دبـــــي الرقمية
DIGITAL DUBAI

UNLOCKING THE POWER OF DATA AND PROTECTING PRIVACY

FRAMEWORK FOR THE IMPLEMENTATION OF SYNTHETIC DATA TECHNIQUES

INTRODUCTION

DIGITAL DUBAI AUTHORITY (DDA)

Digital Dubai was established by His Highness Sheikh Mohammed Bin Rashid Al Maktoum, Vice-President & Prime Minister of the UAE, and Ruler of Dubai, in June 2021 to develop and oversee the implementation of policies and strategies that govern all matters related to Dubai's information technology, data, digital transformation, and cyber-security.

Digital Dubai brings together the expertise of four entities - Dubai Electronic Security Center, Dubai Statistics Center, Dubai Data Establishment, Smart Dubai Government Establishment, - to ensure a collaborative effort towards achieving the vision of the city's leadership to make Dubai a globally leading digital economy.

The entity has been entrusted with four key tasks - accelerate digital transformation of the city through strategic partnerships with governments and private sector entities, increase the Emirate's digital economy contribution to the city's GDP, build and develop digital competencies of national talent, and, maintain and develop Dubai's digital wealth whilst accelerating Dubai's cybersecurity efforts.

www.digitaldubai.ae

INTRODUCTION

DUBAI'S DATA & AI SUB-COMMITTEE

Dubai's Data & AI Sub-Committee was formed as part of the Digital Transformation Committee for the City of Dubai, led by His Excellency Hamad al Mansouri.

It has been established to co-ordinate data and AI related activities across three main areas, as follows:

- AI and automation;
- Data exchange; and
- Data infrastructure.

Membership is drawn from DDA, Dewa, Dubai Customs, Dubai Police, GDRFA, RTA and Dubai Municipality.

ACCOMPANYING RESEARCH

Dubai Digital Authority and Faculty AI conducted research to test the degree of privacy preservation, and the utility of synthetic data.

The Synthetic Data Report shared with this Framework makes it clear that on both counts private synthetic data techniques can deliver datasets that offer a compelling alternative to more commonly used privacy-preserving techniques, and which are suitable across a range of potential uses. In two distinct parts, it makes the case for synthetic data from the management perspective, as well as the data science perspective.

ACKNOWLEDGEMENTS: This framework was created by DEWA and the Digital Dubai Authority, under the auspices of the Data and AI Committee. We'd like to thank all those who have contributed along the way.

1

INTRODUCTION

- ABOUT THIS FRAMEWORK
- WIDER CONTEXT
- SYNTHETIC DATA: A DEFINITION
- THE CASE FOR SYNTHETIC DATA
- SYNTHETIC DATA'S POTENTIAL EXPLAINED
- DETAILED CASE STUDIES IN DIFFERENT INDUSTRIES AND SECTORS

ABOUT THIS FRAMEWORK

WHAT IS THIS FRAMEWORK FOR?

Synthetic data is an emerging field. There is justified excitement about its potential to open up data usage in a way that existing privacy preserving techniques do not. Exploring this potential needs to be done alongside work to understand:

- The best uses of synthetic data, acknowledging that there is a trade-off between data utility and innovation potential, and privacy (synthetic data is not always 100% privacy preserving).
- The most effective, risk-free ways of generating synthetic data, from a range of open and proprietary sources.
- The governance implications of synthetic data as they apply to:
  - Individual organizations; and
  - Dubai Data Establishment in its role as governor of data in the Emirate of Dubai.
- Monetization opportunities, and in a wider sense, how value in the digital economy through synthetic data is both enabled and measured.

Standards work (e.g. in the IEEE) is in its very early stages.

ABOUT THIS FRAMEWORK

WHO IS THIS FRAMEWORK FOR?

This framework is designed for use by:

- **Data professionals,** engaged in the day-to-day management, processing and use of data in tools and analysis.
- **Data governors,** whose job it is to consider the governance checks required around data use, to ensure compliance and coherence with (existing) policies and laws.
- **Data leaders,** responsible for embedding synthetic data practices in their organization.

The framework is designed for direct use by all parties, and to expose each party to the considerations of the others.

A full explanation of the team members and their respective roles around the identification, creation, use and monitoring of synthetic data features later in this framework.

The framework is aimed primarily at government entities, but private sector organizations working either on their own synthetic data use cases or for the government, are invited to use it also.

Chief Data Officers, or their equivalent, should hold accountability for following this framework within their own organization and providing feedback to DDA.

ABOUT THIS FRAMEWORK

HOW AND WHEN SHOULD IT BE USED?

The main feature of the framework is a set of canvasses.
These are designed to drive:

- **Discussion and deliberation;**
- **Development of approaches; and**
- **Ultimately decision making**

around all aspects of individual use cases in which synthetic data is under consideration.
It is hoped that the framework and its canvasses will be used in group settings
(e.g. to guide workshop discussions).

In each stage of the framework, the reader will also find useful resources to prepare for, and to go into more explicit detail on, aspects of synthetic data initiatives.

HOW TO READ THIS DOCUMENT

Because it is aimed at multiple audiences and covers both technical and policy issues. We have attempted to signpost readers to content accordingly.

Further, canvas summaries are aimed at leadership (e.g. Chief Data Officers), whilst the full canvasses and additional guidance materials are to be used by data stewards to guide more detailed discussions across a range of domain experts.

WIDER CONTEXT (I)

HOW DOES IT RELATE TO DDE'S WIDER DATA GOVERNANCE ARRANGEMENTS?

Because synthetic data is still in the very early stages of adoption, this Synthetic Data framework is not to be treated as guidance. Indeed, as we will make clear, we want to establish a feedback loop which generates learning, ultimately allowing us to set policies and establish good practice, to achieve on the wider ambitions for synthetic data.

As with all data activities, existing policies and tools in allied areas like Artificial Intelligence are available for use, as follows:

- Data law, policies and standards (and classifications*)
- Ethical AI guidelines and self-assessment tool
- Data quality tools and compliance tool
- Data Maturity Index

These last two are available to Dubai government entities only.

A SANDBOX FOR SYNTHETIC DATA

Upon the release of this framework and the accompanying research report, a technical and governance sandbox will be hosted in Dubai Data Establishment.

Entities are encouraged to test their own use cases using the canvasses, but to then have these validated in the sandbox. The sandbox will offer advice on the creation, use and monitoring of synthetic data use cases.

As with our work on AI ethics, our intention here is to create an evidence base and shared learning facility. This evidence will then be used to create new policies and adapt data practice (e.g. data classification or labelling systems to account for this new form of data and the data lineage issues it creates).

For more information, please contact
syntheticsandbox@digitaldubai.ae

DISCLAIMER: Implementation of this framework is the responsibility of respective entity and should be done based on criticality of their data and related risk assessment.

* Please note classifications (data classifications) in this document is defined as DDA's own data classification scheme (open, confidential, sensitive and secret) and not that used in machine learning approaches. (see Dubai Data Manual – Module 8 (data classification).

WIDER CONTEXT (II)

A NEW ERA FOR CITY DATA VALUE CREATION

Spurred by the creation of the Dubai Data Law of 2015, the government of Dubai has sought to create a world class data ecosystem in the Emirate. Through a range of initiatives we have built the infrastructure, policies, institutional and workforce capacity, and culture to manage and drive value out of data.

Much of this work has focused on an open data publishing strategy.

Through promoting synthetic data, we are signaling our intention to move into a new phase of data publishing – one that will overcome some of the challenges associated with open data publishing and at the same time create more data that can be shared in this way.

Ultimately, we are in the business of increasing data availability and exchange, purpose-oriented collaboration and data impact. We think synthetic data pushes us further along that route.

WHAT IS SYNTHETIC DATA: A DEFINITION

1 SYNTHETIC DATA

At its simplest, synthetic data is an **innovation-enhancing replacement for real world data**, and a compelling alternative to traditional data anonymization techniques.

