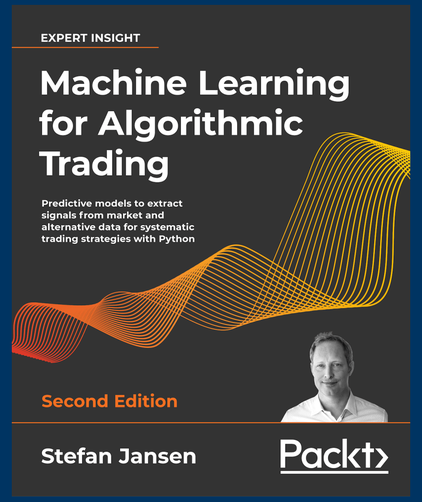
Synthetic Data for Finance

From Theory to Practice



Stefan Jansen

Disclaimer

Agenda

Synthetic Data for Finance

- **Use Cases**: Why would synthetic data be useful?
- **New Methods**: How do ML-based approaches work?
- **Illustration:** Time-Series Generative Adversarial Network
- **Evaluation:** When can we take synthetic for real?
- **Next Steps:** What have we learned, and where do we go?

**What I cannot create,
I do not understand.
Richard Feynman**

Why does finance need synthetic data?

Use Cases from privacy to machine learning for trading

- **Regulation:** enable usage or sharing of information while protecting real data
- **Address gaps and weaknesses of real data:**
  - Create data around rare events from crises to fraud that is scarce by nature
  - Overcome training data shortages that risk model and backtest overfitting
  - Ensure data reflects all current and future customer or market characteristics
- **Off-premise training:** move model data into the cloud or other less safe environments

How does machine learning help generate data?

From model-based simulation to data-driven deep learning

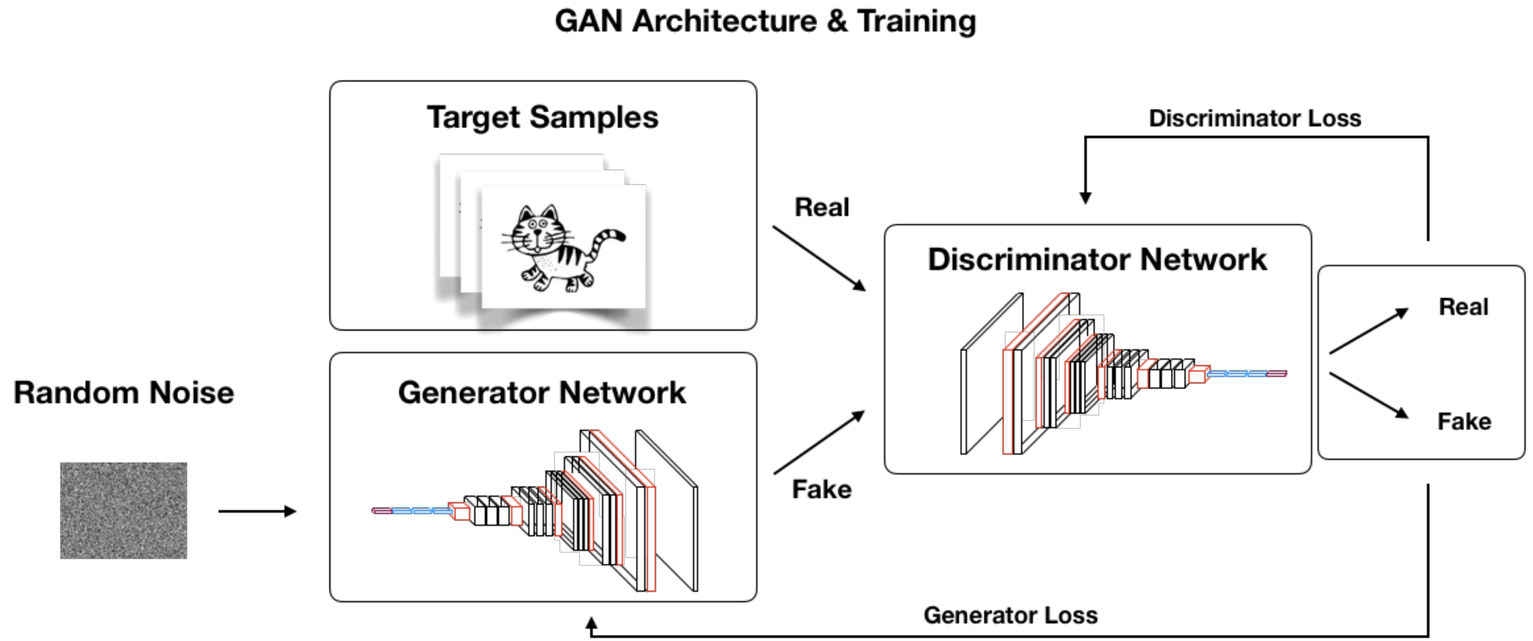
- **Traditional approach:** simulation based on stochastic models fitted to data
  - Downside: often misses stylized facts like fat tails and volatility clustering
  - Challenging to condition results on arbitrary asset or environment attributes
- **Deep Generative Learning to the rescue:**
  - In 2014, Adversarial Networks explode on the scene
  - Famously applied to images, their application has since expanded into various domains

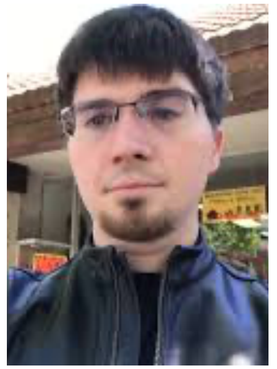
The most interesting idea in ML in the last decade
Yann LeCunn

How do Generative Adversarial Networks work?

