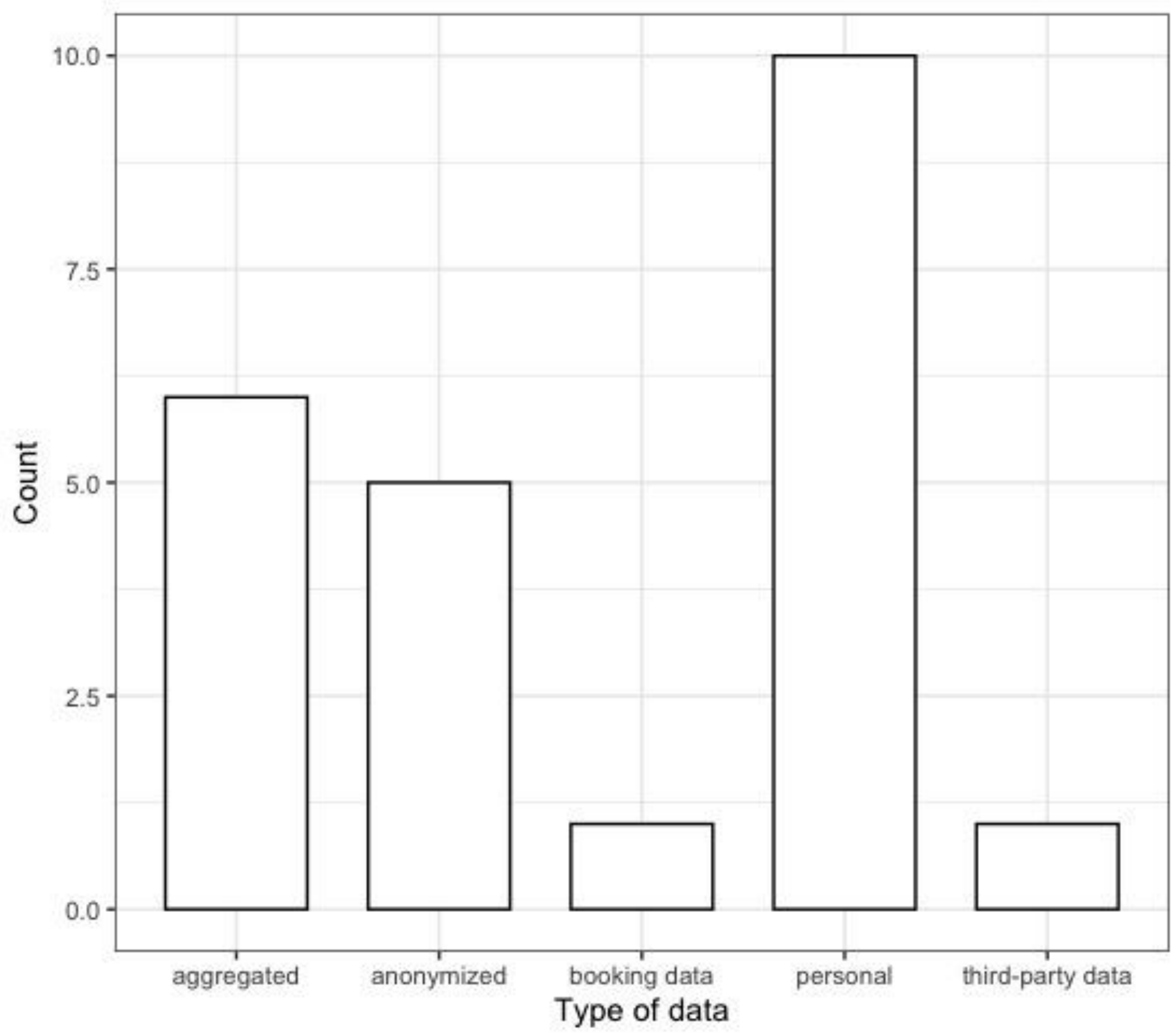

Figure 2: *Types of data that the respondents named to help improve service operations. Note that respondents could nominate more than one data type.*

It becomes obvious that for various industries, various types of data have more importance than others. However, the general picture shows that personal data are still essential for more than half of the companies. Moreover, aggregation seems to have gained a strong foothold in the data processing domain of many industries. After clarifying the types of data, the interview digs deeper into the actual effect of the GDPR.

7.1.3 Impact of the GDPR on companies

The third step is meant to figure out the effects that the GDPR has on industries and specific companies. Here, companies illustrate their changes in business activities and other implementations that were introduced as a result of the introduction of the GDPR. Moreover, the interviewees were asked to rank how strong the GDPR impact is on the company on a Likert scale from not at all (1) to very high (5).

Interviewee A states that the GDPR went hand in hand "with high investment costs. [They] would have to process all [their] internal systems like data output. Also, high administrative effort arose, people had to be asked, and surveys were started." Moreover, interviewee C also states the effect as "quite a challenge," but due to their different sectors, some parts of the company were facing massive restrictions, whereas others were nor really restricted. Interviewee F says, "the measures that had to be implemented due to the GDPR resulted in a very high effort regarding the processes." Interviewees G and H both faced the challenge of data storage and data deletion, which also includes the topic of cloud services. Companies are not allowed to store customer data in cloud services. Interview D and I state that only selected persons have access to the data now. Interviewee J mentions a "great deal of structure and responsibility has been built up and defined." Lastly, interviewee K says, "it has limitations on [their] abilities to serve the customers in a way [they] want, so [they] have to justify every customer relation and customer activity, to be GDPR compliant. So, in some cases, [they] cannot do some of the actions as [they] did before." Considering the question how strongly the GDPR affects the company on a scale from (1) not at all to (5) very high, it turns out that the mean is 4.17 throughout the interview partners. Only for one company from the energy sector there exists a rating that indicates a relatively low impact (2) of the GDPR. Interviewee E was not able to answer this question. In Figure 3, the impact of the GDPR on the industries is illustrated.

Figure 3.