It is **artificially generated** by an AI algorithm that has been **trained on the real data set. Retaining the structure and statistical integrity of original data**, synthetic data has the same predictive power as the original data but replaces it rather than disguising or modifying it.[1]

Synthetic data can be produced at **different levels of accuracy. Its use is not without risk**. Making more accurate synthetic data is more complicated and computationally intensive, and more likely to disclose personal or confidential information. However, it offers more value to the user of the data set because they can learn more from it.[2]

In principle, **risks of disclosing personal or other sensitive information** through synthetic data can be mitigated through the application of advanced privacy preserving techniques like **differential privacy.**[2]

FURTHER PRIVACY PRESERVING STEPS

2

A robust mathematical framework for **limiting statistical disclosure** whilst **controlling privacy risk.**

DIFFERENTIAL PRIVACY

3

The particular **type of synthetic data** that is generated, following the application of **differential privacy techniques.**

Private-synthetic DATA

1. MIT Sloan Management review. The real deal about synthetic data. (2022, link)
2. ADRUK. Accelerating public policy research with synthetic data. (2021, link)

THE CASE FOR SYNTHETIC DATA (I)

For governments around the world, privacy protection is the single most important consideration in sharing and unlocking the value of data. Policies to date have been largely successful in protecting privacy but heavy access restrictions can minimize the chances of data being used to help address public policy challenges, or to create wider value in rapidly digitizing public services and economies. To share data safely is to increase its value exponentially.

It therefore makes sense to look beyond existing approaches - aggregating datasets ensures they are private, but by definition stops them being used for valuable analysis at the individual level, or in creating the pro-active, predictive city (government) services we know can be fed by highly granular, privacy preserving data.

In summary, synthetic data has the potential to greatly increase the supply of data into innovation ecosystems operating within the broader (digital) economy. It can overcome data scarcity and quality issues that can hold back data hungry AI systems and enable data for use in sensitive research and development exercises, that stall because of data governance issues related to the identification of individuals.

CASE STUDY: NORWEGIAN GOVERNMENT'S USE OF SYNTHETIC DATA[1]

Government.no

In their **National AI Strategy**, the Norwegian government advocates the use of synthetic data as an alternative to identifiable or anonymized data. They specifically state that datasets normally considered sensitive can be made open and accessible for use in research and innovation.

The **Norwegian Labour and Welfare Department** has been working with machine-learning models to generate synthetic data with characteristics identical to those of sensitive customer records.

Synthetic data for tens of thousands of customers is then used to test and develop software used in welfare services.

A self-service function also allows for synthetic data to be created in line with specific needs (e.g. amongst a particular group and to reflect an unusual situation).

1. How AI contributes to better privacy in the Norwegian Labour and Welfare Administration. (2020, link)

THE CASE FOR SYNTHETIC DATA (II)

SYNTHETIC DATA'S POTENTIAL IN SUMMARY

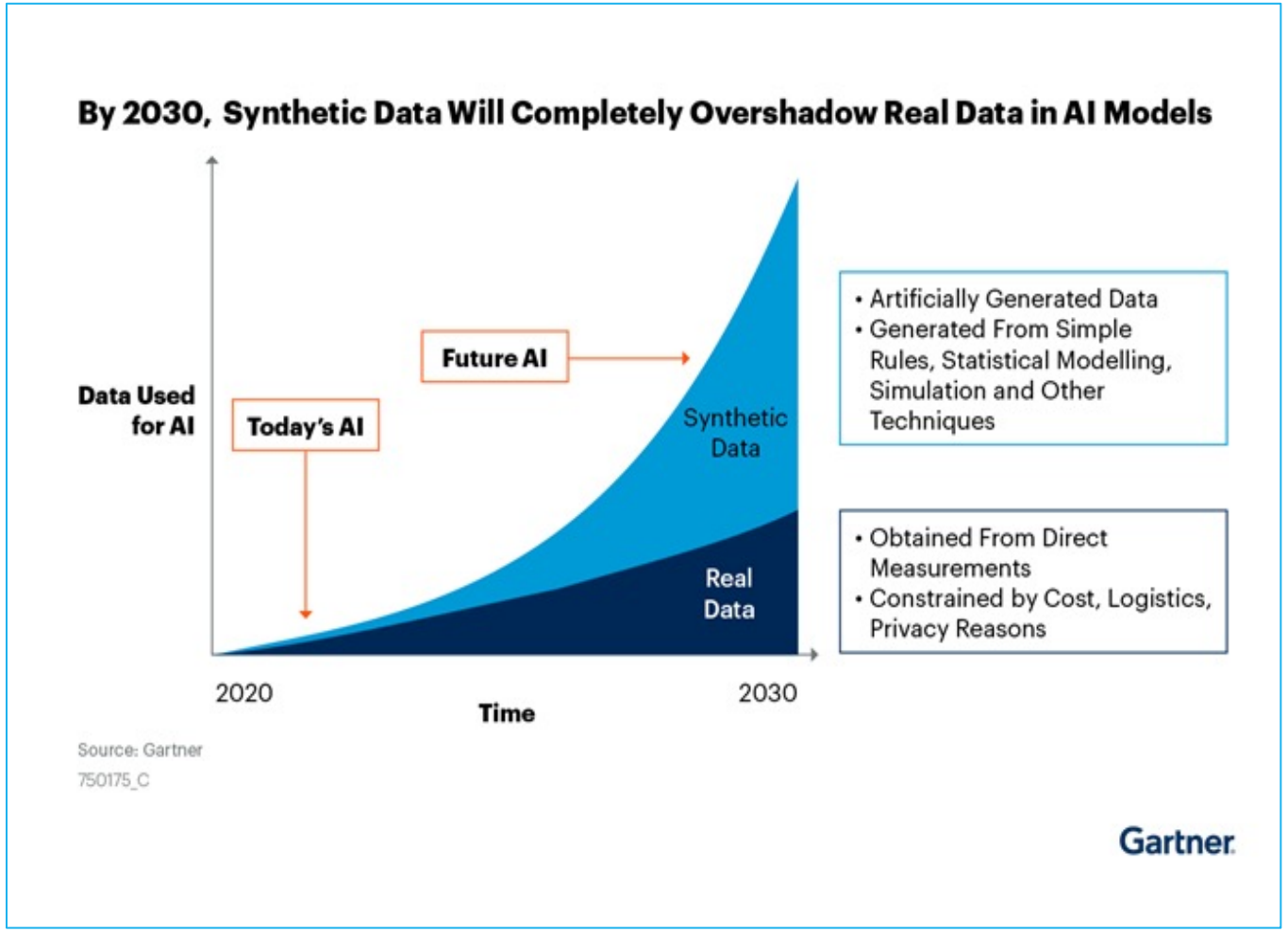
- Training AI and ML models when real world data is lacking in quality and/or quantity.
- Accelerating model development and making this activity cheaper when data is scarce.
- Testing and debiasing AI models using data that copies known demographics.

- Overcoming privacy and security issues in software development and testing.
- Replacing sensitive Personally Identifiable Information (PII) in areas like health and finance.

- Freeing up data for use in research, open innovation exercises (e.g. hackathons) and data collaborations with the start-up community.
- Simulating possible real-world scenarios by creating digital twins (e.g. predictive maintenance, population profiles for policy and marketing use cases).
- Deriving commercial value through synthetic data generation and monetization. (mainly applicable to private sector).

Source: CIO Magazine. What is synthetic data: Generated Data to help your AI Strategy. (2022, link)

THE CASE FOR SYNTHETIC DATA (III)



INDUSTRY VIEW - EVIDENCE FROM GARTNER ON THE RISE OF SYNTHETIC DATA

According to Gartner: "By 2024, 60% of the data used for the development of AI and analytics solutions will be synthetically generated." [1]

According to a 2017 Harvard Business Review study, only 3% of companies' data meets basic quality standards.[2]

Based on a 2020 YData study, the biggest problem faced by data scientists was the unavailability of high-quality data.[3]

1. Gartner. Predict. (2021, link)
2. Harvard Business Review. Article by Tadhg Nagle, Thomas C. Redman, and David Sammon. (2017, link)
3. Ydata. Gonçalo Martins Ribeiro. What we have learned from talking with 100+ data scientists. (2020, link)

DETAILED CASE STUDIES IN DIFFERENT INDUSTRIES AND SECTORS

FINANCIAL SERVICES

Fraud identification is a major part of any financial service, but fraudulent transactions are rare. With synthetic fraud data, new fraud detection methods can be tested and evaluated for their effectiveness.
Customer analytics: Synthetic customer transaction data can be used to perform analysis to understand customer behavior. This is similar to use cases on “internal data sharing” however it is applicable more widely in finance where most customer data is private.

