ANITA

Anonymous big data A project funded by FFG

Use cases and requirements

Deliverable 2.1

Authors: Alexander Kowarik, Johannes Gussenbauer, Michael Jakl, Verena Warringer, Tobias Hann, Thomas Reutterer, Peter Eigenschink, Olha Drozd, Stefan Vamosi, Michael Platzer, Alexandra Ebert, Johannes Bogner

Reviewers: Peter Eigenschink, Klaudius Kalcher, Michael Platzer, Olha Drozd, Thomas Reutterer, Michael Jakl, Alexander Kowarik

Document version: 2.0
Date: 31.03.2020

WU Wirtschaftsuniversität Wien Vienna University of Economics and Business · MOSTLY AI · George Labs · Statistik Austria Die Informationsmanager · FFG Promoting Innovation. · Federal Ministry Republic of Austria Climate Action, Environment, Energy, Mobility, Innovation and Technology

Programme **"ICT of the Future"** - the initiative of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK)

Disclaimer

This deliverable describes the work and findings of the AI-Based Privacy-Preserving Big Data Sharing for Market Research (Anonymous Big Data (ANITA)) project.

The authors of this document have made every effort to ensure that its content was accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this deliverable are responsible for any possible errors or omissions as well as for any results and actions that might occur as a result of using the content of this document.

Table of contents

1 Summary

In the work package two we systematically collected use cases for sharing privacy-sensitive sequential data with third parties as well as captured requirements with respect to accuracy and privacy. This deliverable provides details about the Anonymous Big Data workshop, the use cases of our industry partners, and Master's thesis that is being written in the context of the work package two.

The goal of the Anonymous Big Data workshop was to explore the topic of synthetic data from the following perspectives: opportunity, utility, law, trust, communication, ethics. The participants identified four large groups of potentially privacy sensitive data that could be of interest for (market) research: (i) behavioral tracking data, (ii) demographic and socio-economic data, (iii) attitudinal/preferential data, and (iv) sensor data. From the utility perspective the synthetic data have to be as close to the original data as possible. This contradicts the requirement of privacy. As a result, there has to be a trade-off between privacy and utility. As to the regulations that could be considered for synthetic data generation, the experts named GDPR and their industry-specific legislations as the most important legal frameworks. To gain trust among the data suppliers, data users, and society in general, standard metrics, quality controls, trust frameworks, hands-on trainings, information about benefits for data protection, additional ethical requirements could be introduced. The perception of the synthetic data by the general public clustered into the following topics: (i) motivation to be interested in the synthetic data, (ii) lack of understanding of the methodology itself and its quality metrics, (iii) necessity to disclose the methods, (iv) ways to communicate the synthetic data topic effectively, and (v) general trust issues. The following groups of ethical questions with respect to synthetic data were identified: (i) synthetic data creation, (ii) synthetic data usage, (iii) data ownership, (iv) information disclosure, and (v) fundamental ethical questions.

When collecting use cases, we documented them in terms of number of subjects, frequency / latency for data sharing, accuracy and privacy requirements, technical requirements, etc. The general idea of the Statistics Austria use case lies in sharing the synthetic micro data of the Austrian population with the public and with the scientific community to enable innovative research and to support policy making. George Labs, in their use case, would like to create a customer base that includes representative product associations and corresponding transactions that can be used to shape the product and allow their partners to develop products without endangering any data protection concerns. Additional use cases are being collected in the context of the Master's thesis with a working title "Synthetic data: A new approach for marketing analytics in an increasing environment of data protection". The work aims to provide insights into the requirements of synthetic data for market(ing) analytics. The use cases and requirements described in this deliverable will serve as the basis for other work packages of the Anonymous Big Data project.

2 Introduction

The goal of the work package two is to systematically collect use cases for sharing privacy-sensitive sequential data with third parties, for example for market(ing) research purposes, as well as to capture requirements with respect to accuracy and privacy. As a starting point for the use case and requirements collection we held a workshop where we discussed the synthetic data topic from the perspectives of opportunity, utility, law, trust, communication and ethics. The workshop results serve as a guidance for the ANITA project in general as well as for the work package two with regard to the collection of accuracy, privacy and legal requirements for the use cases.

For this deliverable we collected use cases from our project partners. Statistics Austria provided the use case where they would like to share the synthetic census micro data with the public and with the scientific community. The use case of George Labs is about the synthetic bank customer base that includes representative product associations and corresponding transactions. In addition to these use cases, other use cases from industry representatives, who volunteered them for our research, are being collected for the Master's thesis with a working title "Synthetic data: A new approach for marketing analytics in an increasing environment of data protection" that is written in the context of the work package two.

Deliverable 2.1 is organized as follows: First, we provide details about the workshop and its results. Then, we describe the nonconfidential part of the use cases in detail together with the accuracy and privacy requirements.

3 Anonymous big data workshop

The Anonymous Big Data workshop was organized in the context of the ANonymous bIg daTA (ANITA) project to generate ideas, questions, and concerns that would serve as a guidance for the work package two with regard to the collection of accuracy, privacy and legal requirements, as well as for other work packages of ANITA. In the work package three the results of the workshop could help evaluate the potential of existing synthetization methods. The insights into how companies see and evaluate synthetic data could give an idea of how metrics for existing models could be taken into account. The information gained at the workshop could also influences the final basic structure of the data lab in the work package four. The goal of the workshop was to explore the topic of synthetic data from multiple perspectives and to create a collaborative dialogue around the following questions:

1. Opportunity: Which type(s) of privacy-sensitive data assets are of interest for (market) research?
2. Utility: What are requirements with regard to accuracy and representativeness for synthetic data?
3. Legal: Which legal frameworks are to be considered for synthetic data generation?
4. Trust: What is required to establish trust in synthetic data or other forms of privacy preservation (e.g., data minimization), in terms of accuracy and privacy?
5. Communication: How are data synthetization and other forms of privacy preservation perceived by the general public?
6. Ethics: Are there other ethical questions, aside from privacy, with respect to synthetic data?

3.1 Workshop format

The Anonymous Big Data workshop started with a presentation that provided an overview of the ANITA project and its goals as well as brief introduction of the data synthetization approach and the brainstorming technique of the workshop. The introductory session was then followed by 6 rounds of small group discussions in the form of a carousel brainstorming (also known as rotating review).

Carousel brainstorming technique activates participants' prior knowledge and generates new ideas and solutions via collaborative discussions and movement[1]. For the carousel brainstorming to be effective the participants should be divided into small groups. These groups then rotate through several topic-specific "stations" discussing questions, solving problems or providing feedback at each "station" for a short period of time. Each group writes down their ideas on post-its or flip-chart paper for other groups to read. After a defined period of time the groups rotate to another "station" and follow the same procedure. The ideas that were generated at each station are shortly presented at the end of the brainstorming session.

[1] Fullan, Michael. The taking action guide to building coherence in schools, districts, and systems. Corwin Press, 2016.

