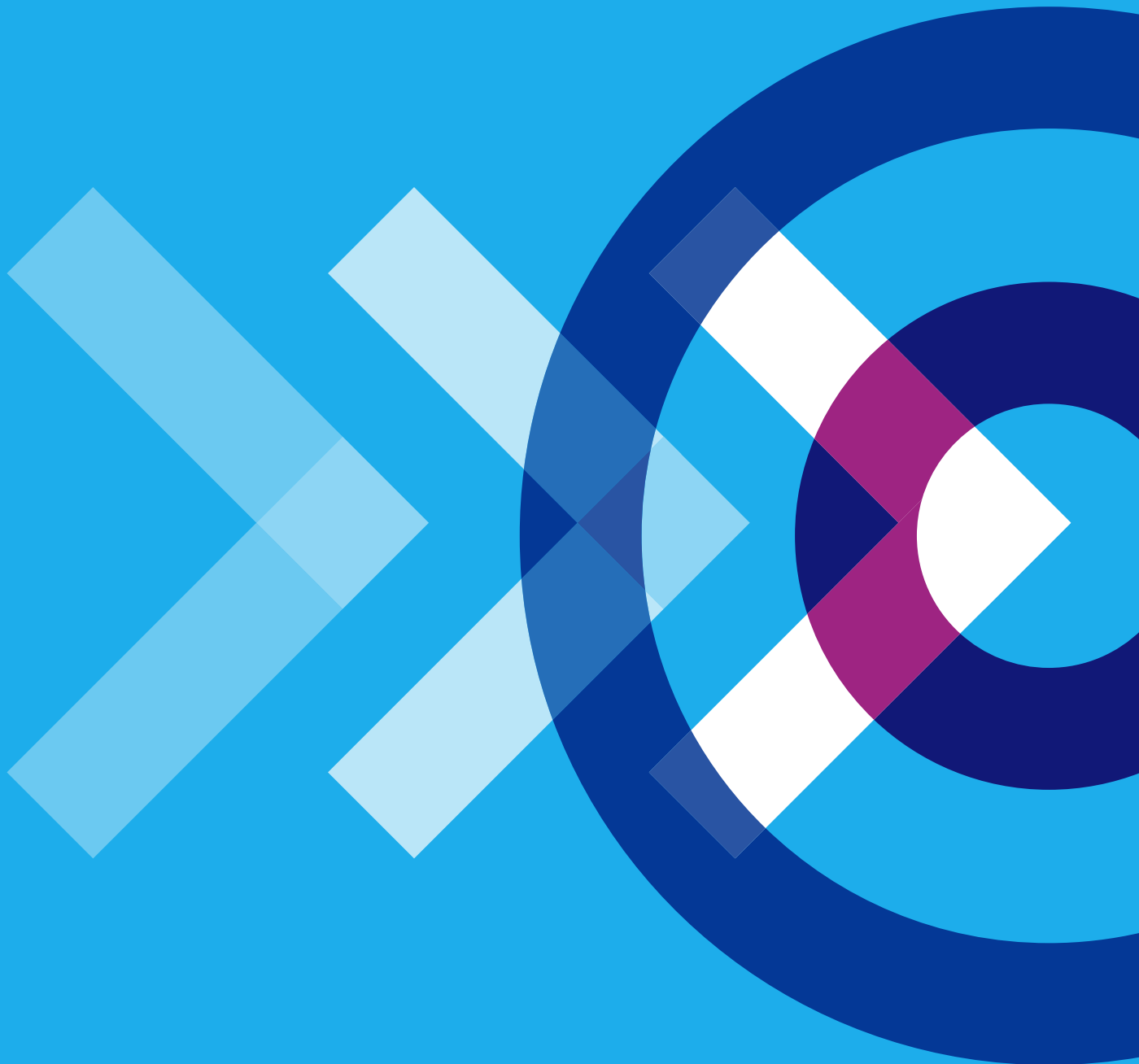




PIONEERING CONFIDENTIALITY AND ACCESSIBILITY

PILOT STUDY ON SYNTHETIC
LABOR FORCE DATA





Synthetic Data Values

Synthetic data holds significant value, particularly in fields like labor force surveys, where data privacy, accuracy, and utility are crucial. These values underscore the transformative potential of synthetic data in labor force surveys, making it a powerful tool for advancing research, enhancing data security, and supporting informed decision-making



Cost Efficiency



**Support for Innovation
and Methodological
Development**



**Improved Data
Accessibility and
Sharing**



**Enhanced
Privacy
Protection**



**Enhanced Quality
and Utility of
Statistical Models**

Page of Contents



Acknowledgments	04
Disclaimer	05
Executive Summary	06
1. Introduction	07
1.1 Background of the Study	
1.2 Importance of Synthetic Data in Labor Force Surveys	
1.3 Overview of the Collaboration: DDSE and BlueGen.ai	
2. Methodology	14
2.1 Data Sources and Selection Criteria	
2.2 Synthetic Data Generation Process	
2.2.1 Model Selection	
2.2.2 Data Simulation Techniques	
2.3 Validation and Testing of Synthetic Data	
2.4 Reporting and Analysis	
3. Results and Analysis	21
3.1 Key Findings from the Pilot Study	
3.2 Comparison of Synthetic Data vs. Real Data	
4. Discussion	26
4.1 Evaluation of the Pilot Study Outcomes	
4.2 Strengths and Weaknesses of Using Synthetic Data	
Conclusion	29



ACKNOWLEDGMENTS

We would like to sincerely thank everyone who has helped us complete this pilot study on AI-generated Synthetic Data using data from the labor force survey.

First and foremost, we would like to express our sincere gratitude to the Dubai Data & Statistics Establishment, the data and statistics arm of Digital Dubai for their forward-thinking leadership, steadfast support, and dedication to the advancement of data innovation. Their contribution of resources and strategic guidance played a pivotal role in molding this research.

We are extremely appreciative to BlueGen.ai for their technical skills and collaborative efforts throughout the project. Their creative approach to synthetic data production, as well as their commitment to assuring the accuracy and dependability of synthetic datasets, have proven crucial to our study.

We would also want to thank our dedicated team members who donated their time, expertise, and insights to this study. Their efforts, from initial concept to final analysis, have been the driving factor behind the project's success.

Finally, we want to thank everyone who contributed to this study, whether directly or indirectly. Your contributions, no matter how large or small, have been critical to the successful completion of this research, and we are grateful.

This work would not have been feasible without the combined efforts of everyone involved, and we hope to continue working together as we improve the use of synthetic data in labor force surveys and beyond.



DISCLAIMER

The authors' findings, interpretations, and conclusions in this research study do not necessarily represent the opinions or official policies of the Dubai Data and Statistics Establishment, or BlueGen.ai.

While every effort has been made to ensure the accuracy and reliability of the information presented, the authors and collaborating entities make no express or implied representations or warranties about the data or methodologies used in this study, including their completeness, accuracy, reliability, suitability, or availability.

This is a pilot study with the goal of gathering information and conducting exploratory research. The synthetic data developed and evaluated in this study is based on specific assumptions and models, hence the conclusions may not be applicable to other situations or datasets. The use of synthetic data has inherent restrictions, which users should consider when interpreting the results.