HEALTHCARE

Healthcare analytics: synthetic data enables healthcare data professionals to allow the internal and external use of record data while still maintaining patient confidentiality. Again, this is similar to internal data sharing use cases, however it is applicable more widely in healthcare where most customer data is private. Clinical trials: synthetic data can be used as a baseline for future studies and testing when no real data yet exists.

SECURITY

Synthetic data can be used to secure organizations’ online & offline assets. Two methods are commonly used:
Training data for video surveillance: to take advantage of **image recognition**, organizations need to create and train neural network models, but this has two limitations: acquiring the volumes of data and **manually tagging the objects**. Synthetic data can help train models at lower cost compared to acquiring and annotating training data.
Deep fakes: these can be used to test face recognition systems. are flexible and can deal with novel attacks.

Source: AI Multiple. Cem Dilmegani. Top 20 synthetic data use cases & applications in 2022. (2022, link)

DETAILED CASE STUDIES IN DIFFERENT INDUSTRIES AND SECTORS

MANUFACTURING

Quality assurance: it is hard to test a system to see whether it identifies anomalies since there are infinite possible anomalies. Synthetic data enables more effective testing of quality control systems, improving their performance.

HR

Employee datasets of companies contain sensitive information and are often protected with data privacy regulations. In-house data teams and external parties may not have access to these datasets but they can leverage synthetic employee data to conduct analyses. It can help companies to optimize HR processes.

MARKETING

Synthetic data allows marketing units to run detailed, individual-level simulations to improve their marketing spend. Such simulations would not be allowed without user consent due to GDPR or similar regulatory constraints. However synthetic data, which follows the properties of real data, can be reliably used in simulation.

AGILE DEVELOPMENT AND DEV-OPS

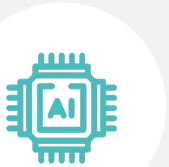
For software testing and **quality assurance**, artificially generated data is often the better choice as it eliminates the need to wait for 'real' data. Often referred to under this circumstance as **'test data'**, This can ultimately lead to decreased test time and increased flexibility and agility during development.

Source: AI Multiple. Cem Dilmegani. Top 20 synthetic data use cases & applications in 2022. (2022, link)

DETAILED CASE STUDIES IN DIFFERENT INDUSTRIES AND SECTORS

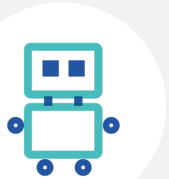
SOCIAL MEDIA

Social networks are using synthetic data to improve products. **Testing content filtering systems:** Social networks are fighting fake news, online harassment, and political propaganda from foreign governments. Testing with synthetic data ensures that the content filters are flexible and can deal with novel attacks.



MACHINE LEARNING

Most ML models require large amounts of data for better accuracy. Synthetic data can be used to increase training data size for ML models.
Prediction of rare events such as fraud or manufacturing defects is hard since small data size leads to inaccuracies for ML models. Generating synthetic instances of such events increases model accuracy.
Synthetic data generation creates **labeled data** instances, ready to be used in training. This reduces the necessity for time-consuming data labeling efforts.



AUTOMOTIVE & ROBOTICS

Autonomous Things (AuT): research to develop autonomous things such as robots, drones and self-driving car simulations pioneered the use of synthetic data. This is because real-life testing of robotic systems is expensive and slow. Synthetic data enables companies to test their robotics solutions in thousands of simulations, improving their robots and complementing expensive real-life testing.

DEEP LEARNING

Computer vision algorithms like object segmentation and semantic segmentation need a mask which is a very time-consuming process to create it. Synthetically creating the images with marks will speed-up the Deep Learning (DL) model building process.
Data augmentation creates more transformed versions of one single image to help the DL model to learn generically. Our model needs to predict “dog” as “dog” irrespective whatever the pose that dog is giving in the test image/real-time image. Data augmentation may help here to get different versions of one dog image.

Sources
AI Multiple. Cem Dilmegani. Top 20 synthetic data use cases & applications in 2022. (2022, link)
AI Multiple. Cem Dilmegani. Synthetic data to improve deep learning models. (2022, link)

2

SYNTHETIC DATA FRAMEWORK

- THE DECISION MATRIX FOR SYNTHETIC DATA: OVERVIEW
- SYNTHETIC DATA GENERATION PROCESS

SYNTHETIC DATA FRAMEWORK

OUR WORK HAS DRAWN ON IN-DEPTH RESEARCH OF POTENTIAL SYNTHETIC DATA APPROACHES AND SCENARIOS ON WHEN TO USE THEM.

THIS RESEARCH HAS BEEN TRANSLATED INTO A FLEXIBLE, PRACTICAL FRAMEWORK THAT COVERS TWO STAGES OF DEVELOPMENT.

A: The decision matrix

Identifies 5 key decision points, prompting and guiding discussions for all the key considerations around the successful use of synthetic data.

We have designed a decision matrix:

- to help build the case for the use of synthetic data;
- to ensure that the right skills feature in your team;
- to make sure that synthetic data is used safely and in a purpose-oriented way.

B: Synthetic data generation process

Identifies a high-level process to be followed in order to generate synthetic data

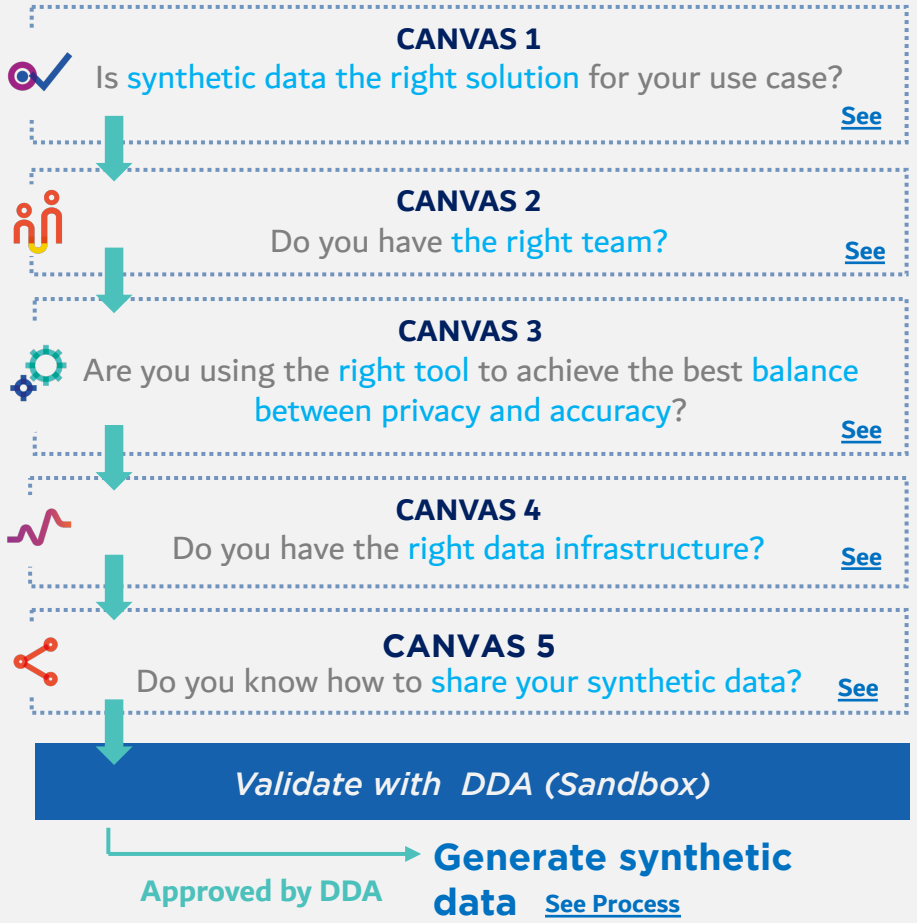
THE DECISION MATRIX FOR SYNTHETIC DATA: OVERVIEW

DO YOU HAVE ALL THE REQUIRED FACTORS FOR SUCCESSFUL SYNTHETIC DATA USE CASE?