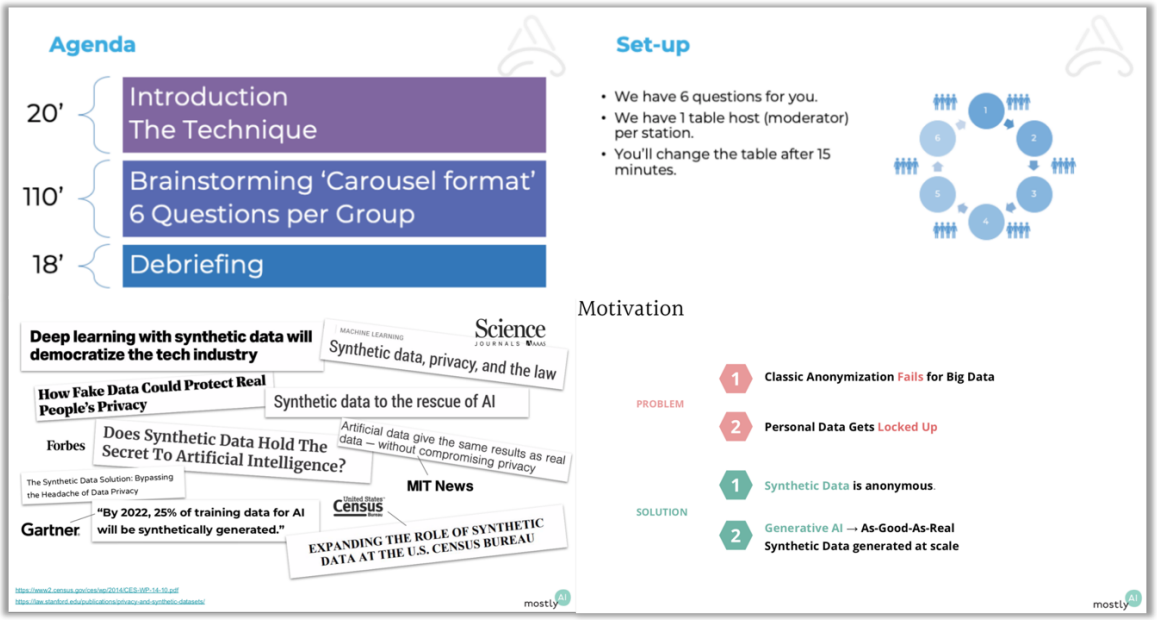


Figure 1: ANITA workshop introduction and set-up

23 experts with data science, marketing, legal, privacy, ethics and philosophy backgrounds attended the workshop. They were organized into 6 groups and colored markers identified the group membership. Each group answered all 6 questions mentioned above. The participants had at least some knowledge of the topics raised in the questions. Each question formed a separate station equipped with a table, the colored sticky notes (76 x 127 mm), and a flip chart. Every group was provided with the sticky notes of the same size, so that the size of the sticky notes did not influence the number of ideas suggested by each group. We assigned one ANITA consortium representative to each station for clarifications and assistance. The consortium representatives participated in the discussion, took notes, assisted with the clustering of ideas on flip charts, briefly shared key insights from the prior rounds of discussion with every new group at the station and debriefed all the participants about the outcomes of all rounds of the discussion at the end of the workshop. They were also instructed to make sure that the participants follow the brainstorming rules of being visual and free with their ideas.

The first round (Figure *2* (left)) began with the participants reading the question of their station and writing their ideas on sticky notes while brainstorming silently for 5 minutes. The groups then had 10 minutes to discuss and cluster their ideas on the flip chart. Finally, the teams moved on to the next station. The next five rounds of discussions (Figure *2* (right)) started with the ANITA consortium representative giving a brief summary of the ideas and clusters developed by other groups in no more than 2 minutes. The teams spent 4 minutes brainstorming and jotting down their ideas that complemented already developed ones. Finally, they shared, discussed and clustered their ideas for 9 minutes and then moved on to the next station.

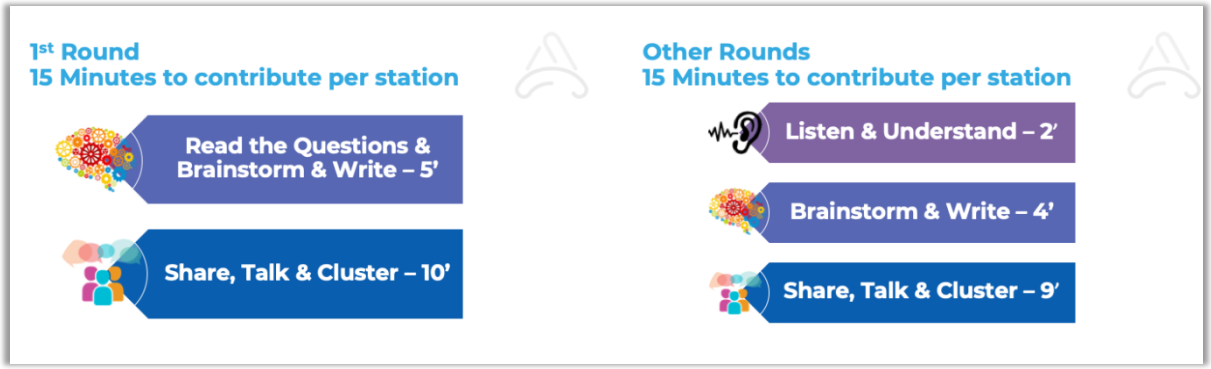


Figure 2: ANITA workshop carousel brainstorming process

3.2 Workshop outcomes

The brainstorming session resulted in a total of 236 sticky notes. The proposed ideas vary in terms of quality and context, including not only the answers to the questions, but also deeper questions, concerns and general comments concerning the topic under discussion. In this report we summarize all collected ideas and provide a list of the most common questions and suggestions. These will serve as a guidance for ANITA's model development and its virtual data lab, as well as for future work beyond ANITA.

3.2.1 Opportunity

The following question was subject to discussion:

Which type(s) of privacy-sensitive data assets are of interest for (market) research?

The question was well received by the workshop participants and stimulated a vivid and fruitful discussion on the various data sources and data assets that are potentially subject to privacy concerns.

One stream of discussion concentrated around the question of what makes data "privacy-sensitive". As a crucial and distinctive characteristic of privacy sensitivity we identified whether a "human interaction" is involved in the data generation process. Thus, data sets like those referring to weather conditions and/or other environmental characteristics, data generated by machines without a link to humans, data on public infrastructure, etc., are considered not to be subject to privacy sensitivity.

The participants identified four large groups of potentially privacy sensitive data, which can be connected and integrated with each other based on common person IDs or semantic identifiers. These data groups and examples for each one can be summarized as follows.

Behavioral tracking data. This large group of data comprises health data (e.g., health records), social interactions and network data, geo-location data, access and movement data, social media and web shop activities, transaction histories, etc.

Demographics and socio-economic data. These data can be further subdivided into static and dynamic data. Examples of static data are gender, ethnic background, place of birth, etc. Most demographic and socio-economic

data are dynamic, but typically evolve at a slower pace as behavioral data. Examples comprise biographic data, education, income, wealth, location, ethnographic and contextual data (such as physical environments).

Attitudinal / preferential data. In contrast to behavioral data (which reflect what individuals are doing), this group of personal data tells us something about what people want, need, or prefer. They are also subject to change and include psychographic profiles, attitudes, interests, political opinions, cultural interests, etc.

