INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

DECEPTIVELY REAL – YET PRIVACY COMPLIANT?

Synthetic Data with Formal Privacy Guarantees

Dr. Marcel Neunhoeffer

DSTS – Two-day meeting Spring 2026

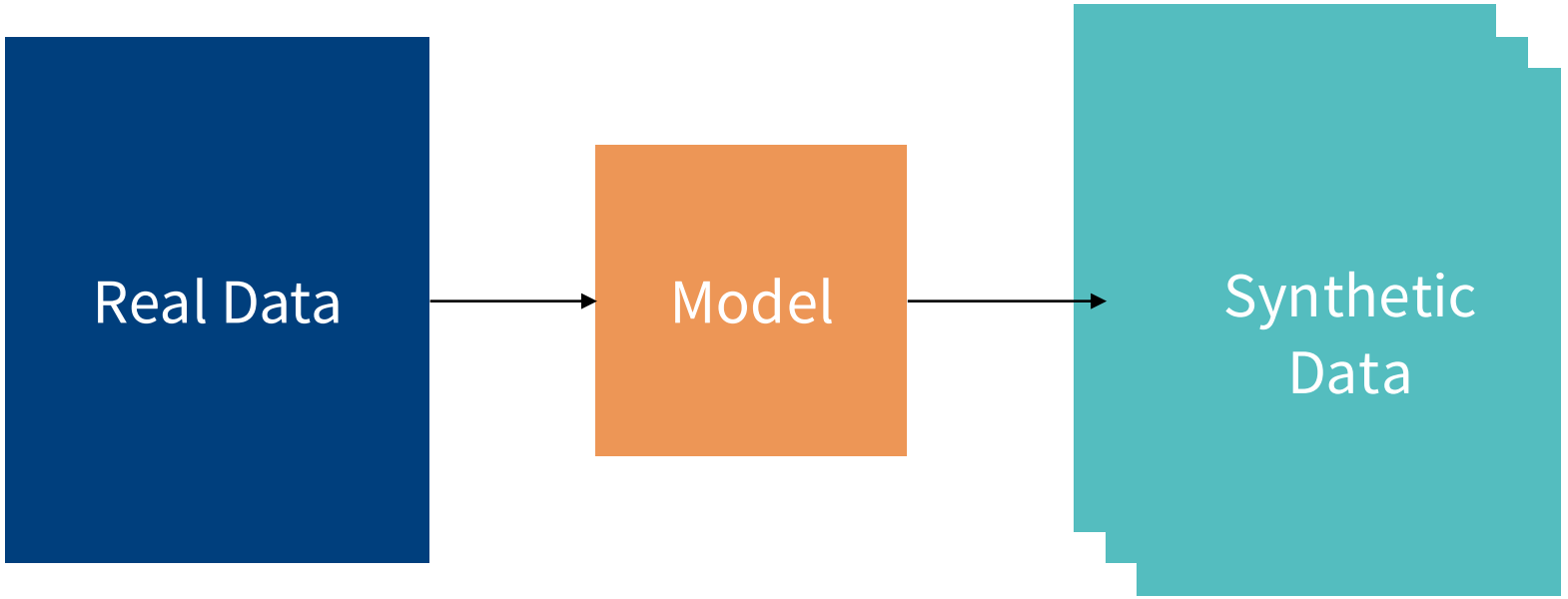
Aalborg, 05 May 2026

SYNTHETIC DATA COULD BE A USEFUL TOOL FOR PRIVACY PROTECTION…

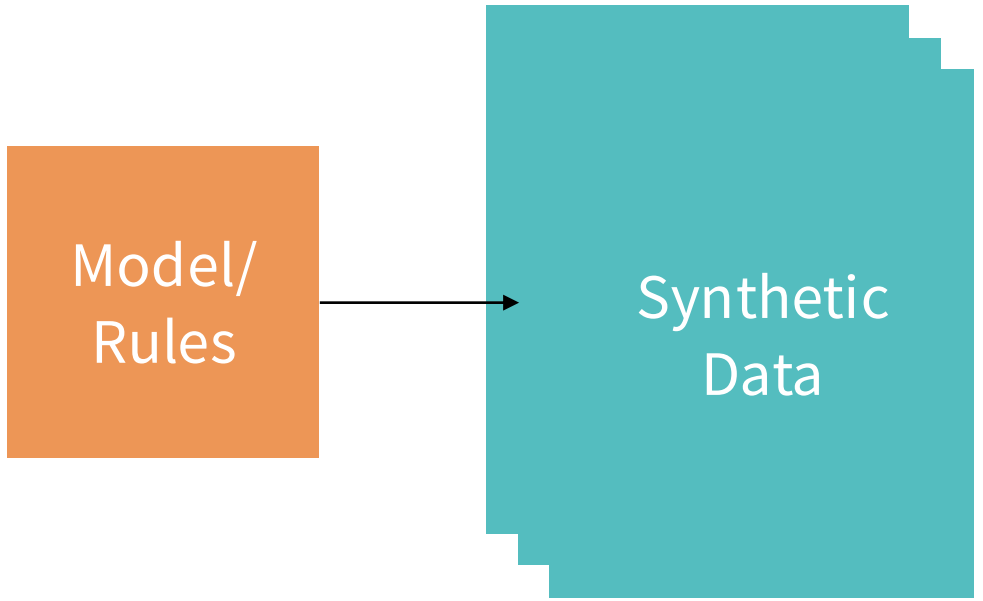
”3. Synthetic data generation: AI systems could be, at least partially, trained using artificially generated training personal data. These synthetic data reflect the real-world data’s statistical properties while not being attributable to an individual. If deemed suitable, this measure should be implemented with due care as it may introduce additional challenges. Thus, if this measure is envisaged, it should be implemented in combination with the additional measures presented above given the possibility of additional attacks (e.g. membership inference attacks).“ (Page 33)

Source: https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2025-11-11-guidance-risk-management-artificial-intelligence-systems_en

WHAT ARE SYNTHETIC DATA?

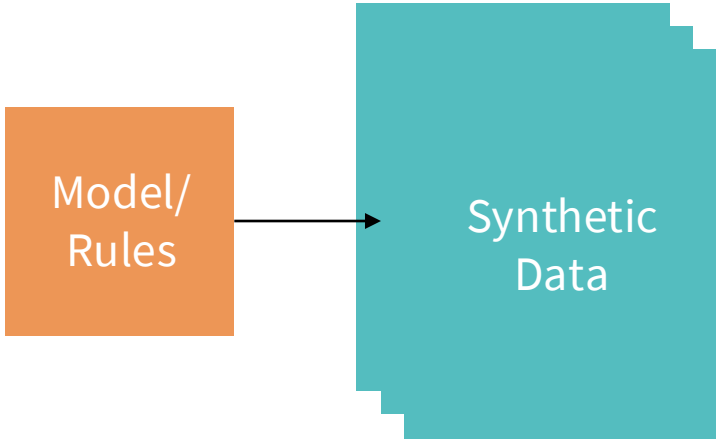


WHAT ARE SYNTHETIC DATA?