With a zero-sum game to diverse, realistic synthetic data

- A GAN sets up a competition between two networks to simultaneously learn the generative process behind some class of data by playing a minimax strategy
- The generator networks receives a reward for tricking the discriminator into mistaking fake data for real, and vice versa





Ian Goodfellow

Applied AI

GAN invention spawns major surge in research

From photos, style transfer, text-to-image and music - to time series

- **Capabilities improve** rapidly through better design and training methods
  - Over 500 GAN architectures by 2018 for wide range of generative applications
  - Conditional models shape the output via additional input signals
- But **challenges remain** as simultaneous, interdependent training lacks robustness
  - Mode collapse: generator output lacks diversity
  - Discriminator dominates generator
  - Models do not converge

GANs for (Financial) Time Series Data

From medical to financial time series

- **Recurrent (Conditional) GAN** for short ICU time-series (ETH Zuerich, 2017)
- Goal: train early warning system on synthetic data generated from 18,000 patients (T=16), conditioned on patient state
- Introduces **Recurrent Neural Networks** (RNN), tailored to translating one sequence into another, for both generator and discriminator
- New task-focused **evaluation method**: train on synthetic data, test on real data => only minor degradation vs train on real, test on real
- Checks the model not just reproduces the training data (thus violating privacy)

GANs for (Financial) Time Series Data

Taking a step back: how similar are financial and image data?

- Images are far from white noise => much higher signal/noise ratio
- Pixel values live in range [0, 255], while financial time series are rarely stationary => only small sample with similar distribution
- To be useful, synthetic data need to reflect key stylized facts:
  - Fat tails imply few samples of extreme values
  - Need to reflect (auto-)correlation present in (multivariate) real time series
- How could GANs adjust the generative process to shape their output accordingly?

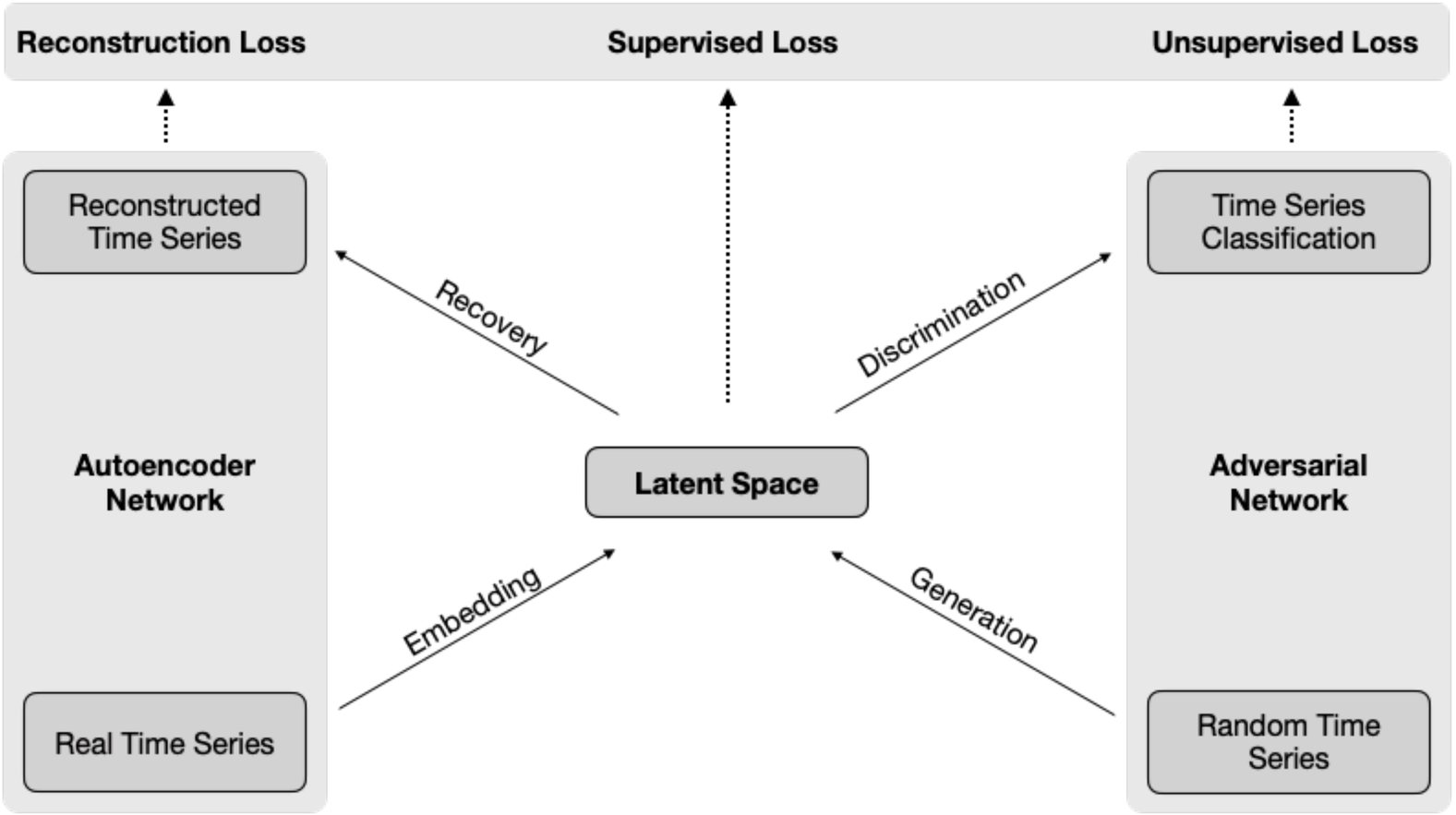
Time-Series Generative Adversarial Network