Sensor data. This group of data is typically generated by machines and/or measurement devices and includes recent developments around the so-called Internet of Things (IoT), but also comprises measurements made by eye tracking devices, fMRI scans, etc. "Inferred data" can also be subsumed to this group of data. For example, internet browsing behavior (measured by web browsing meters) can be utilized to make "inferences" about an individual's gender, opinions, interests, sexual preferences, etc.

The above classification of data sources can be further split into different categories based on their sensitivity levels.

During the workshop other criteria and distinctions of databases were also discussed. These distinctions are listed below:

Internal (e.g., data arising as a part of company's customer relationship management system)	*vs.*	"Pooled" or syndicated data (e.g., data collected and provided by professional market/ing research companies)
Automatic data collection (e.g., sensor machine data)	*vs.*	Semi-automatic and/or manual data collection (e.g., questionnaire data)
Structured data (e.g., machine- or human-generated/internal or external)	*vs.*	Unstructured (e.g., texts, pictures, videos, etc.)
Aggregated data (e.g., segment-level – i.e. less privacy-sensitive)	*vs.*	Disaggregated data (e.g., individual-level – i.e. more privacy-sensitive)
High frequency data	*vs.*	Low frequency data
Length and granularity of time-varying data (e.g., weekly, monthly, yearly)		

Table 1: ANITA workshop, Criteria and distinctions of databases

Another distinction with a potential impact on privacy-sensitivity arises from a distinction of data usage. Based on the discussion, the following three classes of data usage (with increasing privacy-risk) can be distinguished:

- Diagnostic (visualization, exploration, dashboards)
- Inference & prediction (inference of individual behavior)
- Decision automation (data-driven, automized decision-making)

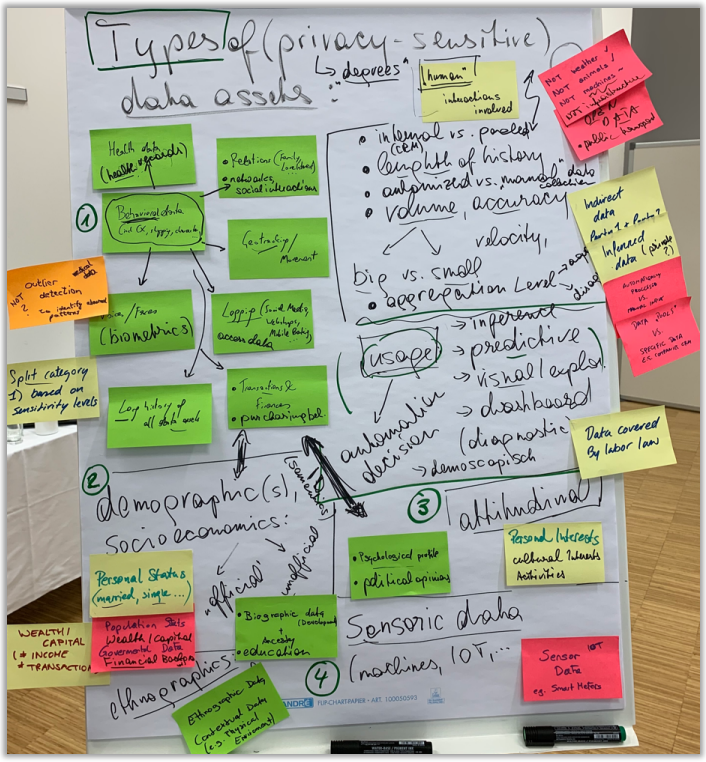


Figure 3: ANITA workshop, Opportunity station

3.2.2 Utility

The following question was central to the discussion:

What are requirements with regard to accuracy and representativeness for synthetic data?

Many ideas and follow-up questions arose during the discussion. These can broadly be put into five categories: (i) prediction/synthetization, (ii) quality, (iii) privacy, (iv) technology, (v) trust. We discuss these categories below.

Prediction/Synthetization. We concluded that it is hard to replace an actual data set by one single synthetic data set; rather it depends on the prediction task at hand. This is due to different requirements for the data, for example, importance of attributes, volume of the data or feasible tests for validation of the synthetic data. Similarly, it is not possible to have a single quality measure for the artificial data sets. It could be necessary to synthesize a data set multiple times either due to an iterative model/software development process or for keeping the data up to date. So, reproducibility of the synthetization and also of the predictions is required.

The topic of the data types that could be subject to synthetization also came up during the discussion. The questions arose (i) if it is even possible to synthesize any given data without losing the required information (e.g., comments) and (ii) how to handle the data that expires after a given date.

Quality. To be of a high quality, the artificial data should be as close to the original data as possible. That is why, it is necessary to establish methods and metrics to measure the accuracy of the synthetic data and the uncertainty

arising from using that data. To gain trust in the data, scientists would need a test environment to benchmark synthetic data against the real ones. The information gained from the outliers and in case of skewed distributions has to be retained, while the privacy must not be neglected. Again, depending on the case, outliers could be more or less important and could be hidden in the crowd to a greater or lesser extent. Besides that, it is also critical to maintain the integrity of the data, and to generate the right amount of data. This should be a scalable process, as some applications will need more data and some will need less.

Privacy. No individual shall be re-identifiable from the information in the synthetic data set (even by chance). This is key to guarantee privacy. As such, this requirement is very much opposed to the quality requirement. Different levels of privacy were discussed: privacy on the individual level, privacy of households or even hierarchies. The meaning of privacy in those cases and how they correlate with each other is still not clear from the participants point of view. All, however, agreed that this should also be addressed in the scientific community.

Even if the data are synthesized, the purpose for collecting, processing and synthetization should still be clear. Requirements regarding privacy will also depend on this purpose.

The presence of biases in the synthetic data, but also in the original sample is a relevant topic. Identifying and exploring them would be beneficial.

Technology. The participants raised questions regarding reliability of the technology, of the models and of the algorithms for synthetic data generation:

- How do we know that the algorithms and models are actually capable of generating data with the specified requirements regarding quality and privacy?
- If there are major errors in the software that undermine quality and privacy requirements, how are they dealt with?

Trust. Lack of trust in both privacy and accuracy of the data could be a big issue. The participants pointed out that the users could think that the model could be unstable with regard to the training data. There also could be doubts that the model could leak training data. These trust issues could be solved by establishing a certification process.

The experts questioned reasonableness of training other models with the synthetic data. They also suggested that the risk of the predictions derived from the artificial data should to be evaluated.

The discussion about trust in synthetic data was accompanied by the concept of transparency. The teams identified a need to know how a certain synthetic data set was generated. One possible solution to increase transparency could be annotation of data sets with accurate metadata about the methods/algorithms used (e.g., how they deal with outliers and other edge

cases?) and information about the original data set (e.g., how was it collected? Are there any known biases?)

Figure 4: ANITA workshop, Utility station

3.2.3 Legal

At the Legal station, the following question was discussed:

Which legal frameworks are to be considered for synthetic data generation?

Most participants didn't think additional legal frameworks for synthetic data generation are needed and that GDPR, research- and industry-specific legislations (e.g., banking) are sufficient. Some wished for a GDPR 2.0, which would give them a say in whether their data could be anonymized/synthesized or not. Others thought, no additional regulation is needed, however, ethical guidelines could be beneficial. The participants mentioned that they would be much more willing to give consent for research purposes instead of marketing optimization use cases.