The authors, Dubai Data and Statistics Establishment and BlueGen.ai accept no responsibility for any losses, damages, or consequences resulting from the use of the information provided in this research paper. Readers and stakeholders are invited to conduct their own analyses and use their own judgment when applying the study's findings.

Furthermore, any mention of specific products, services, processes, or organizations in this article does not constitute or imply an endorsement, recommendation, or favoring by the authors or the cooperating entities.

This study paper is a work in progress, and future investigations may expand or change the findings presented here. The authors seek feedback and recommendations for improvement as part of their ongoing work on synthetic data applications in labor force surveys.



EXECUTIVE SUMMARY

This pilot project, carried out in collaboration with the Dubai Data and Statistic Establishment (DDSE) and BlueGen.ai, investigates the potential of employing synthetic data in labor force surveys.

To improve data protection, accessibility, and informed decision making. The strategic direction of DDSE emphasizes the importance of balancing data privacy with increased accessibility.

This pilot project seeks to resolve these challenges by generating synthetic datasets that replicate real-world labor force data while maintaining full data confidentiality. The synthetic data enables a broader audience, including researchers and decision-makers, to perform detailed analysis and develop insights without compromising privacy.

Key findings show that synthetic data can replicate the statistical characteristics of the entire data set. This includes the key labor force variables such as employment status and demographic patterns, making it appropriate for large-scale labor market analysis.

Initial challenges in effectively recording unusual events and crucial data points were overcome during the project with the extra use of focus columns and deterministic techniques.

This study marks a significant step towards the future of labor force analysis, where synthetic data not only enhances data protection but also empowers policymakers, researchers, and statisticians with the tools needed for deeper, more flexible insights into workforce dynamics. This aligns with global best practices in open data policies and fosters greater transparency, accessibility, and utility of labor market statistics.

INTRODUCTION



→ Introduction

1.1 Background of the Study

Labor force surveys are critical for understanding labor dynamics and informing policy decisions. Traditional data collection methods, such as direct surveys and administrative data, can raise concerns about privacy, security, and accessibility. These methods are time-consuming and costly, which limits their capacity to safely share data for study and analysis.



Labor market data contains sensitive information about individuals and households, making data privacy regulations essential to limit access to the microdata needed for in-depth research. As a result, external parties such as researchers, students, and analysts are restricted to aggregated data, which often lacks the granularity required for advanced academic studies and policy-driven insights. To address this challenge, synthetic microdata can be generated and shared with a wider audience without compromising data confidentiality. This enables thorough academic research and supports the development of meaningful policy recommendations that enhance public policymaking and decision-making.

Synthetic data presents a viable solution to these difficulties. Synthetic data, which generates artificial datasets that mirror the statistical features of real data without exposing sensitive personal information, provides a secure means to communicate and evaluate labor force statistics.

*
The study was motivated by
the need to address **several key**
challenges.



The challenges



Data Privacy and Confidentiality:

Respondent privacy is crucial in labor force surveys, which collect sensitive personal and employment information. Synthetic data allows you to reduce privacy threats by developing datasets that do not reveal individual identities.



Data Accessibility and Sharing:

Researchers, policymakers, and other stakeholders frequently need access to precise labor force data in order to conduct analysis and establish evidence-based initiatives. Synthetic data can promote larger data sharing by providing a secure substitute to real data, hence improving research capabilities while maintaining confidentiality.



Support for Policy Simulation and Scenario Analysis:

Synthetic data enables policymakers to examine the possible effects of various policy initiatives in a risk-free setting. This capacity is critical for making informed decisions and developing strategic plans.



Advancing Statistical Methods:

The use of synthetic data provides a chance to enhance statistical techniques, notably in data integration, model creation, and machine learning applications in labor force analysis.

Purpose and Objectives

Purpose of the Study

The major goal of this pilot study is to investigate the feasibility and practical application of synthetic data in labor force surveys. The study's goal is to address significant difficulties in data privacy, accessibility, and cost efficiency while enhancing the overall utility & accessibility of labor force statistics. This effort aims to offer the team a hands-on experience, promote knowledge transfer from the private sector, and gain a better understanding of the possible effects of synthetic data on operations and future implementation methods.



Objectives of the Study

1

Evaluate the Accuracy and Utility of Synthetic Data

Examine how closely synthetic data can match the statistical features and patterns of actual labor force survey data, ensuring that it remains useful for analysis and decision-making.

2

Enhance Data Privacy and Security Measures

Showcase the ability of synthetic data to protect sensitive personal and employment information, reducing the danger of privacy breaches while allowing for greater data sharing and accessibility.

3

Facilitate Knowledge Transfer and Skill Development

Gain hands-on experience in synthetic data generation and analysis as part of the team, and leverage BlueGen.ai knowledge to develop internal capacities for dealing with future data technologies.

4

Support Implementation of Dubai's Synthetic Data Framework for Privacy-Enhanced AI Solutions

Support implementation of Dubai's synthetic data framework by contributing to the development of rules and best practices for synthetic data use within Dubai's data and statistics ecosystem. This pilot project serves as a real-world example aligned with the strategic objectives of the Dubai Data and Statistics Establishment, showcasing how organizations can adopt AI-driven solutions that leverage synthetic data for analysis and machine learning processes, instead of real data that may involve a violation of privacy.

1.2 Importance of Synthetic Data in Labor Force Surveys

The use of synthetic data in labor force surveys is an innovative approach to tackling some of the most serious issues in data privacy, accessibility and sharing.

The significance of synthetic data in this context lies in its ability to maximize the use and accessibility of labor market information while ensuring the protection of individuals' privacy. By generating realistic yet anonymized datasets, synthetic data enables deeper insights and broader data sharing without compromising sensitive personal information.

Key reasons why synthetic data is vital for labor force surveys:

Protection of Privacy and Confidentiality



Labor force surveys collect sensitive personal information, such as employment status and demographic information. Protecting this data is critical for maintaining public trust and complying with privacy requirements. Synthetic data provides a powerful solution by offering datasets that resemble genuine data without revealing any actual personal information, thus lowering the danger of privacy breaches.

Increased Data Accessibility for Research and Policy Development



Synthetic data enhances access to more detailed labor force statistics, contributing to the facilitation of data-driven policy formulation, strategic workforce planning, and targeted labor market interventions.

Scalability and Flexibility in Data Management



Synthetic data may be scaled up to reflect larger sample populations or more complicated datasets, making it an excellent choice for large-scale labor force surveys.

Promotion of Innovation in Statistical Methodologies



The use of synthetic data fosters statistical methodology innovation, such as novel methods for data integration, analysis, and interpretation. This promotes a culture of continual improvement and adaptability in statistical methods, which aligns with the changing demands of the labor market.



1.3 Overview of the Collaboration: Dubai Data and Statistics Establishment, and BlueGen.ai

This pilot project using synthetic data for labor force surveys is a joint effort by



Each partner contributes unique knowledge, resources, and strategic vision to the project, propelling the investigation of synthetic data as an innovative approach for enhancing the overall utility & accessibility of labor force statistics in Dubai.

Dubai Digital 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 and digital transformation.



