Data science in insurance: leveraging privacy-preserving synthetic data

Statice webinar

Statice

About Statice

Founded in 2018

Synthetic data provider company founded three years ago, with 17 employees.

Berlin-based startup

Berlin-based startup, operating all over the world, with a focus on Europe.

+30 clients worldwide

Serving clients over a large variety of industries: finance, healthcare, insurance, telecoms.

Speakers

Emna Amor, Product & Engineering Manager

Emna is Statice's go-to expert on deep learning matters with a focus on machine learning and model performances.

Dr. Matteo Giomi, Privacy researcher

Matteo leads the research on the privacy side and ensures our synthesization technology remains ahead of the privacy research.

Benjamin Nolan, Head of Business Development

Ben works with Statice's partners to help them overcome their data access and privacy challenges.

Agenda

I.

The need for data agility in insurance

- Opportunities & challenges
- Synthetic data

II.

SD performances for machine learning

- ML applications
- Performance evaluation

III.

SD and privacy preservation

- SD as anonymization
- SD and Differential Privacy
- Measuring SD privacy

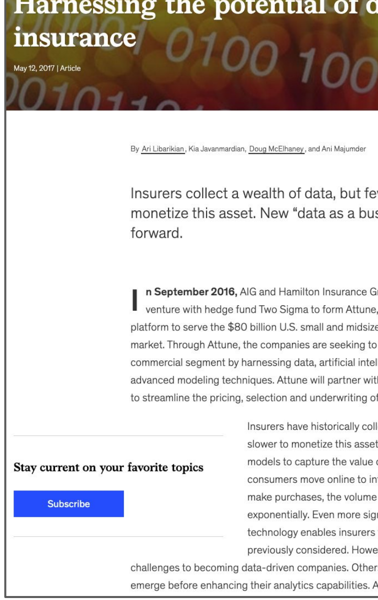
Part 1

The need for data agility in insurance

Data in Insurance

Opportunities

1. **Increased agility of operations**
    - Using modern data methodologies allows insurers to react faster to changes in their customer base and their product risk.
2. **Improved customer experience**
    - Using data enables improved customer experiences like relevant x-sell/upsell, churn prevention, and customer service.
3. **New business models**
    - Using data unlocks new business models like usage-based insurance or behavioral pricing.


Harnessing the potential of d
insurance

May 12, 2017 | Article

By Ari Libarikian, Kia Javanmardian, Doug McElhaney, and Ani Majumder

Insurers collect a wealth of data, but fe
monetize this asset. New "data as a bus
forward.

In September 2016, AIG and Hamilton Insurance G
venture with hedge fund Two Sigma to form Attune,
platform to serve the $80 billion U.S. small and midsize
market. Through Attune, the companies are seeking to
commercial segment by harnessing data, artificial intel
advanced modeling techniques. Attune will partner wit
to streamline the pricing, selection and underwriting of

Stay current on your favorite topics

Subscribe

Insurers have historically coll
slower to monetize this asset
models to capture the value
consumers move online to in
make purchases, the volume
exponentially. Even more sig
technology enables insurers
previously considered. Howe
challenges to becoming data-driven companies. Other
emerge before enhancing their analytics capabilities. A


How Big Data is Revolutionizing Business

Modern society is continuously producing impressive
intelligence, it becomes a valuable source of informa
insurance.

Big Data is mainly used for:

- New distribution models – virtual assistants, robo
interactions and make marketing more targeted;
- Process automation – it substitutes manual labor
workflow;
- New propositions – it enables creating alternative
or digital insurers.

Statice

Big Data for Insurance
Harnessing the potential of data in insurance

Data in insurance

Challenges

1. **Data access**
    - Access to sensitive data is heavily restricted, with long compliance processes associated with getting access, or access not possible at all.
2. **Data usage**
    - Usage of sensitive data is generally limited to the initial collection purpose.
3. **Data sharing**
    - Sensitive data cannot be shared in most circumstances with 3rd parties, or even in many cases with other divisions of the organization.

Data in insurance

Risks

1. **Organizational**
    - Low agility of organization data usage, slow product development
2. **Reputational**
    - Lower consumer trust, potential revenue
3. **Regulatory risk**
    - Mid-single-digit millions for several FS&I companies in the EU in the last 12 months

18 Biggest GDPR Fines of 2020 & 2021 (So Far) | Updated 2021

Data in insurance

What is synthetic data?