The groups also discussed synthetic data quality and privacy-protection measurement criteria. Most participants thought that having quantifiable measures to assess how well a generated synthetic data set protects against the re-identification of the data subjects in the training sample would be desirable. Based on these ideas and arguments participants suggested that certifications, standards and external auditing procedures should be introduced for synthetic data generators.

3.2.4 Trust

The participants discussed the following question at the Trust station of the workshop:

What is required to establish trust in synthetic data or other forms of privacy preservation (e.g., data minimization), in terms of accuracy and privacy?

This topic resulted in exciting discussion and various contributions from the participants.

The experts identified the following groups of stakeholders that are related to this topic: (i) data suppliers (e.g., individuals, customers, etc.), (ii) data users (e.g. institutions, firms, etc.), (iii) society (e.g., public opinion). In terms of trust, these groups have different needs, opinions and fears that have to be addressed individually.

Potential reasons for a lack of trust were also discussed at the "Trust" station. Every stakeholder group might ask different questions that identify trust issues. For example:

Stakeholder group	***Question examples***
Data suppliers	Is my privacy under threat, when my data are used for synthetization? For what purpose are my data used?
Data users	Are synthetic data accurate enough to be used like real data? Can we still violate the GDPR, if we use synthetic data?
Society	Can synthetic data be used for the bad intentions, e.g., fake news?

Table 2: ANITA workshop, Trust issues

6 rounds of discussion produced several approaches for building trust for each stakeholder group. These approaches are described below.

Data suppliers. The following procedures for building trust were identified for the data suppliers group:

- A standard metric could be introduced to measure the risk of re-identification.
- The governance and quality control could be executed by an external accreditation authority, that is independent and trustworthy. Although some legal boundaries are already in place (e.g., GDPR), the above-mentioned mechanisms are not fully established yet.
- The purpose (context) for data processing should be specified.
- A trust framework with explainability algorithms, automatic compliance checking and responsibility algorithms could be introduced.

- The inner mechanisms of synthetization or other privacy preserving algorithms could be explained as a part of a workshop or a hands-on training.
- The legal boundaries that protect the individual and his/her privacy could be highlighted.
- Additional ethical requirements to generate trust could be applied.

Data users. For a data user, the exploitation and dissemination of data are the primary goals. To guarantee a good usability and accuracy, standardized metrics, mostly of statistical nature, should be defined. With such standardized measures, the utility of synthetic data could be shown.

Transparency was mentioned by many participants as another aspect related to trust. Involving people in generation and "playing" with synthetic data could create additional confidence in this technology. Historic trust-building or trust-losing events were mentioned for comparison, for example, trains or self-driving cars in the context of a positive and nuclear power in the context of a negative perception.

Society. At the moment, the public opinion seems to be very critical, when it comes to the processing of personal data. Synthetic data and AI could potentially be negatively associated with fake news. It is very important to highlight the improvements in terms of privacy as well as other potential benefits for the society brought by synthetic data or other privacy friendly techniques. Exemplary sectors where synthetic data could lead to improvements are traffic, transportation, health care, security, etc.

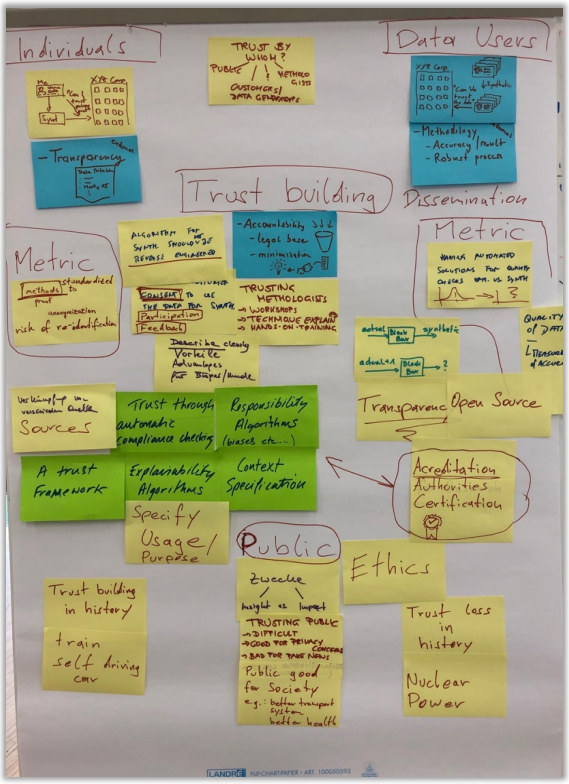


Figure 5: ANITA workshop, Trust station

3.2.5 Communication

The following question led the discussion at the Communication station:

How are data synthetization and other forms of privacy preservation perceived by the general public?

The teams also discussed some related topics:

- How does the public perceive privacy concerns?
- How to communicate effectively towards the general public?
- How to build trust in the methods?

The input of the members clustered into several topics with some ideas being in-between or overlapping other stations' questions/results. The details about each cluster are provided below.

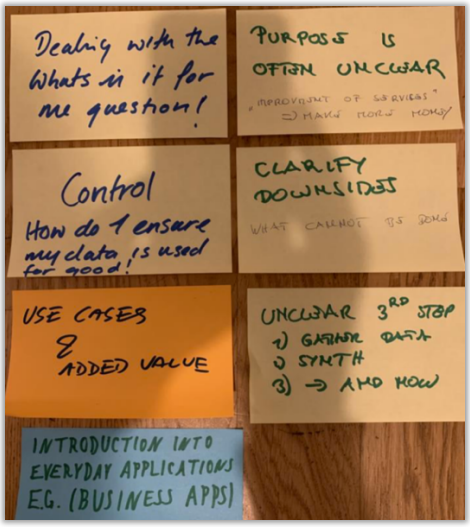


Figure 6: ANITA workshop, Communication station, "What for?" cluster

What for? In this cluster, the groups raised questions related to the individual's motivation to be interested in the topic at hand:

- What does a person asked to provide his/her data for synthetization get out of the process? Why should anyone be willing to help?
- What are the use cases beyond "service improvements", which are often interpreted as "get more money out of our pockets"?
- What are the downsides of working with synthetized data?

The purpose for data processing is often unclear, either to the data controllers / processors or the data subjects (who have to hand over their data and agree to the data processing).

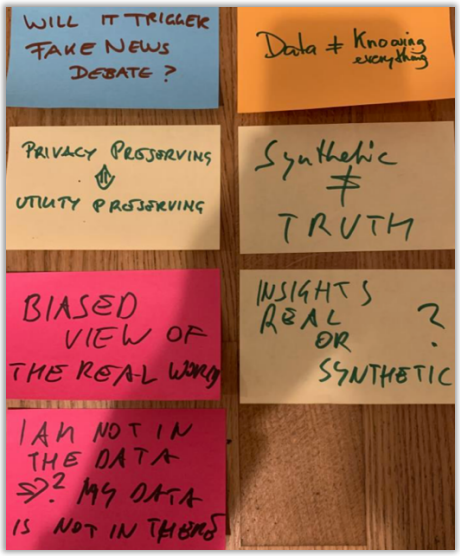


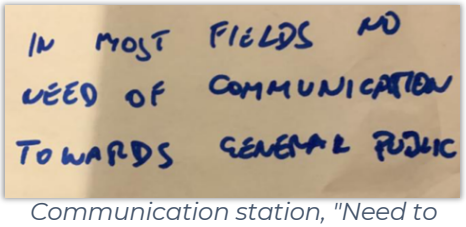
Figure 7: ANITA workshop, Communication station, "Quality issues" cluster

Quality issues / Is it working? The participants assumed that much criticism would be based on little understanding of the methodology, and quality metrics that are hard to communicate. Hyperbolic news articles might be misleading and could create a bad image of the whole data processing industry.

