Synthetic Data

Industry Connections Activity Initiation Document (ICAID)

Version: 1.0, 4 November 2021

IC21-013-01 Approved by the IESS SMDC 13 December 2021

Instructions

- Instructions on how to fill out this form are shown in red. It is recommended to leave the instructions in the final document and simply add the requested information where indicated.
- Shaded Text indicates a placeholder that should be replaced with information specific to this ICAID, and the shading removed.
- Completed forms, in Word format, or any questions should be sent to the IEEE Standards Association (IEEE SA) Industry Connections Committee (ICCom) Administrator at the following address: industryconnections@ieee.org.
- The version number above, along with the date, may be used by the submitter to distinguish successive updates of this document. A separate, unique Industry Connections (IC) Activity Number will be assigned when the document is submitted to the ICCom Administrator.

1. Contact

Provide the name and contact information of the primary contact person for this IC activity. Affiliation is any entity that provides the person financial or other substantive support, for which the person may feel an obligation. If necessary, a second/alternate contact person's information may also be provided.

Name: Alexandra Ebert
Email Address: alexandra.ebert@mostly.ai
Employer: MOSTLY AI Solutions MP GmbH
Affiliation: -

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

2. Participation and Voting Model

Specify whether this activity will be entity-based (participants are entities, which may have multiple representatives, one-entity-one-vote), or individual-based (participants represent themselves, one-person-one-vote).

Individual-Based

3. Purpose

3.1 Motivation and Goal

Briefly explain the context and motivation for starting this IC activity, and the overall purpose or goal to be accomplished.

Data is now at the core of every technological, societal, and economic advance and organizations are under increasing pressure to become data-driven and offer personalized services to meet their customers' expectations. Thus, there is a rising need to utilize customer data. However, by 2023, 65% of the world's population will have its personal information covered under modern privacy regulations,

up from 10% in 2020 (Gartner) and already now EU's GDPR and the US' CCPA present organizations with the challenge to find privacy-preserving ways of utilizing customer data. While anonymization of customer data would be a solution, as fully anonymous data is exempt from privacy legislation, researchers have shown over and over again that legacy anonymization techniques (e.g. masking, obfuscating) fail in the era of big data and are not able to protect individuals from re-identification in supposedly anonymous datasets. Moreover, due to the destructive nature of these approaches (i.e. all parts of a dataset that could be re-identifying need to be deleted), it is not possible to preserve the utility of traditionally anonymized data, which significantly limits its usability for analytical purposes and AI training. This has resulted in customer data being largely locked up, creating a barrier to data-driven innovation.

Recent advancements in deep learning and an increase in computational power facilitated the development and early adoption of an emerging anonymization and privacy protection technique: **synthetic data generation**. Synthetic data is artificial data that is generated based on original customer data. It is highly realistic and statistically representative to the original data and thus suitable to serve as a drop-in replacement for it (e.g. for AI training). Yet – when generated with appropriate privacy mechanisms – synthetic data is fully anonymous and impossible to re-identify. (Short video to learn more). Thus, **synthetic data is as-good-as real data, but it's free and privacy-safe to use**, share and collaborate with in compliance with GDPR, CCPA and all other emerging privacy legislations.

Besides creating replica datasets, **synthetic data is also capable of augmenting data to reduce bias and to correct imbalances.** A research - also by Gartner - estimates that by 2022, 85% of algorithms will be erroneous due to bias. Bias and discrimination of AI systems are problems that are already being taken seriously, and synthetic data can contribute to mitigate bias with **fair synthetic datasets** representing the world, not as it is, but as we would like to see it. For instance, without gender-based or racial discrimination.

Due to synthetic data's immense potential to reconcile data utilization with data & privacy protection, there are already enterprise organizations in the financial services, insurance, healthcare and telco industry, as well as public sector organizations using synthetic data for AI training, analytics, digital product development, cross-border data sharing and testing. However, **there are no commonly agreed criteria for measuring the accuracy and privacy of synthetic data-generating platforms.**

Based on a variety of discussions with customers, regulators, analyst firms and academics, we have seen a need for synthetic data privacy and accuracy standards. However, due to the novelty of this technology and a not yet existent synthetic data community with members from synthetic data producers, synthetic data users, academia as well as from the regulatory side, we were advised by IEEE to first start out with a synthetic data IC activity to build the community and discuss, how standardization of this new technology could best be approached. Besides laying the groundwork for the submission of a proposal for synthetic data standards, the IC activity will also seek to advance the concept of fair synthetic data, as well as to support regulators in their understanding of this new technology and how it can be evaluated.