Digital Dubai brings together the expertise of **two entities**:

Dubai Data & Statistics Establishment

Digital Dubai Government Establishment

Dubai Data and Statistics Establishment (DDSE)

Established by the Dubai Data & Statistics Law of 2023, Dubai Data & Statistics is creating an ambitious strategy where data & statistics is shared seamlessly, safely and securely, using data & statistics **to solve real world problems, create social and economic benefits, and lay the foundations of a truly smart city. Along the way, encouraging participation from the data community as vibrant as it is diverse.**

The Dubai Data and Statistics Establishment is responsible for producing and disseminating high-quality official statistics that inform policy decisions and support the city's development agenda. DDSE plays a crucial role in this study by leveraging its expertise in statistical methodologies, data collection, and analysis. Their commitment to enhancing data quality, accessibility, and security is a driving force behind the exploration of synthetic data as a means to improve the robustness and utility of labor force surveys.

BlueGen.ai

BlueGen.ai is a leading technology company specializing in artificial intelligence and synthetic data solutions. With a strong track record of innovation in data science, BlueGen.ai brings advanced technical expertise and cutting-edge synthetic data generation capabilities to the collaboration. Their role in the pilot study is pivotal, as they provide the technology and know-how to create synthetic datasets that accurately replicate the characteristics of real labor force data while preserving privacy. BlueGen.ai's involvement facilitates the practical application of synthetic data, offering valuable insights into its potential benefits and challenges in the context of official statistics.



Synergy and Shared Vision

This collaboration brings together public and private sector strengths, with a shared vision of using synthetic data in Dubai. By combining DDA's strategic leadership, DDSE's statistical competence, and BlueGen.ai's technology innovation, the alliance hopes to establish a new standard for data protection, accessibility, and analytical capabilities in labor market studies.



METHODOLOGY



→ Methodology

Description of the Pilot Study Design and Main Phases

This pilot project aimed to determine the feasibility and accuracy of using synthetic data in labor force surveys. The process consisted of four major phases: data selection, synthetic data generation process, and dataset validation, reporting & analysis.

Pilot Study Preparation:

After defining the pilot scope, goals, study use case and data set, working teams from both DDSE and BlueGen.ai worked together in an iterative process between October 2023 and August 2024.

Roles involved:

BlueGen.ai: Project Management, AI & ML Engineers

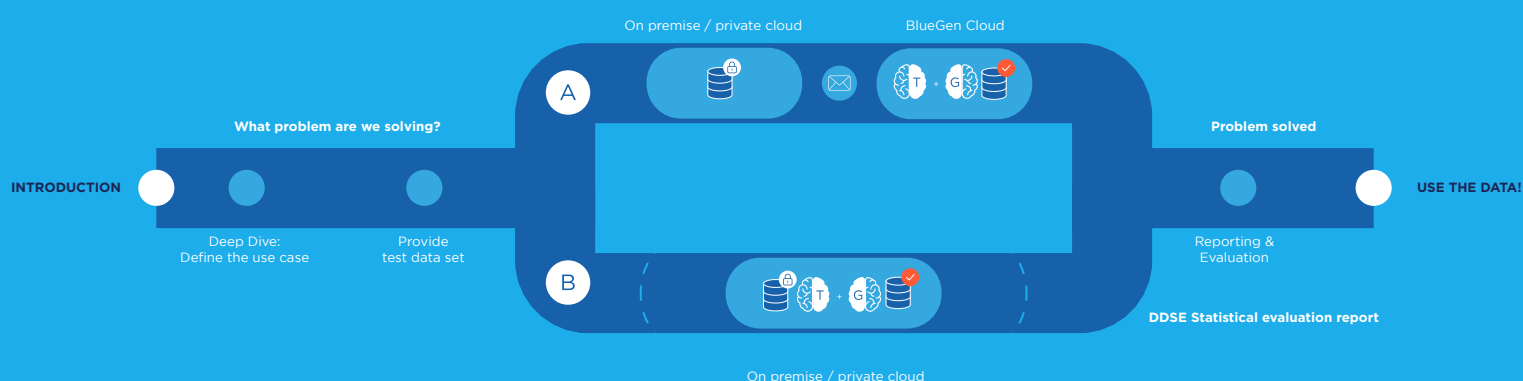
DDSE: Project Management, IT Specialists, Data & Statistics Experts

2.1 Data Sources and Selection Criteria

The study employed a representative sample of real labor force survey data from the Dubai Data and Statistics Establishment. This sample had crucial demographics such as age, gender, employment status, and economic activity, which served as a baseline for analyzing the synthetic data.