The groups identified the following questions relevant to this cluster that might need additional clarifications to improve communication:

- Are the insights real or synthetic?
- Privacy preserving and utility preserving? Is that actually possible?

- Synthetic data are not “the truth”: When to go for synthetic data? When to use real data or other means?
- The person itself is not in the data, but his/her data are in the data set: Data subjects need to give consent, don’t they?
- There’s a lot of data “between the lines”: Is the method able to extract that properly?



Communication station, "Need to communicate" cluster

- The method is bound to be biased by the input data (e.g., observation bias, racist machine-learning): How should that be handle?

Do we actually need to communicate this (differently)? Fundamental questions here are:

- Is it actually necessary to communicate how the data are synthetized?
- The processing itself is already to be disclosed and, at the same time, few/none disclose the methods. Why now?

Public opinion / “Trust issues”. The participants shared their observations regarding the public opinion on data processing and anonymization. The identified trust issues and possible ways of improvement, in the form of questions or statements, are presented below:

Figure 9: ANITA workshop, Communication station, "Trust issues" cluster

- What is privacy in today’s world anyway?
- Why should I trust synthetic data?
- Privacy topics are usually negative in the media: Is it possible to turn that around?
- What are possibilities to prove the value of providing data for synthetization and further processing?
- The high levels of mistrust are caused by the data misuse in the past.
- Black & White thinking (e.g., “I have to give all of my data for processing, otherwise I won’t be able to use the service”).
- Anxiety of the unknown, especially, if there’s no immediate benefit for “me”.
- General awareness problem when it comes to data protection: What can I do and what should I do?
- “Companies just want to make more money, and now they found a new toy…”

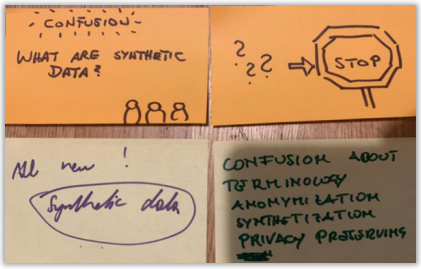


Figure 10: ANITA workshop, Communication station, "Hard to understand" cluster

Hard to understand / Special topic. There's a lot of confusion when it comes to privacy protection and the proposed methods to protect privacy are even more technical and more complex than the current perception of the topic. Such confusion does not build trust.

Compliance & transparency. Companies need to prove that they respect the privacy preferences of the users and the requirements of the regulator as such.

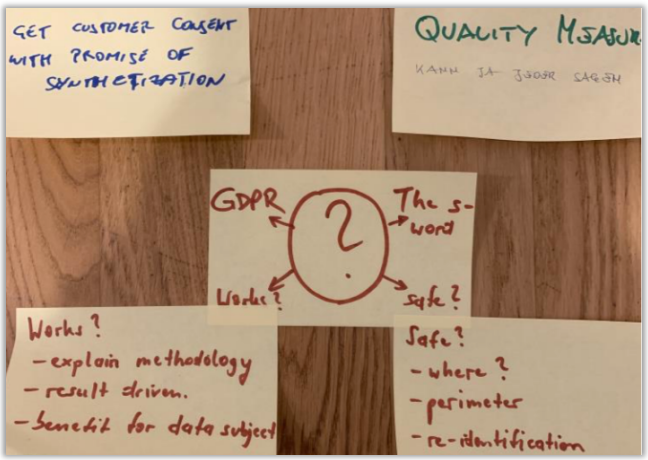


Figure 11: ANITA workshop, Communication station, "Compliance & transparency" cluster

- How to communicate that the synthetization is indeed the correct way to go forward?
- What measures can be taken to ensure that the data has been properly anonymized? Certification?
- Are companies "safe" in terms of GDPR when they ask any synthetization company to generate synthetic data?
- What are the requirements for third-party companies regarding synthetic data generation for their clients?
- What about the regulator? Do we need a central entity that will ensure proper usage of the data and will be the first contact for personal data misuse, doubts, question (e.g., a DPA equivalent institution)?

How to communicate? A big part of the discussion was concentrated on how to communicate the complex topic of synthetic data effectively and what information should be provided:

- Certifications as, for example, "we process only non-personal data" or "anonymous data usage" could build awareness, and trigger interest in the topic.
- The topic needs to be simplified as much as possible to reach the general public. Use case scenarios as well as concrete examples might help.
- Transparency should play the key role. What happens with data before during and after processing & synthetization should be explained. Are they sold to the "Evil Corp."? Or are products shaped to fit the needs of the customer?

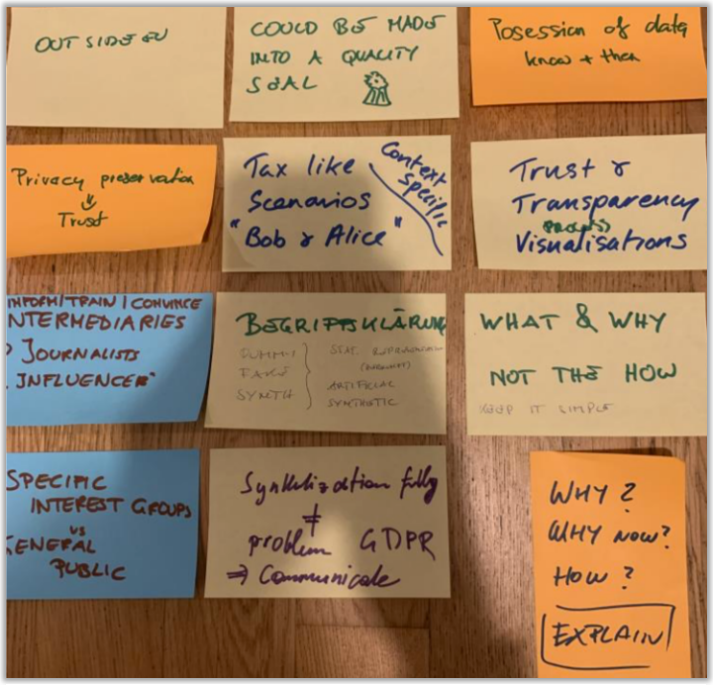


Figure 12: ANITA workshop, Communication station, "How to communicate" cluster

- Companies could communicate (i) what has changed in the landscape over the years (data usage now and then), (ii) why synthetization is such an important topic right now.
- The choice of terminology is very important. Each specific term could carry a connotation of "good", "evil" or "neutral". When communicating, the most effective and correct term should be used. This is likely to differ depending on the target audience (e.g., dummy data, synthetic data, anonymized data, statistically representative data, etc.)
- The "What" and "Why" should be central to communication. The "How" is changing a lot and would, most likely, be too complex when it comes to awareness- and trust-building.