> *Data algorithmically generated approximating original data, which can be used for the same purpose as the original.*

Data in insurance

Synthetic data generation

Employee ID	Commute	Department
0	bike	engineering
1	foot	commercial
2	bike	data
3	bike	data
4	transit	data
5	car	commercial
6	transit	engineering
7	bike	data

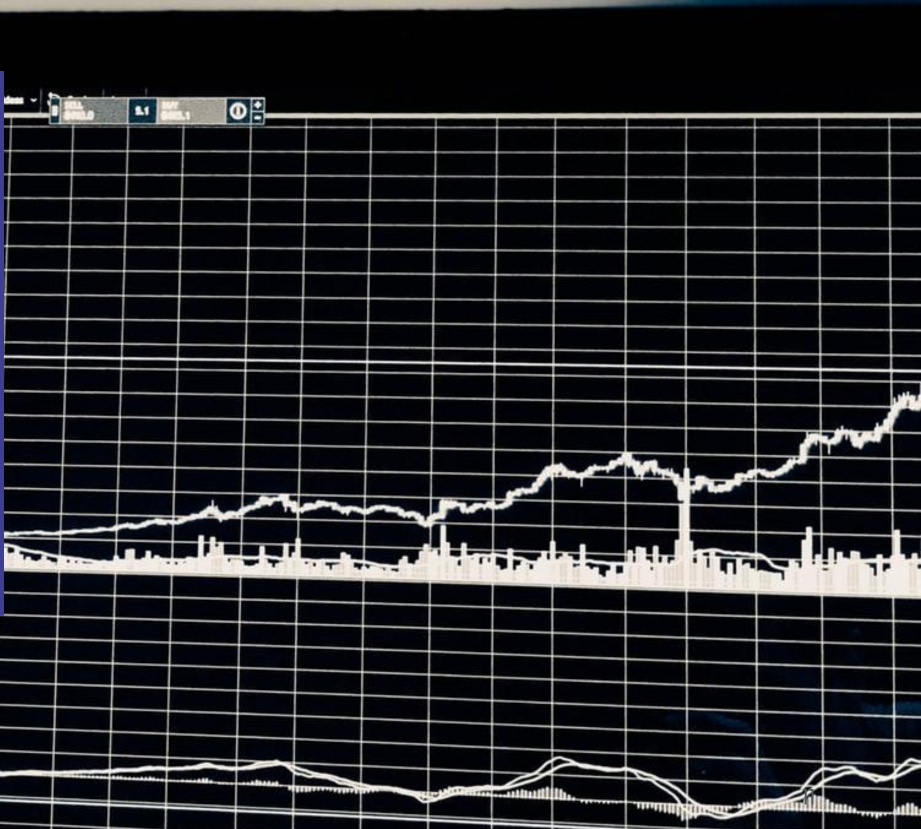
approximates distribution or process behind the data

Employee ID	Commute	Department
0	transit	data
1	bike	commercial
2	car	data
3	bike	engineering
4	transit	commercial
5	car	data
6	bike	data
7	foot	engineering

original data —learn→ **model** —sample→ **synthetic data**

Part 2

Synthetic data performance for ML



Synthetic data for ML applications

Where is synthetic data used in insurance?

1. **Consumer behavior**
   - Churn modeling or purchase behavior analysis

2. **Sales and marketing**
   - Insurance policies and price modelling

2. **Health data**
   - Care pathways or outbreak prediction

3. **Risk, security, and access management**
   - Facial recognition
   - Training of fraud detection models

4. **Automated claims processing**
   - Computer vision

Synthetic data for ML applications

A real-life example: La Mobilière

Context

- Swiss insurance company wanted to anticipate the new privacy regulations expected in the country and implement tools to process data for 2nd purposes.

Project

- Churn prediction models initially relied on customer data (highly sensitive, subject to data protection laws). The data science team needs to maintain utility.