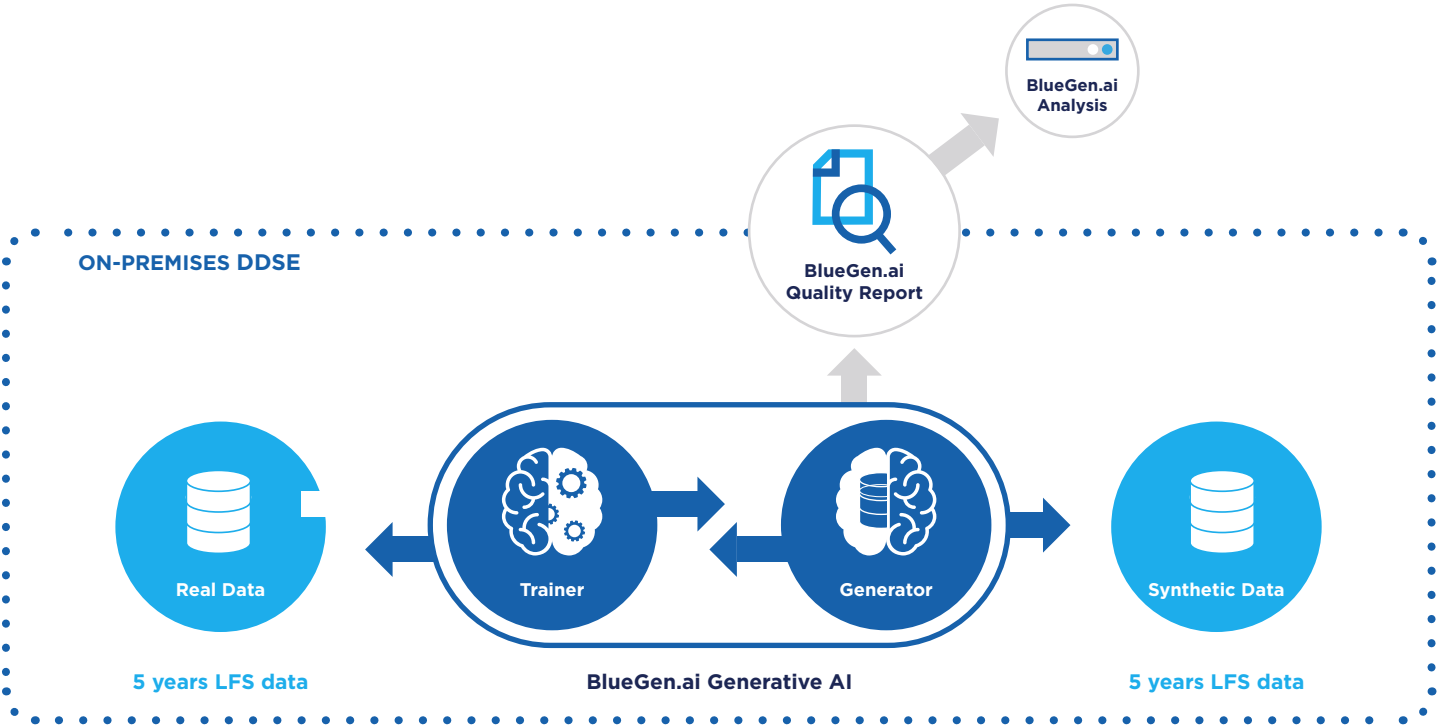
2.2 Synthetic Data Generation Process

2.2.1 Generative Model Development & Configuration



The model describes an organized procedure for solving data-driven problems, beginning with defining the use case and ending with supplying a test dataset. It provides two data processing options: on-premise/private cloud for greater control, and BlueGen.ai Cloud for scalability and flexibility. Following data processing, the results are assessed through reporting, leading to the last stage in which the problem is solved and the insights can be applied.

In this situation, the on-premise/private cloud option was selected, indicating that the firm values greater control over data protection, privacy, and infrastructure. By choosing this approach, the business assures that all data processing occurs within a safe, internal environment, which is appropriate for managing confidential or sensitive data. This option also allows for more customizable configurations, tailored to specific business or technical requirements, and provides complete control over the data processing and analysis phases.



The BlueGen.ai software has been integrated in the DDSE environment, and only internal staff can setup it, train the models, and produce synthetic data. Both actual and synthetic data never leave the ecosystem. BlueGen.ai simply received the evaluation report and logs to provide assistance.

2.2.2 Data Simulation Techniques

The pilot project on synthetic data for labor force surveys used a variety of advanced data simulation techniques to create datasets that closely resembled the features of actual survey data. These techniques are at the cutting edge of statistical and machine learning innovation, allowing for the creation of realistic, synthetic datasets while protecting the confidentiality of individual responders.



The following is a summary of the primary data simulation methodologies used in the study:

Tabular diffusion

At the core of the data simulation process is the tabular diffusion technique, which is based on denoising diffusion probabilistic models (DDPMs) introduced by Ho et al. (2020). This approach is particularly well-suited for generating synthetic tabular data, as it can handle mixed data types (categorical, continuous, datetime, etc.) commonly found in labor force surveys.

BlueGen.ai’s tabular diffusion model builds on the work of Peebles et al. (2023). This method incorporates transformer architectures into the diffusion process. Transformers are particularly effective at capturing complex relationships and long-range dependencies in data. In the context of labor force surveys, this allows the model to better understand and replicate intricate patterns and correlations between different survey variables.

The combination of these tabular diffusion techniques **enables the generation of synthetic data that closely resembles the statistical properties and relationships present in the original survey data**, while ensuring that no individual’s actual responses are directly reproduced.

Fine-tuning Synthesis

To further enhance the quality and utility of the synthetic data, the project configured BlueGen.ai’s built-in functionality to tune the model to specific characteristics of the data:

a) Missing Value Patterns:

This technique is crucial for ensuring that the synthetic data accurately reflects the patterns of missing values in real survey data. In labor force surveys, missing values often occur in specific patterns corresponding to different groups of respondents (e.g., those under 15, employed, unemployed, retired). By integrating these patterns into the model, the synthetic data maintains a realistic representation of data availability across different demographic groups.

b) Focus Columns:

This method allows the model to pay special attention to key characteristics in specified columns. By using probabilistic sampling from the joint distribution of focus groups to prime data synthesis, the model can ensure that critical variables (such as employment status or income brackets) are accurately represented in the synthetic data. This is particularly important for maintaining the utility of the dataset for labor market analysis.

These fine-tuning steps are **essential for ensuring the synthetic data match the nuanced characteristics of real labor force surveys**. The resulting dataset is not only statistically similar but also maintains the specific structural features of the original data.

Probabilistic & deterministic approaches

The project employed a combination of probabilistic and deterministic sampling approaches to balance accuracy and flexibility in the synthetic data generation:

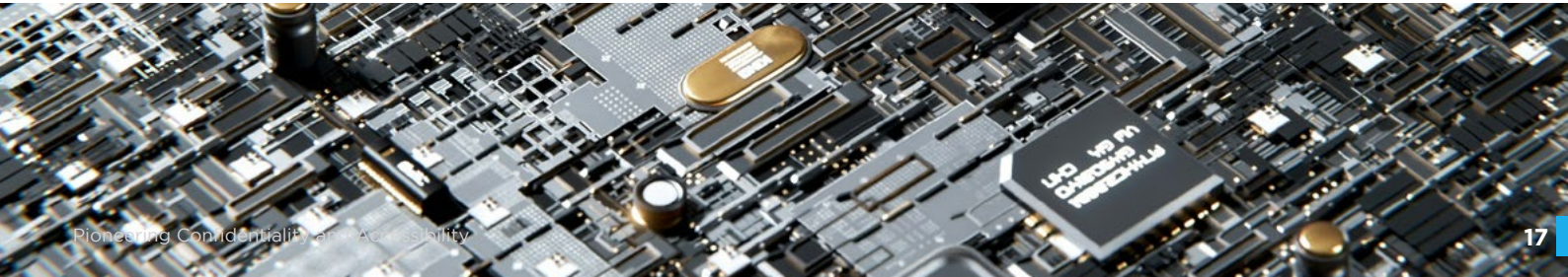
a) Probabilistic Sampling:

This method draws random values according to their probabilities in the original data. While it doesn’t guarantee exact replication of sample counts per subgroup, it allows for natural variation in the synthetic data. This approach is valuable for capturing the overall distribution of variables without overfitting to the exact counts in the original dataset.

b) Deterministic Sampling:

This technique draws an exact proportion of samples per subgroup from a larger pool of options. It guarantees that the subgroup proportions for the considered marginals will be identical to the original data (when even the natural variation from non-deterministic sampling causes too much divergence).

The combination of these approaches allows for a balance between maintaining critical distributional properties (through deterministic sampling) and introducing realistic variability (through probabilistic sampling). This is crucial for creating synthetic datasets that are both statistically robust and resistant to reverse engineering attempts that might compromise privacy.



In conclusion

The integration of these advanced data simulation techniques – tabular diffusion, specialized use-case specific structural input, and balanced sampling approaches – enables the creation of high-quality synthetic datasets for labor force surveys. This synthetic data closely mimics the statistical properties, relationships, and structural features of real survey data. The resulting datasets can be used for a wide range of analyses and policy-making purposes without risking the disclosure of sensitive individual information, nor exact group-level counts for subgroups, thereby offering a powerful solution to the challenge of balancing data utility with privacy protection in official statistics.