Yoon et al. (Google / Cambridge), NEURIPS 2019

- **Goal:** preserve the *temporal dynamics* so that synthetic sequences respect the relationships
  - **between** variables
  - **across** time
- **Approach**: Generate realistic data by combining “the flexibility of the unsupervised paradigm with the control afforded by supervised training”.
- Learn to reduce the dimensionality while optimizing for supervised and adversarial objectives so the network adheres to the dynamics during training

Time-Series Gan Architecture

Joint training: 2 networks, 4 components, 3 loss functions

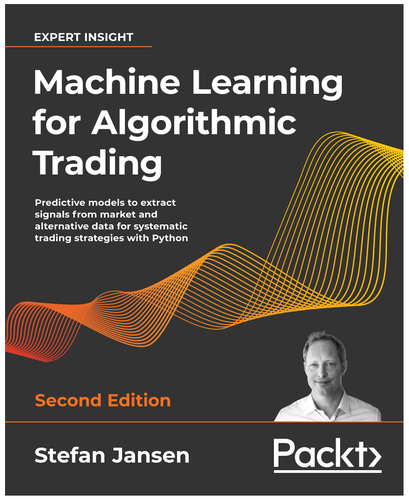


1. Unsupervised adversarial loss on real and synthetic sequences
2. Supervised loss (minimized by training both the embedding and generator networks) captures the stepwise conditional distribution
3. The embedding network reduces the dimensionality of the adversarial learning space, assuming that temporal dynamics are driven by fewer, lower-dimensional factors.

Time-Series GAN in Practice

Code available on GitHub

- One of the few time-series GAN examples with code (also: RCGAN)
- For book chapter 21, port original implementation to TF2 and simplify
- Replicate authors' experiments with short time series (T=24)
  - Closing prices for six stocks instead of OHLCV for one stock
- Add experiments to
  - evaluate scalability and
  - test suitability for returns



https://github.com/stefan-jansen/machine-learning-for-trading

How to evaluate the quality of synthetic time-series data?

Daily close price series (T=24) for six stocks (history since 1990)

The TimeGAN authors use three criteria to assess the generated data:

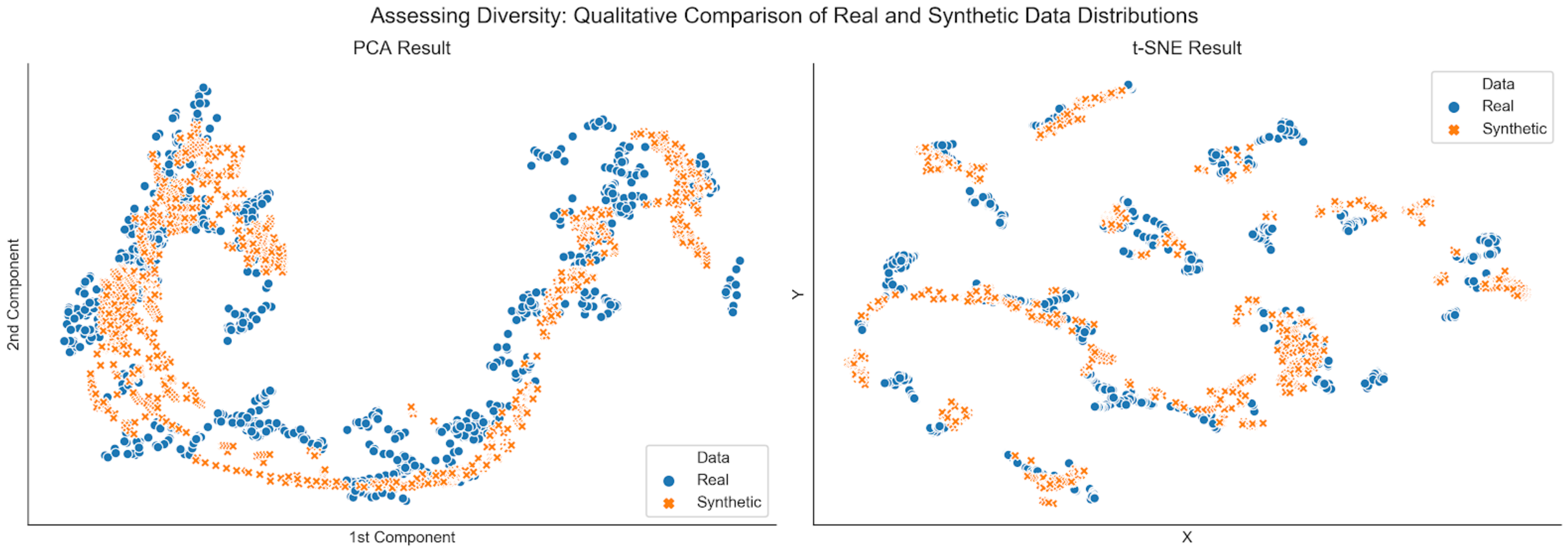
1. **Diversity**: the distribution of the synthetic samples should roughly match that of the real data

2. **Fidelity**: the sample series should be indistinguishable from the real data, and

3. **Usefulness**: the synthetic data should be as useful as their real counterparts for solving a predictive task

Assessing diversity: visualization using PCA and t-SNE

A simple qualitative tool: how similar are the distributions?

