BlueGen ai

Better, safer, and faster than real data ...

Introduction

TU Delft
Delft University of Technology

April - 2023

Why we do what we do.

Data is the lifeblood of digital business, but we believe protecting customers, consumers, citizens and patients is just as important. We believe in a world where data drives prosperity, while the privacy of the individual is preserved and technology operates unbiasedly.

PRIVACY CHALLENGE

Just following the rules is not enough ...

NOS Nieuws • Donderdag, 18:03 • Aangepast donderdag, 23:25

Datalek Nederlandse bedrijven steeds groter: zeker 2 miljoen klanten getroffen

Nando Kasteleijn
redacteur Tech

Julius Moorman
redacteur economie

Zeker 2 miljoen Nederlandse klantgegevens blijken betrokken bij een groot Nederlands datalek. Naast onder meer de NS en VodafoneZiggo zijn er nóg drie bedrijven geraakt: ook klantgegevens van de Nederlandse Golf Federatie, ArboNed en vervoersbedrijf Trevvel blijken te zijn gelekt.

De Autoriteit Persoonsgegevens heeft nog geen totaalbeeld kunnen maken. De toezichthouder start een onderzoek. Het kan overigens zijn dat mensen meerdere keren voorkomen in de data.

Beste reiziger,

Uit voorzorg informeren wij u over het volgende: bij een leverancier van marktonderzoekbureau Blauw waar wij mee samenwerken, is een datalek gevonden. Mogelijk is daarbij een aantal van uw persoonsgegevens gelekt, zoals naam, e-mailadres en telefoonnummer. Het gaat niet om financiële gegevens of wachtwoorden.

Maatregelen genomen om lek te dichten
Het datalek vond plaats bij een leverancier van marktonderzoekbureau Blauw. Dat bureau doet regelmatig onderzoek in opdracht van NS. U bent voor een of meer van deze onderzoeken uitgenodigd. Het zou kunnen dat uw gegevens via hen zijn gelekt, we vinden dat erg vervelend. Er zijn direct maatregelen getroffen om het datalek te dichten en om herhaling te voorkomen. Ook hebben wij melding gedaan bij de Autoriteit Persoonsgegevens.

BlueGen.ai

And the rules can limit innovation and execution ...

Vrijdag 5 augustus 2022, 18:17

'Vervolgonderzoek naar oversterfte moeilijk door beperkte toegang data'

Rudy Bouma
verslaggever
Nieuwsuur

Roel van Niekerk

Het CBS en het RIVM publiceerden in juni hun onderzoek naar oversterfte tijdens de coronapandemie. Op verzoek van de Tweede Kamer doet een groep onafhankelijke wetenschappers een vervolgonderzoek naar die oversterfte. Maar een deel van de benodigde data is om privacy-redenen niet voor hen toegankelijk, zeggen de wetenschappers.

‘Nederlandse privacy-wetgeving hindert innovatie’

oktober 26 2015

How Data Protection Regulation Affects Startup Innovation

PROBLEMS

BlueGen.ai

Data is the lifeblood of digital businesses. However, real data ...

... can't be used or moved because of privacy constraints

... may have imbalances because of lacking “edge cases” or demographic diversity

... is too expensive and time consuming because it's hard to collect at scale

BlueGen.ai

Preventing organisations to privacy-safe ...