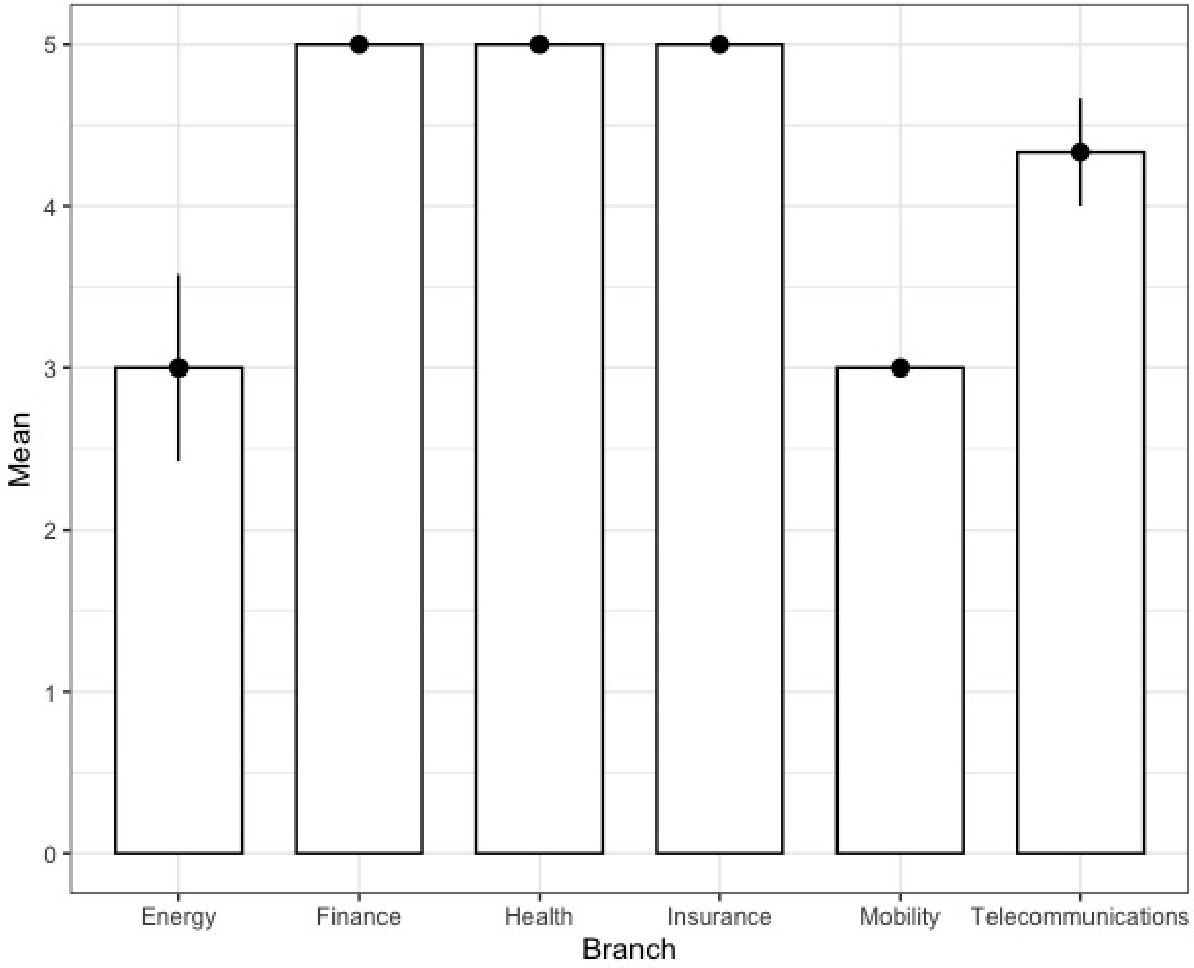

***Figure 3:** Mean values regarding the impact of the GDPR on the various industries on a Likert scale from (1) not at all - (5) very high.*

Table 4.

Branch	Energy (n=3)	Finance (n=1)	Health (n=2)	Insurance (n=3)	Mobility (n=1)	Telecommunications (n=3)
Mean	**3**	**5**	**5**	**5**	**3**	**4.33**
SD	**1**	**NA**	**0.71**	**NA**	**0**	**NA**

***Table 4:** Mean values and standard deviations from figure 3.*

Answers to the questions regarding the impact of the GDPR on companies indicate that companies from every industry were affected rather vigorously by the introduction of the GDPR. Moreover, the challenges that arose can be found across various industries.

7.1.4 Current approaches to deal with the GDPR

Step four tries to evaluate current approaches the companies pursue to deal with the GDPR and its restrictions. This step illustrates the way how companies are currently handling the regulations and already indicates whether a next step could be the synthesis of data.

Half of the interviewed companies state anonymization as one of their approaches to deal with the GDPR. Interviewee A states that "at the moment [they] are trying to make it anonymous, [they] do not have any other technology for that right now. [They] are at the moment still in the status that [they] basically cut away the information." Interviewee D talks about "new solutions that are being developed, such as harmonization or anonymization mechanisms." In the company of interviewee G, a data protection officer was announced to set the exact standards, which also includes the deletion of customers and contacts. Interviewee J has introduced a holistic approach to implementing data protection responsibility, which enables continuous development. Moreover, opt-ins are crucial for companies that want to address their customers personally.

Different industries introduced different approaches to deal with the GDPR, but half of the interviewees state that anonymization is part of their compliance. This is a clear sign that synthesis might be an exciting approach for further development.

7.1.5 Machine learning and artificial intelligence as methods for data privacy

The interview partners were asked if they consider Artificial Intelligence (AI) and machine learning as a way to process their data. On top of various approaches that the companies already use to process data, findings were generated concerning methods such as machine learning and AI which are presented in the following.

These methods seem to be of high interest throughout the whole sample of interview partners. Some of the respondents already use such methods, while others believe that such mechanisms will become a more important part of the future. Interviewee B already has implemented machine learning algorithms and considers it as nothing unusual. Nevertheless, the interviewee states that "many companies also have to work on their homework because they also have to ensure the quality of the data and that sufficient quality-assured data are available to train the models." For interviewee C, an algorithm used to compare the behavior of individuals with historical data can generate larger results faster than human beings. Moreover, the interviewee states that a combination of various data sets with an algorithm can help to place certain promotions and pricing models. Interviewee D states that " he/she appreciates such methods in any case and the more advanced the years, the more mature they become.” This statement is consistent with interviewee L as he/she mentions, "the algorithms are getting better and better and are thus becoming more autonomous." Ten out of the 13 companies are already using such methods and every one of them sees a high potential of these techniques in the future. In conclusion, the participants see the reasons for the rise of such methods in possibilities for process optimization, a better understanding of how to handle or interpret the data, and more precise results than can be achieved by humans.

To sum it up, all participants value methods such as machine learning and AI as very high and the common thread illustrates that it will even increase in the future. These insights are crucial for the next category, which is the processing of synthetic data.

7.1.6 Synthetic data as a means to support marketing activities

As for the insights regarding synthetic data, it is remarkable that almost every participant has heard synthetic data before. Interviewee A has heard of the term, but in a different context than with customer data. Interviewee C and L have never heard of the term before. Even more important is that all interviewees consider synthetic data as a way to foster marketing activities. The possibilities of using data synthesization include use cases such as mailings, customer acquisition, customer behavior, picture synthesis, and innovation.

Although, all of the participants see applications for synthetic data, the mean value across all interviewees on a Likert scale from (1) not at all to (5) very valuable, is only 3.85. This indicates a lack of urgency for its application, but clearly demonstrates a certain awareness and a willingness to look deeper into the subject. Figure 4 illustrates the considered value of synthetic data for different industries.