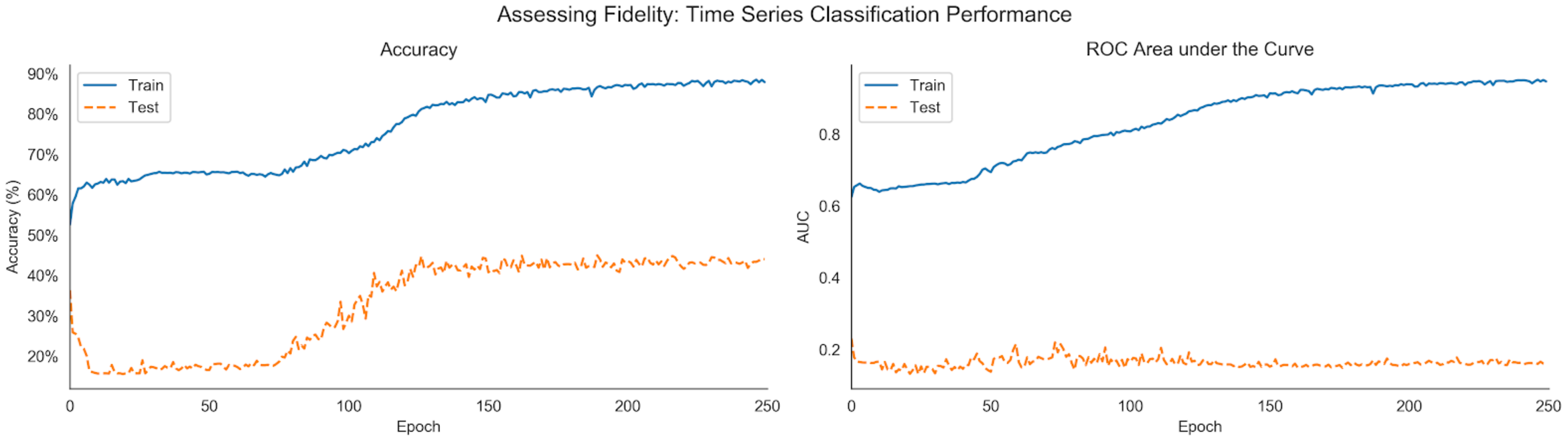


Applied AI

Assessing fidelity: time series classification performance

Can an off-the-shelf classifier differentiate real and synthetic series?

- Performance after training on 80% of labelled real and synthetic data

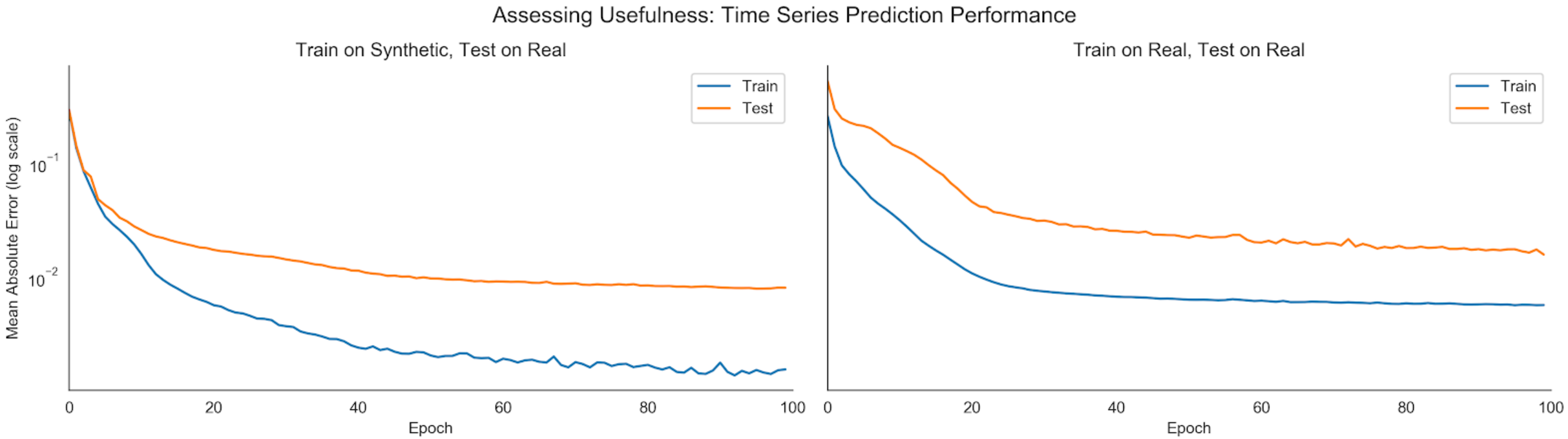


Applied AI

Assessing usefulness: train on synthetic, test on real

Can a model learn to predict a real timestep from synthetic data?