Figure 13: ANITA workshop, Communication station

3.2.6 Ethics

The following question was the starting point of discussion at the Ethics station:

Are there other ethical questions, aside from privacy, with respect to synthetic data?

The discussion results of the Ethics station could be grouped and summarized as follows:

Creation of synthetic data.

- How to integrate/build in existing ethical standards into the process of synthesizing data (e.g., IEEE P7000, "privacy by design", etc.)?
- How to ensure explainability, responsibility & governance mechanisms (GDPR)?

When to use synthetic data?

- Is it ethical to monetize synthetic data without the customer knowing? What if this monetization is necessary to cross-finance a product?
- Are there ethical/non-ethical use cases for synthetic data?
- Is it ethical to try to make synthetic data more fair (e.g., remove gender bias that is present in the original data)?
- When should it be ethically required to work with real data (e.g., public entity plans to base decision around building infrastructure on certain data)?

Data ownership.

- How can it be ensured, that individuals do not contribute to a certain synthetic data set (i.e., "opt out" of personal data being used to synthesize data sets)?

Information disclosure.

- When does a company need to disclose the usage or creation of synthetic data to users?
- What does a company need to proactively disclose in general, when it comes to the processing of personal data?
- Should companies be required to disclose synthetic data sets to the public?
- Should it be required to label synthetic data sets as such?

Fundamental ethical questions

- Who is responsible when synthetization of data goes wrong?
- Is it ethical when large companies become even more powerful through the creation of synthetic data (e.g., a large corporation with many subsidiaries is able for the first time to use a combined data set from all subsidiaries, thus creating a new competitive advantage)?
- Is it ethical to use so much compute power to generate synthetic data?

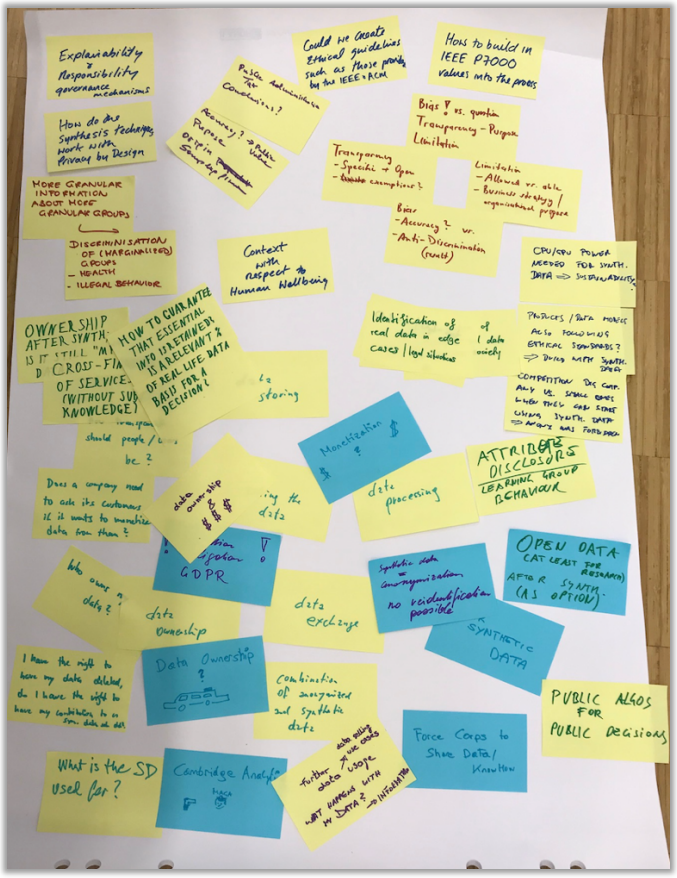


Figure 14: ANITA workshop, Ethics station

3.3 Conclusion

The Anonymous Big Data workshop explored the synthetic data topic from multiple perspectives. During the workshop 23 experts (divided into small groups) discussed, in the form of a carousel brainstorming, the following concepts in the context of synthetic data: (i) opportunity, (ii) utility, (iii) law, (iv) trust, (v) communication, and (vi) ethics. The ideas, generated during the workshop, vary in terms of quality and context and include deeper questions, concerns and general comments. In this report we summarized all collected ideas and provided a list of the most common questions and suggestions for each concept.

The discussion around the opportunity concept led to the identification of four large groups of potentially privacy sensitive data that could be of interest for market research: behavioral tracking data (e.g., social interaction and network data, movement data, etc.), demographic and socio-economic data (e.g., gender, income, biographic data, etc.), attitudinal/preferential data (e.g., psychographic profiles, attitudes, etc.), and sensor data (e.g., IoT, eye tracking, fMRI scans, etc.)

The requirements with regard to accuracy and representativeness of synthetic data were discussed in the context of the utility concept. From the utility point of view, the data have to be as close to the original data as possible. Since this contradicts the requirement of privacy, the trade-off has to be made regarding

the utility and accuracy of the data. The balance between privacy and utility of the synthetic data will highly depend on the prediction task at hand.

As to the legal requirements for synthetic data generation, most of the participants named GDPR and industry-specific legislations as the main legal frameworks to consider. Some of the groups also highlighted the need for ethical guidelines. It was identified that certifications, standards and external auditing procedures could be beneficial for the synthetic data generators.

The trust concept was discussed from the perspectives of different stakeholders, namely, data suppliers, data users, and society. Standard metrics, quality controls, trust frameworks, hands-on trainings, additional ethical requirements could be introduced to gain trust among the data suppliers. The data users are primarily interested in the dissemination of data, so standardized statistical metrics could show them the utility of synthetic data. Additionally, involving data users in the actual synthetic data generation could build extra confidence in this technology. The society should also be informed about the benefits (including data protection perspective) of synthetic data.

The input of the participants regarding the perception of the synthetic data by the general public clustered into several topics such as the motivation to be interested in the synthetic data (e.g., “why should anyone be willing to help?”); the lack of understanding of the methodology itself and its quality metrics (e.g., “are the insights real or synthetic?”); the necessity to disclose the methods (e.g., “the data processing itself is already to be disclosed, however, few/none disclose the methods. Why now?”); the ways to communicate the synthetic data topic effectively (e.g., “simplification of the topic”, “usage of examples or use cases”, “transparency should play the key role”); and the general trust issues (e.g., “anxiety of the unknown, especially, if there’s no immediate benefit for the data subject”).

The experts identified the following groups of ethical questions with respect to synthetic data: synthetic data creation (e.g., “how to integrate existing ethical standards into the process of synthesizing data?”), synthetic data usage (e.g., “Are there ethical/non-ethical use cases for synthetic data?”), data ownership (e.g., “how to opt out” of personal data being used to synthesize data sets?”), information disclosure (e.g., “when does a company need to disclose the usage or creation of synthetic data to users?”), and fundamental ethical questions (e.g., “is it ethical when large companies become even more powerful through the creation of synthetic data?”).

4 Use cases

In order to systematically capture use cases and requirements for sharing synthetic, yet statistically representative data, we created a use case description template (see Annex 5.1 Use case description template) where we ask participants to provide general description of their use cases as well as details about the data processing and requirements concerning accuracy, law, privacy, frequency, latency, etc. In this section we describe the nonconfidential part of the collected use cases.