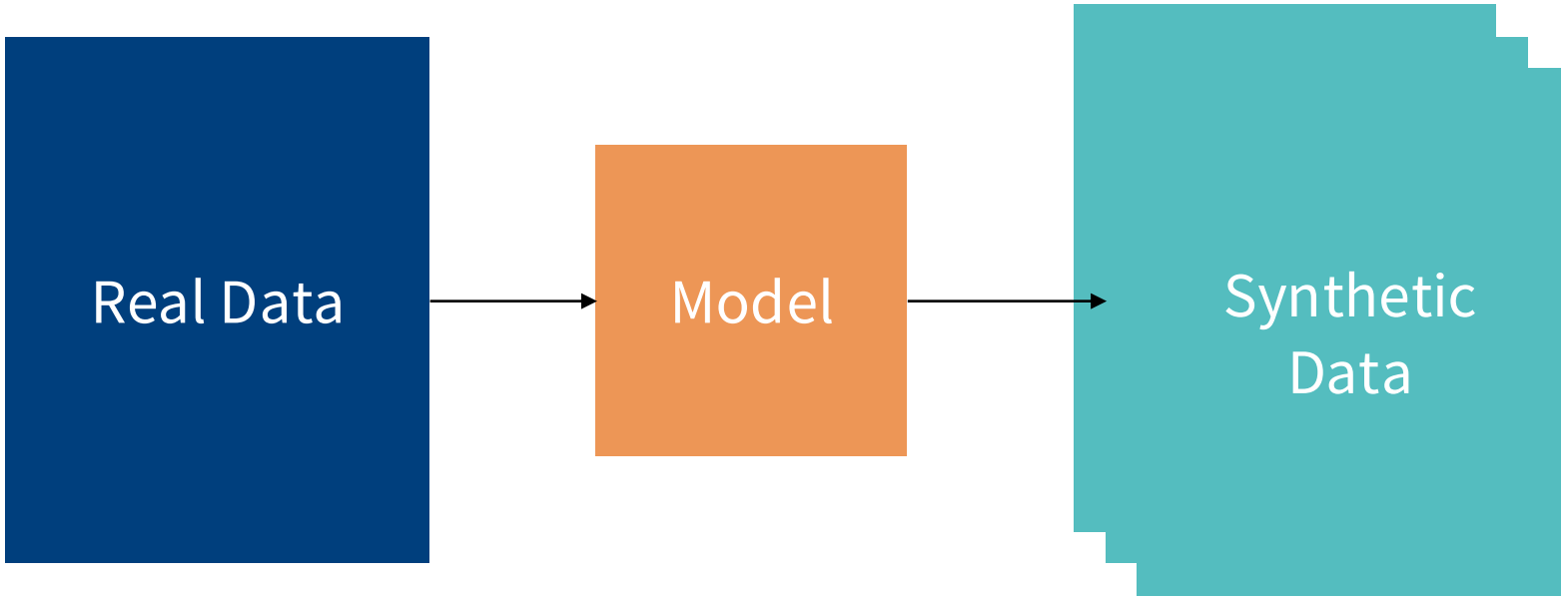


WHAT ARE SYNTHETIC DATA?

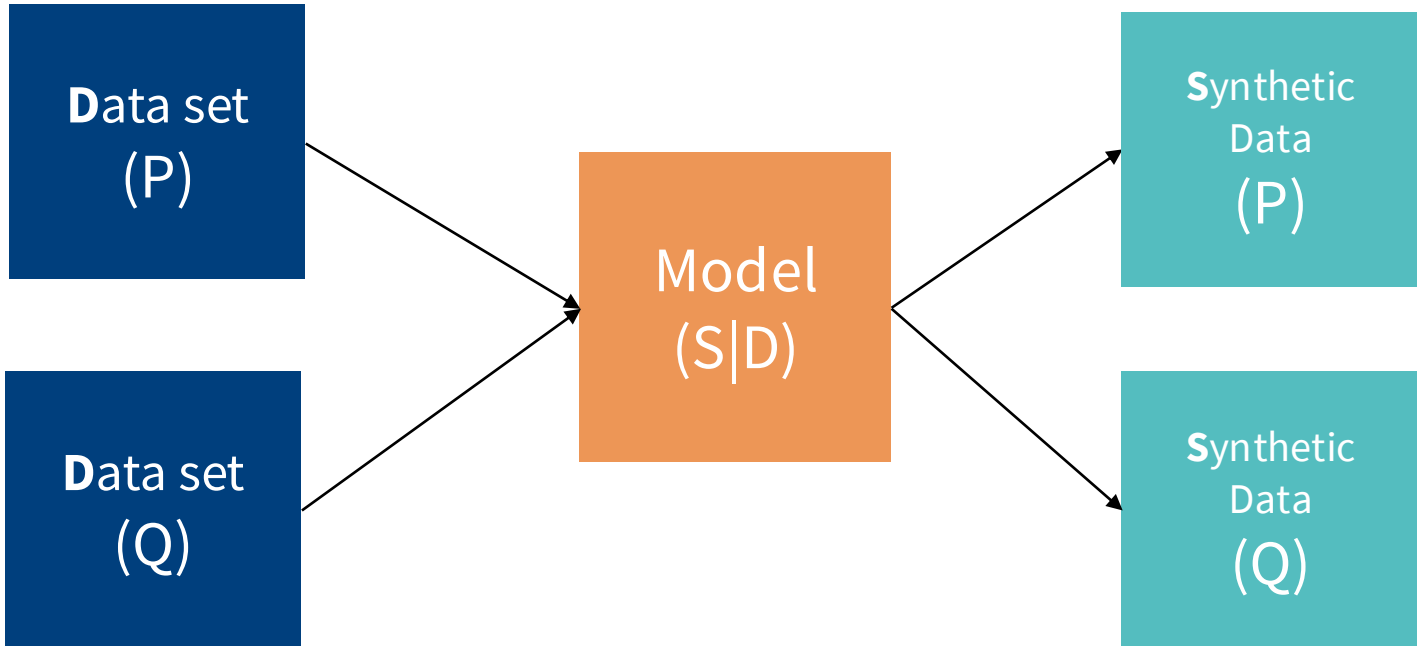
Source: https://synthetichealth.github.io/synthea/



WHAT ARE SYNTHETIC DATA?



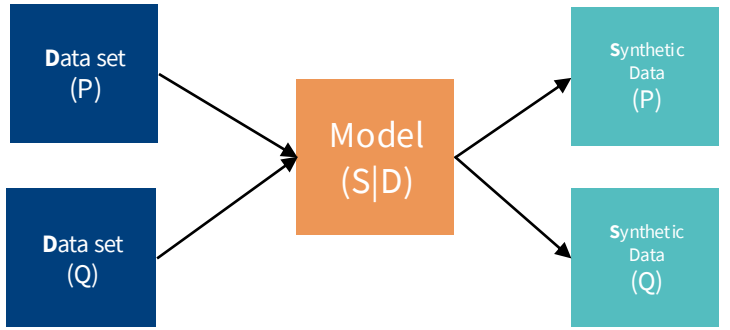
AN INFORMATION THEORETIC VIEW OF SYNTHETIC DATA



$$D_f(P_D||Q_D) \geq D_f(P_S||Q_S)$$

AN INFORMATION THEORETIC VIEW OF SYNTHETIC DATA

- The **statistical utility** of a synthetic data set is higher, the easier it is to distinguish, whether the real data set was P or Q.
- The **privacy protection** of a synthetic data set is higher, the harder it is to distinguish, whether the real data set was P or Q.



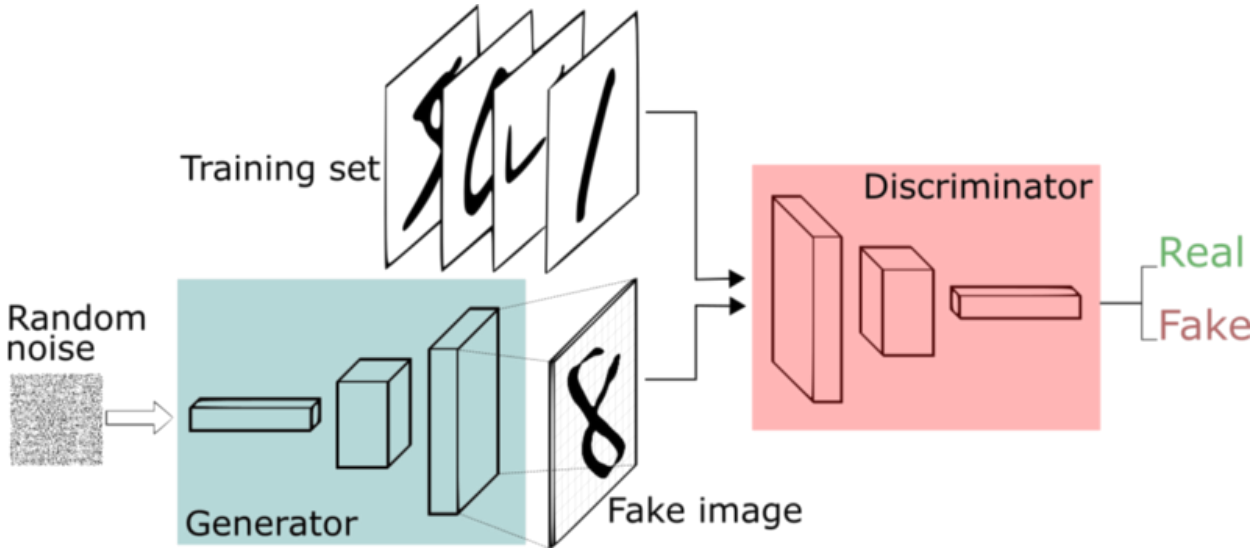
$$D_f(P_D||Q_D) \geq D_f(P_S||Q_S)$$

GENERATIVE AI MODELS AND SYNTHETIC DATA

A VERY BRIEF INTRODUCTION TO GENERATIVE AI MODELS

- **The central promise:** A generative AI model learns the underlying distribution of the real data and then samples new synthetic data from the underlying distribution (mainly images and text).