Ho et al. "Denoising Diffusion Probabilistic Models", Proceedings of the Annual Conference on Neural Information Processing Systems 2020 (NeurIPS 2020)

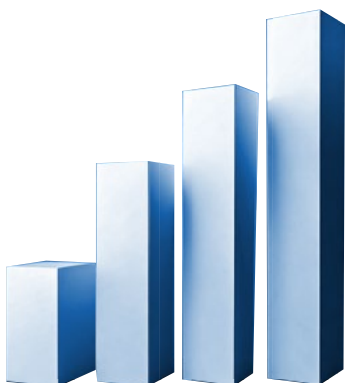
Peebles et al. "Scalable Diffusion Models with Transformers", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV, 2023)

2.3 Validation and Testing of Synthetic Data

Validating and testing synthetic data is an important step in guaranteeing its usefulness and reliability in labor force surveys and other statistical analyses. The pilot study on synthetic data used a comprehensive validation and testing methodology to evaluate the accuracy, reality and privacy protection of the synthetic datasets produced.

2.4 Reporting & Analysis

Here's an outline of the main strategies and approaches employed in this procedure.



Statistical Comparisons

The basic statistical features of the synthetic data are compared to those of the original data to ensure that essential metrics like means, variances, and distributions are correctly recreated.

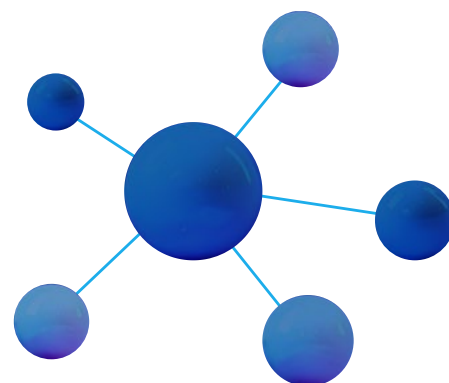
Descriptive statistics, histograms, and frequency distributions are utilized for comparing the synthetic and original datasets. This comparison helps in determining whether the synthetic data retains the overall statistical properties of the real data, hence ensuring its utility for analytical purposes.



Multivariate Analysis

Beyond univariate statistics, synthetic data must keep the linkages and correlations between multiple variables.

Correlation matrices are used to ensure that the synthetic data appropriately represents the complex relationship of variables.



Utility Testing

The utility of synthetic data is evaluated to ensure that it may be utilized for the same goals as the original data, such as predictive modeling and policy analysis.

To forecast crucial outcomes, predictive models (such as logistic regression and decision trees) are trained on both original and synthetic datasets. The performance of these models, as assessed by measures like accuracy, precision, and recall, is compared to see if synthetic data can replace actual data in practical applications.

The fundamental of evaluating synthetic data is the comparison with the real data on various aspects:

RESEMBLANCE

With the use of distribution histograms per column, correlation matrices and multivariate analysis, the statistical resemblance of the synthetic data is assessed. This basically says if the synthetic data looks the same as the real data.

UTILITY

Using various machine learning classifiers and performance indicators such as precision and recall, the utility of the synthetic data is determined. This indicates if the synthetic data behaves the same as the real data when used for training algorithms, forecasting, clustering and more complex analysis.

PRIVACY

Both distance based metrics and privacy attacks are used to assess the privacy leakage risk of synthetic data. The privacy attacks try to simulate an attacker and measure the risk of singling out, likability and attribute inference. The methodology is based on the guidance provided by Article 29 Working Party (WP) of the EU GDPR on the definition of anonymous data.

RESEMBLANCE



UTILITY



The images below are from the BlueGen.ai evaluation report and **demonstrate the primary KPIs for resemblance** which indicate the overall similarity between real and synthetic data.

UNIVARIATE SIMILARITY



Univariate Similarity is based on **the Jensen-Shannon distance***, a symmetric distance measure between the probability distributions of the real and synthetic columns.

BIVARIATE SIMILARITY

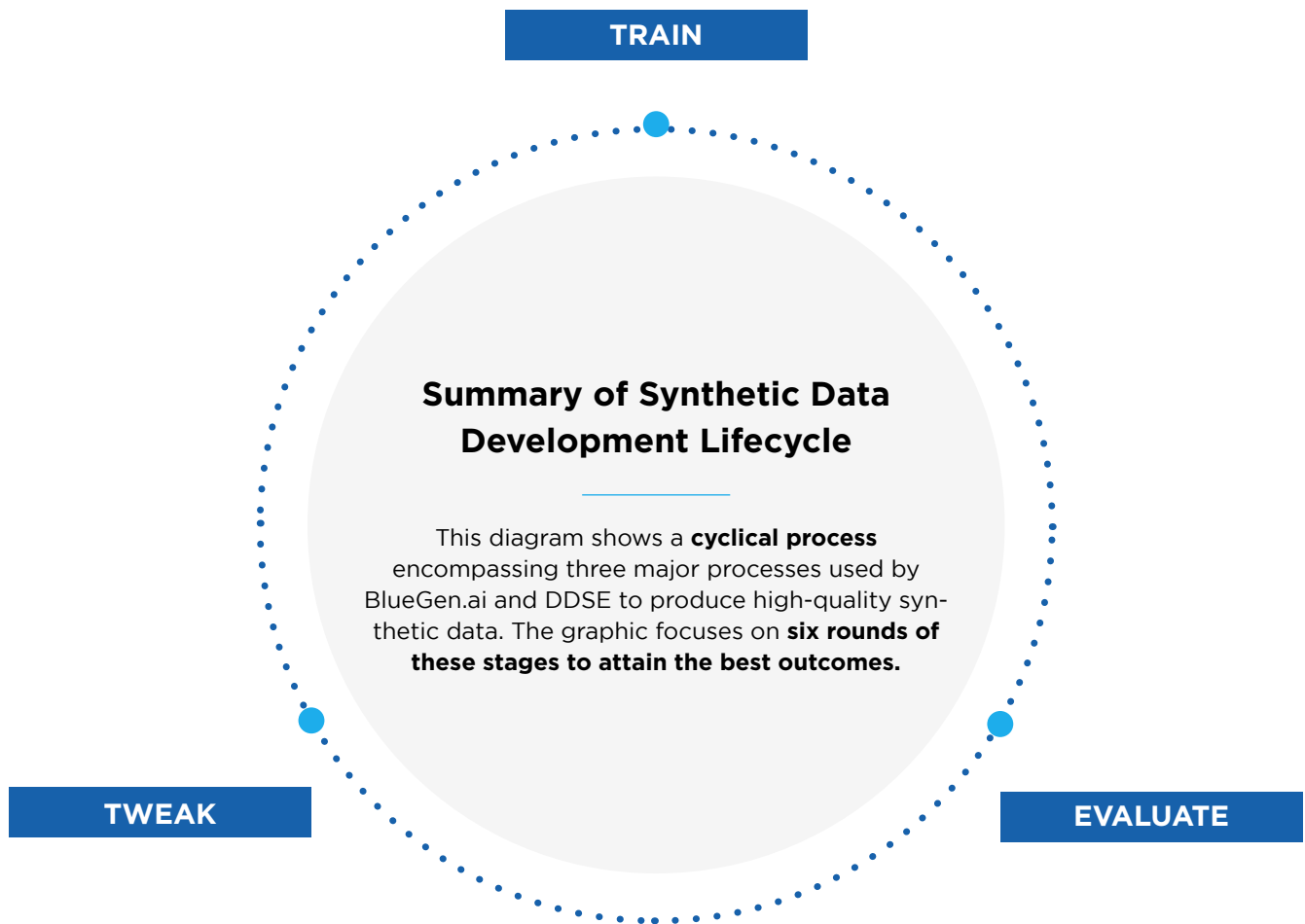


Bivariate Similarity is based on **the Jensen-Shannon similarity** between the correlation coefficients between each pair of columns.