Outcome

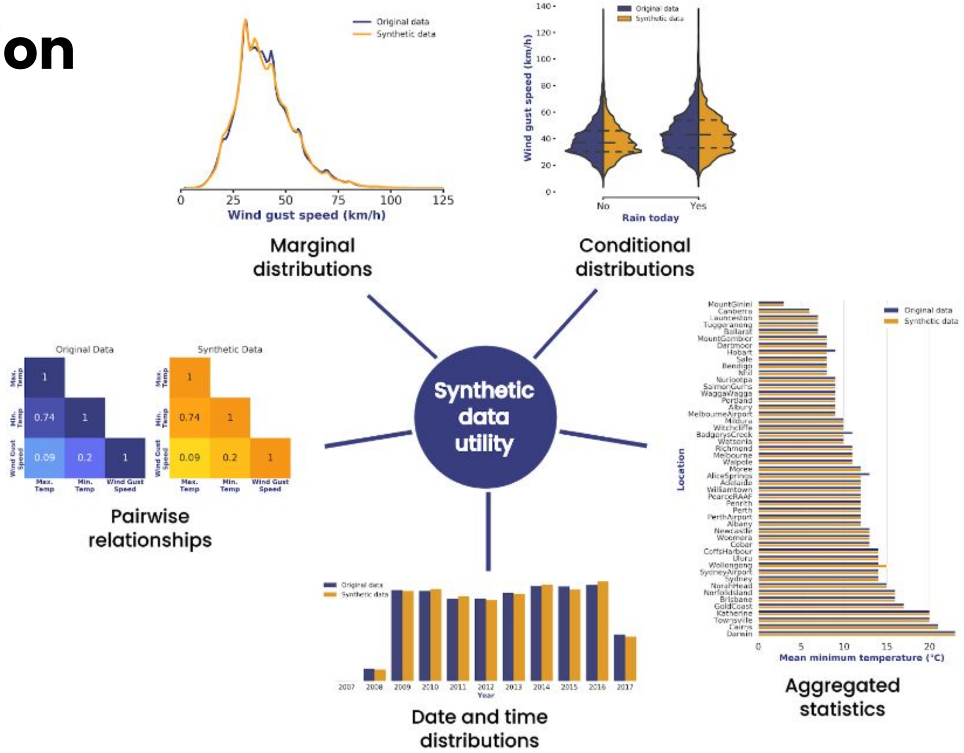
- Validated the use of synthetic data to train churn models, from a utility and privacy point of view.
- In less than 2 weeks, they managed to produce and use highly granular, compliant data that would future-proof this aspect of their data operations.

Performance evaluation

Utility Assessment

Confirming utility

- Marginal distributions
- Conditional distributions
- Aggregated statistics
- Pairwise dependencies:
    1. Correlations
    2. mutual information.
    3. Correlation ratio