Data sharing and collaboration

Machine Learning training and testing

Software product testing and development

Vendor evaluation and procurement

Data & product experimentation

Data Retention

Opportunities for data monetization

Access to third-party analytics consultants

Access to cloud services

Product demonstrations

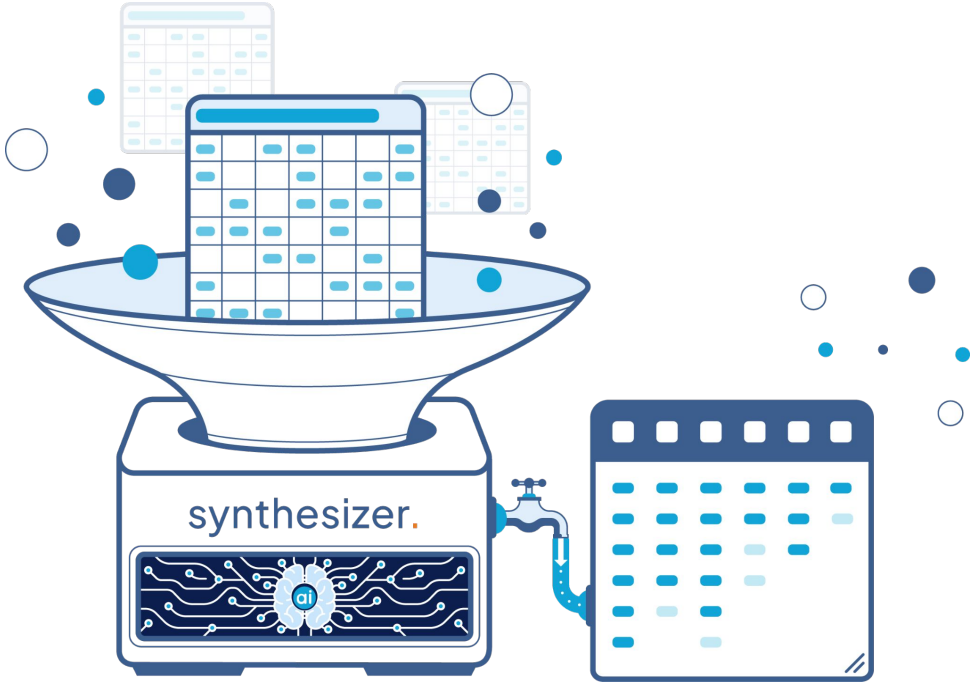
"While in 80% of data innovation use cases, it's not required to identify the individual"

BlueGen.ai Platform .

BlueGen uses AI to learn from the real data to generate a new data set that looks and behaves the same as the real data, without any personal identifiable information

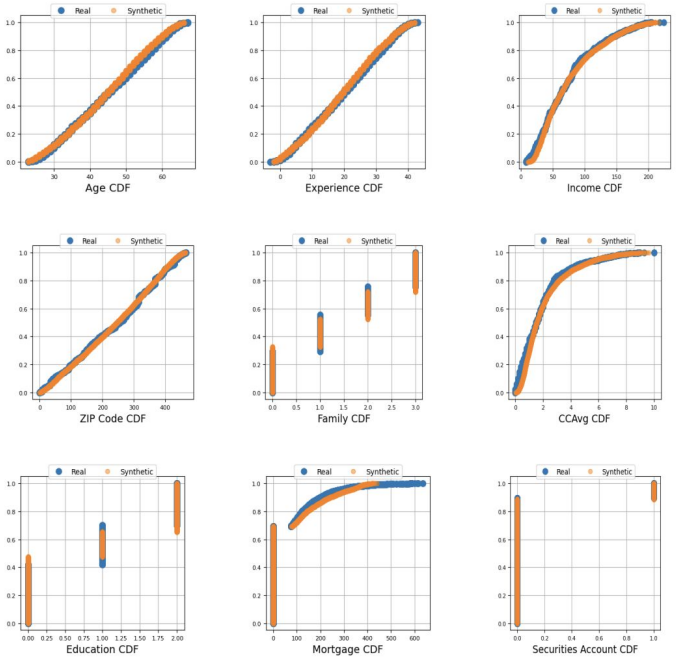
AI generated or “smarter” synthetic tabular data

- retains the statistical properties of the original source data
- is not subject to the rules of GDPR, In accordance with Recital 26 of the GDPR
- is trained decentralized “at the edge” so the data does not leave its location = privacy by design