To help teams consider the complex issues that emerge around synthetic data, we have drawn up a series of 'canvasses' to be filled in by the various team members, preferably in a workshop format. Each canvas contains further information sources and reading.

Please note, due to its experimental nature at the current stage:

1. Synthetic data techniques should not be applied to data classified as secret.
2. Synthetic data should not be openly publicly published in the first instance.
3. Differential privacy techniques should be considered where information disclosure risk is high.



SYNTHETIC DATA GENERATION PROCESS

AFTER USING THE DECISION MATRIX TO DECIDE IF YOUR TEAM IS READY TO USE SYNTHETIC DATA, THE PROCESS OF SYNTHETIC DATA GENERATION CAN BE FOLLOWED. DUE TO THE EXPERIMENTAL NATURE OF SYNTHETIC DATA, THIS SHOULD BE DONE WITH THE SUPERVISION AND APPROVAL OF DDA.

			SECURE ENVIRONMENT	EXTERNAL ENVIRONMENT	VALIDATION SERVER/SECURE ENVIRONMENT		
DESIGNATE TEAM & DETERMINE GOVERNANCE	**IDENTIFICATION & EVALUATION**	**ORIGINAL DATA PREPARATION**	**MODEL TRAINING**	**SYNTHESIZING**	**SYNTHETIC DATA VALIDATION**	**PREPARING & SHARING SYNTHETIC DATA**	**USE & MAINTAIN SYNTHETIC DATA**
• Assign a qualified team • Determine authority	• Identify use cases • Identify datasets • Evaluate datasets • Identify dataset classification	• Clean data • Create features • Determine access and presentation methods based on real data classification	• Train the generator on original data	• Generate the synthetic data	• Validate the model • Validate synthetic data • Evaluate synthetic data (e.g. accuracy and privacy)	• Classify synthetic data • Label synthetic data • Republish to repository	• Data as a service • Data monetization • Data for services • Data for operations • Data for research ...etc • Maintain and monitor synthetic data

3

RELATION TO EXISTING GOVERNANCE

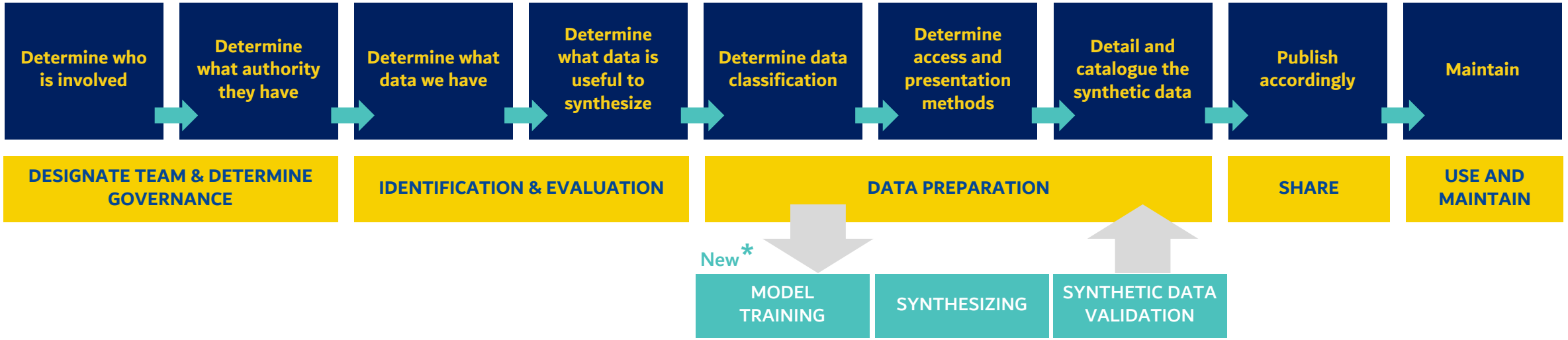
- HOW DOES THIS RELATE TO EXISTING DATA GOVERNANCE PROCESS?
- HOW DOES THIS RELATE TO EXISTING ISR PROCESS?

HOW DOES SYNTHETIC DATA RELATE TO DUBAI'S CURRENT GOVERNANCE MODEL

Existing data governance process:

1. Designate team
2. Determine governance
3. Inventory data
4. Prioritize data
5. Classify data
6. Prepare data release plan
7. Complete data cataloguing
8. Ingest data On platform
9. Maintain data

Relation to synthetic data generation process:



HOW DOES SYNTHETIC DATA RELATE TO ISR PROCESS?

The entities will be required to follow already **defined information governance, information classification, risk assessment, and access control** procedures as part of implementing synthetic data.

Currently, all Dubai Government entities are required to follow ISR for protecting their information as per their information / data classification scheme – which must be aligned with Dubai Data Law.

HOW DOES THIS RELATE TO ISR PROCESS?

EXISTING INFORMATION SECURITY CONTROLS SHOULD BE ADHERED TO AS PART OF BOTH SYNTHETIC DATA PROCESSING FOR BOTH REAL DATA (BEFORE GENERATING SYNTHETIC DATA) AND FOR SYNTHETIC DATA ITSELF:

1

Implement information classification

Define and implement information classification scheme/process to be used within the entity based on information assets criticality, value, legal and protection requirements, etc. in line with applicable laws and regulations.

2

Be sure to define proper classifications of information

Information assets owner has the responsibility of ensuring that proper classifications of information are defined (to include proper access control for the information).

3

Conduct a detailed risk assessment

Conduct and maintain a detailed risk assessment in accordance with the approved risk assessment methodology (entities are allowed to define their risk management framework based on ISR, ISO 31000, ISO 27001 etc.).

4

Determine governance and ISR

Determine and assess governance and information security risks related to its relations with external parties, which includes outsourcing and cloud services providers.

5

Document risk assessment results

6

Secure senior management approval

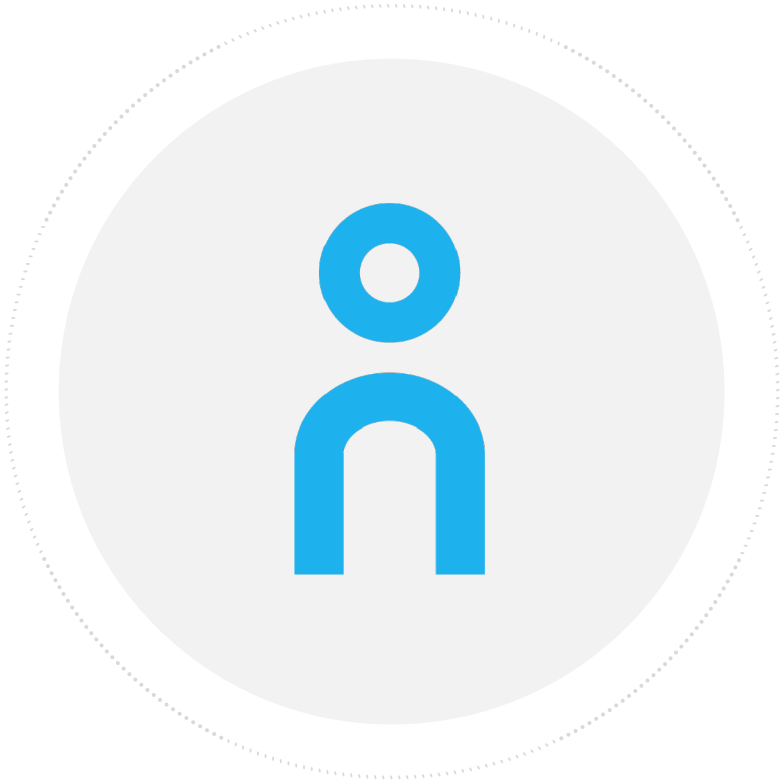
Approve it officially by the Information Security Steering Committee or senior management.

