Can Synthetic Data Replace the Real Thing? An Information-Theoretic Look at Data Fidelity

Brandon Ismalej[1]

Advised by: Kellie M. Evans[2]

[1]Department of Computer Science
California State University, Northridge

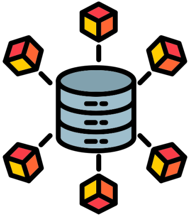
[2]Department of Mathematics
California State University, Northridge

Nov 21, 2025

Contents

Motivation

Motivation: Data is Fuel



- Modern AI requires *massive* amounts of data
- Increasing societal reliance on:
  - Wearables
  - Medical devices
  - Financial systems
  - Smart Systems/Internet of Things (IoT)
- Real data is powerful but risky

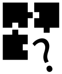
Why Real Data is Risky

- **Privacy**: Real data can reveal sensitive information

- **Bias**: Real data can be biased

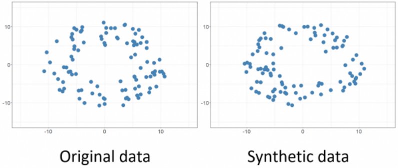
- **Cost**: Real data is expensive to collect and label

- **Scarcity**: Real data is not always available



Synthetic Data as a Potential Solution

- Synthetic data ("fake data") generated to mimic real data
- Benefits:
  - Cost-effective
  - Data augmentation: can fill gaps
  - Democratizes access to high-quality datasets
  - Privacy-preserving?



The synthetic data retains the structure of the original data but is not the same

Source: [Walker, 2020]

A Familiar Concept: Simulation vs. Synthesis

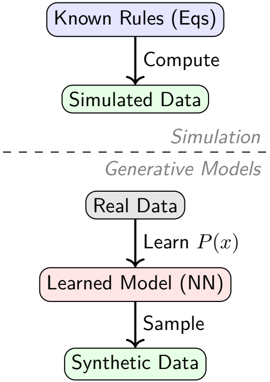
Generating data is not new

Traditional Simulation (Deductive):

- **Premise:** We know the underlying laws (e.g., Physical laws, ODEs, SDEs)
- **Process:** Rules $\rightarrow$ Data
- **Examples:** Monte Carlo integration, N-body simulations

Modern Synthetic Data (Inductive):

- **Premise:** Rules are unknown or too complex (e.g., human behavior, census demographics)
- **Process:** Data $\rightarrow$ Model $\rightarrow$ Synthetic Data
- **Goal:** Approximate distribution $P_{data}(x)$ and sample from it

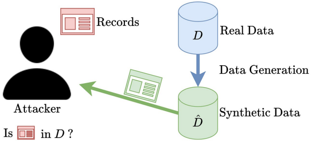


Privacy & Utility of Synthetic Data

Prior work investigates privacy & utility of synthetic data, but not its fidelity [Ismalej et al., 2025]

Privacy:

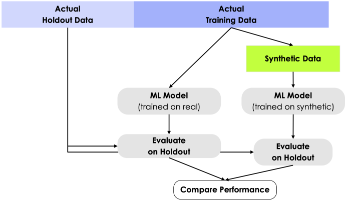
- Can synthetic data protect individuals?
- How vulnerable is it to membership inference?



Source: [Everton Gomede, 2023]

Utility:

- How well do models trained on synthetic data perform on real tasks?
- Can synthetic data replace or augment real datasets effectively?



Source: [MostlyAI, 2019]

Privacy & Utility of Synthetic Data

Findings:

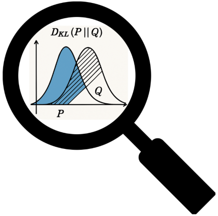
- There exists a nuanced trade-off between privacy and utility of synthetic data

Even if synthetic data is *privacy-preserving* or *useful*, it may still:

- Miss important patterns or relationships
- Distort the underlying data distribution