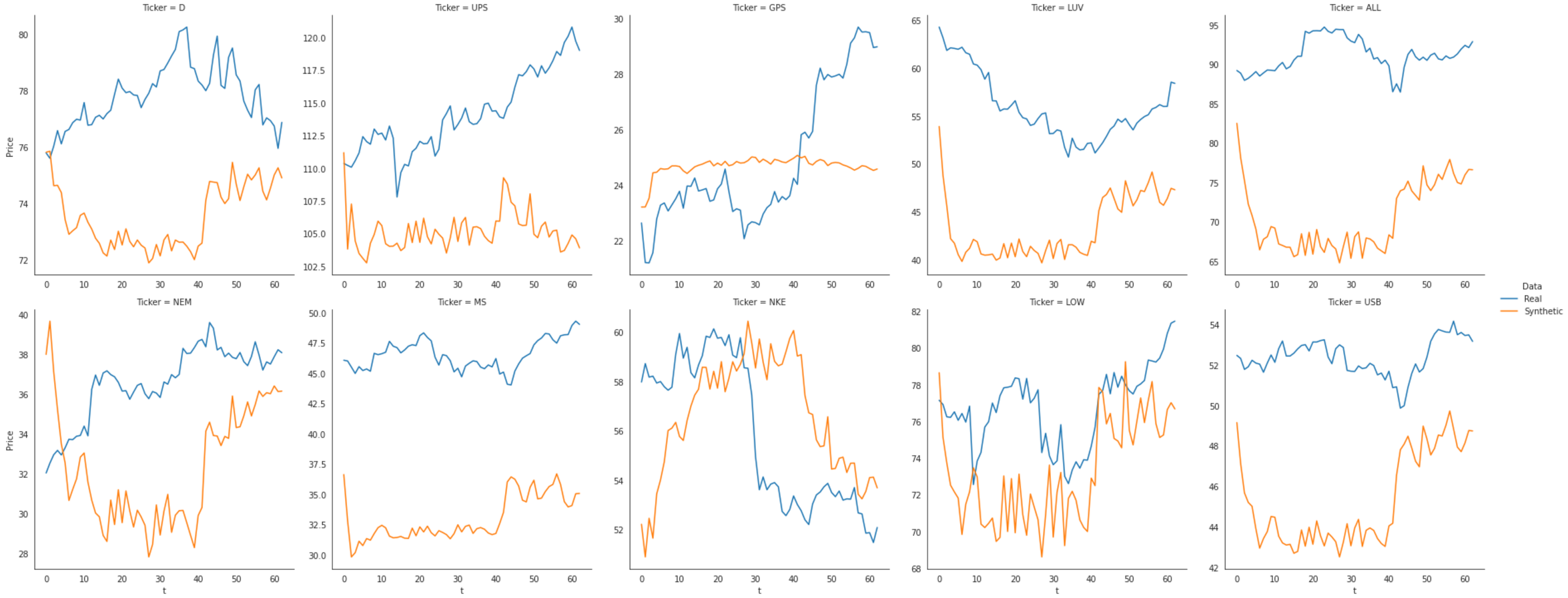
- Performance on predicting the last step in the sequence after training on synthetic and real data, respectively



Does Time-Series GAN scale?

63-day close price series, 50 randomly picked tickers

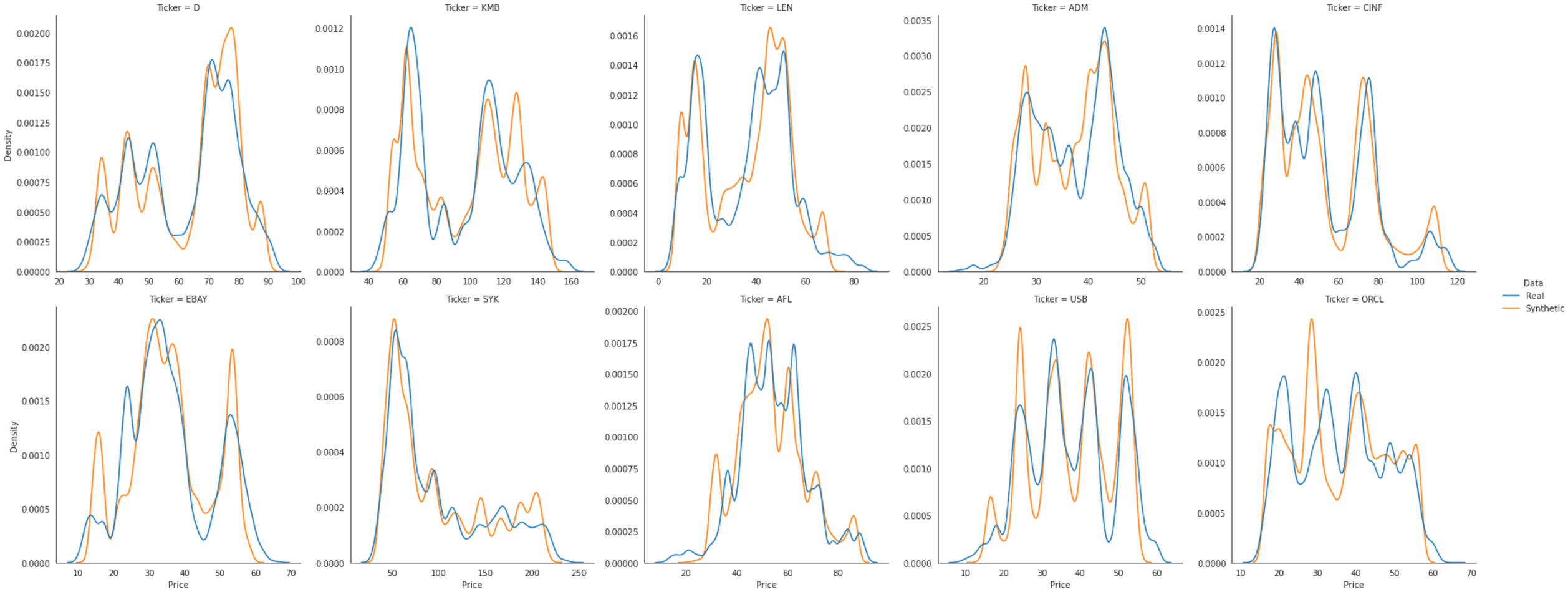
- Training data: S&P500 constituents, daily pricing, 2007-2021, by algoseek THE MARKET DATA COMPANY



Price plots for randomly sampled 63-day periods for 10 tickers

Does Time-Series GAN scale?

