Grup de Recerca en **Bioestadística** i **Bioinformàtica**

Synthetic data generation and its statistical evaluation

Cecilio Angulo

cecilio.angulo@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL

España | digital 2026

Plan de Recuperación, Transformación y Resiliencia

Grup de Recerca en **Bioestadística** i **Bioinformàtica**

Synthetic data generation (using Machine Learning) and its statistical evaluation (you are the experts)

Cecilio Angulo

cecilio.angulo@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL

España digital 2026

Plan de Recuperación, Transformación y Resiliencia

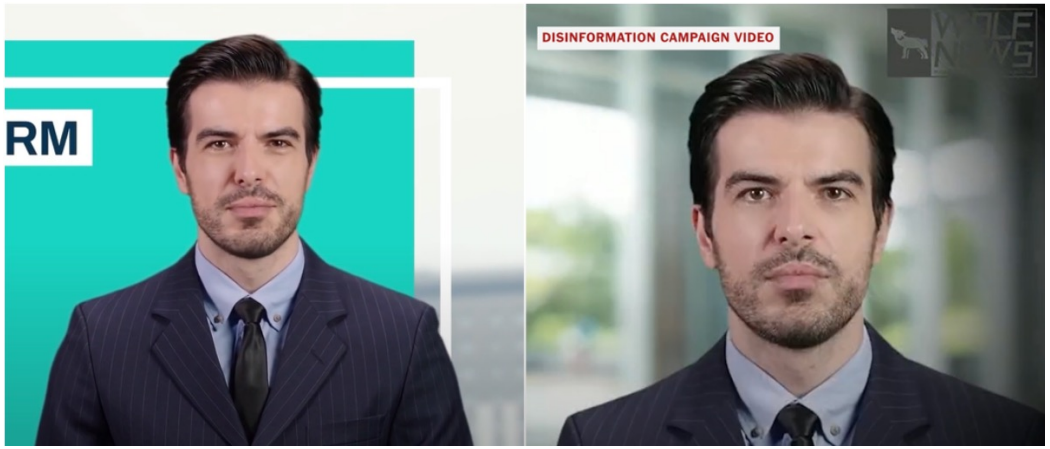
Vull defensar dos missatges

Primer missatge:

La Intel·ligència Artificial NO ha creat, ni popularitzat les fake news!

The New York Times

The People Onscreen Are Fake. The Disinformation Is Real.



Making Deepfakes Gets Cheaper and Easier Thanks to A.I.

Meme-makers and misinformation peddlers are embracing artificial intelligence tools to create convincing fake videos on the cheap.

New software designed to help media detect deepfakes - but it's just a "drop in the bucket"

VOGUE

CULTURE

More and More Women Are Facing the Scary Reality of Deepfakes

New York Times fake news

‘We Are Like Animals’: Inside Greece’s Secret Site for Migrants

The extrajudicial center is one of several tactics Greece is using to prevent a repeat of the 2015 migration crisis.

The New York Times ▪ Satellite image © Maxar Technologies

New York Times crosses limits of fake news, says ‘India blocks naturalization for Muslims.’

by **Akshay Narang** | 10 December 2019 | in **Opinions** | 0


CBS

Vogue magazine slammed for photoshopping disabilities onto able-bodied ambassadors