3.2 Related Work

Provide a brief comparison of this activity to existing, related efforts or standards of which you are aware (industry associations, consortia, standardization activities, etc.).

At the point of submitting this IC proposal, we are not aware of any industry associations, consortia or standardization activities that focus on synthetic data.

3.3 Previously Published Material

Provide a list of any known previously published material intended for inclusion in the proposed deliverables of this activity.

At this point, there are no published materials planned to be included in the deliverables of this IC activity.

3.4 Potential Markets Served

Indicate the main beneficiaries of this work, and what the potential impact might be.

By 2024, according to Gartner, 60% of the data used for the development of AI and analytics projects will be synthetically generated. Moreover, Gartner predicts that by 2025, businesses will avoid 70% of privacy violation penalties by supplementing personal data collection with synthetic data.

Synthetic data is an enabling technology for privacy-preserving data utilization that will impact many industries and AI systems moving forward. However, our IC program intends to begin by focusing on synthetic data for the financial services and insurance sector in the EU, UK and North America, as those industries and geographical regions are currently the most mature when it comes to adoption of this new technology.

3.5 How will the activity benefit the IEEE, society, or humanity?

Synthetic data facilitates privacy-preserving data utilization and has the potential to increase algorithmic fairness. Humanity will benefit from synthetic data, as it is an enabling technology that will unlock valuable data assets for economic and scientific progress (e.g. in healthcare), while securely preserving the privacy of every individual present in the original data.

In regard to the IEEE, we see the work that will emerge out of this IC activity complementing those technical and socio-technical efforts which are already being undertaken in IEEE's AI and data portfolio. Moreover, we look forward to collaborating with all interested parties as our work kicks forward.

4. Estimated Timeframe

Indicate approximately how long you expect this activity to operate to achieve its proposed results (e.g., time to completion of all deliverables).

Expected Completion Date: 12/2023

5. Proposed Deliverables

Outline the anticipated deliverables and output from this IC activity, such as documents (e.g., white papers, reports), proposals for standards, conferences and workshops, databases, computer code, etc., and indicate the expected timeframe for each.

- Q1/2022: **Building the community** by gathering leading privacy and synthetic data researchers, potential synthetic data users from the target industries, various synthetic data vendors and members from regulatory bodies

- Q2/2022: Discussion on how to best approach standardization of synthetic data + recommendation for a standard project authorization request **for a synthetic data privacy and accuracy standard**

Once the above big milestone of the IC activity is reached, the IC activity will dedicate its time to work on other deliverables. The anticipated output during Q3/2022-11/2023 includes:

- Define **typical uses, best practices, and application orientation** for synthetic data
- A series of workshops + final report to advance the concept of **fair synthetic data** (e.g. by answering questions like "Which fairness definitions should fair synthetic data fulfill to ensure fairness of downstream machine learning applications that are trained on top of it?")
- An educational synthetic data **white paper for data protection authorities** and other regulatory bodies
- Creation of a **webinar about synthetic data for AI auditing and explainable AI**
- Defining criteria catalogs for standardized **open synthetic datasets** to enable lesser resourced countries and SMBs to innovate at a more competitive level
- Develop standardized **naming schematics, APIs, and ontological structures** to receive vertically oriented information and support improved interoperability amongst information and AI systems.

However, once the group kicks off, we anticipate refinement and re-prioritization of the work items outlined for the post-Q2/2022 period.

5.1 Open Source Software Development

Indicate whether this IC Activity will develop or incorporate open source software in the deliverables. All contributions of open source software for use in Industry Connections activities shall be accompanied by an approved IEEE Contributor License Agreement (CLA) appropriate for the open source license under which the Work Product will be made available. CLAs, once accepted, are irrevocable. Industry Connections Activities shall comply with the IEEE SA open source policies and procedures and use the IEEE SA open source platform for development of open source software. Information on IEEE SA Open can be found at https://saopen.ieee.org/.

