



ANITA

**Anonymous
big data A**
project funded
by FFG

AI-Based privacy-preserving big data sharing for market research

Anonymous big data workshop report

Authors: Thomas Reutterer (WU), Peter Eigenschink (WU), Olha Drozd (WU), Stefan Vamosi (WU), Michael Jakl (George Labs), Alexandra Ebert (Mostly AI), Tobias Hann (Mostly AI), Michael Platzer (Mostly AI)

Document version: 1.0

ANITA workshop report

1	INTRODUCTION	3
2	WORKSHOP FORMAT	3
3	WORKSHOP OUTCOMES	5
3.1	OPPORTUNITY.....	5
3.2	UTILITY.....	7
3.3	LEGAL.....	9
3.4	TRUST.....	10
3.5	COMMUNICATION.....	12
3.6	ETHICS.....	17
4	CONCLUSION	18

1 Introduction

The Anonymous Big Data workshop was organized in the context of the ANonymous big daTA (ANITA) project¹. ANITA 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 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 (General Data Protection Regulation (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.

2 Workshop format

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?

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 movement². 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

¹ AI-Based Privacy-Preserving Big Data Sharing for Market Research project. <http://anonymousbigdata.net/>

² Fullan, Michael. The taking action guide to building coherence in schools, districts, and systems. Corwin Press, 2016.

ANITA workshop report

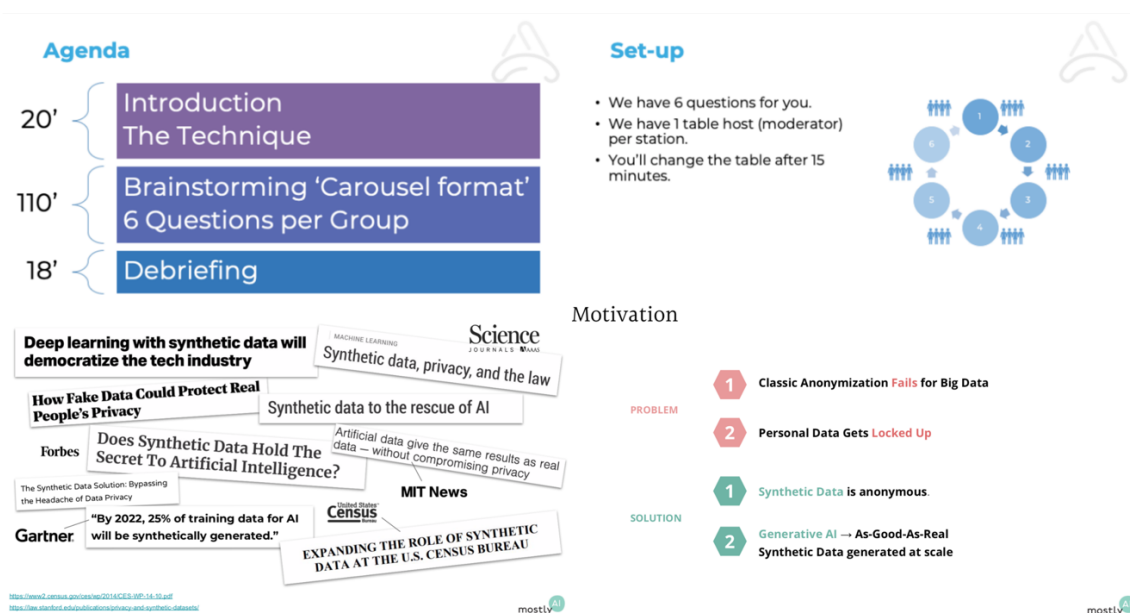


Figure 1: ANITA workshop introduction and set-up

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.

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

ANITA workshop report



Figure 2: ANITA workshop carousel brainstorming process

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.

3 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.

3.1 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

ANITA workshop report

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

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

ANITA workshop report

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?)