So, **what does it mean for synthetic data to be faithful to the real data?**

We can examine fidelity through an information-based lens



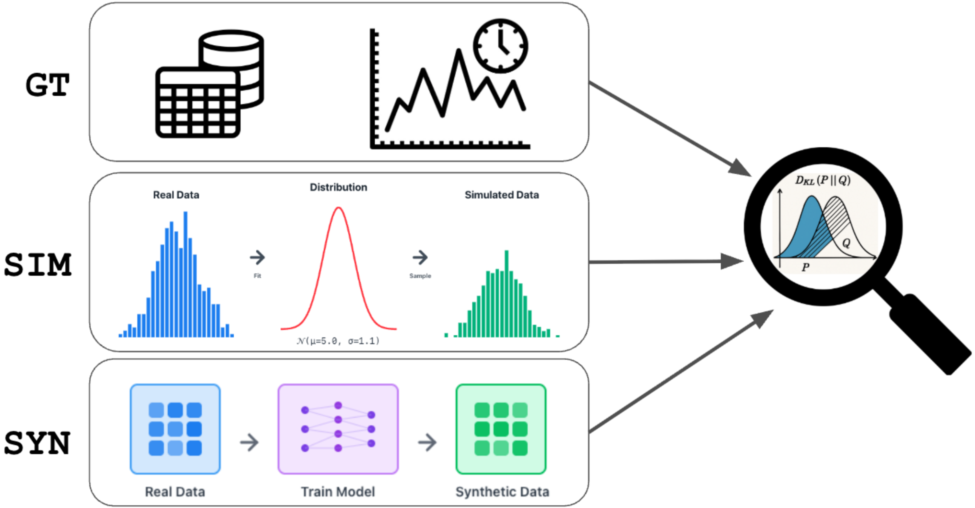
Introduction & Problem Setup

Introduction: What This Study Does

- We investigate the fidelity of two types of synthetic data:
  1. **Tabular** (Adult Census Dataset)
  2. **Time-series** (Human Activity Recognition Dataset)

- Compare three types of data:
  1. Ground Truth (real data)
  2. Simulated (rule-based)
  3. Synthetic (deep learning-based)

Introduction: What This Study Does



Background

Background: Data Formats

- **Tabular data:**
  - Rows and columns
  - Column is a variable/feature
  - Row is an observation/sample

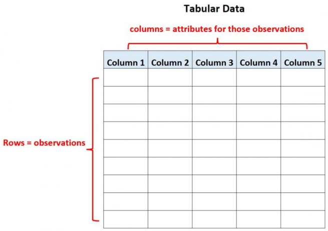


Figure: Source: [Bobbitt, 2022]

- **Time-series data:**
  - Time series of observations
  - Time is the independent variable
  - Observations are the dependent variables

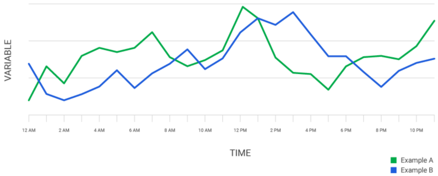


Figure: Source: [Pandita, 2024]

Background: Generative Models

Generative models are a class of machine learning models that can generate new data based on a given dataset

- **Generative Adversarial Networks (GANs)**: uses two neural networks to generate new data
- **Variational Autoencoders (VAEs)**: uses a neural network to encode and decode data

Background: Generative Models

Generative adversarial networks (GANs):

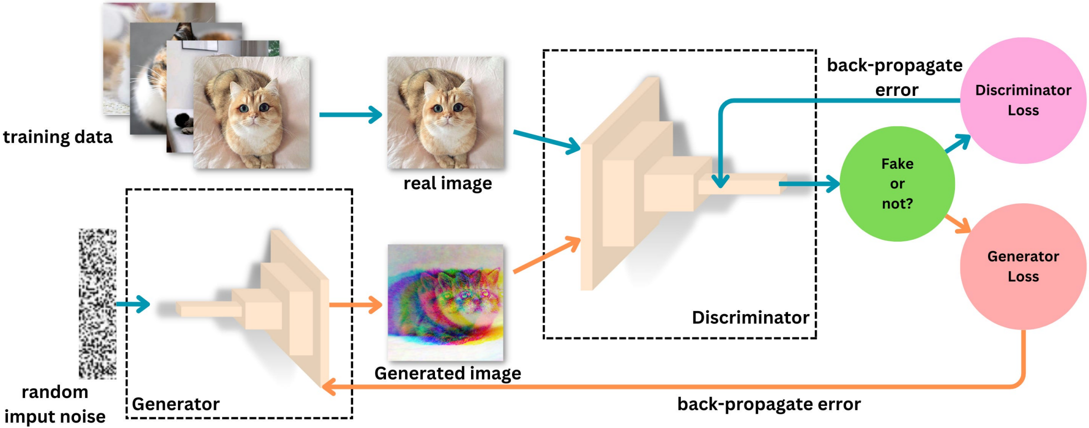


Figure: Source: [Benveniste, 2023]