Will the activity develop or incorporate open source software (either normatively or informatively) in the deliverables? Yes, if APIs are developed.

6. Funding Requirements

Outline any contracted services or other expenses that are currently anticipated, beyond the basic support services provided to all IC activities. Indicate how those funds are expected to be obtained (e.g., through participant fees, sponsorships, government or other grants, etc.). Activities needing substantial funding may require additional reviews and approvals beyond ICCom.

No funding requirements.

7. Management and Procedures

7.1 Activity Oversight Committee

Indicate whether an IEEE Standards Committee or Standards Development Working Group has agreed to oversee this activity and its procedures.

Has an IEEE Standards Committee or Standards Development Working Group agreed to oversee this activity? No

IEEE Committee Name: Committee Name
Chair's Name: Full Name
Chair's Email Address: who@where

Additional IEEE committee information, if any. Please indicate if you are including a letter of support from the IEEE Committee that will oversee this activity.

IEEE collects personal data on this form, which is made publicly available, to allow communication by materially interested parties and with Activity Oversight Committee and Activity officers who are responsible for IEEE work items.

7.2 Activity Management

If no Activity Oversight Committee has been identified in 7.1 above, indicate how this activity will manage itself on a day-to-day basis (e.g., executive committee, officers, etc).

The activity will be managed by the officers.

7.3 Procedures

Indicate what documented procedures will be used to guide the operations of this activity; either (a) modified baseline *Industry Connections Activity Policies and Procedures,* (b) Standards Committee policies and procedures accepted by the IEEE SA Standards Board, or (c) Working Group policies and procedures accepted by the Working Group's Standards Committee. If option (a) is chosen, then ICCom review and approval of the P&P is required. If option (b) or (c) is chosen, then ICCom approval of the use of the P&P is required.

The activity will follow the abridged IC Activity policies and procedures.

8. Participants

8.1 Stakeholder Communities

Indicate the stakeholder communities (the types of companies or other entities, or the different groups of individuals) that are expected to be interested in this IC activity and will be invited to participate.

- Providers of synthetic data/vendors
- Users of synthetic data (mainly from the banking and insurance industry)
- Data protection authorities & other regulators
- Privacy researchers
- Deep Learning/synthetic data researchers
- Potentially members of digital human right activist groups
- Potentially privacy lawyers
- Potentially consulting firms with focus on (Ethical) AI and Privacy

8.2 Expected Number of Participants

Indicate the approximate number of entities (if entity-based) or individuals (if individual-based) expected to be actively involved in this activity.

15-25

8.3 Initial Participants

Provide a number of the entities or individuals that will be participating from the outset. It is recommended there be at least three initial participants for an entity-based activity, or five initial participants (each with a different affiliation) for an individual-based activity.

Use the following table for an individual-based activity:

Individual		**Employer**	**Affiliation**
Alexandra Ebert	Synthetic data vendor	MOSTLY AI	
Michael Platzer	Synthetic data vendor	MOSTLY AI	
Klaudius Kalcher	Synthetic data vendor	MOSTLY AI	
Jochen Papenbrock	AI & synthetic data expert for financial services	Nvidia	
Yves-Alexandre De Montjoye	Privacy researcher	Imperial College London	
Thomas Reutterer	Synthetic data researcher	Vienna University of Economics and Business	
Omar Ali Fdal	Synthetic data vendor	Statice	
Jessie Lamontagne	Data Scientist/User	Kinaxis	
Behrang Raji	Data Protection Authority	Hmb BfDI Germany	

8.4 Activity Supporter/Partner

Indicate whether an IEEE committee (including IEEE Societies and Technical Councils) has agreed to participate or support this activity. Support may include, but is not limited to, financial support, marketing support and other ways to help the Activity complete its deliverables.

Has an IEEE Committee agreed to support this activity? No

If yes, indicate the IEEE committee's name and its chair's contact information.

IEEE Committee Name: Committee Name
Chair's Name: Full Name
Chair's Email Address: who@where

Please indicate if you are including a letter of support from the IEEE Committee.