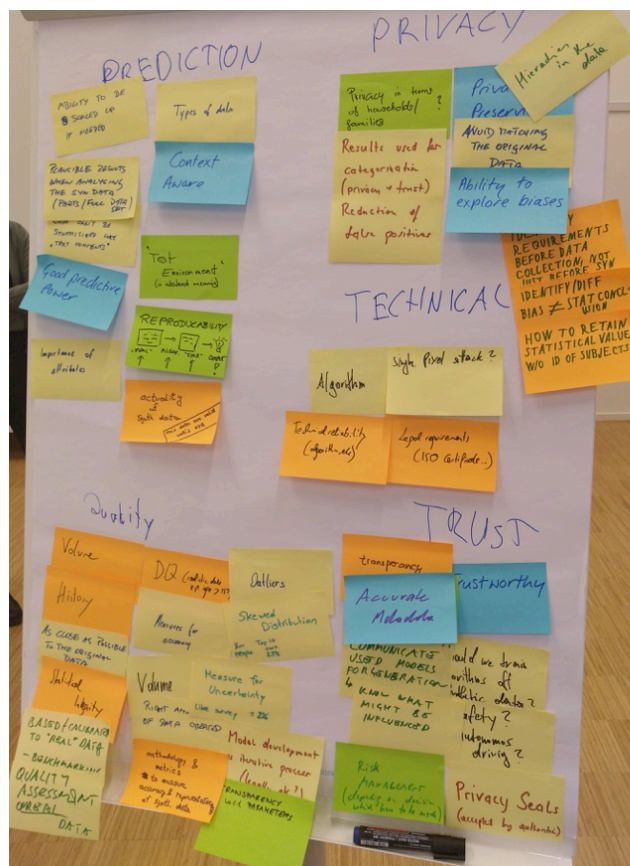


Figure 4: ANITA workshop, Utility station

3.3 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.

3.4 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:

<i>Stakeholder group</i>	<i>Question examples</i>
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:

ANITA workshop report

- 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.

ANITA workshop report

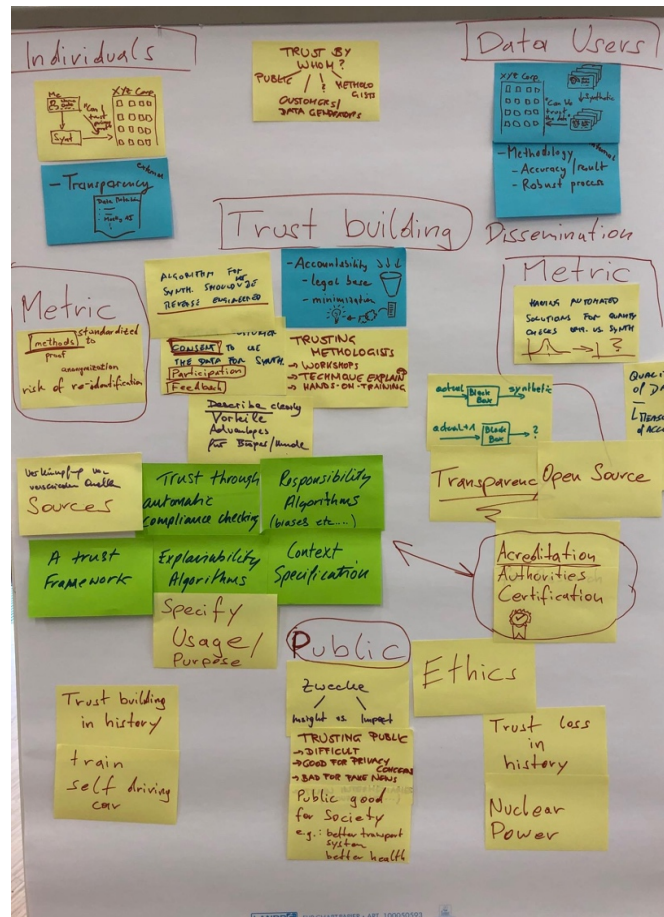


Figure 5: ANITA workshop, Trust station

3.5 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.

ANITA workshop report

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

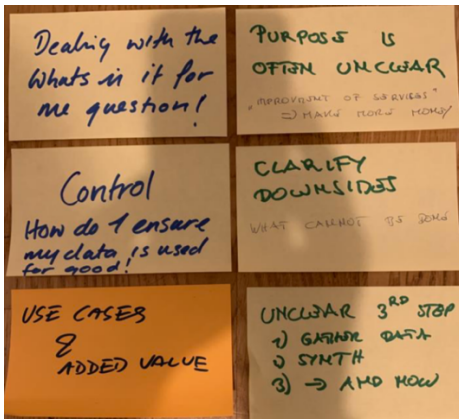


Figure 6: ANITA workshop, Communication station, "What for?" cluster

- 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).

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.