Figure 4

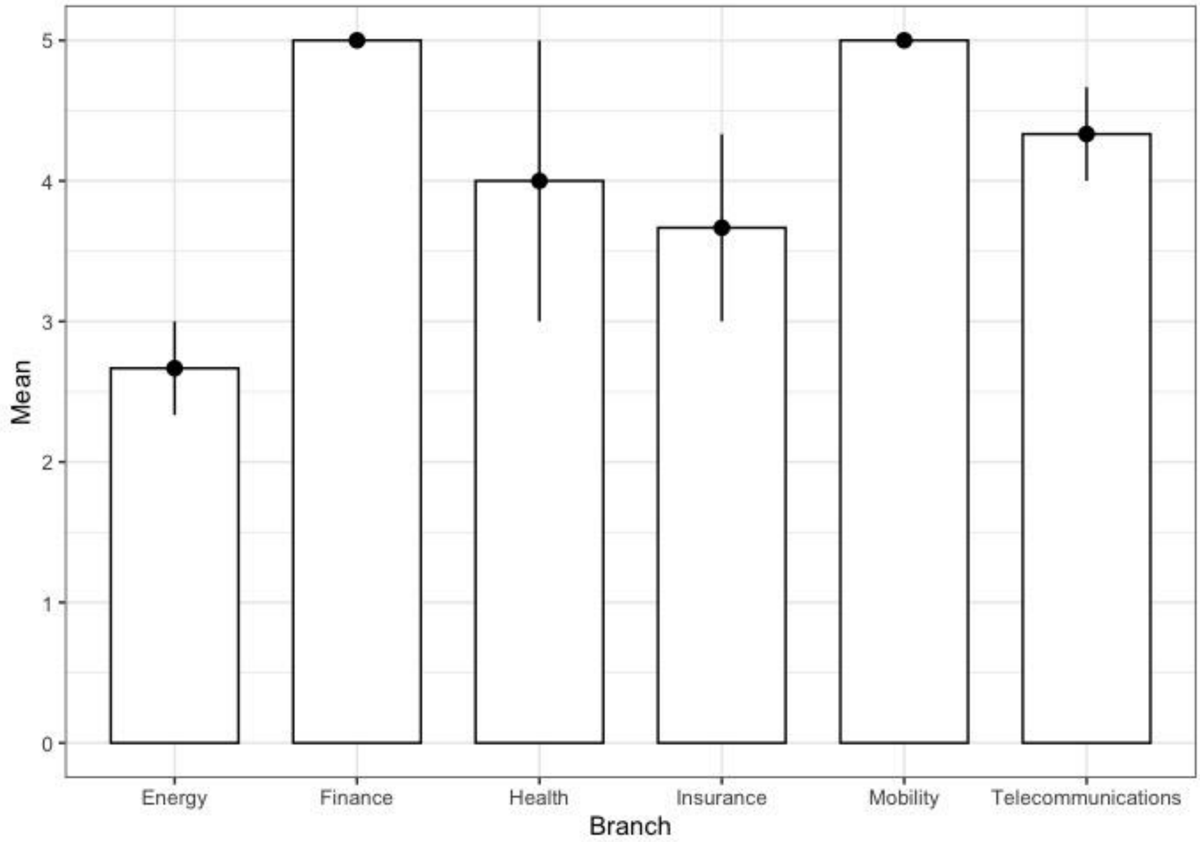

Figure 4: *Interviewees' perceived value of synthetic data for the company as a mean on a Likert scale from (1) not at all - (5) very valuable.*

Table 5.

Branch	Energy	Finance	Health	Insurance	Mobility	Telecommunications
	(n=3)	(n=1)	(n=2)	(n=3)	(n=1)	(n=3)
Mean	**2.67**	**5**	**4**	**3.67**	**5**	**4.33**
SD	**0.58**	**NA**	**1.41**	**1.15**	**NA**	**0.58**

Table 5: *Mean values and standard deviations from figure 4.*

7.2 Use cases for synthetic data as an application in marketing

Participants interested in sharing a use case for the purpose of this research were contacted in a follow-up interview and provided answers to a use case template used as the interview guideline in the second step. In total, four participants were willing to provide their input for the scope of this study and the findings are presented in the following.

To get a clear picture of the use cases that companies see as a way to process synthetic data, the application is illustrated in Table 6, where the different cases with the respective names are visualized.

Table 6.

Use case name	**Case**
Use of cloud services	A
Tourism forecast	B
Analysis of the customer portfolio	C
Generation of new use cases without additional consent	D

Table 6: *Use cases regarding synthetic data.*

As for the presentation of the four cases, each use case is illustrated one after another. In the following, the individual cases are presented further with general information, data processing, and requirements. One can also find the use case template in Appendix 10.2.

7.2.1 Use of cloud services

One application for synthetic data lies in the use of cloud services. Within this section the application is presented and the filled-in form of this use case can be found in the Appendix 10.3.

7.2.1.1 General information

Case A can be described as the development and testing of new models (e.g., cross/up-selling models for marketing purposes) with synthetic training and test data on cloud services. To date, companies are not allowed to store personal data on cloud services. So it is a matter of testing "new" algorithms and methods quickly and above all without data protection conflicts. Here respective stakeholders are an internal group of people - Marketing/Sales. Preconditions that have to be fulfilled are a 100 percent GDPR compliant procedure, ideally confirmed by an “external certificate”. Moreover, a positive 'behavior test' of synthetic data is crucial. The benefit of the use case is the security of the use of cloud services in compliance with the GDPR guidelines, which means no recurring reconciliations (e.g., legal department, etc.). A risk for the case is that there might be no release of the synthetic data. Business impact is a rapid and agile implementation of "new" cloud services. Anonymization and pseudonymization are considered as possible alternatives.

7.2.1.2 Data processing

The company from case A wants to process 1.6 million data subjects, which are each defined as an identifiable natural person in article 4 of the GDPR[8]. The data structure is represented as what they call "a typical customer contract record". Data sources and data targets are both a data warehouse and marketing analysis databases.

7.2.1.3 Requirements

The synthetic data must correspond up to 99 percent to the actual data, which means that all statistical and semantical properties should be preserved. The quality and utility of the synthetic data can be measured with statistical tests such as distributions and key performance indicators (KPIs). These requirements have to be fulfilled concerning *accuracy*.

[8] Art. 4 GDPR

As for *legal and privacy* requirements, the GDPR is the legal framework which must be considered.

The synthetic data generation's *technical* requirement is a local data generation with an operating system of windows server or client. The database system that is supported is an IBM Database 2 (DB2)[9] via Comma-Separated Values files (CSV-file)[10].

No *frequency* requirements have to be met.

Latency requirements allow lapsing between the initial storing of the privacy-sensitive data and the generation of the synthetic version up to 4 months.

7.2.2 Tourism forecast

Another application for synthetic data can be found in a tourism forecast. This chapter further portrays the use case and the filled-in form can be found in Appendix 10.4.

7.2.2.1 General information

Case B aims to develop a forecast model for tourism in Austria that allows conclusions to be drawn from various data sources about the time and number of tourists to be expected. Besides, conclusions about the origin of the tourists and the destinations the tourists are going to travel should be known in advance. Furthermore, it should be known by which route the tourists will arrive. Internal stakeholders are the big-data team and the business department and external stakeholders are the tourism industry Austria, tourism regions, Austria Tourism and Austria Advertising. Availability of actual data, historical data, and an intersection of different data sources such as overnight stay statistics, mobile phone data, ticket data from public transport, or information from airports are considered as preconditions. Internally the company can benefit from monetization that can be generated with such a data model. Moreover, external benefits can be additional information for tourism, which helps for better planning, preparing, and adjusting. Potential risks are false information that might be shared and the failure of data providers for any reason. Moreover, there may be no agreement possible to make the model viable. The legal basis may as well have possible limitations. The sales increase is the

[9] DB2
[10] CSV

predominant business impact. As of alternatives, actual data, imaginary data and absolute numbers without an algorithm or model are mentioned.

7.2.2.2 Data processing

In case B the number of data subjects is 3.9 million and the existing data source predefines the data structure that the interviewee referred to. This source prepares the data as far as possible. Data are continuously collected in the background and columns are cell information (rough geographical description); timestamp; intersection with CM data (sociodemographic, age, gender) and origin (postal code).

7.2.2.3 Requirements

As far as *legal and privacy* requirements are concerned, the GDPR, the Telecommunications Act[11], e-Privacy[12], data protection and other legal requirements must be taken into account. Therefore, anonymization, daily hashing, and guidelines regarding the processing of demographic data are important. Furthermore, it is possible to consider data from the tourism sector further, e.g., accommodation data. Privacy protection can be measured by the methods used to handle the data and through validation, but privacy is already met by anonymizing, hashing and highly aggregating.

To meet the *frequency* requirements, the data must be generated as real-time data, i.e., a continuous generation of synthetic data with 24/7 monitoring.

The time that may elapse between the first storage of the data protection-relevant data and the creation of the synthetic version, i.e., the *latency* requirements of these data, should be as short as possible and technically feasible.

There are no hard restrictions, rules, or fixed relationships in the actual data that must be guaranteed within the synthetic data.

7.2.3 Analysis of the customer portfolio

Another application for synthetic data is the analysis of the customer portfolio. The filled-in form of this use case can be found in Appendix 10.5.