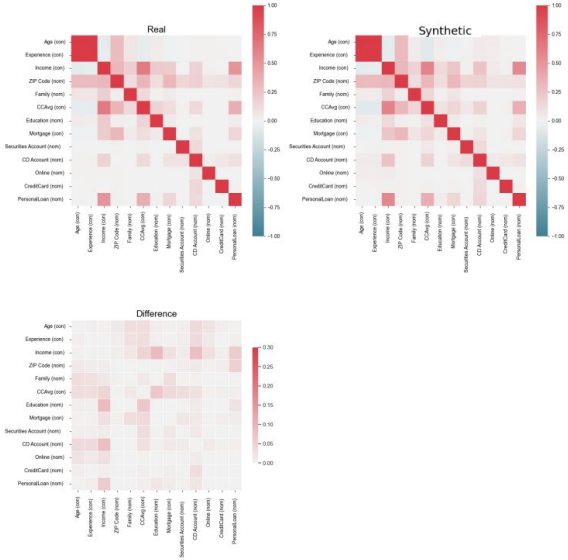
We continuously invest in explainable metrics that prove the quality of our synthetic data for your use case .

Cumulative Distribution Analysis



Correlation Analysis

An analysis of the correlations between columns. Red corresponds to columns whose values are both high together and low together, while blue implies the columns are negatively correlated. The synthetic data should have similar correlations to the real. Therefore the difference plot should have only low values that are randomly divided over all column pairs.



SOLUTION

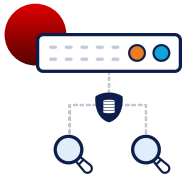
Why BlueGen.ai's synthetic data ?

Real data shortcomings	BlueGen.ai Synthetic Data
Restricted in use due to privacy	Privacy guaranteed by **differential privacy**
Incomplete and biased: not containing all possible scenarios	**Augmented/conditioned** for proper distribution and edge cases
Not enough data	Data can be **multiplied**
Expensive: hard to collect, integrate, store and maintain	**Regenerated** with just a click or integrated into data engineering and CI/CD pipelines
Data needs to stay on-premise and cannot be moved or shared	BlueGen's **decentralized learning** allows data to remain on-premise

BlueGen.ai

SOLUTION

(im)possible ...



Use cases where the exact personal identifiable information (PII's) is needed

Use cases where various sources still need to be matched with unique identifiers

80% of data sharing, analysis, training and development use cases

STATISTICS

According to Gartner, by 2025 .

The use of synthetic data will reduce the volume of real data needed for analytics and machine learning by 70%

Synthetic data will reduce the personal customer data collection, avoiding 70% of privacy violation sanctions

BlueGen.ai

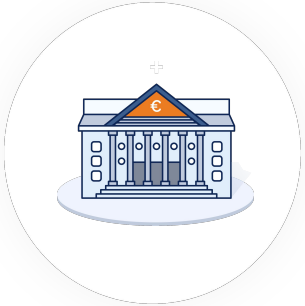
OUR WORKFIELD

Industries .

Healthcare

Test datasets for clinical trials

Synthesizing patient data to train ML



Financial Services

Financial crime and fraud simulation

Innovation sandbox for exploration

Government

Reliable dataset to analyze and improve services

Internal and external data sharing

Utilities

Asset failure prediction

Energy consumption forecasting

Case Study

How NIBC bank uses BlueGen.ai synthetic data to accurately predict loan default and reduce risk .

NIBC is the entrepreneurial asset financier for companies and individuals. We finance assets from private housing to rental property, commercial real estate, vessels, infrastructure, cars and equipment. NIBC employs around 700 people and is headquartered in The Hague.

Problem .

Banks are obliged to manage and model their financial risk. To do this they need to be able to predict the loans that are at risk. Machine learning models are struggling to predict the default of loans due to a data inefficiency. There are not sufficient unhealthy loans in the data which leads to unstable model outcomes. Accessing the privacy sensitive data data is another challenge.

Solution .

The BlueGen.ai synthetic data allows NIBC access the data in a privacy-safe way. The conditioning capabilities of the platform can oversample the unhealthy loans for the model to be trained more robust and accurate.

Results .

Thanks to BlueGen.ai's synthetic data, NIBC is now able to train more robust models that leads to more accurate predictions improving the risk financial modelling. As a result more capital can be reinvested and better interest rates can be offered which will provide a competitive advantage.