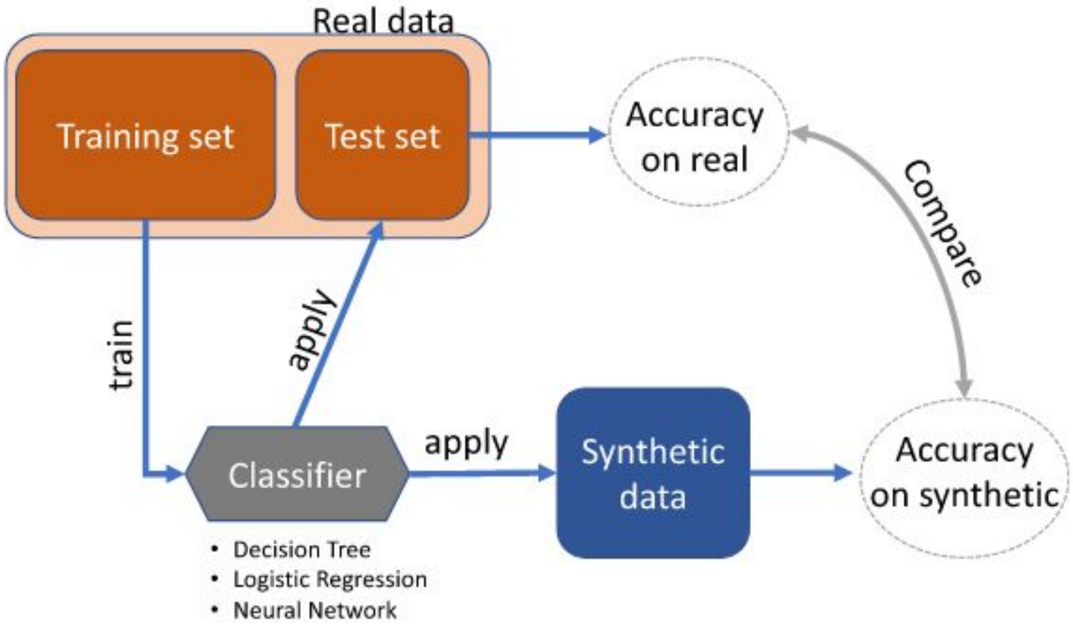


=> Training of machine learning models can be performed on synthetic data with minimal loss in prediction accuracy.

Statice

Performance evaluation

Machine learning



[Generation and evaluation of synthetic patient data]

Part 3

Synthetic data privacy

The quest for anonymization

It all starts with a dataset

phone	race	birth year	sex	zip code	medical condition
015940192	white	1964	f	1203002	chest_pain
010405919	white	1964	f	1203505	obesity
011500159	white	1964	f	1203106	short_breath
010192042	black	1965	m	5403221	heart_disease
015909191	black	1965	m	5403221	heart_disease
015553436	black	1965	m	5403221	heart_disease
016901095	white	1960	f	3003202	ovarian cancer
017497297	white	1960	f	3003555	ovarian cancer
018206810	white	1960	m	3003890	prostate cancer

Statice

The quest for anonymization

PII, quasi identifiers, and secrets

phone	race	birth year	sex	zip code	medical condition
015940192	white	1964	f	1203002	chest_pain
010405919	white	1964	f	1203505	obesity
011500159	white	1964	f	1203106	short_breath
2	black			3221	he
91	black			3221	he
015553436	black	1965	m	5403221	heart_disease
016901095	white	1960	f	3003202	ovarian cancer
017497297	white	1960	f	3003555	ovarian cancer
018206810	white	1960	m	3003890	prostate cancer

Personally identifying information (PII)

"Quasi" identifiers

Sensitive information

Statice

The quest for anonymization

Pseudonymization

~~phone~~	race	birth year	sex	zip code	medical condition
~~[illegible]15940192~~	white	1964	f	1203002	chest_pain
~~01[illegible]04059[illegible]19~~	white	1964	f	1203505	obesity
~~01[illegible]500[illegible]59~~	white	1964	f	1203106	short_breath
~~[illegible]2~~	black	[illegible]	[illegible]	[illegible]3221	he[illegible]
~~[illegible]91~~	black	[illegible]	[illegible]	[illegible]3221	he[illegible]
~~015[illegible]53[illegible]36~~	black	1965	m	5403221	heart_disease
~~01[illegible]59010[illegible]5~~	white	1960	f	3003202	ovarian cancer
~~[illegible]17497297~~	white	1960	f	3003555	ovarian cancer
~~018206810~~	white	1960	m	3003890	prostate cancer

Personally identifying information (PII)