[11] Telecommunications Act
[12] e-Privacy

7.2.3.1 General information

For case C, any analysis of the customer portfolio is undoubtedly of interest to marketing strategies. It is exciting to look at details that are present in the data but are not yet used. Reasons for this are both that the tools are not available or the knowledge of how to process the data best is not available internally and the regulations. There is much data from claims and benefits settlements that have not yet been processed for other purposes. This is where synthetic data can help. For example, a more in-depth analysis of the unique features found in customer data can be used to draw conclusions for marketing activities. For example, whether a specific target group is very strongly represented in the portfolio. In the case of C, there are no external stakeholders, but the marketing department, the departments that provide the data, the contract departments, and the social media team are internal stakeholders. The prerequisite for case C is legally available data from customers of the company that the company wants to process without marketing authorization. These data are based on the entire customer record. This leads to the collection of marketing information from those customers who have not given their consent. The data can be processed extensively to learn from, even without opting in. A risk in case of misinterpretation by the media is the damage to the reputation. There should be no risk to clients, as this should be prevented by synthesis. The risk for investors is derived from reputation risk. Beyond the limitations, the company would like to be able to process the data to do whatever it wants. Only in the course of the analysis and after closer examination of the data would the company find out what could be done with the synthetic data. In contrast, anonymization of data, in many cases, goes hand in hand with a loss of the interesting features. In the case of real data, the company is strictly limited and can only move in the legal grey area. Thus, business impact is the processing of other use cases in further steps. Alternative options are anonymization or aggregation, which limit the usefulness of the data because customer portfolios can only be processed with legal restrictions.

7.2.3.2 Data processing

In case of C, 3.5 million data subjects are represented as a large number of IT systems on different technological bases. Data is processed on host systems, modern architecture, and on an application. The master data depends on the selected products with specific related information, and data sources are, therefore, a data warehouse system and relational databases of all technologies. There is a total of 120 systems with personal data, 40 of which are highly relevant. Historically, the company has a very heterogeneous system landscape. The data target is a data hub as a real link since it contains a lot of data from the source systems. This means that it contains data from different systems in different forms, from real data to anonymized and sampled data to synthetic data in the next step.

The synthetic data must match the real data as closely as possible, have the same statistical significance as the real data and yet remain completely anonymous. A segmentation into specific age groups in a particular size of locations in the customer data would be desirable. Such characteristics should also be found in synthetic data. Less relevant are subjectively direct personal characteristics, e.g. names. Statistically relevant fields should also be present in the target data record. In the case of marketing data, the characteristics must be reproduced correctly so that conclusions can be drawn. Quality can be measured when the synthetic data are available, and the pattern or constellation of the real data has been maintained.

7.2.3.3 Requirements

Concerning *legal and privacy* requirements, the GDPR, DSG2000[13], e-Privacy[14] guidelines, and the Telecommunications Act[15] must be taken into account. There are requirements for compliance with legal regulations and internal guidelines for the entire processing of personal data. Identification, i.e., the appropriate degree of anonymization, must be checked in each case. Synthetic data sets must not have any reference to the real data sets and individual attributes must not allow any conclusions to be drawn about the real data set.

The *technical* prerequisite for generating synthetic data is a data hub technology and a graphics processing unit (GPU) farm that could be made available.

Frequency requirements are not so important because the data does not change quickly. They become more and more but do not change. Therefore, the frequency is monthly, maybe even less frequent.

The time frame that may elapse between the first storage of the personal data and the synthetic version of it rather irrelevant. For *latency* requirements, one-week delay in the monthly frequency is also acceptable.

[13] DSG2000
[14] e-Privacy
[15] Telecommunications Act

Hard constraints, rules, or fixed relationships in the actual data that must be guaranteed within the synthetic data mean that the various source systems would have to be consistent with each other again.

Other requirements that need to be considered are that the data may be passed on to third parties.

7.2.4 Generation of new use cases without additional consent

A further application for synthetic data is the generation of new use cases without additional consent. The filled-in form of this use case can be found in Appendix 10.6.

7.2.4.1 General information

In the case of interviewee D, the company is trying to use data synthesis to perform further, as yet undefined use cases at other times in the future. The data synthesis implies that the company is secure in terms of the GDPR. More specifically, use cases for which explicit consent has not been obtained should be performed. For market research, various use cases arise in the course of data measurement, but for which no explicit consent has been obtained. The synthesis is intended to serve as a seal of approval to implement these use cases and stand out from the competition. Panelists, advertisers, e-commerce, and various websites offering advertising space are external stakeholders. In the case of D, the only requirement is consent, which must be obtained beforehand. Additional protection in the form of certification, where consent only must be obtained once and can be used for other use cases, is advantageous. The risk lies in the possibility of re-identification of the data for the panelists. The business impact is an additional seal of approval that assures the panel that the data is treated confidentially and differs from the competition. Alternatively, the company can continue as before, i.e., obtain consent and maintain a clear separation of the data.

7.2.4.2 Data processing

For case D, there is only the information that they work with a large number of data subjects and data are recorded to the second.

7.2.4.3 Requirements

The synthetic data must be an extrapolation of the panel data and is heavily weighted towards the total online population. Besides, extrapolations must still be possible and consistent. The synthetic data must represent the structure that is mapped in the panel.

Concerning *legal and privacy* requirements, the GDPR and various international market research guidelines such as ESOMAR[16], e-Privacy[17] which contains information on cookie consents, and the prevention of cookies by third parties is relevant.

To meet the *frequency* requirements, the data must be generated with real-time run measurement, i.e., with a continuous frequency.

The time frame that may elapse between the initial storage of the data sensitive data and the generation of the synthetic version of these data must be concise. Thus, *the latency* requirements are stringent.

Hard constraints, rules, or fixed relationships in the actual data, which must be guaranteed within the synthetic data, mean that they must be consistent with the picture after extrapolation.

There are no *other* requirements to be considered.

[16] ESOMAR
[17] e-Privacy

8 Discussion

This thesis aims to give insights into a new approach for marketing activities under the increasing environment of data regulation – synthetic data. Within this section, the formerly mentioned findings are discussed and further connected to already existing literature. The literature on marketing analytics, legal aspects, and literature on current approaches for data synthesis build the foundation for the following analysis and discussion of the findings. Additionally, implications for policymakers, marketing managers, and experts on synthetic data are presented. Finally, some of the limitations that are covered within the scope of this research practice are portrayed and further research opportunities are provided.

8.1 Discussion of results

The findings of the empirical part of the study cover the basic ideas that were identified from the literature review at the beginning of this research project. As Larsson and Teigland (2020) already stated, personal data are relevant for companies from basically every industry. The findings of this thesis confirm this assumption. Besides the health industry, such data are essential for the business and, especially, for marketing activities. With the only exception of the wholesale and retail industries, all predefined sectors were open for at least a short telephone interview. The lack of interest from the retail industry indicates that companies from this branch are rather unwilling to share insights about their perception of privacy regulations and possible applications for synthetic data. However, results from mobility, telecommunications, finance, insurance, health, and energy show a clear relevance of personal data for these industries.

Personal data are still considered a type of data that helps improve service operations—but also anonymized and aggregated data play a crucial role for improvement. Even though according to Drechsler & Reiter (2010) aggregation makes it difficult and often impossible to analyze data at a more detailed level, it is still considered as important. Third-party data are also considered as an essential opportunity to improve service operations. Trusov et al. (2016) conclude that third-party data to enhance customer profile recovery is subject to violate the GDPR, which indicates that there is a potential application for a new method like data synthesis. Taking into account the fact that five respondents mentioned anonymized data as a type of data that improves service operation, it can be concluded that there is a great interest in the methods that work with anonymization. However, Larsson and Teigland (2020) refer to anonymized data sets that are subject to re-identification by combining several data subsets. Considering the alternatives that the participants presented in the various use cases, which are anonymization, aggregation, and obtaining consent, it can again be concluded that there is a lot of potential for the implementation of synthetic data. As mentioned before, anonymization is subject to re-identification and, thus, prone to violate the GDPR. Therefore, data anonymization should be under more scrutiny and instead replaced with other methods.