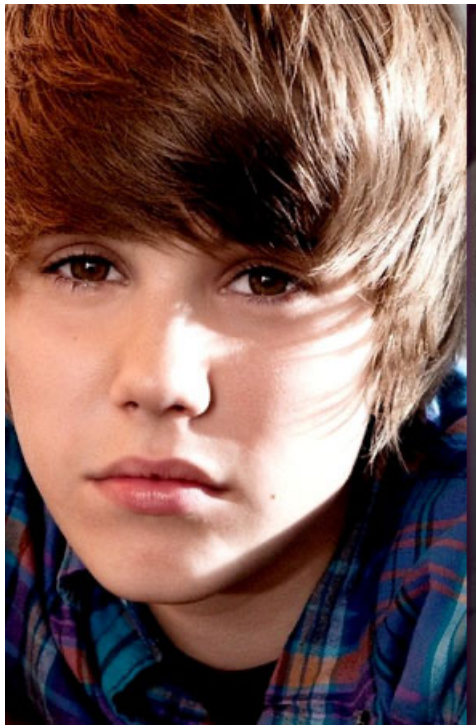
isabelle ⚡
@iswbelle · **Follow**

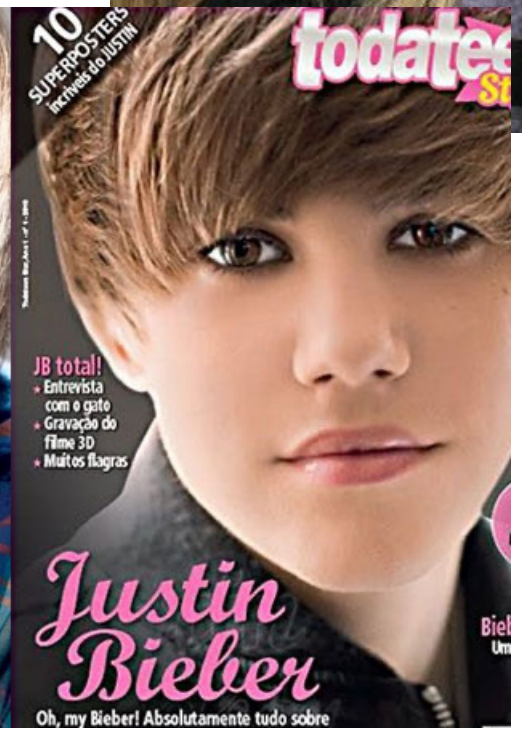
Replying to @voguebrasil and @pires_cleo

Não seria mais fácil chamar um atleta pra fazer a campanha?

2:36 PM · Aug 24, 2016

THE NEWDAILY





Vull defensar dos missatges

Segon missatge:

La Intel·ligència Artificial genera informació sintètica MEGA-útil!

nature

OUTLOOK | 27 April 2023

Synthetic data could be better than real data

Machine-generated data sets have the potential to improve privacy and representation in artificial intelligence, if researchers can find the right balance between accuracy and fakery.

Credit: Janelle Barone

MIT News
ON CAMPUS AND AROUND THE WORLD

In machine learning, synthetic data can offer real performance improvements

Models trained on synthetic data can be more accurate than other models in some cases, which could eliminate some privacy, copyright, and ethical concerns from using real data.

Adam Zewe | MIT News Office
November 3, 2022

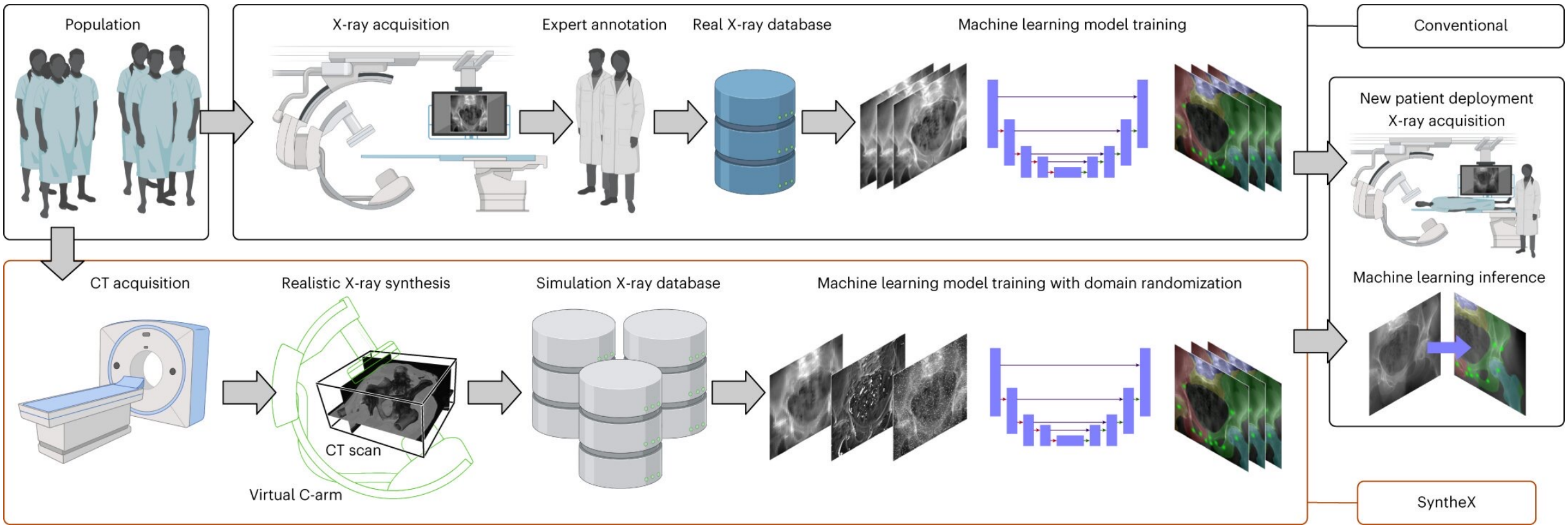
> “Digitally generated data has the same predictive power as real data, as it replicates the statistical characteristics of the existing dataset. It can be generated for unseen conditions and events. Where actual data lacks quality, volume, or variety, synthetic data overcomes these weaknesses, as it is generated for unseen conditions.”