4

DECISION MATRIX CANVAS SUMMARIES

- SUMMARY 1: IS SYNTHETIC DATA THE RIGHT SOLUTION FOR YOUR PROBLEM?
- SUMMARY 2: DO YOU HAVE THE RIGHT TEAM?
- SUMMARY 3: BALANCING PRIVACY AND ACCURACY
- SUMMARY 4: WHAT IS THE APPROPRIATE INFRASTRUCTURE?
- SUMMARY 5: HOW TO SHARE SYNTHETIC DATA

DECISION MATRIX CANVAS SUMMARIES



CHIEF DATA OFFICERS LEADING THE WAY

Chief Data Officers or their equivalents are expected to take overall accountability for work completed in canvasses. The canvas summaries set out high-level questions (to be addressed by your team) to guide the ultimate decision on synthetic data use.

Because of the range of use cases, synthetic data generators, and options for sharing, the final decision on synthetic data use in your organization is effectively a judgement call by a Chief Data Officer or equivalent. This decision should be based on the organization's ability, resources, priorities and appetite for risk. It should then be validated in DDA's sandbox environment, to account for the experimental nature of synthetic data use and to mitigate information disclosure risks.

CDOs are encouraged to share their feedback with DDE for both advice on the decision to use synthetic data and to improve the framework for all.

CANVAS 1 SUMMARY: IS SYNTHETIC DATA THE RIGHT SOLUTION FOR YOUR PROBLEM?

This is the **most complex canvas**. It will help you in the early stages of researching synthetic data. Setting out a range of use cases and resources, it is designed to help you get to the stage of understanding whether using synthetic data is appropriate for you or not.

How to use this canvas?

Review the questions, and bring the team together to define your problem statement. Identify the reasons why you might use synthetic data to achieve the purpose and whether it is the right method to use ahead of other more established approaches.

Questions to ask

- What type of problem are you trying to solve?
- How will synthetic data benefit the use case?
- What other things could you do instead of using synthetic data (e.g. traditional privacy-preserving techniques?)
- Have you identified what data is relevant to the exercise and what questions can and should be answered?
- What is the risk associated with your use case?
- How could additional differential privacy techniques help high-risk cases?

To be filled in by

Business owner or data steward, and involving:

- Data scientist/ statistician
- Data security expert
- Legal/compliance expert

Useful references

- ADRUK, Accelerating public policy research with synthetic data, 2021
- Blog, 10 use-cases for privacy-preserving synthetic data
- Replica Analytics, Synthesis tutorials
- Office for National Statistics, Synthetic data pilot13

FOCUS AREAS

Identify if synthetic data is the right solution for your specific use case and your data challenge, by assessing readiness through the following

1. DEFINE THE PROBLEM/IDENTIFY YOUR DATA CHALLENGE

2. SEEK OUT RELEVANT ASSISTANCE AND SUPPORT

3. IDENTIFY RISKS AND CHALLENGES ASSOCIATED WITH YOUR USE CASE

4. IDENTIFY AVAILABLE ALTERNATIVE SOLUTIONS TO THE PROBLEM

5. EVALUATE THE IDENTIFIED ALTERNATIVES

6. MAKE THE PRELIMINARY DECISION: GO/NO GO

CANVAS 2 SUMMARY: DO YOU HAVE THE RIGHT TEAM?

How to use this canvas?
This canvas helps you form the right team for the successful generation, management and use of synthetic data. Getting knowledge inputs right and knowing where to find help - inside and outside of your organization - is important in striking the right balance between excessive conservatism and ignoring risk.

Questions to ask

- Does your department have the resources to generate synthetic data (for individual use cases and possibly at scale)?
- Does your team have expertise to ensure safety of synthetic data?
- Does your team have the specialized governance expertise to make sound judgements on synthetic data?
- Does it have the technical knowledge to do the same?
- Do you have an ethics oversight function for model building and model uses?
- For high-risk or special cases have you considered seeking advice/consultation from Dubai Data Establishment?
- Does someone in the team have the statistical skills to prove that synthetic data parameter estimations are matching with the original data?

To be filled in by

- Data steward (i.e. the team member most likely to be coordinating an exercise like this in the first instance).

Useful references

- Getting started with synthetic data video tutorial (Gretel AI)
- Why Kainos is hiring a Data Ethicist (Medium blog post)

FOCUS AREAS

TEAM ROLES AND RESPONSIBILITIES

These roles may differ depending on the complexity and sensitivity of your use case overall.

DATA GOVERNANCE

- **Data Protection Officer**
- **ISR champion or security team**
- **Legal Counsel, legal team representative**
- **Data/Digital ethics lead**

TECHNICAL TEAM

- **Data Scientist**
- **Data Engineer**
- **Data/Solutions architect**
- **Statistician**

SIGNIFICANT OTHERS

- **Product Manager/Policy Customer**
- **Chief Data/Chief Information Officer**
- **City Chief Data Officer**

CANVAS 3 SUMMARY: ARE YOU USING THE RIGHT TOOL TO ACHIEVE THE BEST BALANCE BETWEEN PRIVACY AND ACCURACY?

How to use this canvas?

Having framed your policy question and identified supporting datasets this canvas explores in more detail the privacy preserving aspects and accuracy of the approaches available. Understanding the trade-off between the two (the privacy-utility gradient) and how this operates against your overall business objectives is vital in ensuring success.

Questions to ask

- Which attributes within your datasets should be flagged as a privacy risk/sensitive?
- How would the omission of these attributes limit the analysis or your product?
- Can you effectively assess how the different synthetic data approaches (part-synthetic to differential privacy synthetic data) preserve privacy?
- And to what extent do they affect utility of the data?
- Have you considered both low versus high fidelity synthetic versions of data for your case?
- Have you ensured the synthetic data has the same statistical characteristics as the original data (without duplicate samples)?

To be filled in by

- Data steward, with input from:
    - technical team
    - governance lead
    - product owner

Useful references

- Getting started with synthetic data video tutorial (Gretel AI).
- When to synthesize your data video (Replica Analytics).

FOCUS AREAS

TECHNICAL AND GOVERNANCE VIEW

PRIVACY AND ACCURACY SCORING

OTHER ASSESSMENTS

JUSTIFICATION FOR SYNTHETIC DATA APPROACH AND TOOL SELECTION

CHECK BACK AGAINST BUSINESS OBJECTIVES/USE CASE GOALS

CANVAS 4 SUMMARY: DO YOU HAVE THE RIGHT INFRASTRUCTURE?

How to use this canvas?
This canvas helps you identify whether you have the right infrastructure in place for the given synthetic data use case.

Questions to ask

- Do the functional requirements of your system satisfy synthetic data requirements for your use case?
- Based on the risk of the chosen synthetic data method have you considered if the level of security of your environment is appropriate?
- Have you considered verification/validation servers to ensure the system has been appropriately trained (by executing programs developed on synthetic data on confidential data with noise added)?
- What kind of tools are you using (e.g. cloud-based, open-source tools, use of containerization, enterprise/ commercial tools)?

To be filled in by

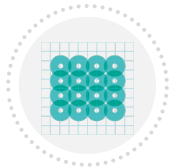
- Data steward and technical team.

Useful references

- ADR UK, Accelerating public policy research with synthetic data, 2021(generating synthetic data at scale).

FOCUS AREAS

SECURITY AND PRIVACY REQUIREMENTS



TEAM ACCESS TO THE REQUIRED TECHNOLOGY RESOURCES

ASSESSMENT OF PRESSURE ON EXISTING INFRASTRUCTURE

OTHER SECONDARY REQUIREMENTS

CANVAS 5 SUMMARY: HOW TO SHARE AND PUBLISH SYNTHETIC DATA?

How to use this canvas?
Having established the case for using synthetic data and understood (the extent of) the use of it in your analysis or product development, this section concerns how you classify synthetic data, maximize its safe use through access permissions, and deal with any further ethical considerations.