Considering the impact that the GDPR has on companies, the empirical part's findings support the current state of the literature. On average interviewed companies from various industries assess the effects of the GDPR on the business of the respective companies as very strongly. Goldberg et al. (2019) already illustrate the difficulty and high cost of privacy regulations for companies. As mentioned in the findings, the interviewees connected the impact of the GDPR with high investment costs and much effort regarding the process. Moreover, Holtrop et al. (2017) raised the question of whether increasing amounts of stored data also increases the difficulty to keep the data private. Taking a closer look into the empirical findings, data storage is a problem that has to be tackled with new methods.

Machine learning approaches and Artificial Intelligence (AI) are approaches that have been researched by extant literature extensively. Surendra and Mohan (2017) already proposed several methods to be relevant for marketing activities. The findings show a clear picture that such methods are of high interest for companies. For many companies, such approaches are already in use, whereas others at least value such methods as a significant opportunity for future development.

Concerning the most crucial purpose of this study, the value of synthetic data, the findings are diverse. Interviewees consider the approach of data synthesis on average as “rather valuable” with a score of 3.85 out of 5 ((1) being not valuable at all and (5) being very valuable). Interestingly, all participants show a keen interest in the topic and see synthetic data as an opportunity for the future. This indicates that the industries are not yet clearly aware of the potential benefits of the approach, but after being introduced to it, they show clear signs of interest in the respective area. The energy industry seems to have the lowest interest in synthetic data, which could also be influenced by the fact that this industry is also the least affected by the GDPR. Industries such as finance and telecommunications consider synthetic data as very valuable, but they are also strongly affected by the GDPR. Therefore, to increase awareness for synthetic data, it is crucial to demonstrate the potential value of data synthesis. Although the general interest seems to be very high, most of the participants were not willing to share a potential use case for data synthesis and its application for marketing activities. Only four out of the 13 participants were willing to share a use case for this study. Moreover, even those willing to share their idea did not answer all of the questions that were part of the use case template. This indicates that they are very cautious about sharing their ideas and data for research purposes.

Possible applications for synthetic data in the marketing domain can be found for the use of cloud services, tourism forecasts, customer portfolio analysis, and a generation of use cases without additional consent. As mentioned above, data storage is quite a challenge after the introduction of the GDPR. Thus, the process of synthesizing the data to be GDPR compliant could be an option to make the data available for cloud services. The point of making data GDPR compliant through synthesis also plays an essential role in sharing the data afterwards with other stakeholders involved in the process. Moreover, for companies, this additional seal of approval helps identify other use cases with fully compliant data.

Digging deeper into the expected business impacts for the different cases, there are some exciting opportunities. Rapid and agile implementation of new cloud services as a possible impact illustrates the need for a faster transition of companies to a digitized organization without being hindered by data regulations. The empirical data clearly shows that such an approach is considered as a way to be monetized and in a further step to increase sales. Moreover, the differentiation from other competitors in the field is a strong argument for the implication of synthetic data. It can be beneficial for companies to take the lead within this domain and step ahead of competition.

Additionally, a basic principle of the approach is considered as sharing the data with other stakeholders to foster innovation and collaboration. Although this business impact has been part of the interview guideline, none of this was considered as a possible business impact. Thus, it can be concluded that the companies are trying to use such methods first and foremost for their own purpose. There may be some reasons concerning their ideas, but especially with regard to the current situation of COVID-19 one could have wished for more collaborative ideas.

Interestingly, requirements for synthetic data sets are quite similar across the various use cases and, thus, across industries. In general, accuracy requirements are rigorous, and the closer the synthesized data set matches the real data, the better it is. Companies seem to strive for almost identical statistical properties in the synthetic data. Taking into account their current approaches, which entails a fraction of their expected accuracy, this seems somewhat irrational. However, aiming high in accuracy also illustrates that companies want to go back to the pre-GDPR conditions. Legal and privacy requirements are mainly the GDPR but in some industries, they take other regulations into account, e.g. the telecommunications act, e-Privacy requirements, ESOMAR, and DSG2000. Technical requirements lack some further investigation, as the use cases did not conclude to specific ideas here. In addition, the frequency is either not considered relevant because the data for industries such as insurance do not change as quickly, or it is very relevant because the data for industries such as market research companies or companies in the telecommunications industry are constantly changing. This is a crucial part that needs to be investigated by companies who offer synthetic data. Somewhat similar are requirements for the latency of data. As mentioned before, insurance companies do not consider latency as necessary, whereas companies from market research or telecommunications see it as very important, which means the lower the latency, the better it is. It is crucial for companies that the synthesized data set can be processed with the source systems which are already in use. Finally, other requirements are related to the collaboration of companies. For market research companies, it is important that the data can be given to third parties without additional consent.

8.2 Limitations

Every qualitative study has its limitations. In the following, these limitations are illustrated and some ideas for future research are portrayed.

First of all, qualitative data are always subject to subjective interpretations of the interviewer. Thus, the discussion of the results is merely the interpretation of the author and there exists no right or wrong. Nevertheless, as the findings are presented, the reader can understand the way how these were interpreted.

Moreover, the study was planned during a very stressful time for many industries. As the COVID-19 pandemic started to spread through Europe and, especially through Austria, many companies did not find the time and resources to take part in the research. Under different circumstances, additional companies had participated at least in the short interviews and the database would have been broader. There are already signs that companies are showing more interest since the measures are getting less and less. Therefore, one opportunity for future research would be a replication of the study after the pandemic is finally settled.

Furthermore, the COVID-19 situation limited the way of how interviews were conducted. It was only possible to conduct interviews via telephone or video call. Although these methods are elaborated and a good alternative, it is hard to capture emotions without sitting face to face with the participants. Therefore, future research potential also exists in personal interviews.

Besides, the participants who were contacted via email in the first step belonged to a specific range of industries. For the purpose of this study and especially with regard to the research project, these industries were considered the most important ones. Nevertheless, as the topic of synthetic data can be extended to other industrial areas as well, there exists an opportunity for future research in other fields.

The structure of the interviews was already given, which was helpful in gaining deep insights in the respective direction. However, future research would benefit from focus groups with experts in the field who discuss the topic in more detail.

As time passes, approaches such as the synthesis of data change and improve. Moreover, the understanding of such approaches receives more awareness. Therefore, this study's outcomes are only relevant for a specific range of time and should be reevaluated in the future.

8.3 Implications

Based on the findings of this study, and even more from the discussion, implications for marketing managers, policymakers, and experts on synthetic data are illustrated. These implications have a wide range of applications and help to raise awareness of the topic of synthetic data.

As the focus of this study is based in the field of marketing, implications for marketing managers are of particular interest. The findings and discussion clearly illustrate that the GDPR and other regulations have a tremendous impact on how such activities can be performed, especially for marketing activities and purposes. Current approaches instead hinder marketing managers as they are either prone to violate the GDPR or delete substantial information from the data. Therefore, marketing managers benefit from the processing of synthetic data for their activities. Not only can companies foster innovative activities, such as the development of new cloud service applications, but also promise synthetic data a new way to analyze customer portfolios and offer new and better solutions. Besides, early adopters of such approaches can benefit from the first-mover advantage and step ahead of the competition. An additional certificate that guarantees GDPR conformity can also lead to joint innovations. Therefore, stakeholders from a variety of industries can use data from other industries to improve and accelerate innovation.

Furthermore, implications for policymakers can also be derived from the findings of this study. Since the introduction of the GDPR in May 2018, companies from every industry are affected in some way. Results from the study show that companies are getting more and more used to the restrictions. However, the general perception about the GDPR is that it will undoubtedly evolve within the next years, but regulations will not increase anymore. There is one opinion that regulations will become even stricter in the future, while another respondent states that GDPR will be useless in a few years as the world moves towards an increasingly digitized state. Policymakers should consider these opinions for further development of data regulations. As synthetic data cannot be traced back to individual customers, it is an interesting topic for policymakers which offers a lot of potential for further discussion.