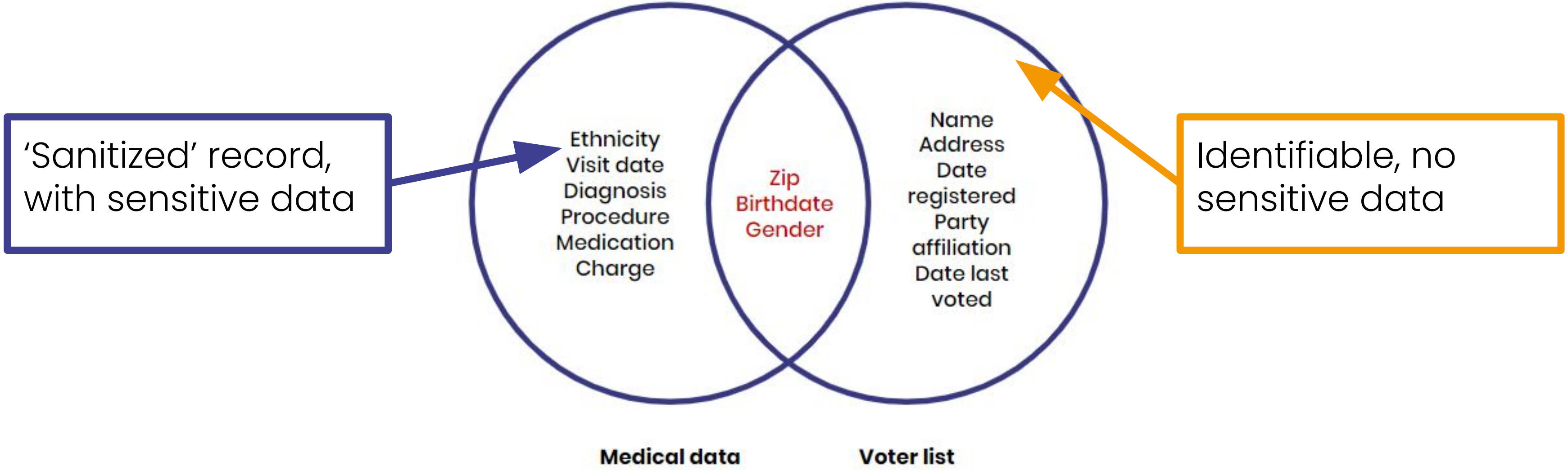
“Quasi” identifiers

Sensitive information

Breaking Pseudonymization

Pseudonymization does not protect from re-identification

An attacker can link pseudonymized records across other datasets to re-identify targets.



Sweeney, Latanya. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine and Ethics, Vol. 25 1997, p. 98-110

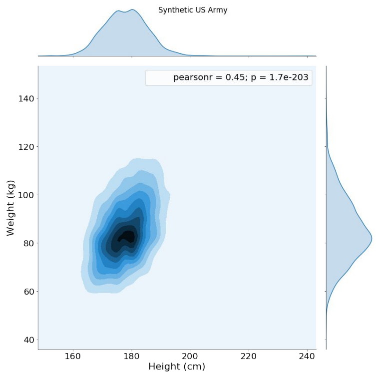
Synthetic data as anonymization

Learn the data generating distribution from the original data and draw samples from it.

	Height (cm)	Weight (kg)
0	165.10	68.03880
1	162.56	88.45044
2	170.18	61.23492
3	172.72	72.57472
4	177.80	88.45044

Original data

inference



sampling

	Height (cm)	Weight (kg)
0	165.059741	67.628771
1	162.493573	89.085417
2	169.926607	50.631068
3	172.567460	72.163759
4	177.599143	88.280858

Synthetic data

This process **breaks the 1-1 relations** between original and synthetic data records.

Statice

Synthetic data and privacy

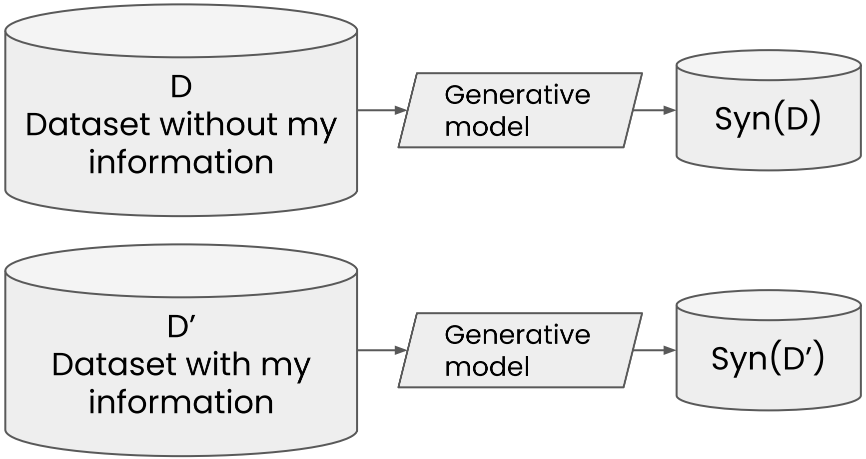
- Generative models come with **big capacity** (i.e,. they have a lot of free parameters).
- These models can "**memorize**" data samples.
- Memorized patterns can be **reproduced** in synthetic data.

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

N. Carlini et al. 2019, The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

Differentially-private synthetic data

Differential privacy (DP) uses **randomness** to mask the presence of any particular individual in the input data.



If the generative model uses DP one would get **"roughly the same" synthetic dataset** whether or not 'your' information is present in the input.