MULTIVARIATE SIMILARITY



Multivariate Similarity is based on **the absolute deviation of prediction probabilities from 50-50 of a binary classifier** trained to discriminate between real and synthetic data samples (also known as the Propensity Mean-Absolute Error).



The breakdown of each step in the process:

Train:

The first stage is to train the model using software. This includes comprehending data distributions, correlations, and linkages within the dataset.

The goal is to ensure that the model understands the underlying patterns and structures in the real data.

Evaluate:

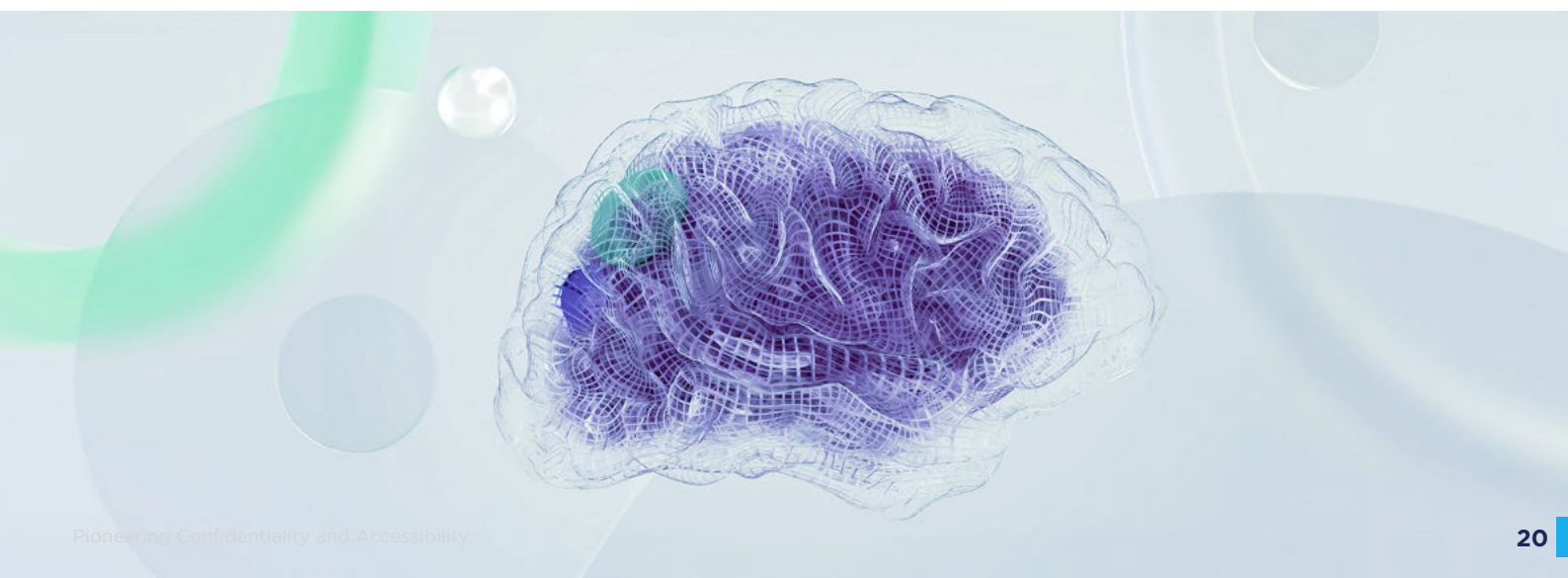
The second phase involves evaluating training data using a synthetic data quality report.

The examination assesses technical and statistical quality, utility, privacy, and the quality of the generated data from a technical and statistical perspective.

Tweak:

The third phase involves optimizing the model for better performance. This could include modifying configurations or tweaking the model to better suit DDSE's specific needs.

This phase is critical for refining the model based on the previous round's output.



RESULTS AND ANALYSIS



→ Results and Analysis



3.1 Key Findings from the Pilot Study

The pilot study on synthetic data for labor force surveys produced several significant findings, providing useful insights into the feasibility, accuracy, and promise of synthetic data for improving utilization & accessibility of labor statistics.

High Fidelity to Real Data

The synthetic data generated in the pilot study closely mirrored the statistical features of the actual labor force survey data, including key metrics like means, variances, and correlations across variables.

Enhanced Privacy and Accessibility

Making data available for research and policy analysis while protecting sensitive information is a major challenge in labor force surveys. As a result, the use of synthetic data enabled a larger sharing of datasets with researchers and other stakeholders while maintaining respondents' confidentiality. This allowed for a more in-depth, collaborative research of labor market trends while maintaining tight privacy and data-sharing rules.

Scalability and Flexibility of Synthetic Data Solutions

One important feature of synthetic data is its capacity to scale for larger datasets or adapt to new data requirements. The synthetic data generation approach proved to be scalable, allowing for the creation of enormous datasets that accurately represented the labor force, thereby accommodating varying research needs.

Challenges in Capturing Rare Events

While synthetic data creation approaches are generally effective at duplicating prevalent trends in labor force data, producing reliable data for **rare and unexpected events was challenging at first.**

These rare and frequently underrepresented events are crucial for a thorough labor market analysis, particularly when focused on specific demographic segments or unusual labor force characteristics

The pilot study discovered that synthetic data often struggled to capture unusual events or extreme outliers. Replicating variables like unemployment rates by household type and the age and gender distribution of unemployed individuals within each household type proved especially challenging when these occurrences were rare and represented by only a few samples in the dataset. These unusual patterns are critical for understanding specific labor market dynamics, especially in the context of focused policymaking and social welfare interventions.

To address this challenge, the study implemented **two key approaches:**

Focus Columns Approach:

Key factors linked with unusual events, such as household type, age, and gender structure of the unemployed, were used as focus columns. By giving these variables more weight and priority throughout the synthetic data generation process, the model was able to catch unusual events and ensure that these critical data points were preserved.

Deterministic Approaches:

When probabilistic models proved insufficient for obtaining accurate rare event data, deterministic methods were used. These methods entailed actively conditioning the synthetic data to retain the presence of rare occurrences, such as specific unemployment rates by household type, ensuring that the synthetic data adequately represented these key components of labor force data.

3.2 Comparison of Synthetic Data vs. Real Data

The comparison of synthetic and real data shows that, **while synthetic data has many advantages, such as privacy protection, cost effectiveness, and scalability, real data continues to excel in capturing unusual events and providing the best accuracy for thorough labor force analysis.** Synthetic data is a solid option for most broad-based labor market evaluations and policy simulations. However, for high-precision jobs that necessitate precise representations of rare events, real data remains essential.

This pilot study demonstrates that



Synthetic data has enormous potential for improving labor market research and operations while finding a balance between data utility and privacy protection.