- Today, various generative AI models are used, such as those based on
  - Generative Pre-TrainedTransformers (GPT) (Radford et al., 2018),
  - Diffusion models (Sohl-Dickstein et al., 2015) or
  - **Generative Adversarial Nets** (GAN) (Goodfellow et al., 2014)

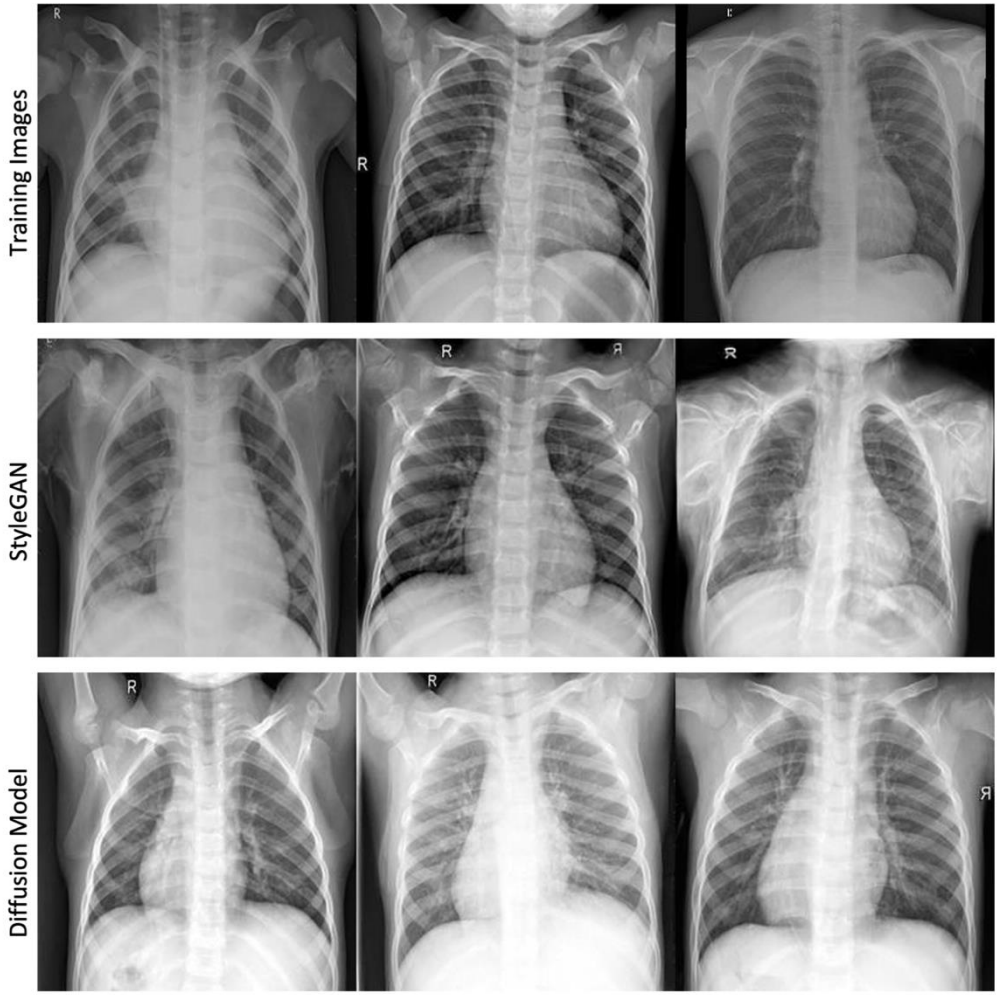


Source: https://www.freecodecamp.org/news/an-intuitive-introduction-to-generative-adversarial-networks-gans-7a2264a81394

GENERATIVE MODELS CAN PRODUCE STRIKINGLY REALISTIC DATA

Source: Karras et al. 2019

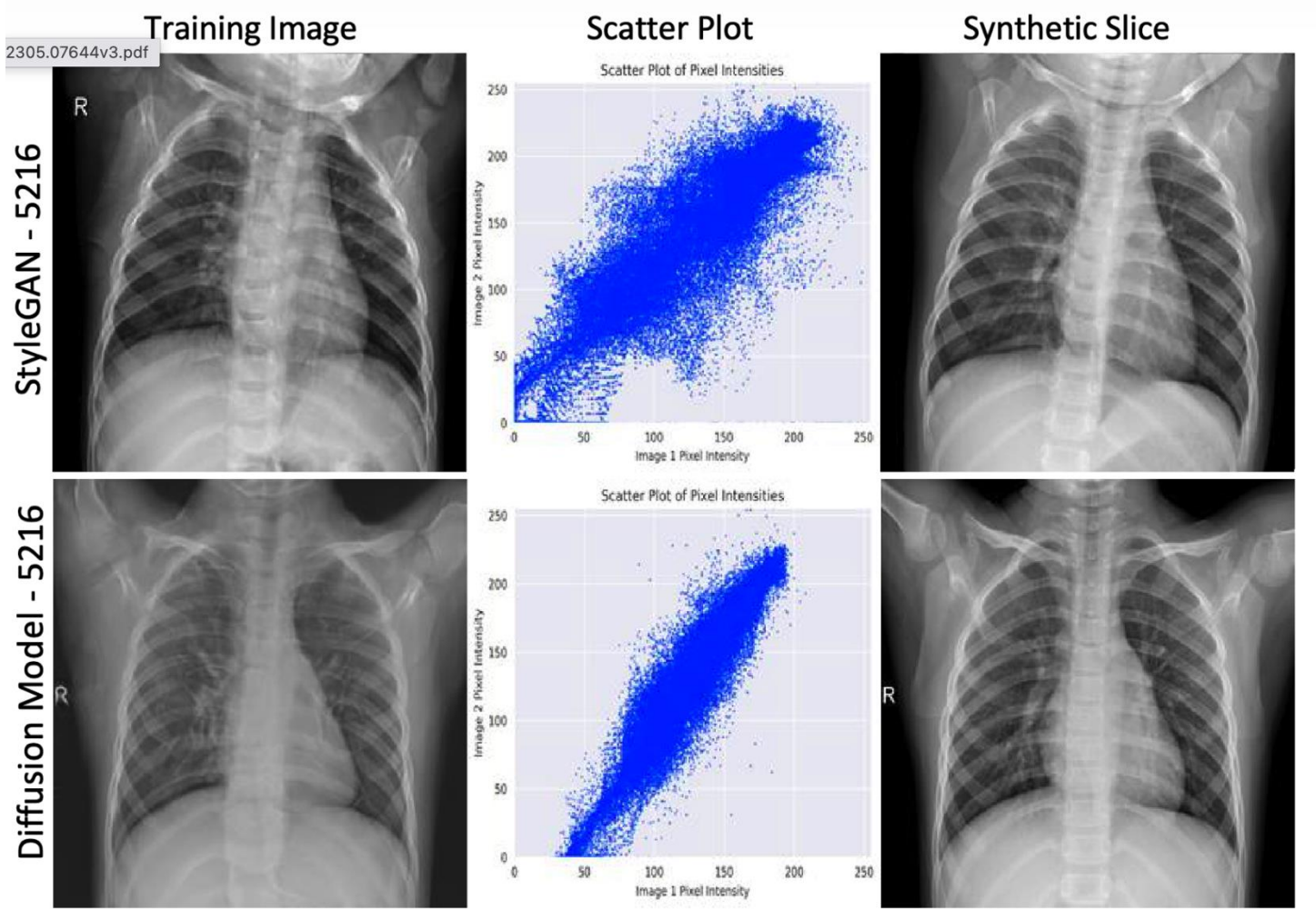
GENERATIVE MODELS CAN PRODUCE STRIKINGLY REALISTIC DATA



Source: Akbar et al. 2024

ARE SYNTHETIC DATA PRIVACY-PRESERVING?