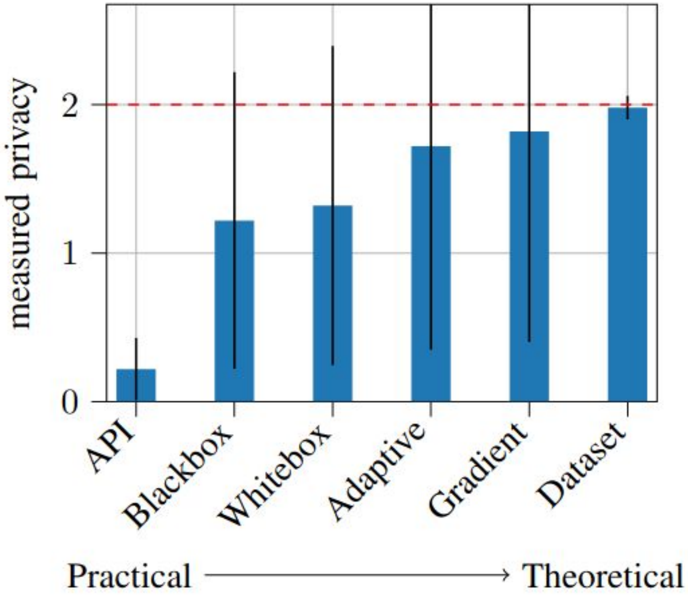
Parameter **ε** quantifies the strength of the privacy (smaller is better).

[Dwork C., et al. (2006) Calibrating Noise to Sensitivity in Private Data Analysis](#)
[M. Abadi et al, (2016) Deep Learning with Differential Privacy](#)

Understanding the ε of DP

- **It is not "black or white".** There is always a risk of information disclosure. If ε is small, this risk is small in all cases. If it's large the mathematical guarantee offer little reassurance. -> **Utility / privacy tradeoff.**

- It is a worse case guarantee. The attack model of DP is often unrealistic. It can provide better levels of protection in practice.



[Nasr et al. 2021, Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning]

Measuring privacy

How can we prove compliance?

A working anonymization technique must protect against:

- **Singling out**: the ability to isolate some of the records which identify an individual.
- **Linkability**: the ability to link 2+ records concerning the same data subject.
- **Inference**: the ability to deduce value(s) of a set of attributes.

We follow three directions as guidelines to develop privacy evaluations that complement the DP guarantee of the synthetic data.

Article 29 working party, Opinion 05/2014 on Anonymisation Techniques

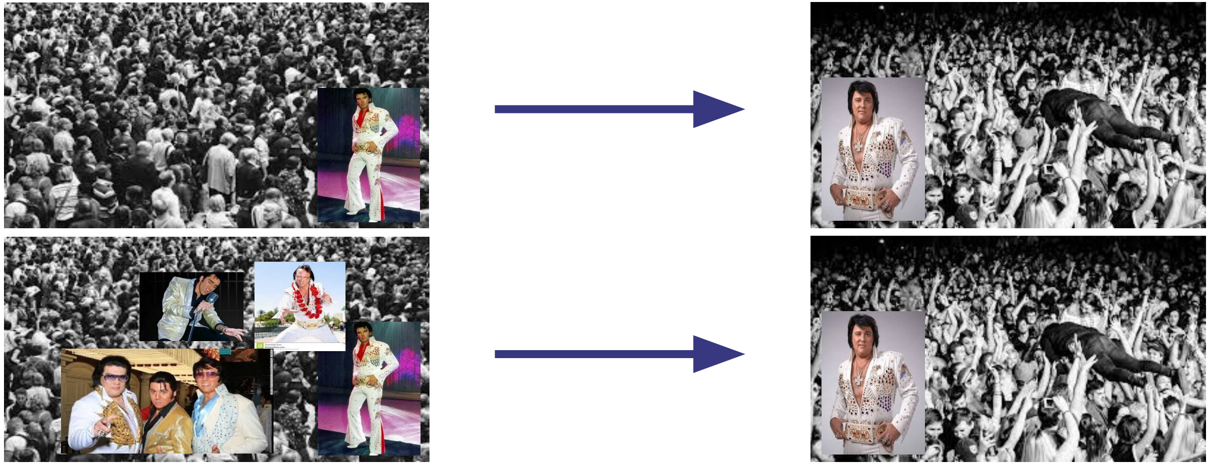
Statice Privacy Evaluations

Linkability analysis

Detect synthetic records that could be linked to original records.