Finally, there are also implications for experts in the field of synthetic data. Results and observations show that companies have a high interest in the topic. However, the general understanding of such approaches is still limited. Thus, companies still fear to implement synthetic data which can hinder the successful development of the approaches. Experts in the field can take advantage of the findings evaluated and presented in this study to improve their approaches in the right way that marketing activities benefit the most from the synthesis of data.

9 References

An, D., Little, R.J.A., 2007. Multiple imputation: an alternative to top coding for statistical disclosure control. J Royal Statistical Soc A 170, 923–940. https://doi.org/10.1111/j.1467-985X.2007.00492.x

Ascarza, E., Hardie, B.G.S., 2013. A Joint Model of Usage and Churn in Contractual Settings. Marketing Science 32, 570–590. https://doi.org/10.1287/mksc.2013.0786

Boone, H.N., Boone, D.A., 2012. Analyzing Likert Data 5.

Booth, A., Sutton, A., Papaioannou, D., 2016. Systematic Approaches to a Successful Literature Review, 2nd ed. SAGE.

Darra Hofman, Luciana Duranti, Elissa How, 2017. Trust in the Balance: Data Protection Laws as Tools for Privacy and Security in the Cloud. Algorithms 10, 47. https://doi.org/10.3390/a10020047

Drechsler, J., Reiter, J.P., 2010. Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata. Journal of the American Statistical Association 105, 1347–1357. https://doi.org/10.1198/jasa.2010.ap09480

Dwork, C., Roth, A., 2013. The Algorithmic Foundations of Differential Privacy. FNT in Theoretical Computer Science 9, 211–407. https://doi.org/10.1561/0400000042

European Parliament, 2016. Regulation (EU) 2016/679 of the european parliament and of the council. Official Journal of the European Union 88.

France, S.L., Ghose, S., 2019. Marketing analytics: Methods, practice, implementation, and links to other fields. Expert Systems with Applications 119, 456–475. https://doi.org/10.1016/j.eswa.2018.11.002

Goldberg, S., Johnson, G., Shriver, S., 2019. Regulating Privacy Online: The Early Impact of the GDPR on European Web Traffic &amp; E-Commerce Outcomes. SSRN Journal. https://doi.org/10.2139/ssrn.3421731

Hart, J., 2018. Data breaches are taking a toll on customer loyalty. Data breaches are taking a toll on customer loyalty. URL https://www.csoonline.com/article/3250836/data-breaches-are-taking-a-toll-on-customer-loyalty.html (accessed 7.3.20).

Holtrop, N., Wieringa, J.E., Gijsenberg, M.J., Verhoef, P.C., 2017. No future without the past? Predicting churn in the face of customer privacy. International Journal of Research in Marketing 34, 154–172. https://doi.org/10.1016/j.ijresmar.2016.06.001

Jarmin, R.S., Louis, T.A., Miranda, J., 2014. Expanding the Role of Synthetic Data at the U.S. Census Bureau. SSRN Journal. https://doi.org/10.2139/ssrn.2408030

Johnson, E.J., Bellman, S., Lohse, G.L., 2001. Defaults, Framing and Privacy: Why Opting In-Opting Out 11.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv:1710.10196 [cs, stat].

Kennickel, A., Lane, J., 2006. Measuring the Impact of Data Protection Techniques on Data Utility: Evidence From the Survey of Consumer Fi- nances, in: Privacy in Statistical Databases. Springer- Verlag, New York, pp. 291–303.

Larsson, A., Teigland, R. (Eds.), 2020. The digital transformation of labor: automation, the gig economy and welfare. Routledge, Abingdon, Oxon ; New York, NY.

Li, H., Xiong, L., Jiang, X., 2014. Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions. https://doi.org/10.5441/002/EDBT.2014.43

Marketing Post-GDPR: Two Tribes of Marketing, 2018. URL https://www.insightsforprofessionals.com/marketing/marketing-technology/marketing-post-gdpr-two-tribes-of-marketing/download (accessed 4.5.20).

Martin, K.D., Murphy, P.E., 2017. The role of data privacy in marketing. J. of the Acad. Mark. Sci. 45, 135–155. https://doi.org/10.1007/s11747-016-0495-4

Menon, M., 2019. GDPR and Data Powered Marketing: The Beginning of a New Paradigm. JMDC 13. https://doi.org/10.33423/jmdc.v13i2.2010

Mitra, R., Reiter, J.P., 2006. Adjusting Survey Weights When Altering Identifying Design Variables via Synthetic Data, in: Privacy in Statistical Databases. Springer- Verlag, New York, pp. 177–188.

Müller, N.M., Kowatsch, D., Debus, P., Mirdita, D., Böttinger, K., 2019. On GDPR Compliance of Companies' Privacy Policies, in: Ekštein, K. (Ed.), Text, Speech, and Dialogue. Springer International Publishing, Cham, pp. 151–159.

Nabbosa, V.L., Iftikhar, R., 2019. Digital Retail Challenges within the EU: Fulfillment of Holistic Customer Journey Post GDPR, in: Proceedings of the 2019 3rd International Conference on E-Education, E-Business and E-Technology - ICEBT 2019. Presented at the the 2019 3rd International Conference, ACM Press, Madrid, Spain, pp. 51–58. https://doi.org/10.1145/3355166.3355170

Patki, N., Wedge, R., Veeramachaneni, K., 2016. The Synthetic Data Vault, in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Presented at the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Montreal, QC, Canada, pp. 399–410. https://doi.org/10.1109/DSAA.2016.49

Reiter, J.P., 2005. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. J Royal Statistical Soc A 168, 185–205. https://doi.org/10.1111/j.1467-985X.2004.00343.x

Reiter, J.P., 2004. Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology 20.

Saunders, M.N.K., Lewis, P., Thornhill, A., 2012. Research methods for business students, 6th ed. ed. Pearson, Harlow, England ; New York.

Schneider, M.J., Jagpal, S., Gupta, S., Li, S., Yu, Y., 2018. A Flexible Method for Protecting Marketing Data: An Application to Point-of-Sale Data. Marketing Science 37, 153–171. https://doi.org/10.1287/mksc.2017.1064

Schneider, M.J., Jagpal, S., Gupta, S., Li, S., Yu, Y., 2017. Protecting customer privacy when marketing with second-party data. International Journal of Research in Marketing 34, 593–603. https://doi.org/10.1016/j.ijresmar.2017.02.003

Schweigert, V.-A., Greyer-Schulz, A., 2019. The Impact of the General Data Protection Regulation on the Design and Measurement of Marketing Activities: Introducing Permission Marketing and Tracking for Improved Marketing & CRM Compliance with Legal Requirements. JMDC 13. https://doi.org/10.33423/jmdc.v13i4.2352

Sirur, S., Nurse, J.R.C., Webb, H., 2018. Are We There Yet?: Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR), in: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security - MPS '18. Presented at the the 2nd International Workshop, ACM Press, Toronto, Canada, pp. 88–95. https://doi.org/10.1145/3267357.3267368

Surendra, H., Mohan, H.S., 2017. A Review Of Synthetic Data Generation Methods For Privacy Preserving Data Publishing. International Journal of Scientific and Technology Research 6.

Trusov, M., Ma, L., Jamal, Z., 2016. Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting. Marketing Science 35, 405–426. https://doi.org/10.1287/mksc.2015.0956

Wedel, M., Kannan, P.K., 2016. Marketing Analytics for Data-Rich Environments. Journal of Marketing 80, 97–121. https://doi.org/10.1509/jm.15.0413

Wieringa, J., Kannan, P.K., Ma, X., Reutterer, T., Risselada, H., Skiera, B., 2019. Data analytics in a privacy-concerned world. Journal of Business Research S0148296319303078. https://doi.org/10.1016/j.jbusres.2019.05.005

Winkler, W.E., 2005. Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Research Report Series 14.