MARCH 21, 2023

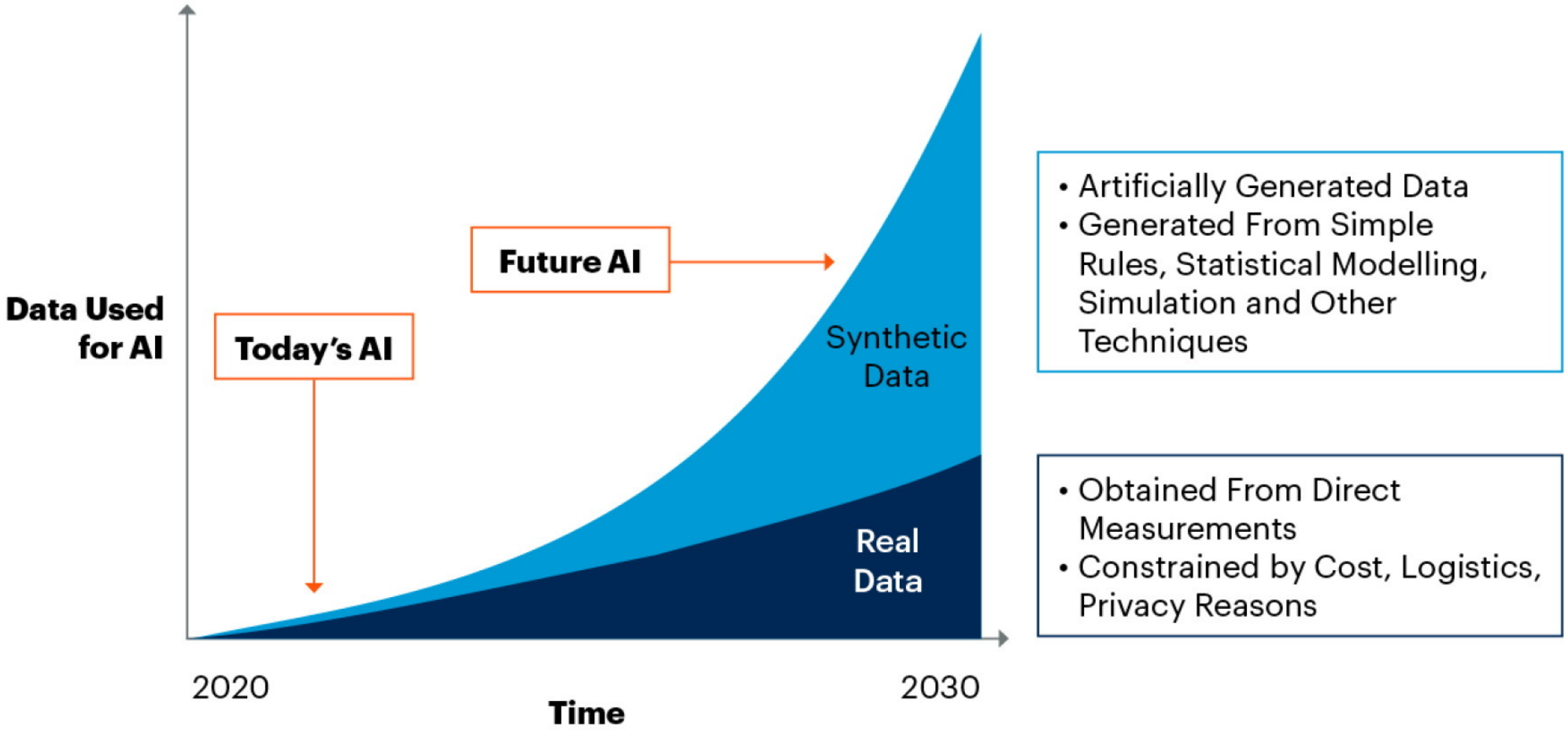
Editors' notes

Synthetic data for AI outperform real data in robot-assisted surgery

by Catherine Graham, Johns Hopkins University



By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



Source: Gartner

750175_C

Gartner

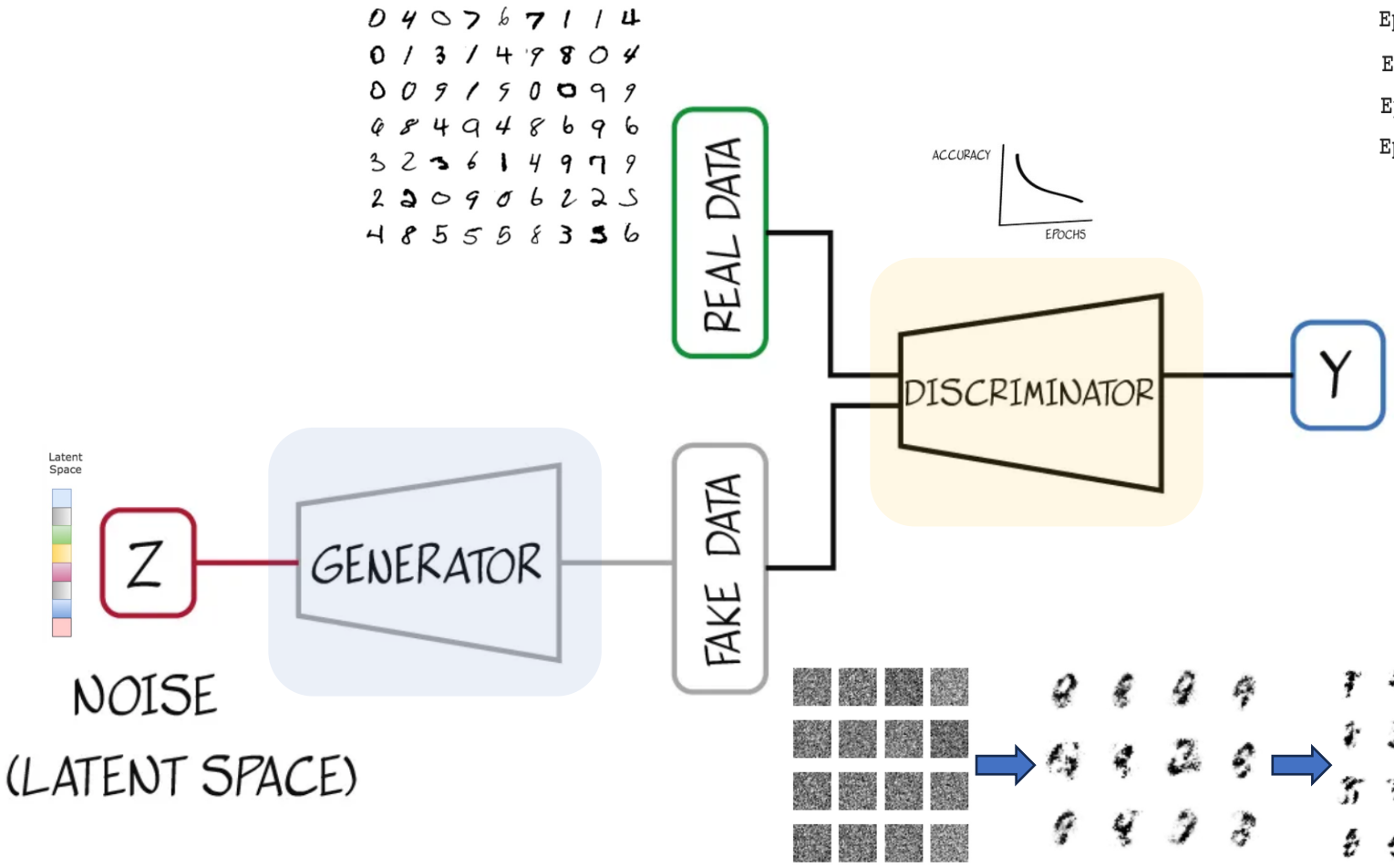
Anem per feina

Researchers have proposed synthetic data (SD) as an alternative to data transformation. Synthetic data is created artificially, possesses the same statistical characteristics as the original, and yet shows better resilience to privacy attacks [12]. SD is required to exhibit the same distribution as well as correlational structure as the original data also known as "realism" or "resemblance". Realistic synthetic data can replace the original data in many applications [12] and can be of utmost significance for medical research where real datasets are unavailable. Synthetic data is being used for testing and evaluation [13], [14], and statistical disclosure control [15]. However, generating synthetic medical data is surrounded by unique challenges because of its inherent complexity and longitudinal nature [16]. There has been a sharp rise in research publications in the field of synthetic medical data generation in the past few years and wider adoption of SD is expected in the future [17]. Researchers are experimenting with different methods to generate realistic synthetic data and have proposed many different quality evaluation metrics to assess its plausibility as a substitute for real data. This has introduced an array of new concepts, terms, techniques, and metrics in the literature. There is a need to consolidate this body of knowledge for better understanding.