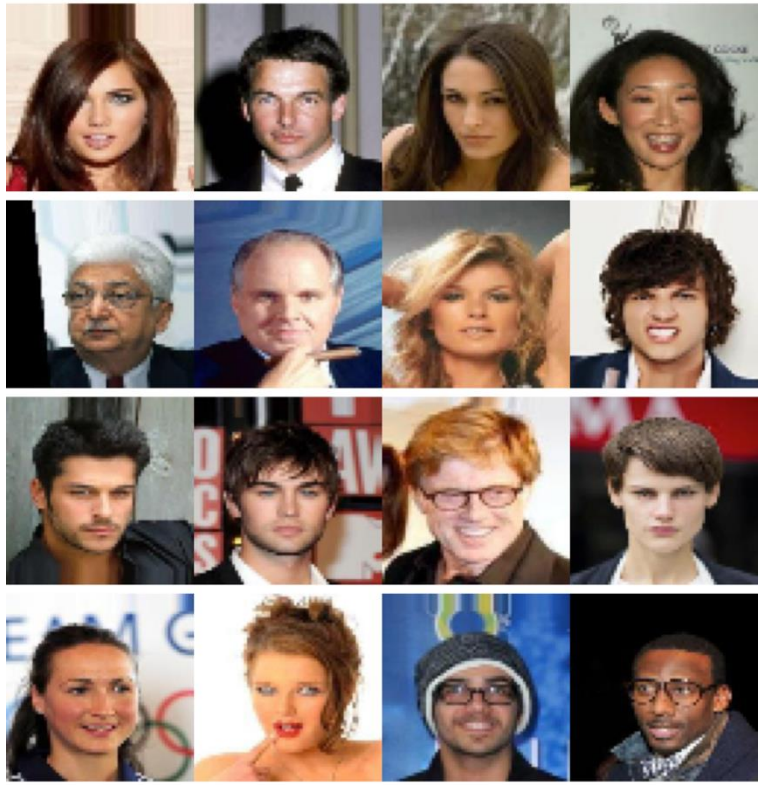
GENERATIVE MODELS CAN MEMORIZE REAL DATA



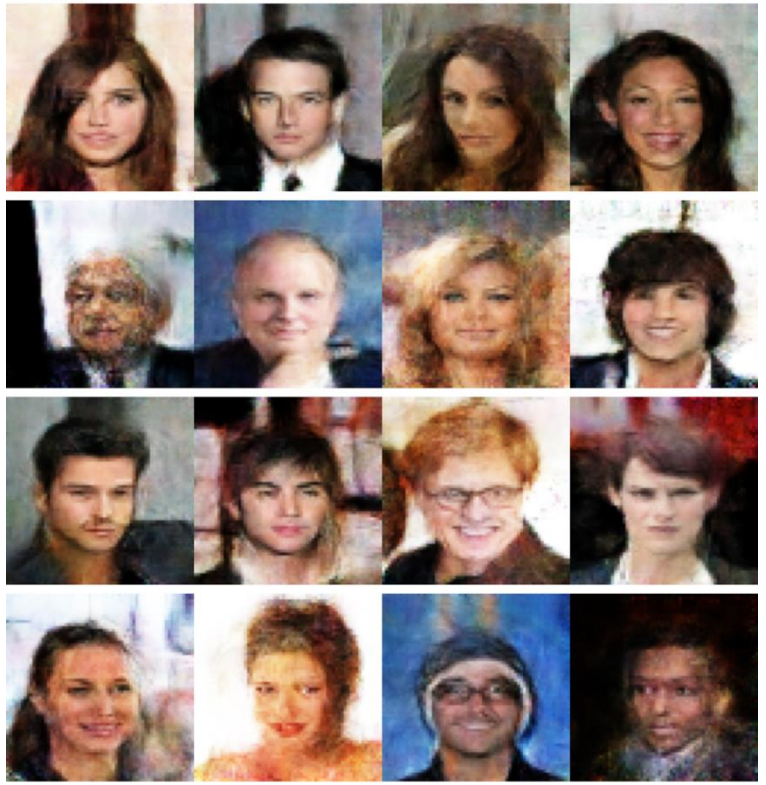
Source: Akbar et al. 2024

GENERATIVE MODELS CAN MEMORIZE REAL DATA

Images from the Training Data



Reconstructed generated Images



SYNTHETIC DATA COULD BE A USEFUL TOOL FOR PRIVACY PROTECTION…

„3. Synthetic data generation: AI systems could be, at least partially, trained using artificially generated training personal data. These synthetic data reflect the real-world data's statistical properties while not being attributable to an individual. If deemed suitable, this measure **should be implemented with due care as it may introduce additional challenges**. Thus, if this measure is envisaged, it should be implemented in combination with the additional measures presented above given the possibility of additional attacks (e.g. membership inference attacks).“ (Page 33)

Source: https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2025-11-11-guidance-risk-management-artificial-intelligence-systems_en

SYNTHETIC DATA COULD BE A USEFUL TOOL FOR PRIVACY PROTECTION…

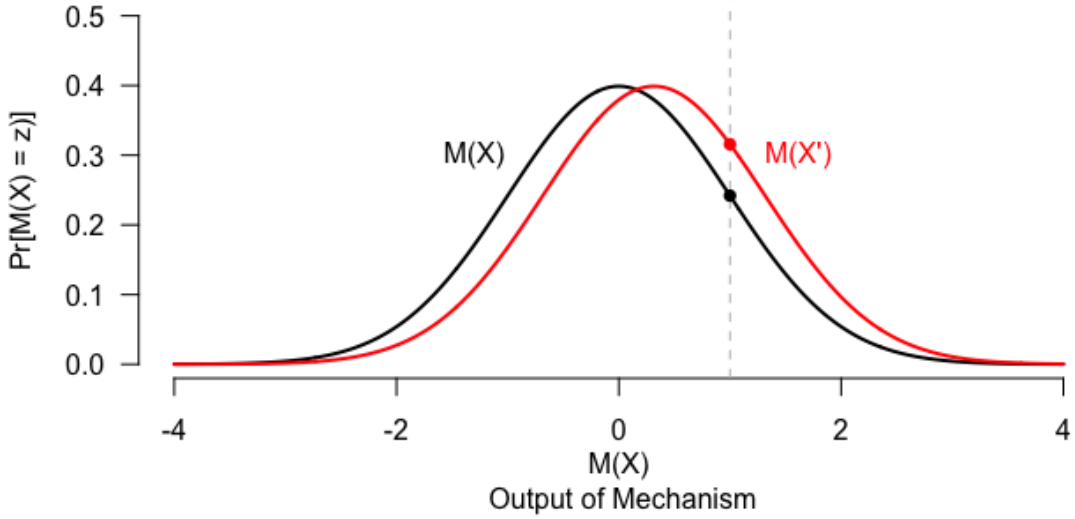
„3. Synthetic data generation: AI systems could be, at least partially, trained using artificially generated training personal data. These synthetic data reflect the real-world data's statistical properties while not being attributable to an individual. If deemed suitable, this measure should be implemented with due care as it may introduce additional challenges. Thus, if this measure is envisaged, it **should be implemented in combination with the additional measures presented above** given the possibility of additional attacks (e.g. membership inference attacks).“ (Page 33)

Source: https://www.edps.europa.eu/data-protection/our-work/publications/guidelines/2025-11-11-guidance-risk-management-artificial-intelligence-systems_en