Background: Generative Models

Variational autoencoders (VAEs):

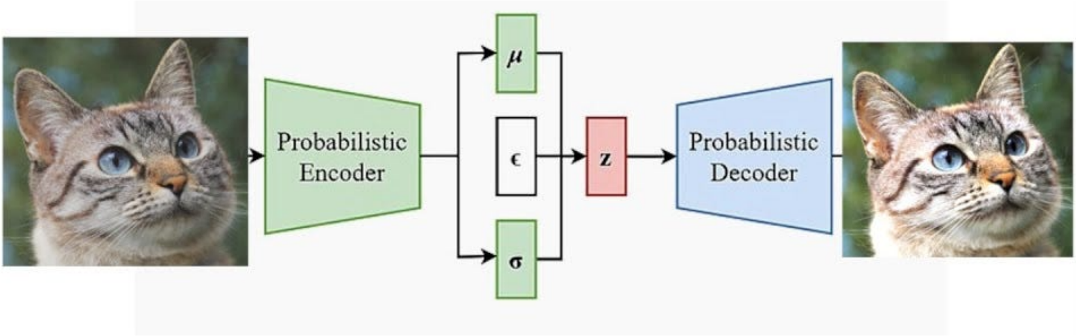


Figure: Source: [MacFarquhar, 2024]

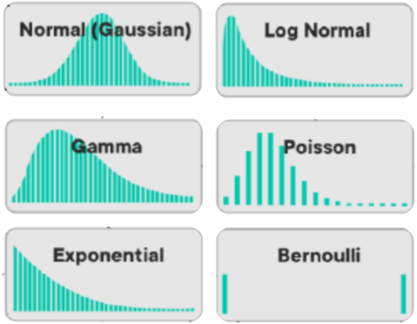
Datasets & Generative Pipelines

Dataset 1: Tabular Data - Adult Census

- Dataset: Adult Census
- Type: Tabular
- Source: UCI Machine Learning Repository [Becker and Kohavi, 1996]
- Description:
  - Number of samples/observations: 48,842
  - Number of features/variables: 14
  - Number of classes: 2
  - Target variable: income

Simulating Tabular Data: Overview

- **Fit univariate distributions per feature**
  - For each feature, select and fit a statistical distribution to match real data
  - Best fit distributions are chosen based on Akaike Information Criterion (AIC)
- **Sample from fitted marginals**
  - Generate synthetic values by sampling each feature independently from its fitted distribution
- **Limitation: No feature interactions**
  - Each feature is generated independently, so correlations and dependencies between features in the real data are not preserved



Synthetic Tabular Data: CTGAN & TVAE

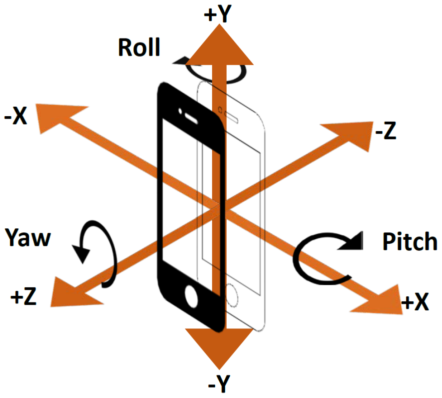
We utilize the **Synthetic Data Vault (SDV)** ecosystem [Montanez, 2018] to deploy deep learning models adapted for tabular constraints:

- **CTGAN (Conditional Tabular GAN):**
  - Adapts GANs to handle **discrete/categorical** variables (which usually block gradient flow)
  - Uses mode-specific normalization for non-Gaussian features
- **TVAE (Tabular Variational Autoencoder):**
  - An optimized VAE specifically for **mixed data types** (continuous numerical + categorical)