On visual inspection, synthetic series and distributions match source

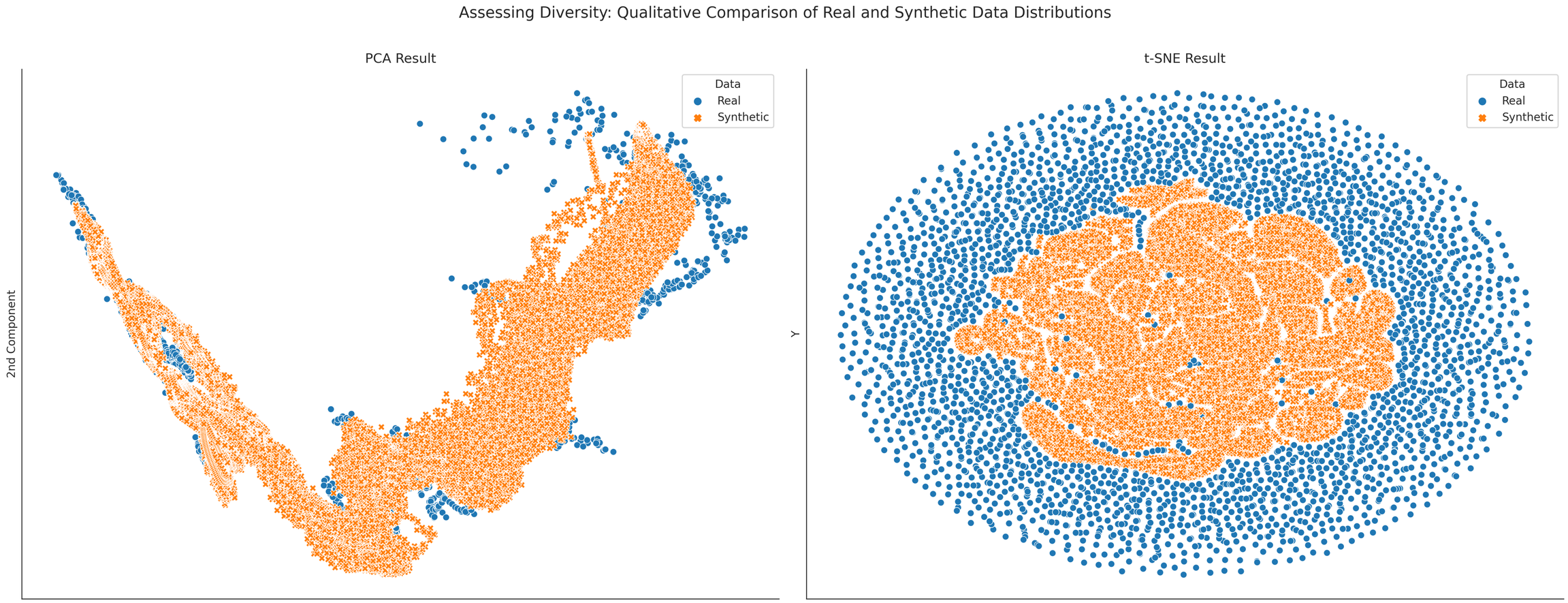


Price distributions for random sample of 2,500 63-day periods (~8%)

Applied AI

Does Time-Series GAN scale?

However, the synthetic data do not cover the entire real sample space

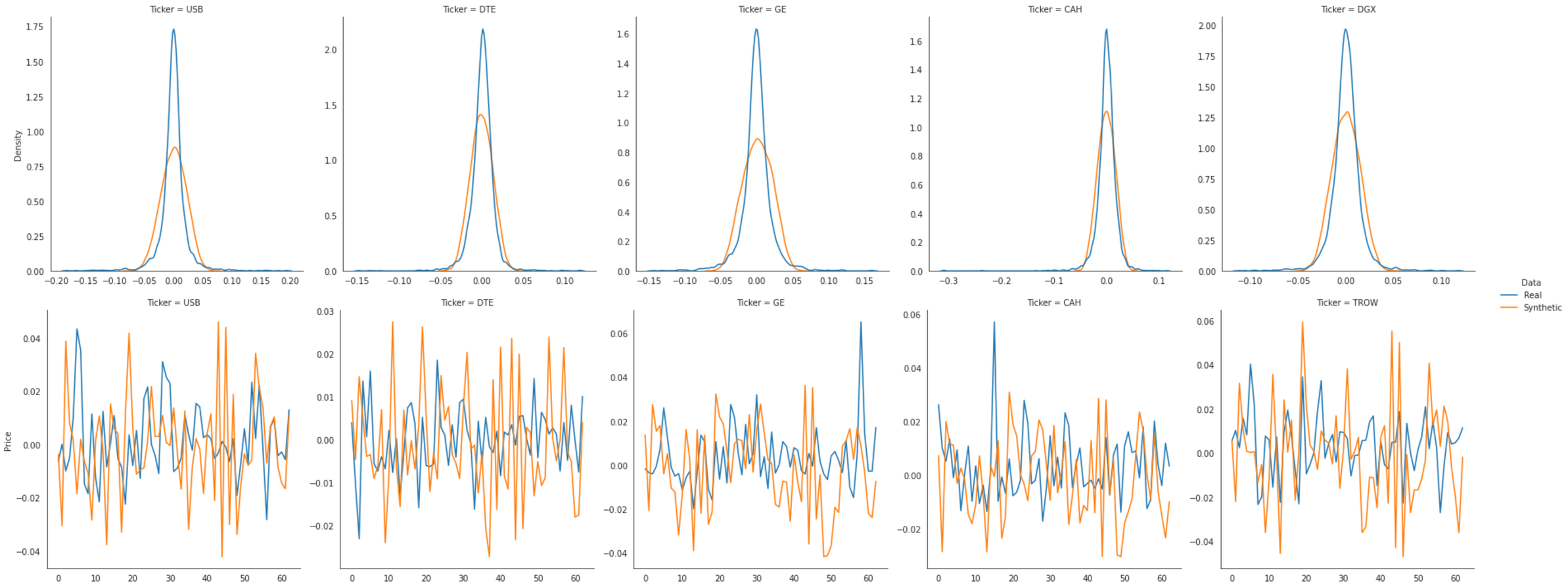


Visualizations for random sample of 2,500 63-day periods (~8%)

Applied AI

Does Time-Series GAN scale?