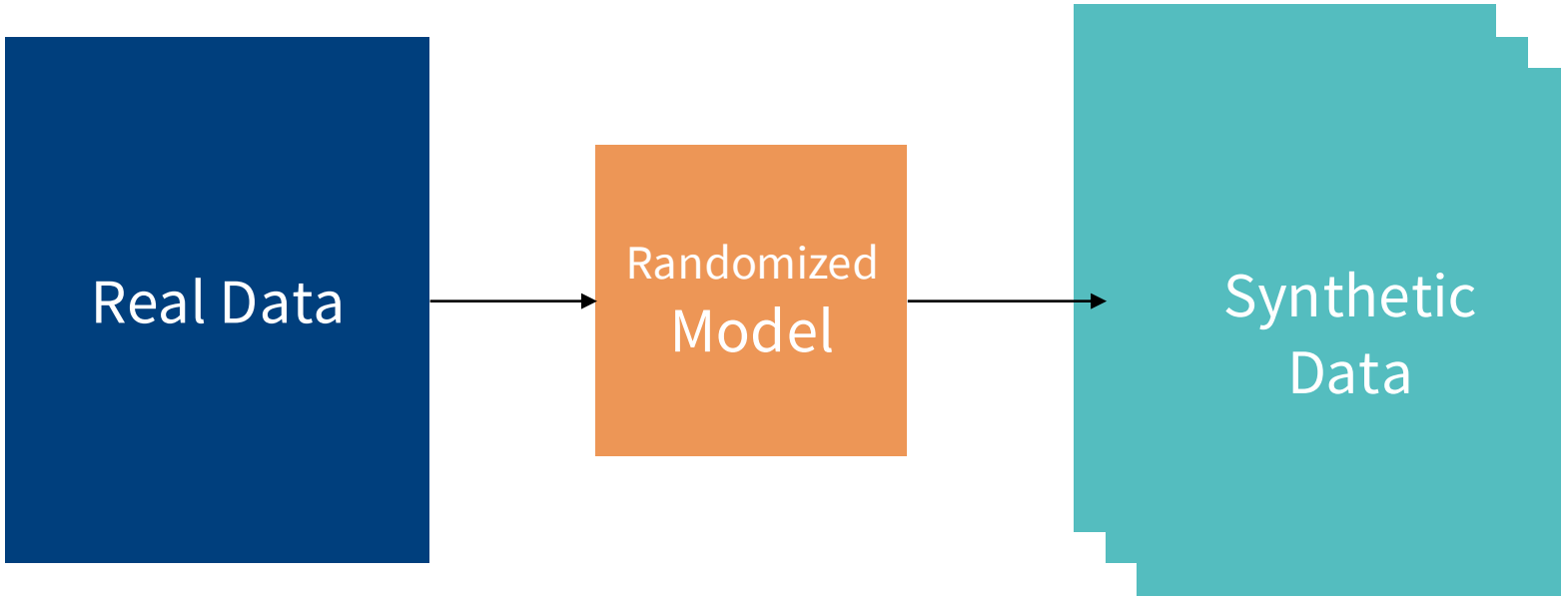
WHAT ARE FORMAL PRIVACY GUARANTEES?

- A mathematically formal privacy guarantee is known as Differential Privacy (Dwork et al., 2006).

- "A randomized algorithm M is differentially private if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$, the random variables $M(X)$ and $M(X')$ are similarly distributed."



WHAT ARE SYNTHETIC DATA?

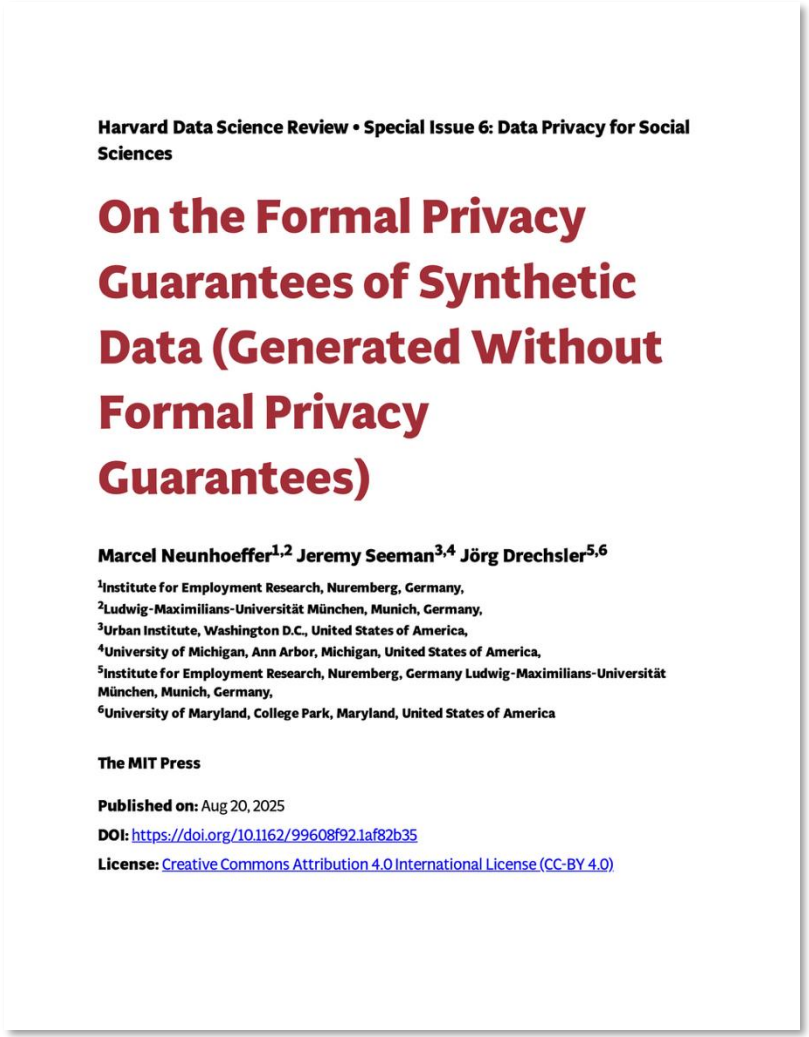


SYNTHETIC DATA AND FORMAL PRIVACY GUARANTEES

Our research question:

(**Neunhoeffer**, Seeman & Drechsler, HDSR 2026)

- Can synthetic data that was generated without formal privacy guarantees still satisfy some form of privacy guarantees?



Harvard Data Science Review • Special Issue 6: Data Privacy for Social Sciences

On the Formal Privacy Guarantees of Synthetic Data (Generated Without Formal Privacy Guarantees)

Marcel Neunhoeffer[1,2] Jeremy Seeman[3,4] Jörg Drechsler[5,6]

[1]Institute for Employment Research, Nuremberg, Germany,
[2]Ludwig-Maximilians-Universität München, Munich, Germany,
[3]Urban Institute, Washington D.C., United States of America,
[4]University of Michigan, Ann Arbor, Michigan, United States of America,
[5]Institute for Employment Research, Nuremberg, Germany Ludwig-Maximilians-Universität München, Munich, Germany,
[6]University of Maryland, College Park, Maryland, United States of America

The MIT Press

Published on: Aug 20, 2025

DOI: https://doi.org/10.1162/99608f92.1af82b35

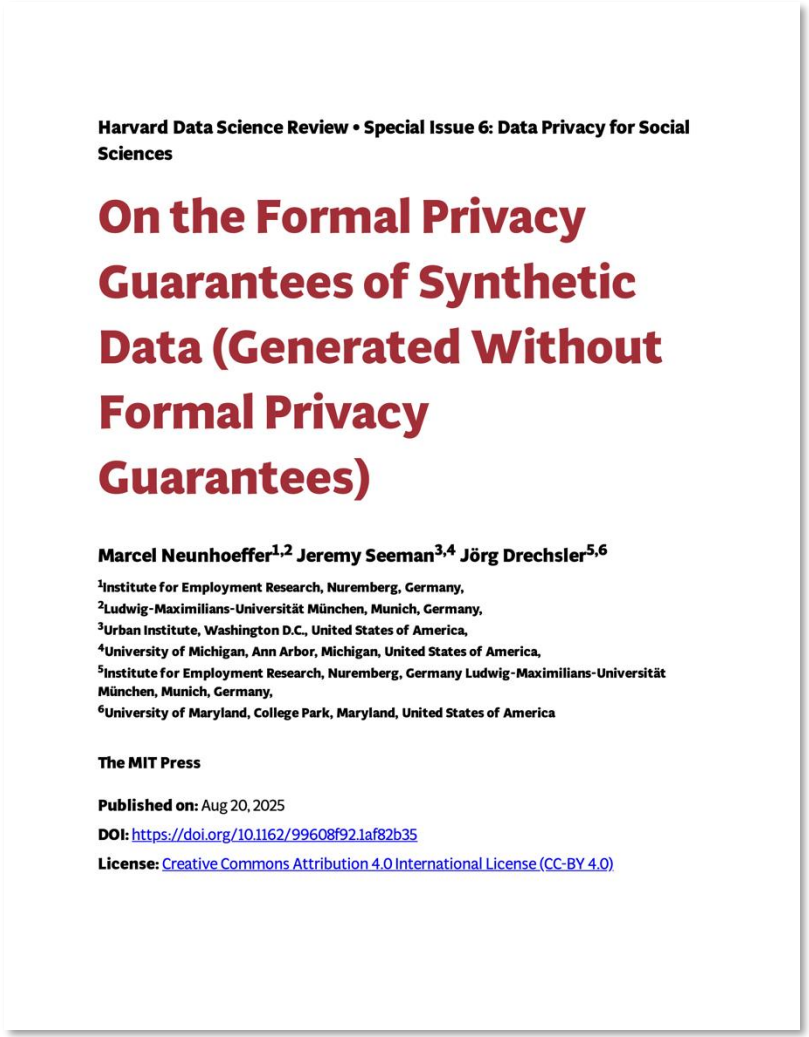
SYNTHETIC DATA AND FORMAL PRIVACY GUARANTEES