SDV
The Synthetic Data Vault

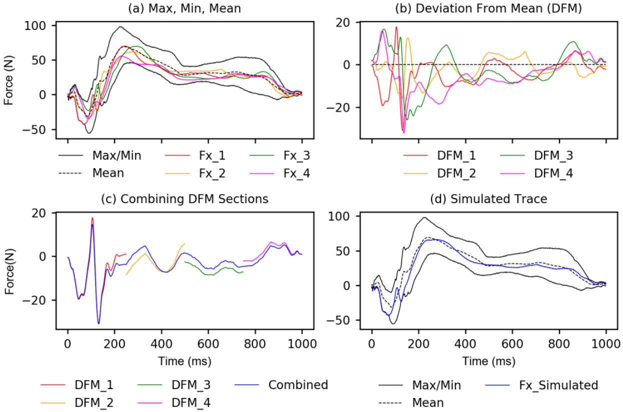
Dataset 2: Time-series Data - Human Activity Recognition

- Dataset: Human Activity Recognition Using Smartphones
- Type: Time-series
- Source: UCI Machine Learning Repository [Reyes-Ortiz et al., 2013]
- Description:
    - Number of samples/observations: 10,299
    - 128-length windows (2.56 seconds) sampled at 50 Hz
    - 9 channels (body/total acceleration and gyroscope)
    - Number of classes/activities: 6 (six human activities)



Simulating Time Series Data: Overview

- **Goal:** Generate synthetic time series data that reflects the key patterns of real human activity sensor data
- **Process** [Yeomans et al., 2019]:



Synthetic Time Series Data: TimeGAN

- **Core Idea:** Combines the Generative Adversarial Network (GAN) framework with a recurrent neural network (RNN) embedding
- Learns and preserves temporal dynamics in sequential (time-series) data
- Generates realistic multi-channel time series that capture both feature dependencies and temporal patterns

Fidelity Metrics

Tabular Fidelity: Per-Feature Histograms (JS)

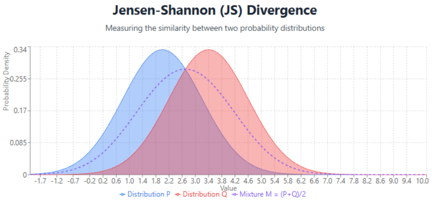
Goal: Do the marginal distributions match?

For feature X with real histogram P and synthetic Q:

$$\mathrm{JS}(P, Q) = \tfrac{1}{2}\mathrm{KL}(P \,\|\, M) + \tfrac{1}{2}\mathrm{KL}(Q \,\|\, M)$$

$$M = \tfrac{1}{2}(P + Q).$$

- Computed per continuous feature (e.g., age, fnlwgt).
- **Interpretation:** Lower JS $\Rightarrow$ histograms match.

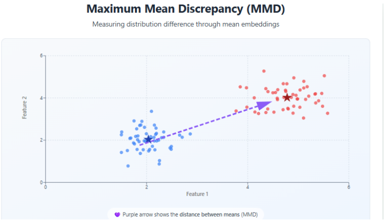


Tabular Fidelity: Global Geometry (MMD)

Goal: Measure global similarity using kernel geometry.
Flatten datapoints to vectors x, y. With kernel $k(\cdot,\cdot)$:

$$\mathrm{MMD}^2(X,Y) = \frac{1}{n^2}\sum_{i,j} k(x_i,x_j)$$

$$+\frac{1}{m^2}\sum_{i,j} k(y_i,y_j) - \frac{2}{nm}\sum_{i,j} k(x_i,y_j).$$



- **Interpretation:** Lower MMD $\Rightarrow$ more overlap between global feature distributions.

Tabular Fidelity: Classifier Two-Sample Test (C2ST)

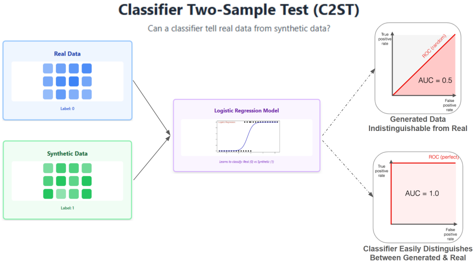
Goal: Can a machine distinguish Real from Fake?

Method:

- Train a **logistic regression** classifier.
- Label: Real $= 1$, Generated $= 0$.
- Report **ROC AUC** on held-out test set.

Interpretation:

- **AUC $\approx$ 0.5:** Perfect Fidelity (indistinguishable).
- **AUC $\approx$ 1.0:** Low Fidelity.



Time-Series Fidelity: Frequency (PSD)

The Goal: Do the synthetic signals have the same "rhythm" and oscillations?