Industry: Banking

Use Cases: Accurate training of ML models

BlueGen.ai

The quality of our synthetic data for NIBC is recognized by downstream use cases from Amsterdam Data Collective (ADC) .

Case Study

BlueGen.ai

How EDF uses BlueGen.ai synthetic data to accurately predict electricity consumption and reduce cost .

EDF (Électricité de France) is a global, integrated energy company, one of the world's largest electricity producers, and the largest renewable energy producer in Europe. EDF specialises in electricity, from engineering to distribution, has 37.6m customers with a turnover of €85b

EDF

Problem .

Electricity is hard store store and as a result it is extremely important to predict how much electricity is going to be consumed in order align the supply. . Due to privacy regulation EDF is not allowed to store energy consumption for more than two years, making it almost impossible to make accurate predictions of future energy consumption.

Solution .

BlueGen.ai's synthetic data allows EDF to maintain the statistical properties of the energy consumption and the characteristics of the household without any personal identifiable information. Events in the past can be investigated and with the extended baseline accurate predictions can be made. Thanks to the conditioning capabilities of the BlueGen.ai platform, EDF will also be able to create what if scenarios, e.g. increase in the use if EV's.

Results .

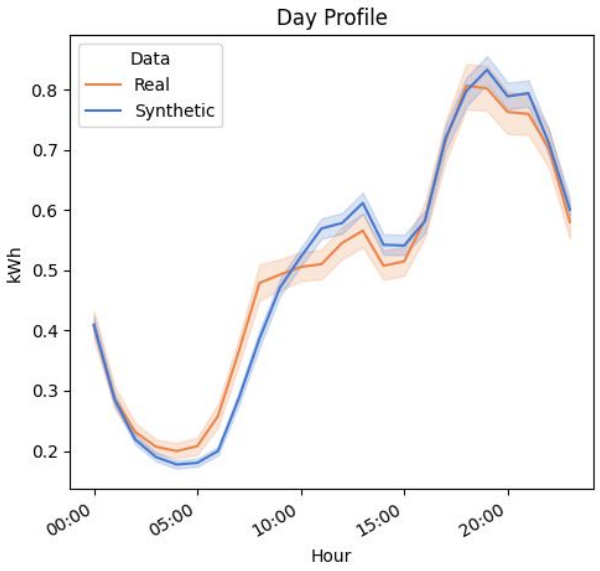
The accuracy of the energy consumption predictions will significantly increase allowing EDF to better match their supply in a more cost-effective manner.