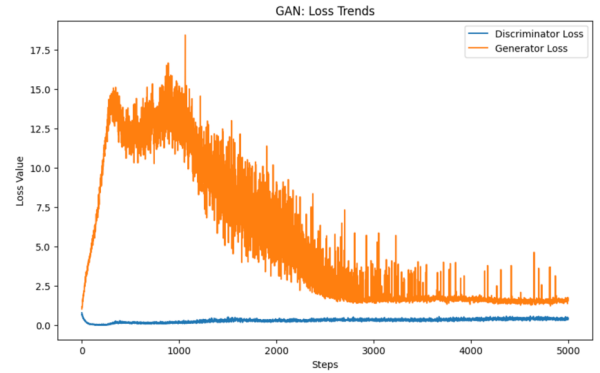
Synthetic data generation: State of the art in health care domain
https://www.sciencedirect.com/science/article/pii/S1574013723000138

Algorismes i mètriques

Generative Adversarial Networks




```
Epoch:0, Step:0, D-Loss:0.791, D-Acc:61.000, G-Loss:0.804
Epoch:19, Step:450, D-Loss:0.479, D-Acc:80.000, G-Loss:1.471
Epoch:39, Step:450, D-Loss:0.603, D-Acc:71.000, G-Loss:1.076
Epoch:59, Step:450, D-Loss:0.629, D-Acc:63.000, G-Loss:0.860
Epoch:109, Step:450, D-Loss:0.671, D-Acc:59.000, G-Loss:0.840
```



Generative Adversarial Networks

A brief technical introduction of the GANs model is as follows: given a real sample ($\mathbf{x}$) and some random noise vector ($\mathbf{z}$), the following terms are defined:

- $D(\mathbf{x})$ is the output of the Discriminator when a real sample $\mathbf{x}$ is processed.
- $G(\mathbf{z})$ is the output of the Generator from the noise $\mathbf{z}$, that is, the synthetic data.
- $D(G(\mathbf{z}))$ is the prediction from the Discriminator on the synthetic data.
- m is the size of samples.
- $P_{\mathbf{x}}$ and $P_{\mathbf{z}}$ are the distribution of real and noise data, respectively.
- $E_{\mathbf{x}}$ and $E_{G(\mathbf{z})}$ are the expected log likelihood from the different outputs of real and generated data.
- θ^D and θ^G are the weights of the Discriminator and Generator model, respectively.