Metric: PSD Jensen-Shannon Divergence

- Compute Power Spectral Density (PSD) histograms $P^{(c)}, Q^{(c)}$ for each channel c.

$$\mathrm{JS}_{\mathrm{PSD}}^{(c)} = \mathrm{JS}\big(P^{(c)}, Q^{(c)}\big).$$

Interpretation:

- **Lower is better.**
- Smaller JS means generated data matches the frequency content of ground truth.

Time-Series Fidelity: Temporal Dependence (ACF)

The Goal: Does the past predict the future in the same way?

Metric: Autocorrelation Absolute Difference

- Compute autocorrelation $r^{(c)}(\tau)$ for lags $\tau = 1, \ldots, L$.
- Measure average absolute difference:

$$\Delta_{\mathrm{ACF}}^{(c)} = \frac{1}{L} \sum_{\tau=1}^{L} \left| r_{\mathrm{GT}}^{(c)}(\tau) - r_{\mathrm{GEN}}^{(c)}(\tau) \right|.$$

Interpretation:

- **Lower is better.**
- Means generated data has temporal dependence structure closer to ground truth.

Time-Series Fidelity: Cross-Channel Coupling

The Goal: Are the physical relationships between sensors preserved?

Metric: Cross-channel correlation matrix difference

- Compute $C \times C$ correlation matrix Σ across channels.
- Compare via Frobenius norm:

$$\Delta_{\text{corr}} = \|\Sigma_{\mathrm{GT}} - \Sigma_{\mathrm{GEN}}\|_F.$$

Interpretation:

- **Lower is better.**
- Cross-channel relationships (e.g. coupling of axes) are preserved.

Time-Series Fidelity: Global Window Geometry

The Goal: Do the full windows occupy the same region of the high-dimensional manifold?

Metric: MMD on full windows

- Flatten each window to a vector in $\mathbb{R}^{T \cdot C}$ (here, $128 \times 9 = 1152$).
- Compute RBF-kernel MMD between GT windows and generated windows.

$$\mathrm{MMD}^2(X_{\text{win}}, Y_{\text{win}}) = \ldots \text{(Same Kernel formulation as Tabular)}$$

Interpretation:

- **Lower is better.**
- Generated windows occupy the same region as real windows.

Results

Adult (Tabular): Global Fidelity

Experimental Setup:

- **Real:** UCI Adult (104-dim vectors).
- **SIM:** Per-feature marginals + Gaussian copula.
- **SYN:** CTGAN and TVAE (Deep Learning models).

Global Metrics (Lower is better for MMD, C2ST $\approx$ 0.5 is best):

Metric	GT vs **SIM**	GT vs CTGAN	GT vs TVAE
MMD (RBF)	**0.0030**	0.0067	0.0088
C2ST AUC	**0.576**	0.746	0.741

Interpretation:

- **MMD:** SIM is significantly closer to real data geometry.
- **C2ST:** The classifier struggles most to distinguish SIM from Real.
- Verdict: Simple Simulation (SIM) is globally more faithful.

Adult (Tabular): Per-feature Histogram Match (JS)

Metric: Jensen-Shannon Divergence (Lower is better)

Feature	**SIM**	CTGAN	TVAE
age (Continuous)	**0.004**	0.018	0.132
capital_gain (Continuous)	**0.001**	0.018	0.005
education_num (Discrete)	0.554	**0.024**	0.950
hours_per_week (Discrete)	0.337	**0.182**	0.622
fnlwgt (Mixed)	0.165	0.281	**0.068**

Interpretation:

- **SIM wins on smooth/continuous features** (e.g., Age, Capital Gain).
- **Deep Learning (CTGAN) wins on discrete/categorical features.**
- Verdict: No single winner. SIM captures simple shapes; GANs capture complex discrete modes.

HAR (Time-Series): Global Fidelity

Experimental Setup:

- **Real:** HAR windows ($N = 2947, T = 128, C = 9$).
- **SIM:** DFM-mosaic simulator (Physically-inspired smoothing).
- **SYN:** TimeGAN (Recurrent GAN).

Global Metrics (Lower is better):

Metric	GT vs **SIM**	GT vs TimeGAN
MMD (RBF)	**0.0237**	0.3863
Correlation Δ_{corr} (Frobenius)	**0.399**	4.157

Interpretation:

- **Correlation:** SIM preserves channel coupling (axes physics) far better.
- **Geometry:** TimeGAN windows are distant from the real manifold.
- Verdict: DFM-mosaic clearly outperforms TimeGAN on global structure.