Our results:

(**Neunhoeffer**, Seeman & Drechsler, HDSR 2026)

- We'll start with a very simple case: We generate synthetic data from a sample of a univariate normal distribution with known variance using a model for normal synthetic data.

- In this case, we can show that the generation of the synthetic data corresponds to a randomized mechanism with a formal ρ-zCDP (Bun & Steinke, 2016) guarantee.



Harvard Data Science Review • Special Issue 6: Data Privacy for Social Sciences

On the Formal Privacy Guarantees of Synthetic Data (Generated Without Formal Privacy Guarantees)

Marcel Neunhoeffer[1,2] Jeremy Seeman[3,4] Jörg Drechsler[5,6]

[1]Institute for Employment Research, Nuremberg, Germany,
[2]Ludwig-Maximilians-Universität München, Munich, Germany,
[3]Urban Institute, Washington D.C., United States of America,
[4]University of Michigan, Ann Arbor, Michigan, United States of America,
[5]Institute for Employment Research, Nuremberg, Germany Ludwig-Maximilians-Universität München, Munich, Germany,
[6]University of Maryland, College Park, Maryland, United States of America

The MIT Press

Published on: Aug 20, 2025

DOI: https://doi.org/10.1162/99608f92.1af82b35

SYNTHETIC DATA AND FORMAL PRIVACY GUARANTEES

Additive noise

$$\tilde{x} \sim \bar{x} + \mathcal{N}(0, \frac{\Delta^2}{2\rho})$$

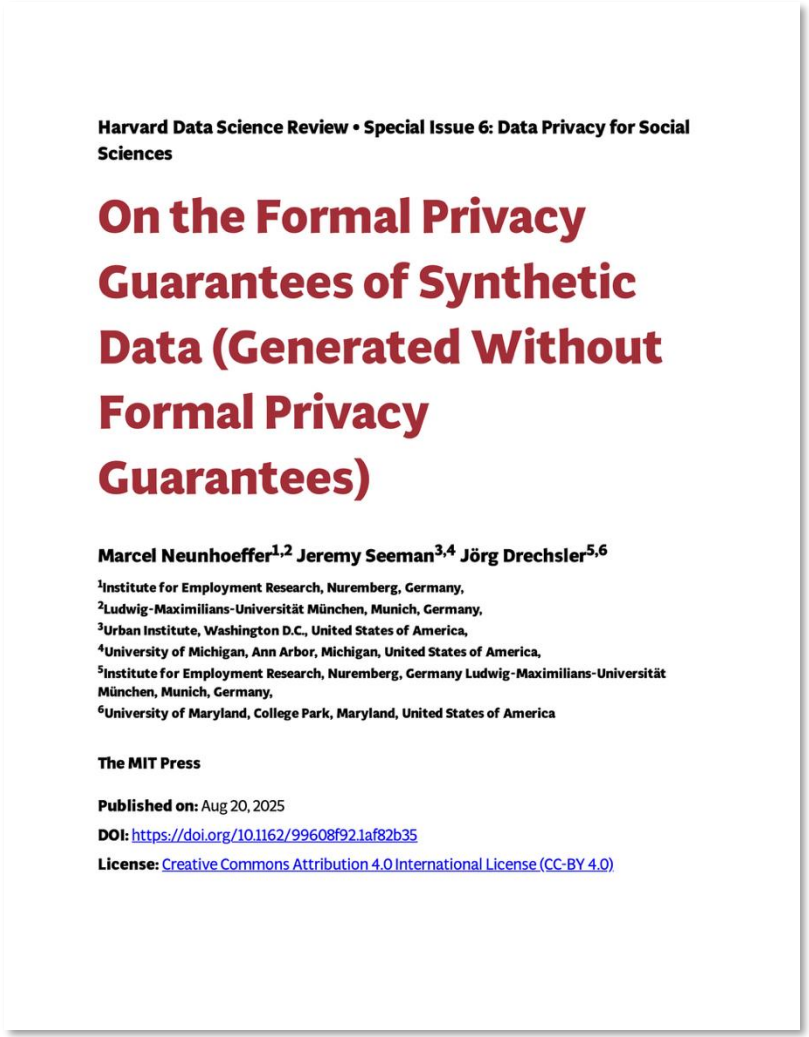
with $\Delta = \frac{|2d|}{n_{orig}}$.

Synthetic Data

$$\tilde{x} \sim \bar{x} + \mathcal{N}(0, \frac{\sigma^2}{n_{syn}})$$

Then the number of synthetic data points n_{synth} can be chosen such that the synthetic data satisfies a formal ρ-zCDP guarantee:

$$n_{syn} = \lfloor \frac{n_{orig}^2 \rho \sigma^2}{2d^2 m} \rfloor$$



Harvard Data Science Review • Special Issue 6: Data Privacy for Social Sciences

On the Formal Privacy Guarantees of Synthetic Data (Generated Without Formal Privacy Guarantees)

Marcel Neunhoeffer[1,2] Jeremy Seeman[3,4] Jörg Drechsler[5,6]

[1]Institute for Employment Research, Nuremberg, Germany,
[2]Ludwig-Maximilians-Universität München, Munich, Germany,
[3]Urban Institute, Washington D.C., United States of America,
[4]University of Michigan, Ann Arbor, Michigan, United States of America,
[5]Institute for Employment Research, Nuremberg, Germany Ludwig-Maximilians-Universität München, Munich, Germany,
[6]University of Maryland, College Park, Maryland, United States of America

The MIT Press

Published on: Aug 20, 2025

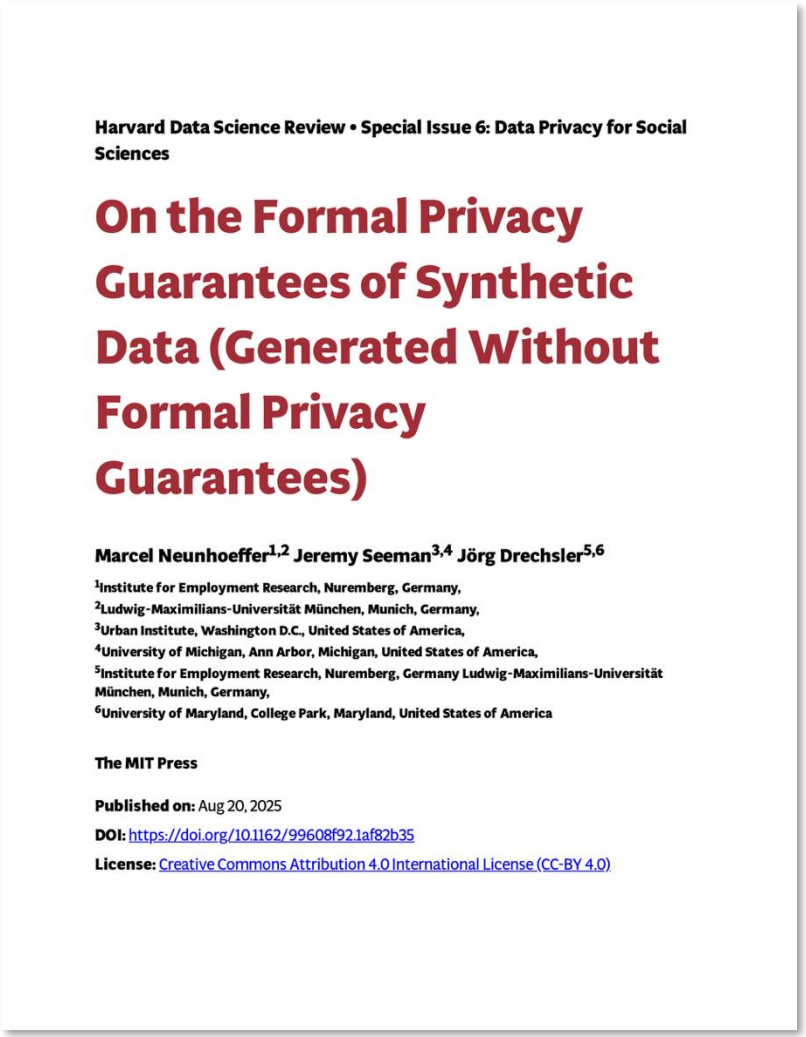
DOI: https://doi.org/10.1162/99608f92.1af82b35