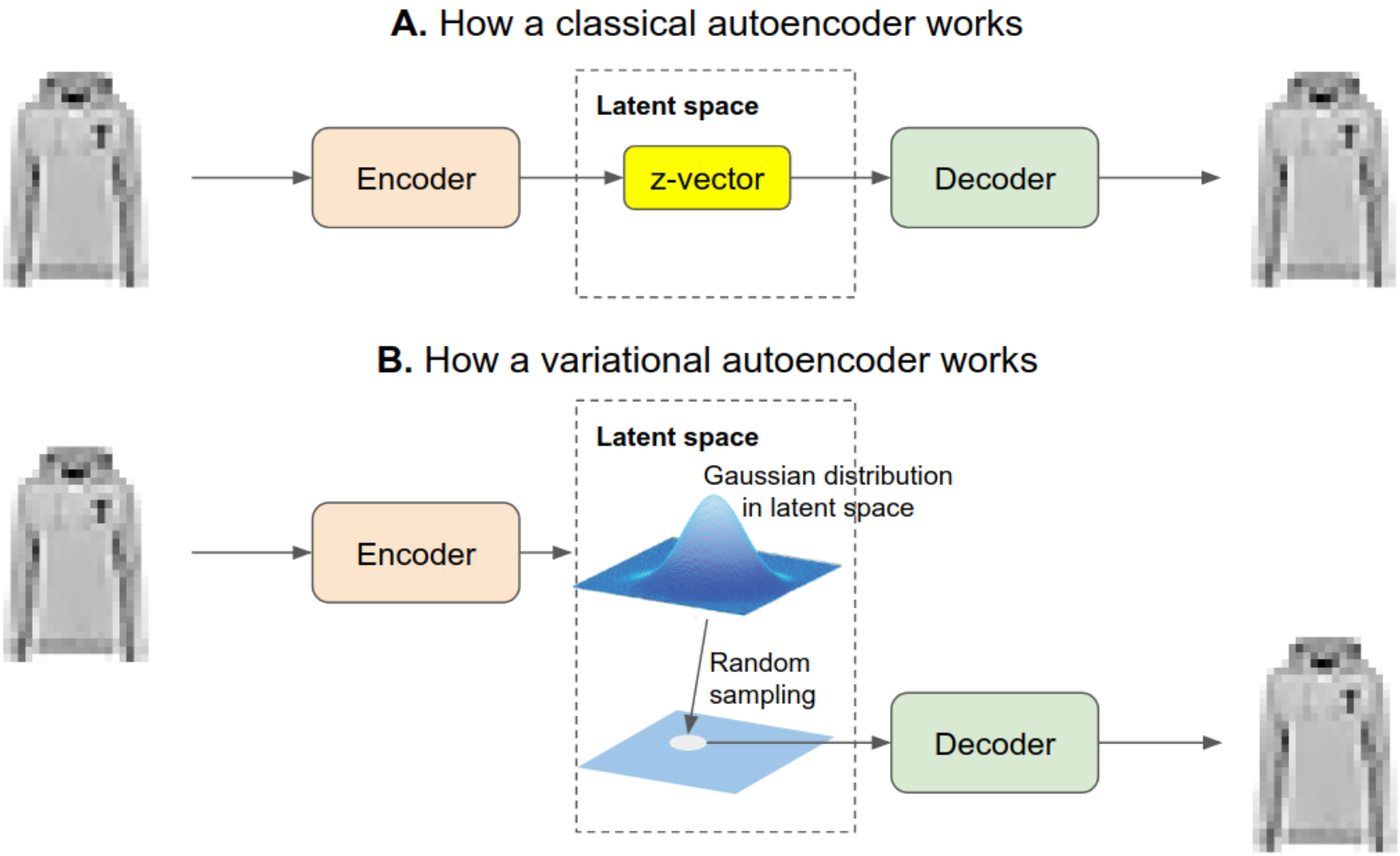
The expression to be considered for the complete network, Discriminator and Generator, is the following, and represents a value, *V*,

$$V\left(\theta^D, \theta^G\right) = E_{x\sim P\mathbf{x}}\left[\ \log\ D\left(\mathbf{x}\right)\right] + E_{z\sim P_{\mathbf{z}}}\left[\ \log\ \left(1 - D\left(G\left(\mathbf{z}\right)\right)\right)\right]. \tag{2}$$

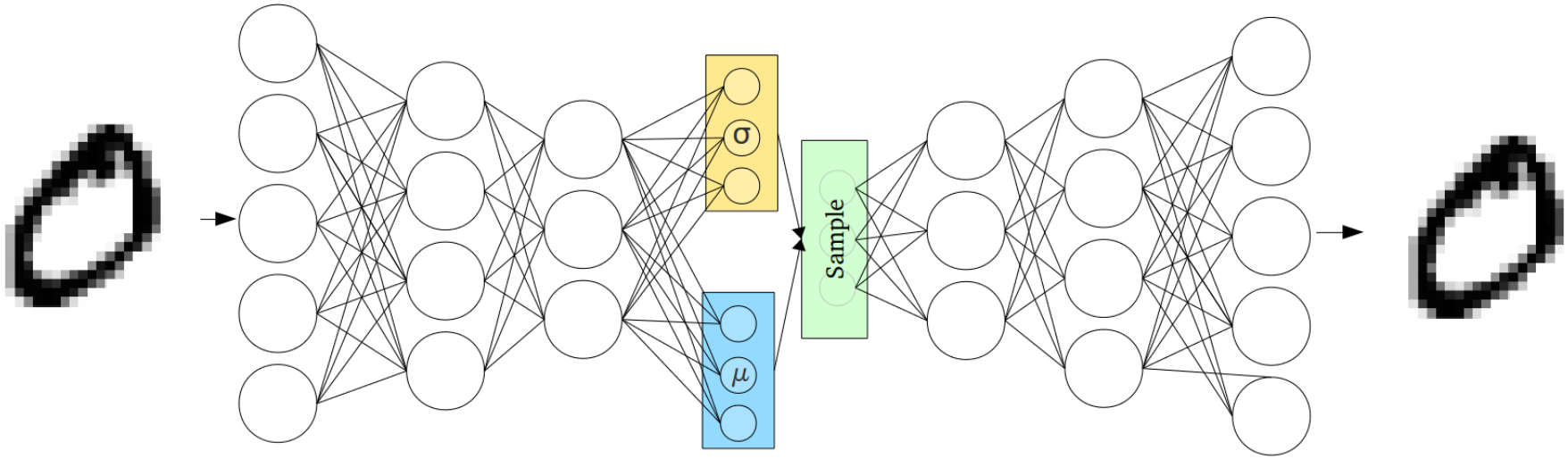
This value function is submitted to a min–max strategy with the goal to maximize the Discriminator loss and minimize the Generator loss,