4.1 Synthetic census micro data

The synthetic census micro data use case was provided by Statistics Austria.

4.1.1 General description

The general idea of the use case lies in creating synthetic micro data of the Austrian population with the aim to share these data with the public and with the scientific community. The synthetic data will be based on the Rich Frame which is a pseudonymized micro data set that contains every person registered in the housing and living register as main residence within private (non-institutional) households, including personal and household specific attributes, as well as income data.

Sharing data with the public is one of the core principles of Statistics Austria. The access and use of micro data are, however, restricted due to privacy protection principles and regulations. The synthetic micro data could be safely shared with the public and the scientific community to enable innovative research and to support policy making.

A stronger collaboration with the scientific community and the aim to share more data with them is also mentioned in the new government programme 2020-2024[2]., p. 165.

Stakeholders. Statistics Austria, researchers and universities could be named among the stakeholders.

Preconditions. A dummy/structural data set must be generated to see the variables and the possible values. The actual execution on the real data set must be possible at Statistics Austria.

Benefits. Sharing of the synthetic census micro data with universities and the scientific community will facilitate innovative research and policy making.

Risks. If the synthetic data set is not “safe” but is still disseminated, data about individuals could be potentially disclosed.

Expected business impact if the data are anonymized. Statistics Austria will have a possibility to share full (synthetic) population data sets with the research community. This will result in more and closer collaborations with the researchers.

[2] The government programme 2020-2024 is available at https://short.wu.ac.at/programme

Alternatives. In the context of this use case the following alternatives could be applied:

- Sampling.
- Application of traditional statistical disclosure control (SDC) methods:
  - recoding,
  - noise,
  - post randomisation method (PRAM),
  - suppression, etc.

4.1.2 Data processing

Number of data subjects. Roughly 8.8 Mio people who live in 3.9 Mio households for each quarter from 2015 until 2019 could be considered as data subjects in this use case.

Data structure. Each row in a data table corresponds to a single person at a certain time point and contains columns which refer to personal and household attributes. These attributes include (but are not limited to):

- `Year`
- `Quarter`
- `Personal identifier`
- `Household identifier`
- `Age` (non-negative integer)
- Personal variables (categorical):
  - `employment status`
  - `education`
  - `citizenship`
  - `country of birth`, etc.
- Geographical variables (categorical):
  - `county`
  - `municipality`
  - `degree of urbanization`
- Income components (numerical):
  - `cash from employment`
  - `self-employment`
  - `pension`, etc.

Data source. The consolidated data are stored in house in a Db2[3] data base. Source data comes from multiple different registries.

Data target. The synthetic data are to be stored on premises – Statistics Austria.

[3] Db2. https://www.ibm.com/products/db2-database

ANITA
Anonymous
big data

Data entry sample.

Year	Quarter	Hid	Pid	County	Municipality	Urban	Age
2017	3	3505	350501	8	80207	2	50
2017	3	3505	350502	8	80207	2	72
2017	3	3505	350503	8	80207	2	42
2017	3	3505	350504	8	80207	2	23

Gender	Citizenship	Education	Employment	IncEmployment	IncSelfemployment	Pension
male	AT	tertiery	full time	32058.860	2033.220	0.000
female	AT	primary	pension	0.000	0.000	19178.750
male	Other	secondary	full time	17876.360	0.000	0.000
female	AT	secondary	student	1937.150	0.000	0.000

4.1.3 Requirements

Accuracy requirements. The hierarchal structure (household) should be represented in a realistic way in the synthetic data (e.g., no households where the oldest person is younger than 16 years old).

The *quality* of the synthetic data should be evaluated using the *total variation distance for each pairwise variables*.

Additionally, the following variable combinations should be considered for evaluation:

Employment status • Age group • Gender • District
Employment status • Citizenship • Gender • District
Employment status • Country of birth • Gender • District
Employment status • Education • Gender • District
Urban • Education • Personal income
Age group • Education • Personal income
Employment status (self-employed/employed) • Personal income • Urban • Age group
Equivalised household income[4] • District • Urban
District • Median age in household
District • Number of people living in household older/younger 16 years
District • Gender of main earner in household • Employment status (self-employed/employed)
District • Citizenship of main earner in household

Legal requirements. The disclosure of individual persons or households must not be possible. The legal requirements details are described in Bundesstatistikgesetz[5]:

> (2) Die Statistiken sind in solcher Weise zu veröffentlichen, dass ein Rückschluss auf Angaben über bestimmte oder bestimmbare Betroffene ausgeschlossen werden kann, es sei denn, dass der Betroffene an der Geheimhaltung der Angaben kein schutzwürdiges Interesse hat. Kann ein Rückschluss nicht ausgeschlossen werden, so darf nach vorheriger ausdrücklicher schriftlicher Zustimmung des Betroffenen die Veröffentlichung vorgenommen werden.
>
> (3) Bei der Veröffentlichung sind insbesondere konkrete Hinweise der Betroffenen über die Möglichkeit von Rückschlüssen auf Angaben, an deren Geheimhaltung ein schutzwürdiges Interesse des Betroffenen besteht, zu berücksichtigen.
>
> (4) Die Organe der Bundesstatistik sind verpflichtet, ihre Tätigkeitsberichte und Arbeitsprogramme im Bereich der Bundesstatistik unverzüglich der Bundesanstalt „Statistik Österreich" zur Kenntnis zu bringen.

Privacy requirements. Privacy can be addressed by computing k-anonymity on a limited set of variables and comparing the high-risk individuals in the real

[4] Equivalised household income. https://short.wu.ac.at/income
[5] The Bundesstatistikgesetz is available at https://short.wu.ac.at/bundesstatistikgesetz

data set with the synthetic data set. The set of variables for identifying high risk persons should be: `age group` • `gender` • `district` • `education` • `personal income >100k`

Technical requirements. Currently Db2 is used as a database system. However, Statistics Austria could easily transfer to PostgreSQL[6] or MariaDB[7] for processing. In general, if open database connectivity (ODBC) or Java database connectivity (JDBC) is used, it should be easy to connect to the Db2.

Ubuntu Linux Server[8] with current versions of R[9], Python[10] and relevant machine learning packages is available. Additional necessary packages can be installed if needed. The server is currently limited to 150GB of memory and 14 cores.

Frequency requirements. The original data set is updated quarterly. The same would probably apply to the synthetic generation.

Latency requirements. There are no specific requirements regarding latency.

Constraint requirements. There are a lot of implicit constraints in the data. For example, for the employment status, being employed is only possible above a certain age and when a person is employed a minimum income should be found in the data. However, at the moment, a complete list concerning employment status (or other variables) stating these constraints is not available.

4.2 Synthetic bank customer base

The synthetic bank customer base use case was provided by George Labs.

4.2.1 General description

Working with financial data is severely restricted. For product development or when collaborating with partners, George Labs must, most of the time, rely on best guess instead of hard facts based on the data. George Labs would like to create a customer base that includes representative product associations and corresponding transactions that can be used to shape the product and allow partners to develop without endangering any data protection concerns.

The synthetization engine should run in the background and provide up-to-date data on demand in any quantity desired. Having up-to-date data is very important, since financial data fluctuates and shifts over time.