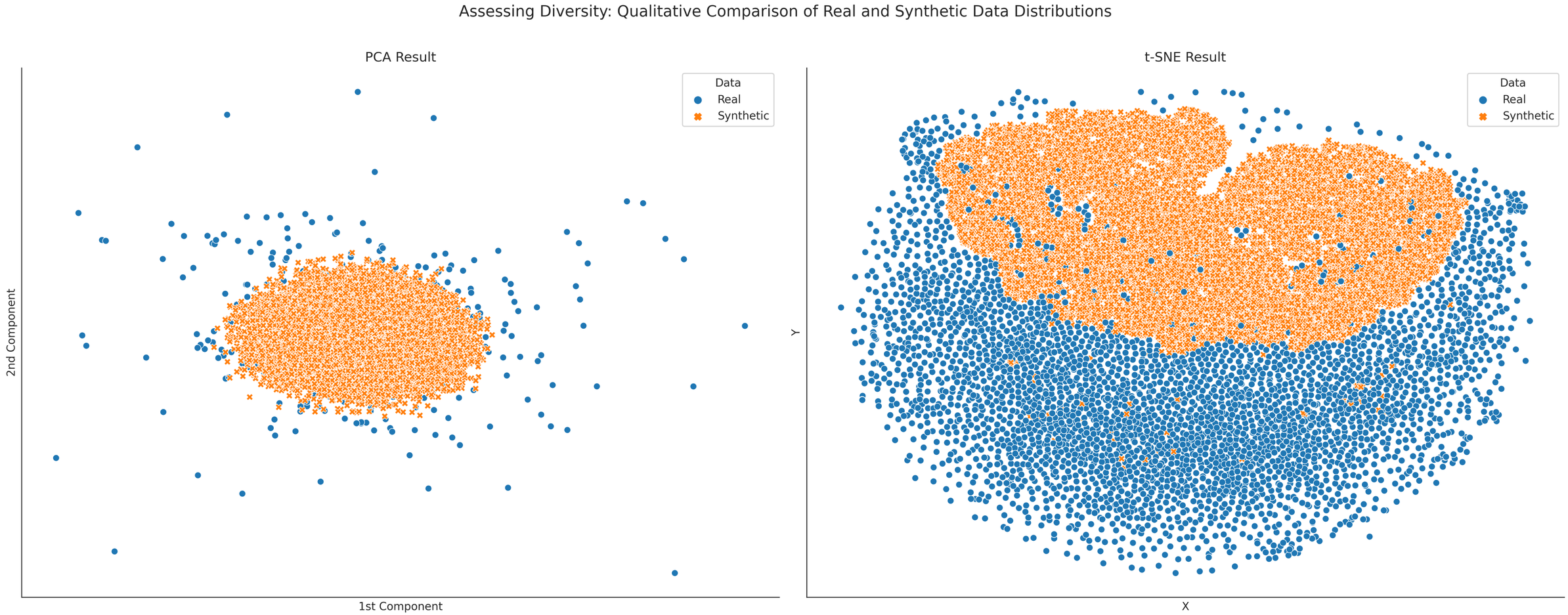
How about daily returns?



Applied AI

Does Time-Series GAN scale?

The lower temporal dynamic of returns further reduces the fidelity of the data.



Time-Series GAN Takeaways

Promising, but not yet ready for prime time (out of the box)

- Using a small dataset, TimeGAN creates synthetic data that, to some extent, mimic actual stock price series.
- While far from conclusive about our ability to create artificial data that is, in fact, useful for model training and backtesting, it’s a promising first step.
- There are numerous avenues to build on this architecture by refining the configuration and the training process

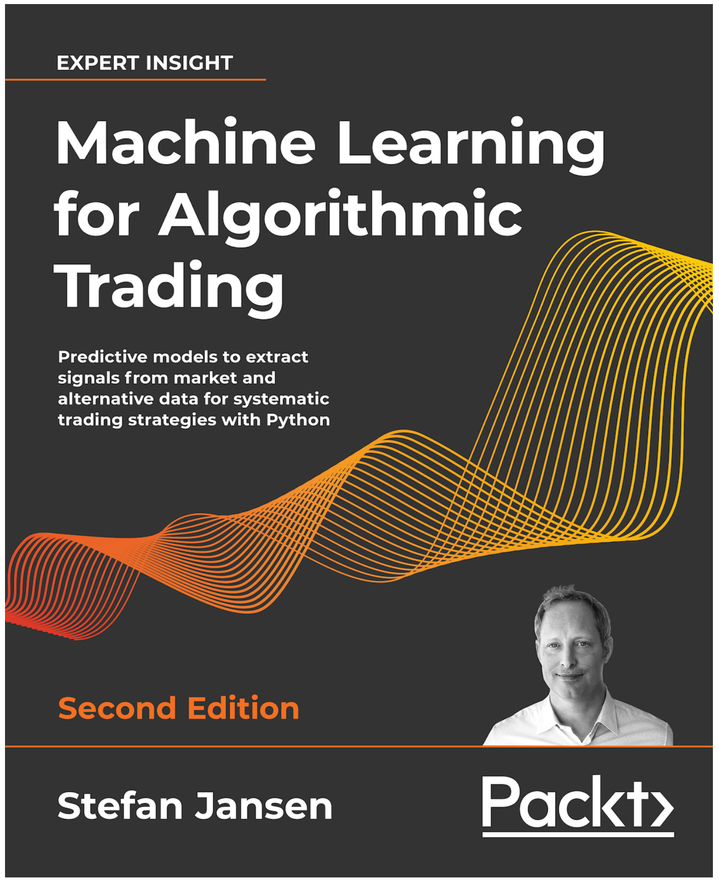
What have we learned and what's next?