Questions to ask

- Have you carried out an information disclosure risk assessment?
- Is your synthetic data labelled?
- Have you classified your synthetic data?
- Where original data is classified as confidential or sensitive are you content that access rights and security processes allow for safe yet effective analysis/development work?
- Is your use of synthetic data consistent with other data protection laws and ethical considerations?
- Are you sharing your data with a third party (for synthetic data creation)?
- Have you considered enabling internal experimentation between departments through using synthetic data?

To be filled in by

- Data steward and technical team.

Useful references

- Dubai Data Manual – Module 8 (data classification)
- ADR UK, Accelerating public policy research with synthetic data, 2021(specifically on digital watermarking)
- UN Centre for Human Data – disclosure risk assessment overview

FOCUS AREAS

CLASSIFYING SYNTHETIC DATA

LABELLING SYNTHETIC DATA

MEETING LEGAL REQUIREMENTS

THIRD PARTY ACCESS

ACCESS PERMISSIONS FOR SYNTHETIC DATA

PUBLISHING SYNTHETIC DATA ON DUBAI'S SHARED DATA PLATFORM

DDE INVOLVEMENT

5

DECISION MATRIX CANVASES

- CANVAS 1: IS SYNTHETIC DATA THE RIGHT SOLUTION FOR YOUR PROBLEM?
- CANVAS 2: DO YOU HAVE THE RIGHT TEAM?
- CANVAS 3: BALANCING PRIVACY AND ACCURACY
- CANVAS 4: WHAT IS THE APPROPRIATE INFRASTRUCTURE?
- CANVAS 5: HOW TO SHARE SYNTHETIC DATA

DECISION MATRIX CANVAS WORKBOOKS

DATA STEWARDS ORGANISING A TEAM EFFORT

Work on these detailed canvas workbooks is most likely to be organized by the data steward or product owner.

Answers to the questions they contain must come from experts in data science, infrastructure, management, legal and governance domains.

Significant decisions and sign-off for use must come from the senior data lead (e.g. Chief Data Officer or similar).

While the workbooks provide the guiding questions to be considered during the discussion of whether to proceed with synthetic data or not the team should also rely on their own professional judgement to use and enhance the framework according to their needs. Teams should not be restricted by them.

CANVAS 1: IS SYNTHETIC DATA THE RIGHT SOLUTION FOR YOUR PROBLEM?

DEFINE YOUR BUSINESS GOAL PROBLEM STATEMENT

Define your challenge with data in the use case

Select the reasons you might need to use synthetic data to achieve your purpose
(see more on Potential uses for synthetic data)

- ☐ **Cannot use real data due to confidentiality/ privacy restriction/ concerns/ regulations**
  Make confidential/sensitive data available for use
- ☐ **Do not have real data or the real data is lacking**
  Missing data, fix biases, real data doesn't exist etc.

Get support from the right people

Did you make sure to involve the right people in this decision?

- ☐ **Data scientist & data analyst**
- ☐ **Legal compliance**
- ☐ **Security specialist**
- ☐ **Steering Committee/senior management**

Identify risks associated with using synthetic data

☐ Did you conduct and document a detailed risk assessment associated with synthetic data in accordance with the approved risk assessment methodology (entities should define their risk management framework based on ISR, ISO 31000, ISO 27001 etc.)?

Did you also conduct an evaluation or verification process, covering the following?

- ☐ Synthetic data represents the original data to the level of accuracy required (determined by entity) but cannot be traced back to the original data in any way. Consider that the higher the accuracy level, the higher the risk of disclosing information and the higher the complexity level. If the risk is high - differential privacy techniques may need to be considered.
- ☐ Synthetic data generation process should be in alignment with original data (real data) classification. (The algorithm will be supervised by the data owner to trace any issues).
- ☐ Security policy and controls requirements should be in place - when handling of synthetic data is outsourced / handed to third party (wherein, the access to original data may be required to be provided to third party)
- ☐ Data quality level of original data is at minimum level 3 according to Dubai Data Standards. Data quality of synthetic data is dependent on data quality of real (original) data it is generated from.
- ☐ Complexity of data model for synthetic data generation. Ensure you have resources and capacity for potential effort, time and infrastructure required.
- ☐ Bias in original data. Synthetic data will inherit bias in original data so make sure that you have evaluated and verified the potential bias and prepared for it.

CANVAS 1: IS SYNTHETIC DATA THE RIGHT SOLUTION FOR YOUR PROBLEM?

Identify and evaluate alternatives

What are the advantages of using synthetic data over other traditional privacy preserving methods? Specify some other potential privacy preserving methods to be used and their risks and benefits. (see more on Comparison with traditional privacy preserving methods) Evaluate all the alternatives including synthetic data method.
(see more on Evaluate all privacy preserving solutions)

	Advantages	Potential risks and limitations
Synthetic data		
Alternative 1		
Alternative 2		
Alternative 3		

Make the decision

Justify your preliminary decision based on your best judgement

CANVAS 2: WHO TO INVOLVE 1/3

DATA GOVERNANCE

ROLE	INPUT	NOTES / OTHER OPTIONS	Nominee (& organization)
DATA PROTECTION OFFICER	Ensuring compliance and proper handling of personal data. Providing advice on avoiding data breaches from a policy and legal perspective and likely consequences in the event of a disclosure.	Advice available from external data privacy consultancy or Dubai Data Establishment.	
ISR CHAMPION OR SECURITY TEAM	Examining the use case from a technical perspective by validating tools and assessing suitabliity of existing departmental IT infrastructure. Also advising on access permissions and compliance with existing Dubai Data Law.		
LEGAL COUNSEL	Tests compliance with regional, international and sector-specific legislation (e.g. health, finance).	Advice can be sought from Legal Affairs Department or private law firms with data expertise.	
DATA/DIGITAL ETHICS LEAD	Adding another dimension to governance, but also considers how consumer trust should be protected and indeed boosted by the use of synthetic data (ref. Replica Analytics assessment framework).	Not a standard job at all. But consider how unbiased data especially is at the heart of responsible AI and machine learning. Consider academic institutions as possible source of advice.	

CANVAS 2: WHO TO INVOLVE 2/3

TECHNICAL TEAM

ROLE	INPUT	NOTES / OTHER OPTIONS	Nominee (& organization)
DATA SCIENTIST	Reviewing suitability of commercial solutions or pre-built packages available in open source libraries in terms of both the utility of synthetic data in the use case as well as privacy preserving effectiveness. Generating requirements for Data Engineer (columns, values, range of variables). Making sure that the characteristics of original and synthetic data are the same statistically (but not really traceable),	Dubai Data Establishment can provide limited data science support. The limited number of commercial entities providing synthetic data as a service may be able to provide consultancy support, but doing so may involve security clearance, especially for those operating outside of the national jurisdiction.	
STATISTICIAN	Providing statistical evidence to prove that the synthetic data parameter estimations are matching with the original. Framing hypotheses and proving it statistically.	Sometimes a data scientist can have required statistics skills. If not, it is preferable to have one statistician on the team.	
DATA ENGINEER	Analyzing the data scientist's requirements and subsequently, building the scripts to generate the synthetic data. Testing (and deploying) synthetic data.	Support available from Dubai Data Establishment.	
DATA/ SOLUTIONS ARCHITECT	Planning and implementing data infrastructure to accommodate the synthetic data lifecycle.		

CANVAS 2: WHO TO INVOLVE 3/3

SIGNIFICANT OTHERS

ROLE	INPUT	NOTES / OTHER OPTIONS	Nominee (& organization)
PRODUCT MANAGER/POLICY CUSTOMER	Providing early stages input on the optimal 'end user' business benefits for the use case (e.g. how will making up for data scarcity enhance software testing, analytical precision etc.) Involved in closing discussions, especially with the data steward, about how synthetic data achieves the maximum trade-off between privacy and utility whilst delivering on the use case aims.	This role will work most closely with the data steward.	
SENIOR DATA LEAD (CHIEF INFORMATION OR DATA OFFICER)	Signing off on any software or data infrastructure requirements, as well as data governance arrangements.	Advice from Dubai Data Establishment is available.	
CITY CHIEF DATA OFFICER	Inputting across any outstanding points of contention or ambiguity in the technical and governance domains, once the framework and processes in this guide have been followed. Leading the longer-term development of strategy and data infrastructure to account for synthetic data generation, storage, sharing and usage.	Effectively the Chief Executive Officer of Dubai Data Establishment.	