As George Labs is working with several entities in the Bank, they want to use the data without major adaptions in all markets George Labs (Erste Group Bank) is active. The system should be flexible enough to learn and produce usable data when given data from different markets (i.e., Austria, Czechia, Slovakia, Hungary, Romania, Croatia, Serbia).

[6] PostgreSQL. https://www.postgresql.org/
[7] MariaDB. https://mariadb.org/
[8] Ubuntu Linux Server. https://ubuntu.com/server
[9] R programming language. https://www.r-project.org/
[10] Python. https://www.python.org/

The concept of the use case is depicted in Figure 15. Bank data store (i.e., the authoritative data store or a data lake) contains all relevant data (customers, products, transactions including all relations). George Labs entities work with very different stores/systems, so there is a need to unify the data to make it at least similar in structure. The data are then fed into the synthetization engine and should update the model incrementally to prevent regular, very expensive, large-data exports from the authoritative systems. Synthetic data are stored in the storage that could be less restrictive and accessible by authorized persons.

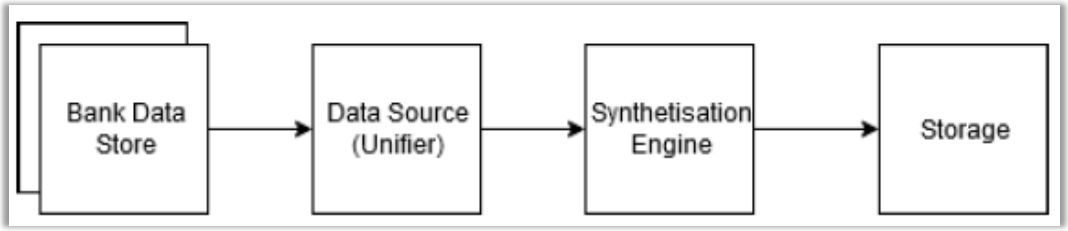


Figure 15: The synthetic bank customer base use case concept

Stakeholders. Maurizio Poletto (George Labs), Stefan Häbich (George Labs), and Verena Warringer (George Labs) are the stakeholders of this use case.

Benefits. The main benefit of using synthetic data is that the product development could be based on realistic data. Data intensive ideas (e.g., pro-active features, data analytics, and digital advisory) would be realized and shaped early in the process, instead of doing that with friendly customers.

Partners could test their products on realistic data as well, and George Labs would be able to verify the utility of various approaches/vendors without large investments and expensive proof of concept projects.

Alternatives. The alternative to using synthetic data would be custom creation of test data as George Labs would expect the data to be. However, the research has shown that product owner's expectation of user data is very biased.

Another alternative is to get prototypes, with production quality code but maybe not the full feature set, into production and to test them with friendly customers.

One more alternative is to use only aggregated data/statistics.

4.2.2 Requirements

Accuracy requirements. The data should be realistic and offer a good idea of how customers in the real Bank systems look and behave. The best result would be achieved if for every generated user, a user, who has the same characteristics (e.g., number of products, distribution of transactions in terms of volume, amount, number of partners) could be found in the real data set within a certain spread.

Regularity in the data should be consistent. For instance, credit cards are always charged on a specific date; certain payments are always done at the beginning/middle/end of the month; some payments have quarterly or yearly intervals.

Stable associations (e.g., brand loyalty, family ties, etc.) should be detectable and representable in the data. However, brands themselves should not be represented in the data, at most, some abstract representation in categories could be present (e.g., discounter or premium stores).

The data should have the same representative quality in terms of the sociodemographic and behavioral data that we know of. For example, relation between app and web usage, between gender, age, location, family status, education, and income.

The evaluation of synthetic data should integrate realistic data about tracked event patterns (i.e., user journeys) as, for example, login-payment-logout pattern.

Legal and privacy requirements. George Labs is dealing with data that identify individuals and their behavior, as a result, GDPR applies to those data. Additionally, since George Labs is regulated by Bankwesengesetz (BWG), which, is in many cases, even stricter than GDPR, there is also a requirement to satisfy the constraints defined by BWG.

Frequency requirements. The data set should be updated monthly.

Latency requirements. There are no specific requirements regarding latency.

Constraint requirements. There are no specific constraint requirements.

4.3 Master's thesis

A Master's thesis with a working title "Synthetic data: A new approach for marketing analytics in an increasing environment of data protection" is being written in the context of the work package two. The goal of the thesis is to find out whether companies of various industries see potential benefits or downsides of synthetic data for innovative and market(ing) research activities. The work aims to also provide deeper insights into the requirements of synthetic data for marketing analytics via additional use case collection.

In order to collect additional use cases and requirements, semi-structured expert interviews were conducted. The relevant interview partners were identified in a three-step approach. Firstly, a cross industry search for potential interviewees was conducted. Based on this search, potential partners were contacted via email introducing the ANITA project. Secondly, companies, that were interested in the project, were asked to participate in a short telephone interview to find out if there were potential use cases that could be shared with ANITA for the purpose of our research. Lastly, the follow-up interview was organized, where companies described their use cases for sharing privacy-sensitive sequential data with third parties.

The thesis is still work in progress and will be made available separately as soon as it is finished.

4.4 Conclusion

In the work package two we systematically collected use cases for sharing privacy-sensitive sequential data with third parties as well as captured

requirements with respect to accuracy and privacy. We documented both the general description of use cases and detailed information about each use case (i.e., number of subjects, frequency / latency for data sharing, accuracy & privacy requirements, technical requirements, etc.) In this section we described the nonconfidential part of the use cases of our project partners.

The general idea of the Statistics Austria use case lies in sharing the synthetic micro data of the Austrian population with the public and with the scientific community to enable innovative research and to support policy making.

In their use case, George Labs would like to create a customer base that includes representative product associations and corresponding transactions that can be used to shape the product and allow their partners to develop products without endangering any data protection concerns.

Additional use cases are being collected in the context of the Master's thesis with a working title "Synthetic data: A new approach for marketing analytics in an increasing environment of data protection". The work aims to provide insights into the requirements of synthetic data for market(ing) analytics.

The use cases and requirements described in this deliverable will serve as the basis for the work packages four and five. The goal there is to set up and run a data lab, that can be used for experimentation by generating artificial datasets to be used for validating accuracy and privacy. The existing model architectures could also be refined in order to meet the requirements established by work package two.

5 Annexes

5.1 Use case description template

ANITA Use Case Template
Version 1.2

Dear Experts,

The ANonymous bIg daTA (ANITA) project aims to systematically examine and validate the feasibility of using artificial intelligence and advanced machine learning to generate synthetic data that preserve individual privacy as well as to retain enough substantive and statistical information to ascertain its usefulness for market(ing) research purposes. In the face of stricter data protection regulations within Europe (GDPR), the success of this approach would allow safe cross-organizational data sharing and thus facilitate data-driven innovation and research processes distributed across industries.

In order to systematically capture use cases and requirements for **sharing synthetic, yet statistically representative data**, we kindly ask you to fill in this template.

Please mark / stipulate if any piece of information provided about the use case or existing systems needs confidential treatment (i.e., project internal only, not for dissemination purposes).

Thank you for your collaboration.

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.]