Comparing Synthetic Data Versions Across the Pilot Lifecycle (2018–2022/2023)

The comparison of real data to various versions of synthetic data generated from different runs and changes between 2018 and 2022/2023 demonstrates a high degree of consistency across situations. **These synthetic datasets, which were created utilizing advanced approaches such as “NORMAL” and “LARGE” datasets, were subjected to many modifications.** The outcomes of these changes consistently tracked real-world trends in the majority of key labor force indicators.

While the synthetic data models were highly accurate to the real data, some minor deviations were detected. **These discrepancies primarily reflect rare-event dynamics,** which are naturally more difficult to replicate using synthetic methods. However, the model’s adaptability and performance throughout these periods demonstrate ongoing improvements in accuracy and data representation.



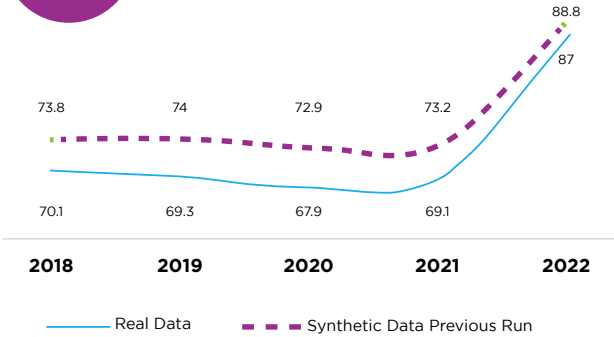
The improvements made to synthetic data models reflect significant progress in the development of synthetic data methods.

These developments have not only improved the reliability of synthetic datasets for labor force analysis, but also proven their potential for broader applications in policy simulation and trend analysis.

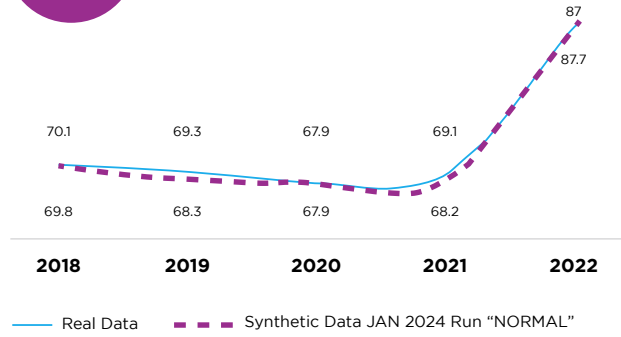
Synthetic data’s ability to continuously develop and adapt to new data patterns makes it a promising tool for future data analysis, with further improvements projected to decrease the gap between synthetic and real data in rare-event representation.

As an illustration, the graphs below provide six comparative charts that examine the link between real and synthetic data obtained from various runs and changes spanning the years 2018-2023. The synthetic data is generated using various approaches and tweaks. For privacy reasons, the chart titles have been masked. However, the comparison between real data and synthetic data across six different scenarios remains the focus of analysis.

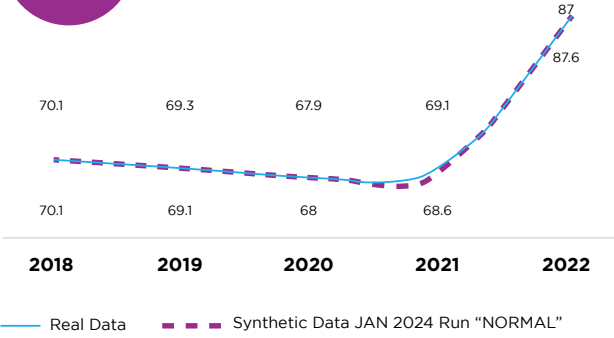
1st



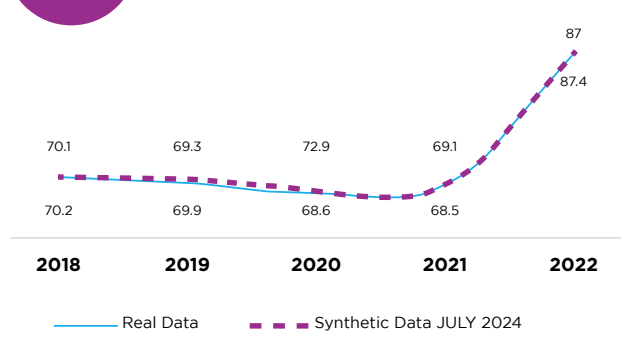
2nd



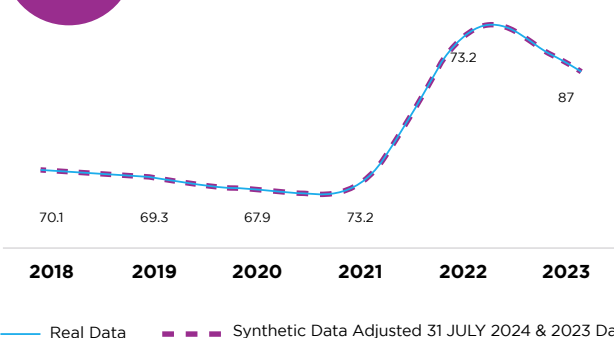
3rd



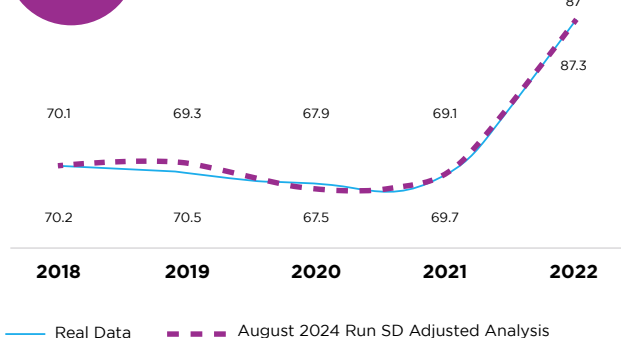
4th



5th



6th



DISCUSSION



→ Discussion

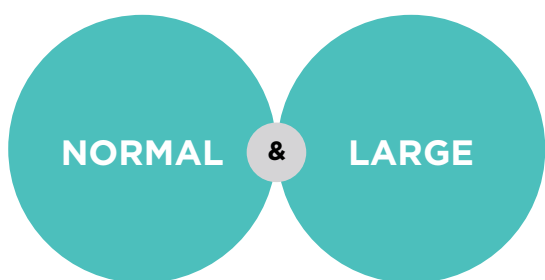
4.1 Evaluation of the Pilot Study Outcomes

The pilot study on synthetic data for labor force surveys was evaluated based on its application in various analytical contexts, the accuracy and reliability of the created datasets, and the larger implications for data privacy and cost-efficiency. This evaluation sheds light on the potential of synthetic data for future use in labor market research and decision-making.

Overall Effectiveness of Synthetic Data

The pilot study demonstrated that synthetic data is an effective instrument for conducting a variety of labor market evaluations. The study's synthetic datasets successfully captured key labor-force features such as employment status and demographic distribution. The data proved to be valuable in common labor force measures such as workforce participation and sectoral employment trends, as it closely matched real-world data. Across several versions and time periods, the generated synthetic data remained consistent with genuine data..