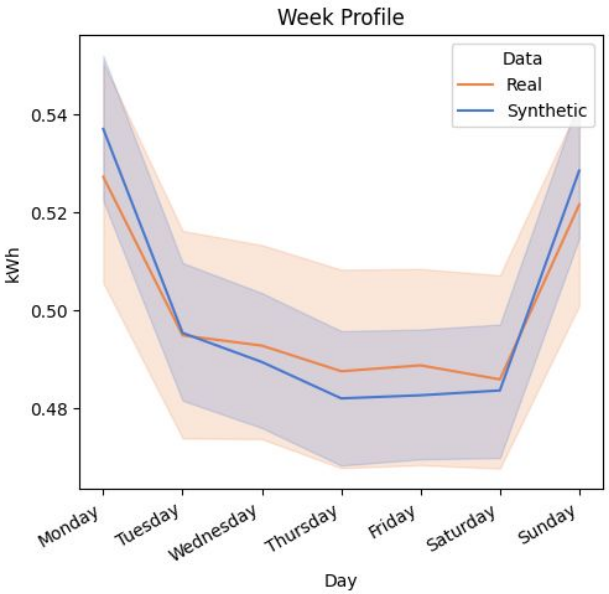
Industry: Energy

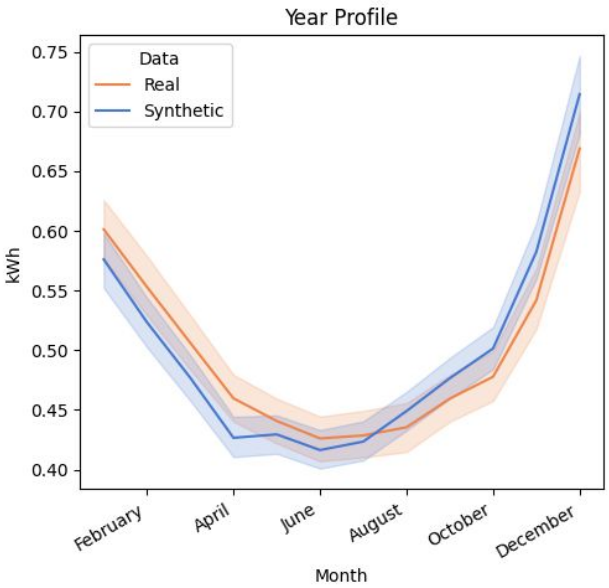
Use Cases: data retention, Accurate training of ML models

BlueGen.ai

The accuracy of our synthetic time series data over a whole year is spot on for various analytics and ML use cases .







BlueGen.ai

Government .

- interdepartmental analysis and collaboration
  EXAMPLE: RIVM, Municipalities
- unbiased and fair algorithm training
  EXAMPLE: UWV, Belastingdienst
- open data research and experimentation
  EXAMPLE: DUO, Statistics Centres

BlueGen.ai

We guarantee privacy .

Our synthetic data does not inherit personal identifiable information

Our privacy evaluations are based on transparent metrics

We do not need or want to have access to your real data

SOLUTION

Enabling ...

Privacy-compliant data sharing and collaboration

Machine Learning training and testing

Software product testing and development

Vendor evaluation and procurement

Data & product experimentation

Data Retention

Opportunities for data monetization

Access to third-party analytics consultants

Access to cloud services

Product demonstrations

BlueGen.ai

BlueGen.ai Publications .

CTAB-GAN: Effective Table Data Synthesizing, 2021 (published in ACML 2021)

Permutation-Invariant Tabular Data Synthesis, 2022 (published in BigData 2022)

FCT-GAN: Enhancing Global Correlation of Table Synthesis via Fourier Transform, 2022

GDTS: GAN-based Distributed Tabular Synthesizer, 2022 on horizontal federated learning

GTV: Generating Tabular Data via Vertical Federated Learning, 2023 on vertical federated learning (under submission to VLDB)

On Auditing Stolen Risk of Generative Adversarial Networks, 2023 on knowledge extraction of GAN (under submission to ICML)

Differential Privacy & Synthetic data in collaboration with CWI - in progress

ChatGPT & Synthetic data - in progress

Diffusion Models & Synthetic data - in progress

Thank you for your time.

NICPET APRIL 2023

BlueGen

ABOUT US

Meet our founding team.

Iman Alipour

CHIEF EXECUTIVE OFFICER

Investor, Former CEO Ymor and MD EMEA Lakeside Software

Edwin Kooistra

CHIEF STRATEGY AND MARKETING OFFICER

Growth strategist, ex-Gartner, founder and CEO Chasm

Vincent Campfens

CHIEF PRODUCT OFFICER

Digital Transformation and strategy specialist

Hans Brouwer

LEAD ENGINEER

Master in Artificial Intelligence Technology

Lydia Chen

TECHNICAL ADVISOR

Associated professor Computer Science Technical University Delft

ABOUT US

TU Delft: impact with AI research .

Highly active in education, research and innovation in AI, Data & Digitalisation. Highest ranked in NL for Computer Science.

Unique combination of disciplines, relevant to AI research: computer science, engineering, ethics and design

Large AI scientific community. Over 700 in-AI & 700 with-AI scientists. 24 AI Labs, 5 ICAI Labs

Leadership in national, regional and local AI impact. Academic networks of excellence, research & development networks, startup communities, professional education.

TU Delft

Delft University of Technology

[1] Rocher, L., Hendrickx, J.M. & de Montjoye, YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10, 3069 (2019). https://doi.org/10.1038/s41467-019-10933-3