Suspicious

Not suspicious



Original crowd

Synthetic crowd

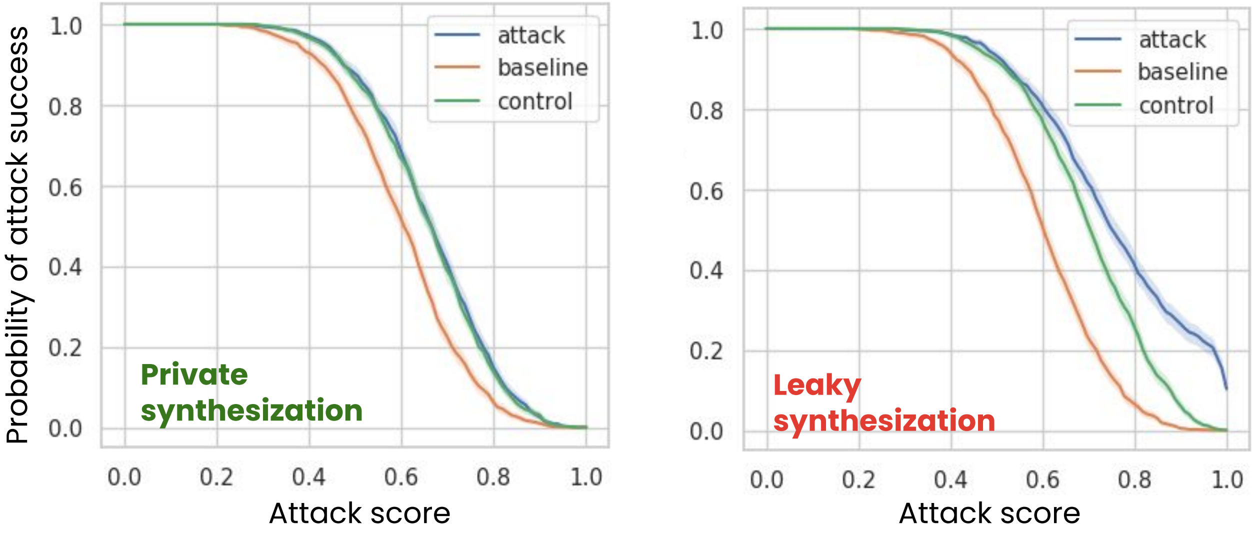
[Article 29 working party, Opinion 05/2014 on Anonymisation Techniques]

Statice

Statice Privacy Evaluations

Inference analysis

How much knowledge on specific records does an attacker gain by seeing the synthetic data?



Statice

Synthetic data and privacy

Takeaways

- Synthetic data is a promising technology for anonymization.
- The sole fact that the data is synthetic does not mean that it's private.
- We can combine SD with differential privacy for state-of-the-art privacy protection.
- Additionally, we assess the privacy of the SD along with the directions of the GDPR.

Dive deeper

Read & watch

Read more:

- Statice's blog
- On privacy matters: "The Machine Learning Revolution in Data Privacy", V. Shmatikov | "The Algorithmic Foundations of Differential Privacy", C. Dwork, A. Roth | "Deep learning with differential privacy" M. Abadi *et al*

Watch more:

- [On-demand] Synthetic data generation methods - Statice webinar

Learn more

- [On-demand] Statice technical white paper
- Evaluate Statice: book a demo with us

Sources

1. Big Data for Insurance
2. Harnessing the potential of data in insurance - McKinsey
3. Biggest GDPR fines in 2020 - Tessian
4. Generation and evaluation of synthetic patient data
5. Sweeney, Latanya. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine and Ethics, Vol. 25 1997, p. 98-110
6. Dwork C., et al. (2006) Calibrating Noise to Sensitivity in Private Data Analysis
7. M. Abadi et al, (2016) Deep Learning with Differential Privacy
8. Nasr et al. 2021, Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning
9. Article 29 working party, Opinion 05/2014 on Anonymisation Techniques

www.statice.ai

Visit our website for more information

Statice