THE APPROACHES USED, NOTABLY



produced datasets that closely **reflected real-world labor force patterns.**

SYNTHETIC DATA

demonstrated great reliability in general analysis, highlighting its potential as a trustworthy substitute for real data in a variety of scenarios.

Initially the deep learning model **struggled to fully capture rare events** (such as unemployment rates by household type and their age and gender structure). BlueGen.ai's model and software enhancements successfully improved the accuracy for these categories. However, **real data remains indispensable to evaluate the quality of synthetic data and verify important analysis results.**

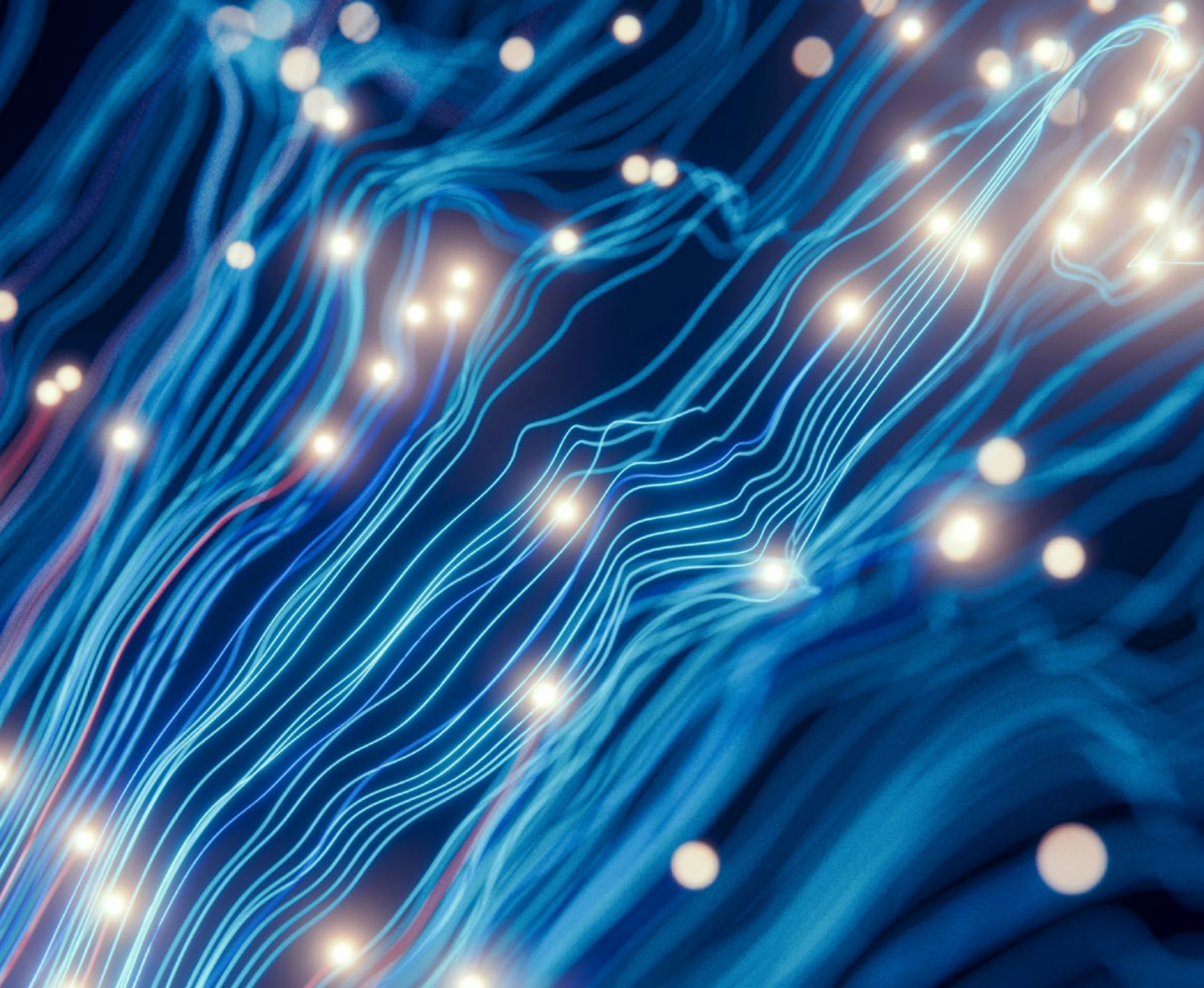
Synthetic data provides numerous benefits, including

▶ Improved Privacy Protection

▶ Scalability

▶ Data Accessibility

allowing for the production of enormous datasets while maintaining confidentiality



4.2 Strengths and Weaknesses of Using Synthetic Data

The pilot study using synthetic data for labor force surveys reveals both its advantages and disadvantages.

This makes it perfect for more widespread sharing and collaboration across sectors. However, there is always a trade-off between privacy and utility. This means that **high quality privacy-safe synthetic data will have minor deviations from the real data, especially in the sparse features**. This is to prevent overfitting and unwanted privacy-leakage risks.

Besides that, the model needs to be maintained over time to capture recent and unexpected labor market changes. **Re-training the model every year with the newest real-world data is advised.**



BlueGen.ai has shown that by using various methods it's able to optimize the trade-off between privacy and utility

Users of synthetic data, like policymakers, should be aware of this and learn how and when to apply it



CONCLUSION

This pilot study illustrates the promising potential of synthetic data to improve labor force survey statistics accessibility

By offering scalable and privacy-protected data utilization and analysis solutions, synthetic data offers a pathway to more efficient and secure data sharing.

The tight collaboration between BlueGen.ai and DDSE shows that deep learning models enriched with domain knowledge can offer tremendous results on recapturing all kinds of behaviour and events in the data. While there are still challenges, they can easily be tackled by the continuous development of the software and the enrichment with new data.

The overall conclusion is that this synthetic data is suitable for sharing under well-governed arrangements with partner organizations. This represents a significant step forward for city data sharing approaches. That said, further research of the privacy implications and accompanying governance arrangements is needed before we are ready to consider open publishing of synthetic data. The study establishes a foundation for future innovations in data-driven policymaking, ensuring that privacy concerns are adequately managed while enhancing the utility of labor statistics. Ultimately, this approach fosters the integration of modern technologies, promotes learning, and expands the vision for building robust databases that support both large-scale and targeted data-driven initiatives.

Next Steps for the Collaboration Between Dubai Data and Statistics Establishment and BlueGen.ai

The collaboration between Dubai Data and Statistics Establishment (DDSE) and BlueGen.ai has been central to the successful execution of this pilot study.

Each partner brought complementary strengths: DDSE contributed deep expertise in statistical methodologies and labor force analysis, while BlueGen.ai offered cutting-edge synthetic data technology.

Looking ahead, the collaboration among DDSE, and BlueGen.ai will focus on refining synthetic data models to improve their accuracy, especially in detecting unusual events. The partnership aims to expand the application of synthetic data beyond labor force surveys to include other domains such as households living conditions, social and economic research. Continuous knowledge exchange and model enhancement will be priorities as this collaboration works towards positioning Dubai as a leader in data-driven innovation, ultimately supporting the city's broader digital transformation goals.