CANVAS 3: BALANCING PRIVACY AND ACCURACY 1/2

1. What is the overall justification of your choice of synthetic data approach and tool selection (please insert)?

There is a range of existing open source and proprietary tools you can use for synthetic data generation.
Evaluate and assess the suitability and safety of the tools and justify your choice (see more on Types of synthetic data).

2. For all approaches and tools:

- [ ] **Please list tools you used to generate and test synthetic data generation. If there is a high risk of information disclosure, have you considered additional differential privacy techniques (i.e. private synthetic data) and how have these affected the risk?**

- [ ] **Please list in each case the privacy and utility/quality score for the synthetic data generated.**

	PRIVACY	UTILITY/QUALITY

CANVAS 3: BALANCING PRIVACY AND ACCURACY 2/2

3. Use of other assessments

☐ **Please describe how any other assessments have been considered in the selection (e.g. consumer trust and cost as set out in Canvas #1)**

PROVIDE SUPPORTING FILES

4. Technical and governance view

☐ **Please include input from technical and governance colleagues, taking a particularly risk-based perspective (e.g. disclosure of personal private information).**

PROVIDE SUPPORTING FILES

5. Check back against business goals as outlined in Canvas #1

☐ **With privacy and accuracy adequately assessed, and other considerations made, does the use of synthetic data in the use case still make sense and deliver against business goals (e.g. overcoming data scarcity for software design and testing)?**

CANVAS 4: WHAT IS THE APPROPROPRIATE INFRASTRUCTURE 1/2

1. Does your synthetic data generation tool meet the following security and privacy requirements:

- [ ] **It is capable of being run on existing departmental IT infrastructure (describe)**

- [ ] **It has undergone a strong independent QA process (describe and provide supporting documentation)**

PROVIDE SUPPORTING FILES

- [ ] **It has equivalent or better security than the original data platform (describe and provide supporting documentation)**

PROVIDE SUPPORTING FILES

- [ ] **It gives clear guidance on what privacy protections it does and does not provide (describe and provide supporting documentation)**

PROVIDE SUPPORTING FILES

2. Does the whole team involved in the use case have access to the required technological resources?

- [ ] **What do they need (e.g. access to source data platform?**
Describe requirements and set out any alternative arrangements such as an intermediary secure environment.

CANVAS 4: WHAT IS THE APPROPROPRIATE INFRASTRUCTURE 2/2

3. What assessment has been made of the pressure on existing resources?

- [ ] **Has existing infrastructure been assessed as sufficient, and if not what additional resources are required? (describe)**

- [ ] **Does the integration/installation of this tool represent an impact on existing functions? (describe)**

- [ ] **How much data storage does this tool require? (describe)**

CANVAS 4: WHAT IS THE APPROPROPRIATE INFRASTRUCTURE 2/2

4. Are there any secondary requirements that this tool would represent that need to be factored in?

- [ ] **Does the tool use any external or cloud resources that are antithetical to your security policies (describe)?**

- [ ] **Is any intermediary infrastructure required to generate/store synthetic data while it is under review (e.g. validation servers)? (describe)**

- [ ] **What level of access/security should the intermediary and end infrastructure have compared to the source data? (describe)**

CANVAS 5: HOW TO SHARE SYNTHETIC DATA 1/4

1. Have you reclassified your synthetic data?

- [ ] **Have you used the Dubai Data Classification Standards (Open -> Confidential -> Sensitive -> Secret) independently of the source data? (describe)**

- [ ] **Has this been approved by Dubai Data Establishment? (describe)**

- [ ] **Does it now reach the level of classification required for the exercise? e.g. a public hackathon using open data or a small number of government researchers using confidential data in a collaborative environment. Usually if synthetic data was created to solve a data privacy challenge, then the original classification should have been lowered to allow more freedom in working with data. (describe)**

- [ ] **Did you establish a regular security control and governance process according to the synthetic data classification? (describe)**

CANVAS 5: HOW TO SHARE SYNTHETIC DATA 2/4

2. Have you clearly labelled your data as synthetic? (This section contains suggested good practice only)

- ☐ **Potential solutions include but are not limited to:**
- ☐ **Digital watermarking – to identify the data itself without changing any statistical properties. Specific tools would be needed to show that the watermark is present.**
- ☐ **Adding the synthetic label to metadata, dataset name, dataset description, attribute description etc. and ensuring processes that regularly check for this label to stay in place.**
- ☐ **Producing a clear and easy-to-understand explanation of synthetic data.**
- ☐ **Ensuring that the method for generating the synthetic data is clearly stated in the data lineage.**

(Please describe)

CANVAS 5: HOW TO SHARE SYNTHETIC DATA 3/4

3. Have you clearly identified who can have access to your classified synthetic data?

- [ ] **Have you used the Dubai Data Classification Standards (Open -> Confidential -> Sensitive -> Secret) independently of the source data? (describe)**

- [ ] **Have you completed an information disclosure risk exercise and does this show a) a justification if disclosive risk is high and b) access rights for individuals? (describe)**

PROVIDE SUPPORTING FILES

- [ ] **Did you prepare data sharing agreements, if required, for the generated synthetic data?**

PROVIDE SUPPORTING FILES

4. What measures are in place to assess third party access risk to synthetic data?

- [ ] **How do these measures mitigate against unauthorized third-party access to synthetic data? (describe)**

CANVAS 5: HOW TO SHARE SYNTHETIC DATA 4/4

5. Does any law (e.g. sector specific laws) regulate or affect the resulting synthetic data? (describe)?

6. Have you shared your synthetic data with DDE in order to validate and publish the synthetic data on DDA's sandbox as part of Dubai's data initiative?

7. Based on your answers to the above questions, please indicate where (further) involvement from Dubai Data Establishment is required?

WHAT'S NEXT? DDA'S REGULATORY SANDBOX FOR SYNTHETIC DATA

After filling in all the canvasses please contact DDA (syntheticsandbox@digitaldubai.ae) in order to validate the methodology and to ensure safe synthetic data generation and usage.

Due to synthetic data's experimental nature, please avoid publishing synthetic data openly without discussing with DDA and validating in the sandbox first.

SUPPLEMENTARY MATERIAL

This section contains additional information and self-serve assessment tools to help fill in canvasses

CANVAS 1: POTENTIAL USES FOR SYNTHETIC DATA 1/2

This page consists list of potential cases where synthetic data could be used. However please consider this is not a comprehensive list but can be used to guide the brainstorming process.
* Consider traditional privacy preserving techniques depending on cost, privacy, consumer trust and data utility priorities.

1

NO OR LACK OF REAL DATA

When synthetic data is easier to produce than collecting real data, real data doesn't exist, or real data is lacking

- Overcome scarcity – simulate not yet encountered events (manufacturing or infrastructure failure) – to e.g., train algorithms
- Correct bias in data, or balance out the dataset (to account for under-represented groups, you wish to target).
- Simulate the future to keep AI models relevant to latest changes and trends; simulate alternate futures to be prepared for different scenarios.
- Simulate "black swan events".
- Other examples (to build digital twins, metaverse applications).

Sources:
Elise Devaux. 10 use-cases for privacy-preserving synthetic data. (2020, link)
CIO. Maria Korolov. What is synthetic data? Generated data to help your AI strategy. (2022, link)

SEE CANVAS 1

CANVAS 1 : POTENTIAL USES FOR SYNTHETIC DATA 2/2

This page consists list of potential cases where synthetic data could be used. However please consider this is not a comprehensive list but can be used to guide the brainstorming process.