$$min_{\theta^G}\ \max_{\theta^D}\ V\left(\theta^D, \theta^G\right). \tag{3}$$

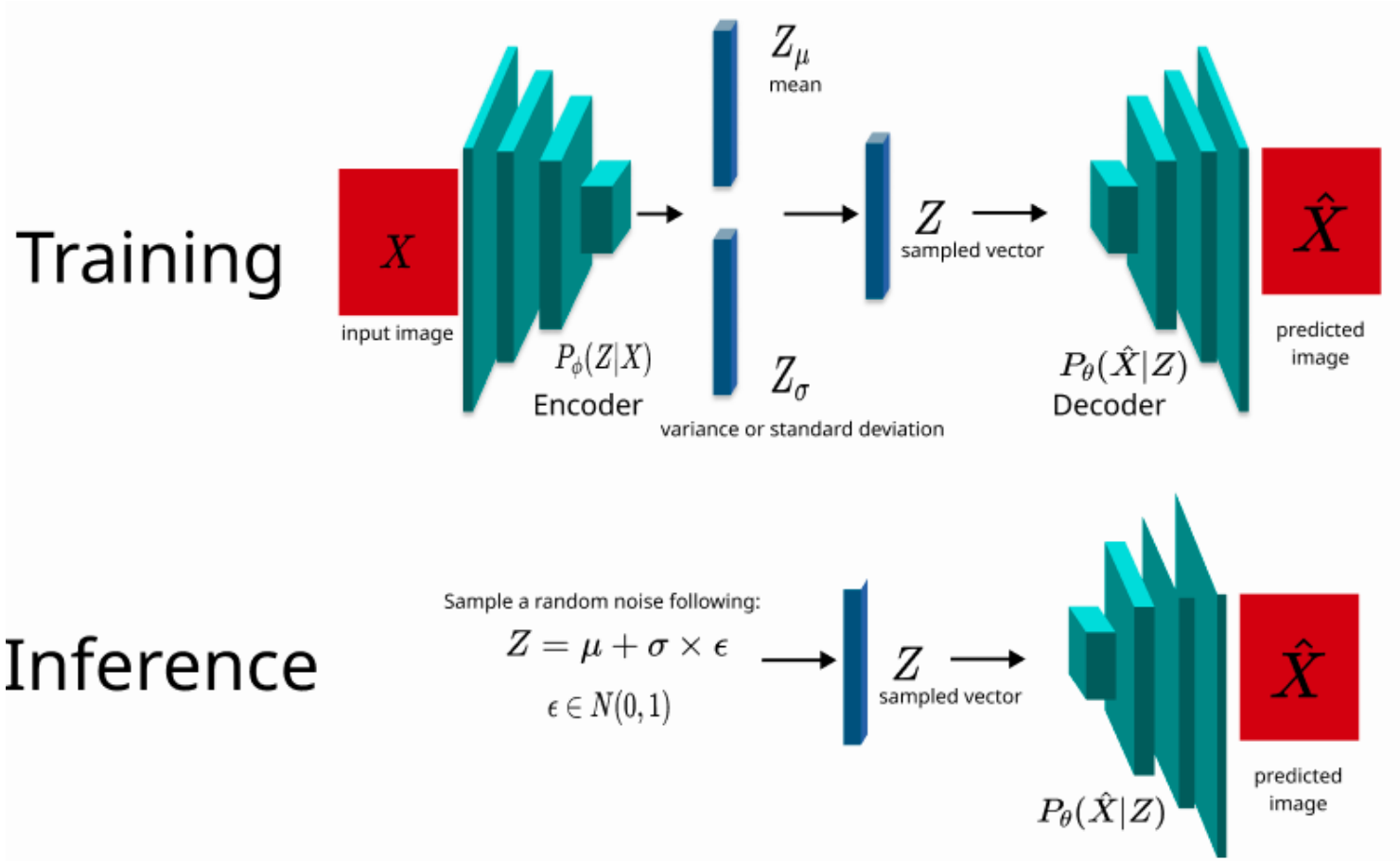
Variational AutoEncoder



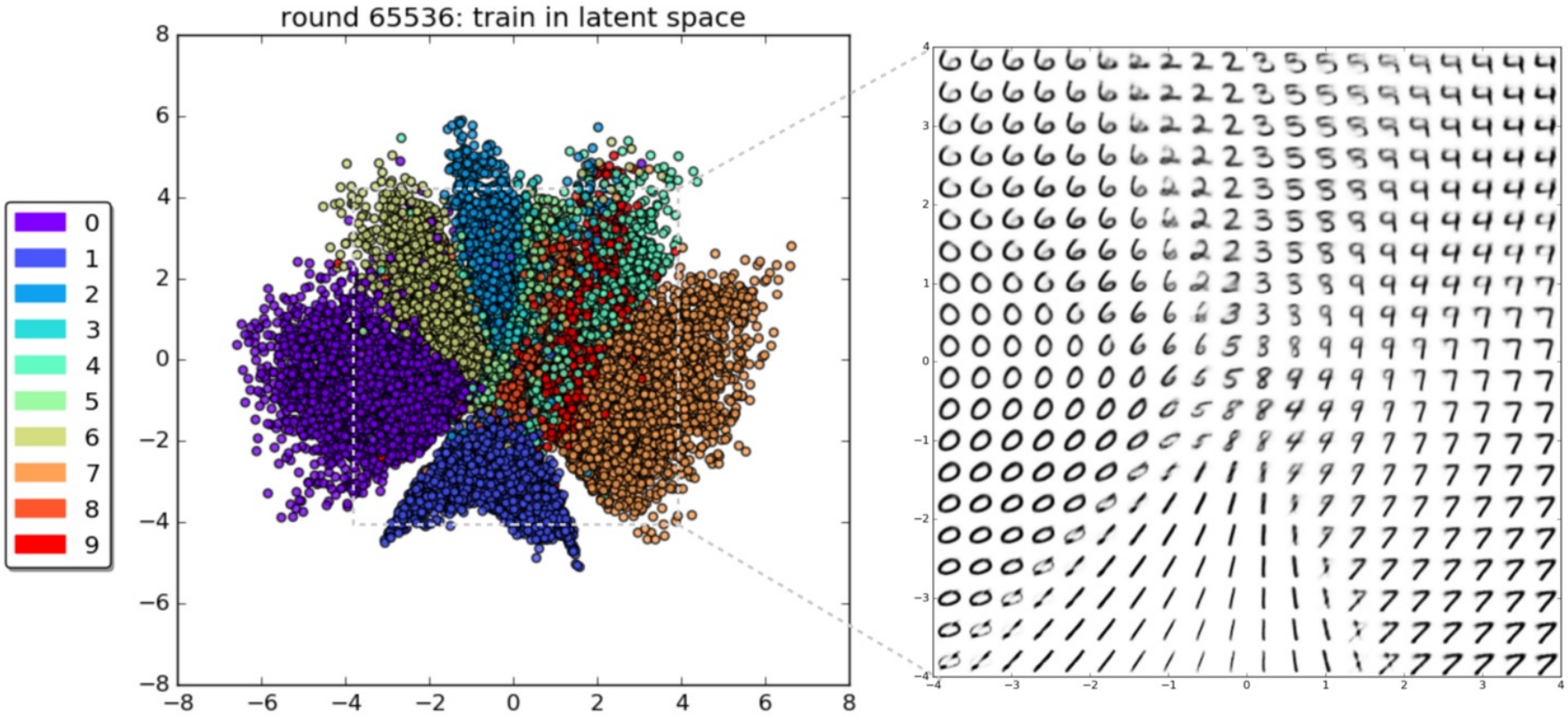
Variational AutoEncoder



Variational AutoEncoder



Variational AutoEncoder



Mètriques

Són útils les dades sintètiques?

- ‘per se’ (mesura directa)
- en el context d’ús (mesura indirecta)

Mètriques

- ‘per se’ (mesura directa): statistical validation or/and…

Existing generation methods have focused on properties such as,

- Fidelity
- Quality.
- Diversity
- Privacy
- Fairness

Mètriques

- en el context d'ús (mesura indirecta): accuracy on indirect tasks

Cecilio Angulo

cecilio.angulo@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH · UNIVERSITAT DE BARCELONA

Grup de Recerca en **Bioestadística** i **Bioinformàtica**

CCIA 2023

25è CONGRÈS INTERNACIONAL DE L'ASSOCIACIÓ CATALANA D'INTEL·LIGÈNCIA ARTIFICIAL

Món Sant Benet

25-27/10/2023

Financiado por la Unión Europea NextGenerationEU · Gobierno de España · Ministerio de Asuntos Económicos y Transformación Digital · España digital 2026 · Plan de Recuperación, Transformación y Resiliencia