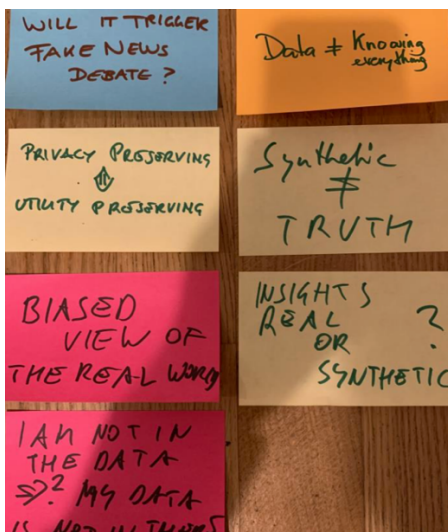


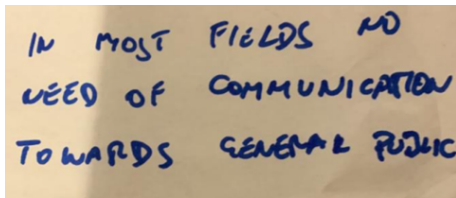
Figure 7: ANITA workshop, Communication station, "Quality issues" cluster

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

ANITA workshop report

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



IN MOST FIELDS NO
NEED OF COMMUNICATION
TOWARDS GENERAL PUBLIC

Figure 8: ANITA workshop, Communication station, "Need to communicate" cluster

- Is it actually necessary to communicate how the data are synthesized?
- 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:

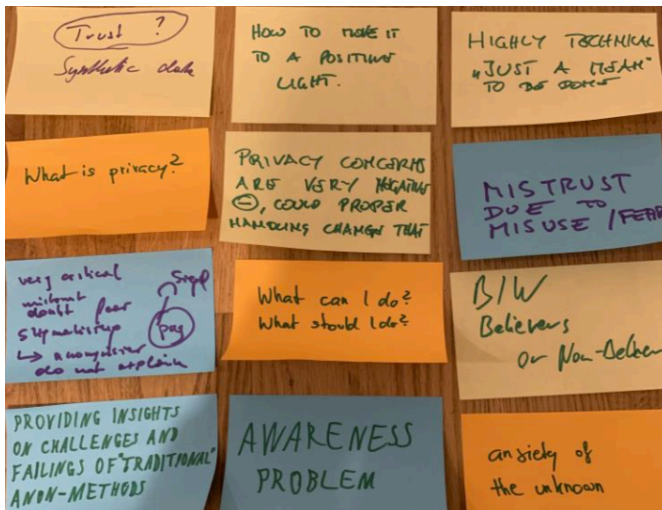


Figure 9: ANITA workshop, Communication station, "Trust issues" cluster

- 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 synthesis 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..."

ANITA workshop report

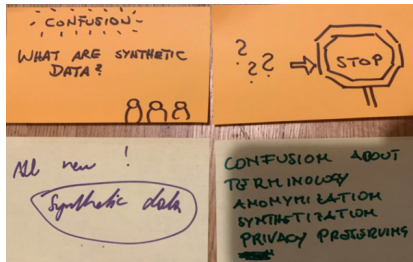


Figure 10: ANITA workshop, Communication station, "Hard to understand" cluster

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.

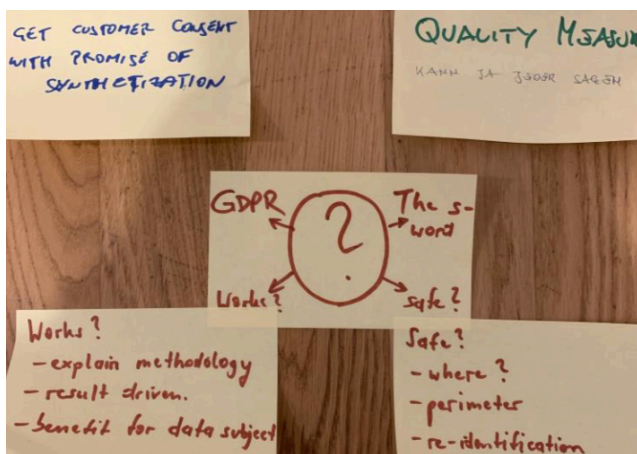


Figure 11: ANITA workshop, Communication station, "Compliance & transparency" cluster

- 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?

3.6 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?

ANITA workshop report

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?”).

All the proposed ideas, questions, and concerns will serve as a guidance for ANITA project in general as well as for future work beyond ANITA.