SYNTHETIC DATA AND FORMAL PRIVACY GUARANTEES

- **Two observations:**

1. To meet a formal data protection guarantee, the sensitivity of the sufficient statistic (Δ^2) must be limited.

2. The number of generated synthetic data points determines the level of protection. The more artificial data points are published, the weaker the protection becomes.



Harvard Data Science Review • Special Issue 6: Data Privacy for Social Sciences

On the Formal Privacy Guarantees of Synthetic Data (Generated Without Formal Privacy Guarantees)

Marcel Neunhoeffer[1,2] Jeremy Seeman[3,4] Jörg Drechsler[5,6]

[1]Institute for Employment Research, Nuremberg, Germany,
[2]Ludwig-Maximilians-Universität München, Munich, Germany,
[3]Urban Institute, Washington D.C., United States of America,
[4]University of Michigan, Ann Arbor, Michigan, United States of America,
[5]Institute for Employment Research, Nuremberg, Germany Ludwig-Maximilians-Universität München, Munich, Germany,
[6]University of Maryland, College Park, Maryland, United States of America

The MIT Press

Published on: Aug 20, 2025

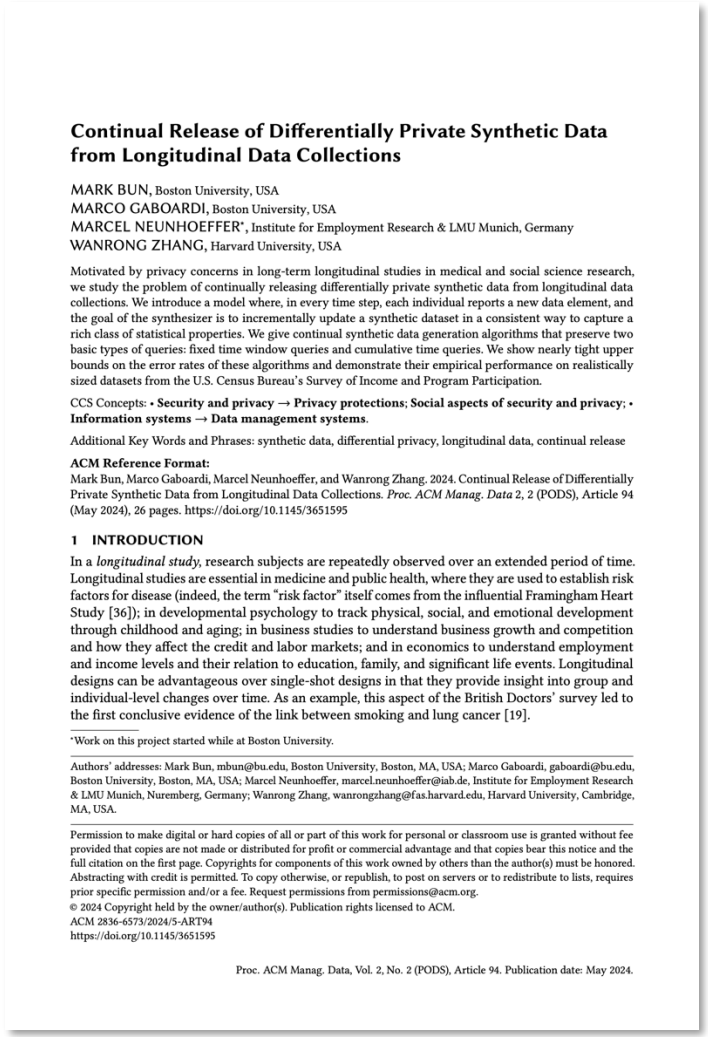
DOI: https://doi.org/10.1162/99608f92.1af82b35

SYNTHETIC DATA AND FORMAL PRIVACY GUARANTEES

Our research question:

(Bun, Gaboardi, **Neunhoeffer** & Zhang, PODS 2024)

- How can we release differentially private synthetic data from longitudinal studies (where individuals are repeatedly observed over time)?
- We need to maintain privacy while preserving individual-level trends and longitudinal consistency across continual data releases
- Key challenge: synthetic individuals must persist over time, with records updated incrementally



Continual Release of Differentially Private Synthetic Data from Longitudinal Data Collections

MARK BUN, Boston University, USA
MARCO GABOARDI, Boston University, USA
MARCEL NEUNHOEFFER*, Institute for Employment Research & LMU Munich, Germany
WANRONG ZHANG, Harvard University, USA

Motivated by privacy concerns in long-term longitudinal studies in medical and social science research, we study the problem of continually releasing differentially private synthetic data from longitudinal data collections. We introduce a model where, in every time step, each individual reports a new data element, and the goal of the synthesizer is to incrementally update a synthetic dataset in a consistent way to capture a rich class of statistical properties. We give continual synthetic data generation algorithms that preserve two basic types of queries: fixed time window queries and cumulative time queries. We show nearly tight upper bounds on the error rates of these algorithms and demonstrate their empirical performance on realistically sized datasets from the U.S. Census Bureau's Survey of Income and Program Participation.

CCS Concepts: • **Security and privacy → Privacy protections**; **Social aspects of security and privacy**; • **Information systems → Data management systems**.

Additional Key Words and Phrases: synthetic data, differential privacy, longitudinal data, continual release

ACM Reference Format:
Mark Bun, Marco Gaboardi, Marcel Neunhoeffer, and Wanrong Zhang. 2024. Continual Release of Differentially Private Synthetic Data from Longitudinal Data Collections. *Proc. ACM Manag. Data* 2, 2 (PODS), Article 94 (May 2024), 26 pages. https://doi.org/10.1145/3651595

1 INTRODUCTION

In a *longitudinal study*, research subjects are repeatedly observed over an extended period of time. Longitudinal studies are essential in medicine and public health, where they are used to establish risk factors for disease (indeed, the term "risk factor" itself comes from the influential Framingham Heart Study [36]); in developmental psychology to track physical, social, and emotional development through childhood and aging; in business studies to understand business growth and competition and how they affect the credit and labor markets; and in economics to understand employment and income levels and their relation to education, family, and significant life events. Longitudinal designs can be advantageous over single-shot designs in that they provide insight into group and individual-level changes over time. As an example, this aspect of the British Doctors' survey led to the first conclusive evidence of the link between smoking and lung cancer [19].

*Work on this project started while at Boston University.

Authors' addresses: Mark Bun, mbun@bu.edu, Boston University, Boston, MA, USA; Marco Gaboardi, gaboardi@bu.edu, Boston University, Boston, MA, USA; Marcel Neunhoeffer, marcel.neunhoeffer@iab.de, Institute for Employment Research & LMU Munich, Nuremberg, Germany; Wanrong Zhang, wanrongzhang@fas.harvard.edu, Harvard University, Cambridge, MA, USA.