2

CANNOT USE REAL DATA

When privacy needs to be overcome as a barrier in order to make confidential/sensitive data available

Analytical value

- AI/ML model training – get significant volumes of compliant data, for training models when real data is lacking.
- Data analysis* - masked data could impact quality of analysis and have re-identification risks; attaining systematic consent is drawn out too.

Technical

- Internal data sharing/product development - enable collaboration, give teams easy access*.
- Cloud migration – safely avoid compliance processes for upcoming uses. *Testing cloud migrations by ensuring the same app running on two infrastructures generates identical results.
- Data retention – enable long term analysis when regulation doesn't allow to store data for long periods.
- Software testing – realistic testing for faster time to production.

Data sharing

- 3rd party sharing – share with government, vendors, clients, send offshore*.
- Data monetization – build revenue from data streams that are too sensitive to use.
- Data publication for hackathons, etc.

Others

* Consider traditional privacy preserving techniques depending on cost, privacy, consumer trust and data utility priorities.

Sources:
Elise Devaux. 10 use-cases for privacy-preserving synthetic data. (2020, link)
CIO. Maria Korolov. What is synthetic data? Generated data to help your AI strategy. (2022, link)

SEE CANVAS 1

CANVAS 1 : COMPARISON WITH TRADITIONAL PRIVACY PRESERVING METHODS 1/3

Traditional privacy preserving methods were compared with synthetic data method based on the following high-level criteria (for more information please refer to the source):

1

Re-identification Risk:

Highlights the chances of re-identifying the original data from the obscured data by multiple means.

2

Statistical Features:

Highlights the value of statistical features retained in the generated/obscured data

3

Feature Correlations:

Highlights the value of correlations that is retained within the generated/obscured data

4

ML Performance:

Highlights the performance aspects of the generated data

Source: Mostly AI. Ivona Krchova. 3 reasons why organizations are moving away from legacy data masking. (2020, link)

SEE CANVAS 1

CANVAS 1 : COMPARISON WITH TRADITIONAL PRIVACY PRESERVING METHODS 2/3

Different types of data obscuration methods are mentioned below in comparison to synthetic data:

METHOD	RE-IDENTIFICATION RISK	FEATURE STATISTICAL VALUE	FEATURE CORRELATIONS VALUE	ML PERFORMANCE	WHY / WHEN WOULD YOU USE THIS METHOD	DESCRIPTION
SYNTHESIZING	**Low**	**High**	**High**	**High**	When high confidentiality is required, with no sacrifice of granularity. Often requires large existing datasets, a lot of processing and pre-existing understanding of the dataset.	Algorithmically generate dataset mimicking the original data without losing it statistical features.
RANDOMIZATION	**High**	**Medium**	**Low**	**Low**	When only the specific values are confidential but not the identifiable information. Requires no additional knowledge of the dataset.	Changing attributes to make them less precise while maintaining the overall distribution.
PERMUTATION	**High**	**High**	**Very low**	**Very low**	When linking the subject of the data isn't needed but you require all attributes to be real.	Mixing attribute values in a table in such a way that some of them are artificially linked to different data subjects.

Source: Mostly AI. Ivona Krchova. 3 reasons why organizations are moving away from legacy data masking. (2020, link)

SEE CANVAS 1

CANVAS 1 : COMPARISON WITH TRADITIONAL PRIVACY PRESERVING METHODS 3/3

Different types of data obscuration methods are mentioned below in comparison synthetic data:

METHOD	RE-IDENTIFICATION RISK	FEATURE STATISTICAL VALUE	FEATURE CORRELATIONS VALUE	ML PERFORMANCE	WHY / WHEN WOULD YOU USE THIS METHOD	DESCRIPTION
GENERALIZATION	**Low**	**Low**	**Low**	**Low**	Often aggregation. If the end requirements are at a lower level of granularity, this is often much easier to handle for everyone involved, and at no point gives visibility on confidential specifics.	Changing the scale of dataset attributes, or the order of magnitude to ensure that they are common to set of people.
PSEUDONYMIZATION	**Very High**	**High**	**High**	**High**	When designing data protection for production systems use pseudonymization. Then only authorized users will have access to personal data.	Switching original sensitive data with an alias or pseudonym. The personal data can no longer be attributed to a data subject without the use of additional information.

Source: Mostly AI. Ivona Krchova. 3 reasons why organizations are moving away from legacy data masking. (2020, link)

SEE CANVAS 1

CANVAS 1 : EVALUATE ALL PRIVACY PRESERVING SOLUTIONS (INCLUDING SYNTHETIC DATA) 1/2

The following methodology by Replica Analytics can help identify the best privacy preserving technology to use, considering other factors. A full online version could be found on Replica Analytics website.

Follow the steps:

1. Identify your organization's priorities between the following 4 factors:
    1. Privacy
    2. Data Utility
    3. Vendor specific and operational Costs
    4. Consumer Trust
2. Rank the potential privacy preserving techniques of your choice by how well they satisfy the four criteria based on your experience. Example by Replica Analytics in the figure – value 1 (higher rank)-6 (lower rank), lower value for the rank means that a particular method ranks higher or is better on that criteria.
3. Give weights to the 4 factors based on your organization's priorities for the use case. The weights should add up to 1.
4. Compute the final scores. The scores are a normalized total weighted rank. The score are also ranks but scored between 0 and 1.
5. The highest scored is the technology that is most aligned with your organization's explicitly stated priorities along the 4 dimensions.

Source: Replica Analytics. Synthesis tutorials. When to synthesize. (2022, link)

SEE CANVAS 1

CANVAS 1 : EVALUATE ALL PRIVACY PRESERVING SOLUTIONS (INCLUDING SYNTHETIC DATA) 2/2

	WEIGHT	Synthetic data	Alternative 1	Alternative 2	Alternative 3	Alternative 4
PRIVACY						
CONSUMER TRUST						
OPERATIONAL COST						
DATA UTILITY						

	WEIGHT	TRANSFORM DIRECT IDENTIFIE	HIPAA LDS	GDPR PSEUDONYMIZATI	HIPAA SAFE HARBOR	RISK-BASED DE-IDENTIFICATIC	DATA SYNTHESIS
PRIVACY	0.25	6	3	3	5	1	1
CONSUMER TRUST	0.25	6	3	3	5	2	1
OPERATIONAL COST	0.25	1	5	6	2	4	3
DATA UTILITY	0.25	1	1	1	4	5	5
SCORE (HIGHER = BETTER)		0.3	0.7	0.5	0	0.7	1

Replica Analytics

Source: Replica Analytics. Synthesis tutorials. When to synthesize. (2022, link)

SEE CANVAS 1

CANVAS 3: TYPES OF SYNTHETIC DATA

This document will help with filling in Canvas 3 by giving an idea on the types of synthetic data and methods of generating it. For more information, please refer to references below.

TYPES OF SYNTHETIC DATA[1]

FULLY SYNTHETIC	PARTIALLY SYNTHETIC	HYBRID SYNTHETIC
This data does not contain any original data. This means that re-identification of any single unit is almost impossible, and all variables are still fully available.	Only data that is sensitive is replaced with synthetic data. In this situation, genuine values are only changed if there is a substantial risk of disclosure.	Hybrid synthetic data is derived from both real and synthetic data. A near-record in the synthetic data is chosen for each record of real data, and the two are then joined to generate hybrid data.

The AI Multiple website sets out a range of methodologies for creating synthetic data (ex. generating according to a distribution, fitting real data to a known distribution and using deep learning, please see the link.

1. AI Multiple. Cem Dilmegani. in-Depth Synthetic Data Guide: What is it? How does it enable AI? (2022, link)

More Synthetic data research and development papers:

Statice. Dr. Christoph Wehmeyer. How do you generate synthetic data? (2021, link)

ONS. ONS methodology working paper series number 16 - Synthetic data pilot. (2022, link)

Scott McLachlan. Realism in Synthetic Data Generation (2017, link)

Statice. Elise Devaux & Dr. Christoph Wehmeyer. An Overview of synthetic data types and generation methods. (2021, link)

SEE CANVAS 3

THANK YOU