10 Appendix

10.1 Telephone interview questionnaire

Questionnaire (EN)

	Topic	**Question**
1	Relevance	How important is personal data for your company on a scale from 1 to 5? (1) Not at all □ to (5) very high □
2	Kind of data	What kind of data do improve your service operations? (Personal, anonymized, pseudonymized, aggregated, big data, quantitative, qualitative, descriptive)
2	Impact GDPR	Does the increasing environment of data protection regulations affect your industry and why?
3	Impact GDPR	How strongly is the GDPR affecting your company on a scale from 1 to 5? (1) Not at all □ to (5) very high □
4	Impact GDPR	Do you think the impact of the GDPR on your industry will change in the near future and why?
5	Current approach	How do you currently comply with the GDPR (anonymization, opt-ins, aggregation, not yet, others)?
6	Current approach	Do you consider artificial intelligence and machine learning as a way to handle your data?
7	Synthetic data understanding	Have you ever heard of the term “synthetic data”?
8	Value of synthetic data	How valuable do you consider synthetic data for dealing with issues concerning data processing and its regulation?

		(1) Not at all □ to (5) very high □
9	Possible application	Can you think of a way to use synthetic data within your company for research or maybe innovative activities?
10	Interest	Would you be interested in sharing a use case in form of an interview to support research in the field of synthetic data?
11	Interest	Would you like to arrange a meeting to conduct an interview?

10.2 Use case template

GENERAL			
Use case name:	[Enter a short title to identify this use case.]		
Created by:	[Company name, person.]	**Date created:**	[ ]
Description:	[Briefly describe this use case. Describe the purpose and the context of the use case. What is the "story" behind this use case?]		
Stakeholders:	[Persons/entities (internal and/or external) that have an interest in the outcome of the use case.]		
Preconditions:	[Conditions that must be true / activities that must be finished before the use case can be executed.]		
Benefits:	[The benefits of the use case to the company, actors, and stakeholders (e.g., investors, customers, data subjects, etc.)]		
Risks:	[The potential risks of the use case to the company, actors, and stakeholders (e.g., investors, customers, data subjects, etc.)]		
Expected business impact if the data is anonymized:	[Describe expected impact (financial and non-financial) if the data is anonymized.]		

Alternatives:	[What are (potential) alternatives to sharing synthetic data for the given use case?]
Use case concept: ***Figure***	[A figure depicting components, data flows, and involved entities, etc.]

Data Processing	
Number of data subjects:	[Number of data subjects available to learn patterns for synthetic data.]
Data structure:	[Description of data tables, their attributes (i.e. columns) with their variable types (e.g. numeric, date, identifier, categorical, text, etc.), that are to be shared.]
Data entry sample:	[Provide a representative record with made-up attributes for a single data subject.]
Data source:	[Where is the original privacy-sensitive data stored, and in what type of a database system?]
Data target:	[Where is the synthetic data to be stored, resp. to be delivered to?]

Requirements	
Accuracy requirements:	[How close does the synthetic data need to be to the actual data?] [Which statistical properties of the actual data need to be retained in the synthetic data?] [How can the quality & utility of the synthetic data be measured?]
Legal & Privacy requirements:	[What are the legal frameworks to consider for the given use case?] [What are the requirements in terms of privacy?] [How can the privacy of a dataset be measured?]
Technical requirements:	[What are the technical requirements for the generation of the synthetic data?] [Are there restrictions in terms of computing environment, operating system or hardware?] [What database systems are to be supported?]
Frequency requirements:	[How often will the synthetic data need to be generated in order to meet requirements of this use case?]
Latency requirements:	[How much time is allowed to lapse between initial storing of the privacy-sensitive data, and generating a synthetic version thereof?]
Constraint requirements:	[Are there any hard constraints, rules or fixed relations in the actual data, that need to be guaranteed within the synthetic data? (e.g., if age < 10 then employment status = ‘student’)]
Any other requirements:	[Are there any other requirements present for this use case?]

ANITA
[Where does the ANITA project come in for the use case?] [What would the specific advantage(s) be?] [Is test data available?]
Issues
[Issues related to the definition of the use case.]
Other
[Any other remaining remarks related to the use case.]

10.3 Use of cloud services

GENERAL			
Use case name:	Use of Cloud Services		
Created by:	-	**Date created:**	-
Description:	Use of cloud services (AI algorithms, ...) without data protection conflicts. Developing and testing new models (e.g. cross/up selling models for marketing purposes) with "synthetic" training and test data. Rapid testing of "new" algorithms and methods.		
Stakeholders:	Internal group of persons - Marketing/Sales		
Preconditions:	100 percent GDPR compliant "procedure", ideally confirmed by an "external certificate". Positive "behaviour test" of the synthetic data.		

Benefits:	Security of the use of cloud services in compliance with the GDPR guidelines. No recurring reconciliations. (Legal department, ...)
Risks:	No release of the synthetic data.
Expected business impact if the data is anonymized:	Rapid and agile implementation of "new" Cloud Services.
Alternatives:	Anonymisation and/or pseudonymisation
Use case concept: ***Figure***	-
Data Processing	
Number of data subjects:	1,6 Mio.
Data structure:	"Typical" customer, contract record
Data entry sample:	-
Data source:	Data Warehouse, Marketing Analysis DB
Data target:	Data Warehouse, Marketing Analysis DB
Requirements	
Accuracy requirements:	Data needs to be as close as 99 percent to the actual data. All statistical properties of the actual data need to be retained in the synthetic data quality & utility of the synthetic data can be measured with statistical tests (distributions, KPI, ...)

Legal & Privacy requirements:	Legal framework and requirements in terms of privacy to consider for the given use case is the GDPR
Technical requirements:	Technical requirements for the generation of the synthetic data: Generation should be done locally Restrictions in terms of computing environment, operating system or hardware: Windows Server or Client Database systems that are to be supported: DB2 via csv-Files
Frequency requirements:	1x the test.
Latency requirements:	4 months
Constraint requirements:	-
Any other requirements:	-
ANITA	
Test data available: In case of need YES.	
Issues	
-	
Other	
-	

10.4 Tourism forecast

GENERAL			
Use case name:	Tourism Forecast		
Created by:	-	**Date created:**	-
Description:	The aim is to develop a forecast model for tourism in Austria, which allows conclusions to be drawn about the time and number of expected tourists using various data sources. In addition, conclusions about the origin of the tourists and to which tourist destinations they are going to travel should be known in advance. Furthermore, it should be known by which route the tourists will arrive		
Stakeholders:	Internal: Big-Data Team and Business Department for Monetization External: Tourism industry Austria, tourism regions, Austria Tourism, Austria Advertising.		
Preconditions:	Actual data, historical data and intersection of different data sources. Overnight stay statistics, mobile phone data, ticket data from public transport or information from airports. Weather data are always exciting, hotspots of leisure facilities It is still possible to use data from tourist facilities, but this could be too much work for the basic product.		
Benefits:	Internal: Earn money with such a data model External: Information for tourism. You can better prepare, adjust and plan for it, etc.		
Risks:	False information could be shared. Failure of data providers for any reason. There may be no agreement possible to make the model viable. The legal basis may have possible limitations.		
Expected business impact if the data is anonymized:	Sales increase		

Alternatives:	actual data, fictitious data, absolute numbers, without an algorithm, no model.
Use case concept: ***Figure***	-
Data Processing	
Number of data subjects:	3.900.000
Data structure:	"Categorical The existing data source prepares the data as far as possible. Data is continuously collected in the background. Anonymized data and no personalized data. For example: Cell information (rough geographical description); time stamp; intersection with CM data (sociodemographic, age, gender) Origin (postal code)
Data entry sample:	-
Data source:	-
Data target:	Data Center
Requirements	
Accuracy requirements:	-
Legal & Privacy requirements:	Data handling -> DSGVO compliant, Telecommunications Act and e-Privacy and Data Privacy, other legislation requirements: Anonymization, daily hashing, guidelines concerning the use of demographic data,