Generating realistic time-series data is hard - but there is progress

- More recent research adds to the Time-Series GAN results:
  - **QuantGAN** (Wiese, et al 2019): Temporal Convolutional Network generates a daily log-S&P500 series (T=2,400) that outperforms a GARCH(1, 1) model
  - **CorrGAN** (Mart*í, 2020):* sampling realistic **correlation matrices**
  - Models using the **signature of a *rough path*** and learn with fewer samples:
    - Generating Financial Markets with Signatures (Buehler, et al 2020)
    - Conditional Sig-Wasserstein GAN (Ni, et al 2020)

ML for Trading: end-to-end strategy workflow

On over 800 pages and 150 notebooks, the book demonstrates how ML can add value to algorithmic trading.



Four parts and 24 chapters cover:

- key aspects of data sourcing, financial feature engineering, and portfolio management,
- the design and evaluation of long-short strategies based on a broad range of ML algorithms,
- how to extract tradeable signals from financial text data like SEC filings, earnings call transcripts or financial news,
- using deep learning models like CNN and RNN with financial and alternative data, how to generate synthetic data with GANs, and training a trading agent using deep reinforcement learning.

Stefan Jansen, Founder of Applied AI

Strategy, research, implementation and training around ML for Trading

- Stefan is the founder and Lead Data Scientist at Applied AI. He advises Fortune 500 companies, investment firms and startups across industries on data & AI strategy, building data science teams, and developing machine learning solutions.
- Before his current venture, he was a partner and managing director at an international investment firm where he built the predictive analytics and investment research practice. He also was a senior executive at a global fintech company with operations in 15 markets.
- Earlier, he advised Central Banks in emerging markets, consulted for the World Bank, helped raise $35m from the Gates Foundation to cofound the Alliance for Financial Inclusion, and has worked in six languages across Asia, Africa, and Latin America.
- Stefan holds Master degrees in Computer Science from Georgia Tech and in Economics from Harvard and Free University Berlin and is a CFA Charterholder. He has also been teaching data science at Datacamp and General Assembly.

A.i Applied AI

Bibliography

Key Papers on Synthetic Data Generation for Finance

Balch, T., Assefa, S., Dervovic, D., Mahfouz, M., Reddy, P., Veloso, M., 2019. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls, in: NeurIPS'19 Workshop on Robust AI in Financial Services. Presented at NeurIPS 2019.

Bellovin, S.M., Dutta, P.K., Reitinger, N., 2018. Privacy and Synthetic Datasets (SSRN Scholarly Paper No. ID 3255766). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3255766

Buehler, H., Horvath, B., Lyons, T., Perez Arribas, I., Wood, B., 2020. Generating Financial Markets With Signatures (SSRN Scholarly Paper No. ID 3657366). Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3657366

Esteban, C., Hyland, S.L., Rätsch, G., 2017. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv:1706.02633 [cs, stat].

Franco-Pedroso, J., Gonzalez-Rodriguez, J., Cubero, J., Planas, M., Cobo, R., Pablos, F., 2018. Generating virtual scenarios of multivariate financial data for quantitative trading applications.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. arXiv:1406.2661 [cs, stat].

Kang, Y., Hyndman, R.J., Li, F., 2020. GRATIS: GeneRAting TIme Series with diverse and controllable characteristics. Statistical Analysis and Data Mining: The ASA Data Science Journal 13, 354–376. https://doi.org/10.1002/sam.11461

Bibliography

Key Papers on Synthetic Data Generation for Finance

Koshiyama, A., Firoozye, N., Treleaven, P., 2019. Generative Adversarial Networks for Financial Trading Strategies Fine-Tuning and Combination. arXiv:1901.01751 [cs, q-fin, stat].

Lin, Z., Jain, A., Wang, C., Fanti, G., Sekar, V., 2020. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. Proceedings of the ACM Internet Measurement Conference 464–483. https://doi.org/10.1145/3419394.3423643

Marti, G., 2020. CORRGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. https://doi.org/10.1109/ICASSP40776.2020.9053276

Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets. ArXiv.

Ni, H., Szpruch, L., Wiese, M., Liao, S., Xiao, B., 2020. Conditional Sig-Wasserstein GANs for Time Series Generation. arXiv:2006.05421 [cs, stat].

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., Xu, H., 2020. Time Series Data Augmentation for Deep Learning: A Survey.

Wiese, M., Knobloch, R., Korn, R., Kretschmer, P., 2019. Quant GANs: Deep Generation of Financial Time Series. arXiv:1907.06673 [cs, q-fin, stat]. https://doi.org/10.1080/14697688.2020.1730426

Yoon, J., Jarrett, D., van der Schaar, M., 2019. Time-series Generative Adversarial Networks. NeurIPS 2019.