HAR (Time-Series): Channel-level Example

Deep Dive: Channel `body_acc_x` (x-axis acceleration)

Metric (ch0)	GT vs **SIM**	GT vs TimeGAN
PSD JS (Frequency)	**0.0625**	0.2783
ACF Δ (Time Dep.)	**0.0639**	0.1398

Interpretation:

- **Frequency:** SIM matches the spectral "rhythm" of walking.
- **Time:** SIM tracks the autocorrelation decay accurately.
- Verdict: SIM preserves the physical dynamics (“how the phone moves”); TimeGAN distorts them.

Summary: Tabular Data (Adult)

The Battle: Simple Simulation (SIM) vs. Deep Learning (CTGAN/TVAE)

- **Global Geometry (MMD) & Indistinguishability (C2ST):**
  - **SIM wins** It achieved the lowest MMD and was hardest for a classifier to distinguish from real data
- **Feature-Level Detail:**
  - **Deep Learning wins on complexity** CTGAN and TVAE fixed specific, challenging features (like `education_num` and `capital_loss`) where the simple simulation failed to capture the distribution shape

Tabular Takeaway

Simple simulation captures the *global structure* effectively, but deep generative models are necessary to capture complex, non-standard *marginal distributions*

Summary: Time-Series Data (HAR)

The Battle: Mechanistic Simulation (DFM-mosaic) vs. Recurrent GAN (TimeGAN)

- **Global Structure (MMD & Correlations):**
  - **SIM dominates** It preserved cross-channel relationships (e.g., how x-axis acceleration relates to y-axis) far better than TimeGAN
- **Temporal Dynamics (PSD & ACF):**
  - **SIM dominates** It matched the frequency content (rhythm of walking) and temporal dependencies closer than the GAN

Time-Series Takeaway

A physically-inspired simulator (even a simple one) preserves *dynamics* better than a generic deep learning model that attempts to learn physics from scratch

Conclusion

Conclusion: Implications for Synthetic Data

1. **Fidelity is Multi-Dimensional**
   - A model can have perfect marginals but broken geometry, or perfect geometry but broken temporal dynamics We must measure all axes
2. **"State-of-the-Art" Requires Auditing**
   - Deep generative models are powerful, but they are not magic They can distort underlying physics or correlations if not carefully tuned
3. **Don't Discard Simulation**
   - As shown in the Time-Series results: Sometimes, *knowing the rules* (Simulation) yields more faithful data than *trying to learn them* (Synthesis)

Thank You!

Any Questions?

Contact

`brandon.ismalej.671@my.csun.edu`

Acknowledgments

Special thanks to my advisor:
Dr. Kellie Evans,
Department of Mathematics

References

References I

Becker, B. and Kohavi, R. (1996).
Adult.
UCI Machine Learning Repository.
DOI: https://doi.org/10.24432/C5XW20.

Benveniste, D. (2023).
How generative adversarial networks work!

Bobbitt, Z. (2022).
What is tabular data? (definition & example).

Everton Gomede, P. (2023).
Exploring the frontiers of synthetic data: A comprehensive overview of ctgan and its applications.

Ismalej, B., Ruan, X., and Jiang, X. (2025).
Evaluating privacy and utility of synthetic tabular data with membership inference attacks.
In *IEEE Future Machine Learning and Data Science (FMLDS)*, Los Angeles, CA.

MacFarquhar, M. (2024).
The art of encoding: Building a variational autoencoder for mnist digits.

Montanez, A. M. (2018).
Sdv : an open source library for synthetic data generation.
MIT DSpace.
hdl:1721.1/121631.

References II

MostlyAI (2019).
Validate synthetic data via train-synthetic-test-real.

Pandita, A. (2024).
Unveiling the essence of time series analysis.

Reyes-Ortiz, J., Anguita, D., Ghio, A., Oneto, L., and Parra, X. (2013).
Human Activity Recognition Using Smartphones.
UCI Machine Learning Repository.
DOI: https://doi.org/10.24432/C54S4K.

Walker, K. (2020).
Synthetic data: Unlocking the power of data and skills for machine learning – data in government.
Blog.gov.uk.

Yeomans, J., Thwaites, S., Robertson, W. S. P., Booth, D., Ng, B., and Thewlis, D. (2019).
Simulating time-series data for improved deep neural network performance.