<table>
<tr><td></td><td>For example: at least 5 SIM cards with the same attributes from one region to be allowed to use data point
Further consideration from the tourism sector and, for example, accommodation data
Measurement of privacy, with the procedures how the data is handled. Through validation, but privacy is already met by anonymizing, hashing and highly aggregating.</td></tr>
<tr><td><b>Technical requirements:</b></td><td>-</td></tr>
<tr><td><b>Frequency requirements:</b></td><td>real-time data -> continuous generation of synthetic data
24/7 monitoring
Historical data and real-time data</td></tr>
<tr><td><b>Latency requirements:</b></td><td>The faster the better
Close to real-time
What is technically possible to implement</td></tr>
<tr><td><b>Constraint requirements:</b></td><td>no</td></tr>
<tr><td><b>Any other requirements:</b></td><td>For example, if you extend the time window.
You can also show how many nights (nights of stay) there are, with the tourists. Possible via the overnight stay statistics, for example as an average.
Data from tourism, request for overlap of the different data sets</td></tr>
<tr><td colspan="2"><b>ANITA</b></td></tr>
<tr><td colspan="2">Provide stakeholders with better information than is now possible.</td></tr>
<tr><td colspan="2"><b>Issues</b></td></tr>
<tr><td colspan="2">-</td></tr>
</table>

Other
-

10.5 Analysis of customer portfolio

GENERAL			
Use case name:	Analysis of the customer portfolio		
Created by:	**-**	**Date created:**	-
Description:	Every analysis of the customer portfolio is certainly of interest for marketing strategies. It is interesting to look at details that are present in the data but are not yet used. Reasons for this are that the tools are not available or the knowledge of how to best use the data is not in-house. A lot of data from claims and benefit settlements exist that have not been used for other purposes so far. Here synthetic data can help. Deeper analysis of specifics found in the customer data to draw conclusions for marketing activities. E.g. certain target group is very strongly represented in the portfolio. (Targeting single older ladies in particular)		
Stakeholders:	Internal: Marketing department, departments that supply the data, contract departments, social media team External: none		
Preconditions:	Legally available data of customers, which he company would like to process without marketing consent. Based on the entire customer record.		
Benefits:	Collect marketing information from those customers who have not given consent. Not to contact them, but data may be used for other purposes. Data may be used extensively to learn from, even without opt-in.		

Risks:	Reputational risk, in case of wrong interpretation by the media this can damage the reputation. No risk for customers, as this should be prevented through synthesis. Investors risk is derived from the reputation risk.
Expected business impact if the data is synthesized/anonymized:	Business impact: Beyond the restrictions, everything that the company wants, can be done with the data. The exciting thing is that the company does not know exactly what is interesting and it is only in the course of the analysis that it finds out where it can take a closer look, and that is what it can do with the synthetic data. With anonymized data, the features that are interesting have been lost in many cases. With real data you are strictly limited or move in the legal grey area. With the data, further use cases may be processed in further steps.
Alternatives:	Anonymization or aggregation. Data of the customer portfolio may be used in a legal way. Perform analyses within the scope of the legal requirements
Use case concept: ***Figure***	Multiple source systems for the data Damage data and contract data sources: These would have to be brought together in a data warehouse - or they already are. The data would have to be synthesized on the basis of this. Own data warehouse instance with synthetic data only.
Data Processing	
Number of data subjects:	3.5 Million data subjects
Data structure:	There is a large number of IT systems on different technological bases, from host systems to modern architecture and data via an application. Not to be named exactly on the detailed level. Relational databases or just the data warehouse. Master data, depending on the selected insurance products, specific information in this regard E.g. Household insurance -> basic data of the apartment, where it is, how big it is, how high is the sum insured for the objects in it, risk area for natural disasters, flooding or earthquake, lightning strike statistics etc. Looking at claims, the insurance has a lot of data specific to products and product history.

Data entry sample:	-
Data source:	Data warehouse system. Relational databases of all technologies Host databases 120 systems containing personal data of which 40 are very relevant. History of the company has a very heterogeneous system landscape. (MySQL, SSQL, Datahub (Hadoop))
Data target:	Datahub Would be the right link, because it has a lot of data from the source systems. Data from different systems in different forms from real data, to anonymized and sampled data and in a next step also synthetic data.
Requirements	
Accuracy requirements:	Get as close as possible and still remain completely anonymous, but have the same statistical significance as the real data Certain age groups in a certain size of locations in my customer data. Such characteristics should also be found in the synthetic data. Less relevant are subjectively directly personal characteristics, such as name Statistically relevant fields should be available in their constellation, also in the target data set For marketing data, the characteristics must be reproduced correctly so that conclusions can be drawn. Measurement of quality: If the synthetic data are available, whether the pattern or constellation has been retained.
Legal & Privacy requirements:	GDPR, DSG2000, e-browsing directive, telecommunications law

	Requirements that all processing operations involving personal data are compliant with the legal requirements, but also with internal guidelines. The identification, appropriate level of anonymization, must be assessed in each individual case. Synthetic data records must not have a connection to the real data records. Individual attributes must not allow conclusions to be drawn about the real data record.
Technical requirements:	In principle, there are options available, but the costs must be weighed up. GPU farm is not there ad hoc, but could be made available. HADOOP ->Datahub technology
Frequency requirements:	Not so important, because the data does not change quickly. They are becoming more and more, but do not change. Monthly, maybe even less frequently.
Latency requirements:	Not so relevant. One week of delay in the monthly frequencies is also okay.
Constraint requirements:	The different source systems would have to be consistent with each other again.
Any other requirements:	The data should be allowed to be given to third parties.
ANITA	
Providing synthetic data for three possible use-cases 1. Use for open marketing purposes 2. DataHub loading 3. Test data purposes Test data is not available overnight, only in cooperation.	

Issues
-
Other
-

10.6 Generation of new use cases without additional consent

GENERAL			
Use case name:	Generation of new use-cases without additional consent		
Created by:	-	**Date created:**	-
Description:	The company seeks to use the synthesis of data as a way to carry out further use cases, that have not yet been defined, at other points in time in the future. The synthesis of data implies that the company is secured with regard to the GDPR. To be more precise, use cases are to be run for which consent was not explicitly obtained. For market research, various use cases arise in the course of the measurement of data, but for these use cases consent was not explicitly obtained. The synthesis should serve as a seal of approval in order to implement these use cases and to stand out from other market research providers.		
Stakeholders:	Panelists, advertisers, e-commerce, various websites offering advertising space		
Preconditions:	Consent that must be obtained		
Benefits:	Additional protection in the form of a certification Consent only has to be obtained once and can be used for further use cases		
Risks:	Possibility of re-identification of data (for panelists)		
Expected business impact if the data is anonymized:	Additional seal of approval Panel can be sure, that data is treated confidentially Differentiation from the competition		

Alternatives:	Continue as before -> obtain consent and maintain a clear separation of data
Use case concept: ***Figure***	-
Data Processing	
Number of data subjects:	Very large number, as data is recorded to the second
Data structure:	-
Data entry sample:	Demographic data, questionnaire with 300 attributes Age, sex, place of residence, media use, interests, acquisition requests
Data source:	-
Data target:	-
Requirements	
Accuracy requirements:	- Extrapolation of data in the panel will be heavily weighted to the whole online population Extrapolations must still be possible and consistent Synthetic data must represent the structure that is mapped in the panel
Legal & Privacy requirements:	GDPR and various guidelines of international market research ESOMAR E-privacy and cookies Prevention of cookies by third parties

Technical requirements:	-
Frequency requirements:	Real-time run measurement Ongoing frequency
Latency requirements:	Measurement of structural data of websites -> very promptly
Constraint requirements:	It has to correspond to the image after the extrapolation
Any other requirements:	-
ANITA	
Test data is available, but the company does not fully understand their benefits of giving away a data set	
Issues	
The use-case treats the generation of new use-cases	
Other	
-	