ACM 2836-6573/2024/5-ART94
https://doi.org/10.1145/3651595

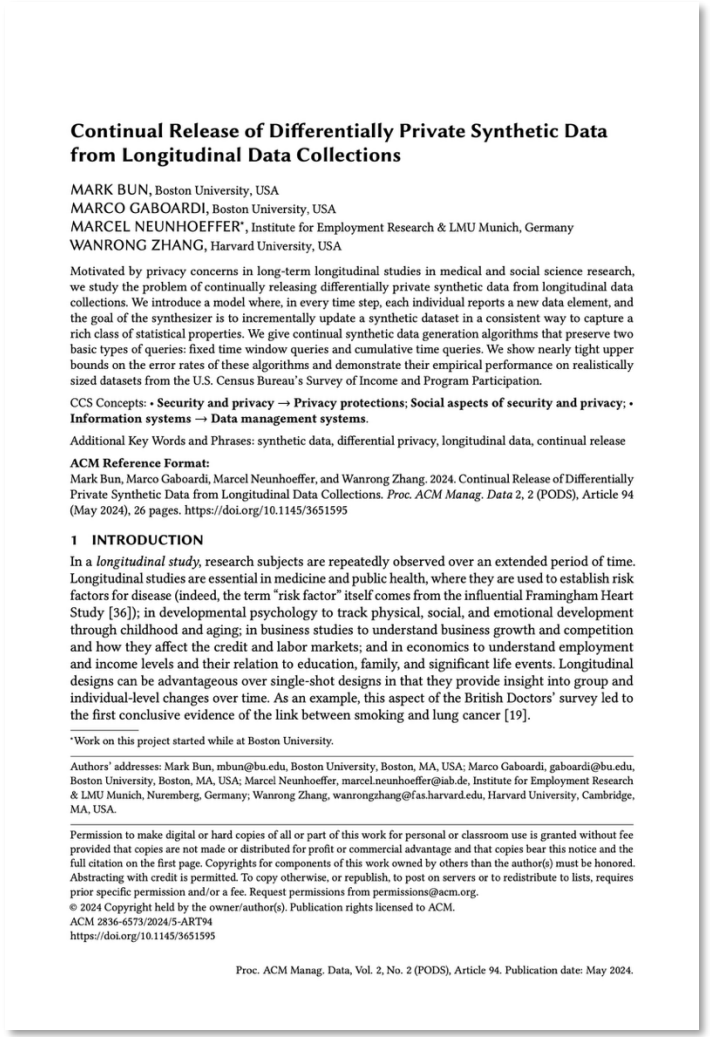
Proc. ACM Manag. Data, Vol. 2, No. 2 (PODS), Article 94. Publication date: May 2024.

SYNTHETIC DATA AND FORMAL PRIVACY GUARANTEES

Our results:

- Two algorithms for two query classes: fixed time window queries (k-month windows) and cumulative time queries (Hamming weight tracking)

- Nearly tight error bounds: $O(\sqrt{(kT)}/n)$ for windows and $O(\sqrt{T}/n)$ for cumulative, matching single-shot lower bounds despite continual release



Continual Release of Differentially Private Synthetic Data from Longitudinal Data Collections

MARK BUN, Boston University, USA
MARCO GABOARDI, Boston University, USA
MARCEL NEUNHOEFFER*, Institute for Employment Research & LMU Munich, Germany
WANRONG ZHANG, Harvard University, USA

Motivated by privacy concerns in long-term longitudinal studies in medical and social science research, we study the problem of continually releasing differentially private synthetic data from longitudinal data collections. We introduce a model where, in every time step, each individual reports a new data element, and the goal of the synthesizer is to incrementally update a synthetic dataset in a consistent way to capture a rich class of statistical properties. We give continual synthetic data generation algorithms that preserve two basic types of queries: fixed time window queries and cumulative time queries. We show nearly tight upper bounds on the error rates of these algorithms and demonstrate their empirical performance on realistically sized datasets from the U.S. Census Bureau's Survey of Income and Program Participation.

CCS Concepts: • **Security and privacy → Privacy protections**; **Social aspects of security and privacy**; • **Information systems → Data management systems**.

Additional Key Words and Phrases: synthetic data, differential privacy, longitudinal data, continual release

ACM Reference Format:
Mark Bun, Marco Gaboardi, Marcel Neunhoeffer, and Wanrong Zhang. 2024. Continual Release of Differentially Private Synthetic Data from Longitudinal Data Collections. *Proc. ACM Manag. Data* 2, 2 (PODS), Article 94 (May 2024), 26 pages. https://doi.org/10.1145/3651595

1 INTRODUCTION

In a *longitudinal study*, research subjects are repeatedly observed over an extended period of time. Longitudinal studies are essential in medicine and public health, where they are used to establish risk factors for disease (indeed, the term "risk factor" itself comes from the influential Framingham Heart Study [36]); in developmental psychology to track physical, social, and emotional development through childhood and aging; in business studies to understand business growth and competition and how they affect the credit and labor markets; and in economics to understand employment and income levels and their relation to education, family, and significant life events. Longitudinal designs can be advantageous over single-shot designs in that they provide insight into group and individual-level changes over time. As an example, this aspect of the British Doctors' survey led to the first conclusive evidence of the link between smoking and lung cancer [19].

*Work on this project started while at Boston University.

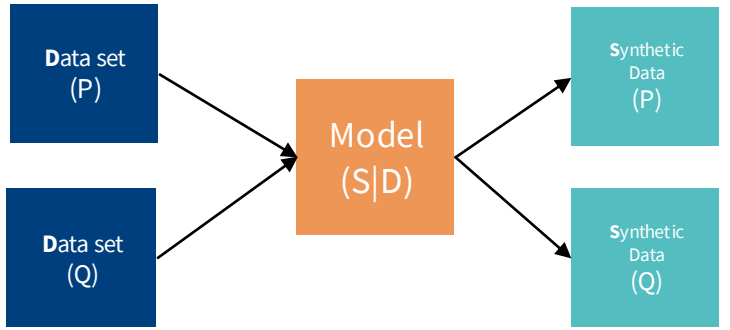
Authors' addresses: Mark Bun, mbun@bu.edu, Boston University, Boston, MA, USA; Marco Gaboardi, gaboardi@bu.edu, Boston University, Boston, MA, USA; Marcel Neunhoeffer, marcel.neunhoeffer@iab.de, Institute for Employment Research & LMU Munich, Nuremberg, Germany; Wanrong Zhang, wanrongzhang@fas.harvard.edu, Harvard University, Cambridge, MA, USA.

ACM 2836-6573/2024/5-ART94
https://doi.org/10.1145/3651595

DECEPTIVELY REAL AND PRIVACY PRESERVING?

SYNTHETIC DATA THAT IS BOTH DECEPTIVELY REAL AND PRIVACY-PRESERVING HAS SO FAR BEEN DIFFICULT TO PRODUCE…

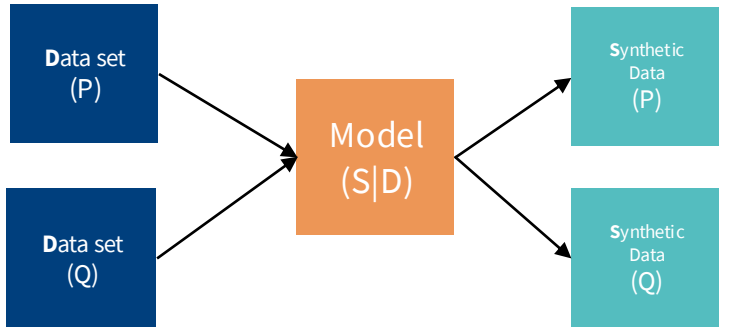
- Generative AI models can generate deceptively real synthetic data.
- Synthetic data is not automatically privacy preserving.



$$D_f(P_D||Q_D) \geq D_f(P_S||Q_S)$$

SYNTHETIC DATA THAT IS BOTH DECEPTIVELY REAL AND PRIVACY-PRESERVING HAS SO FAR BEEN DIFFICULT TO PRODUCE…

- Randomized generative models **can** satisfy formal privacy guarantees under narrow conditions.
- We **can** build generative models with formal privacy guarantees, however, these are only useful for restricted query classes.
- Bridging that gap is the open research frontier: formal guarantees for realistic, high-dimensional synthesizers, and richer longitudinal patterns (spell lengths, multi-attribute joins, cross-temporal queries).



$$D_f(P_D||Q_D) \geq D_f(P_S||Q_S)$$

CONTACT

Dr. Marcel Neunhoeffer
marcel.neunhoeffer@iab.de
http://marcel-neunhoeffer.com/