Boosting Data Analytics Through High-Fidelity Synthetic Data

Xiaotong Shen
xshen@umn.edu
School of Statistics
University of Minnesota

The 5th NATIONAL BIG DATA HEALTH SCIENCE CONFERENCE, Columbia
Joint with Yifei Liu and Rex Shen [Shen et al., 2023]

Generative AI and Synthetic Data

- Synthetic data generation, propelled by generative AI, promotes paradigm shift for data analytics.
- Synthetic data: artificially created to closely mirror the characteristics and distribution of real data.
- MIT-gartner report [Gartner, 2022, Eastwood, 2023]: 60% of data utilized in AI and analytics will be synthetically generated by 2024, and synthetic data will surpass real data in AI models by 2030.
- As synthetic data gains prominence, questions arise concerning our data analytics paradigm: (1) how to utilize synthetic data; (2) its connection with raw data.
- Can we benefit from synthetic data for any analytic task?

Example

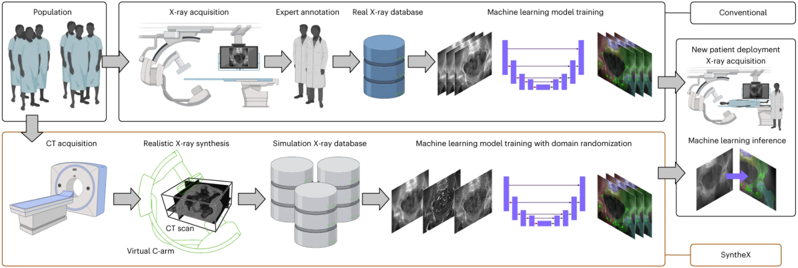


Figure 1: [Gao et al., 2023]: Machine learning models trained on synthetic data achieves state-of-art performances compared with real-data-trained models for medical imaging.

Challenges for Health Care Data

- Two importance aspects for healthcare data and medical research
  - Compliance—storage must be compliant with regulations–role based access control.
  - Efficacy.
- Data sharing becomes difficulty due to concern of security and privacy.
- Focus on the potential impact of generative AI: Can we effectively utilize synthetic data to enhance data privacy & efficacy.

Overview

- Synthetic data: produced by a generative model to replicate raw data, trained on raw data via pre-trained models with knowledge transfer from similar studies.
- Benefits
  - **(1)** privacy: privacy leakage when sharing real data...
  - **(2)** scarcity: limited size; expensive trials; time-consuming; imbalance...
- Generative models:
  - GANs [Goodfellow et al., 2014, Karras et al., 2019, Liu et al., 2020].
  - Normalizing flows [Dinh et al., 2016, Kingma and Dhariwal, 2018].
  - Diffusion models: DDPM for images [Ho et al., 2020, Rombach et al., 2022] and models for tabular data [Kotelnikov et al., 2023, Zhang et al., 2023]
  - LLMS such as OpenAI gpt family [Bubeck et al., 2023, OpenAI, 2023], Meta's llama, google's bard, anthropic's claude ...
- **Q1: Privacy.** Can synthetic data satisfy data privacy standard?
- **Q2: Efficacy.** Does a method gain accuracy on synthetic compared to raw data?
  - Diverging viewpoints: [Gao et al., 2023, Kotelnikov et al., 2023]
  - **Key:** trade-off between generation error and synthetic size.

Outline

1. **Q1: Privacy. Can synthetic data satisfy data privacy standard?**

2. **Q2: Efficacy. Does a method gain accuracy on synthetic compared to raw data?**

Data privacy

- Methods for privacy protection:
  (1) Methods (noise injection, sampling) satisifying differential privacy–gold standard: 2020 u.s. decennial census;
    - Adversarial attacks: membership, linkage, attribute inference, reverse engineering, aggregate, temporal, query-based...
    - Simple, low cost, effective.

  (2) Federated learning: secure multi-party computation;
  (3) Homomorphic encryption;
  (4) De-identification: still has high risks of disclosing due to Linkage, small size, data combination.
- Use of synthetic data may change way of protecting privacy.
  - Less privacy risk except reversed engineering attack.
  - No trade-off between statistical accuracy and level of protection.

Differential Privacy

- Differential privacy [Dwork, 2008] quantifies amount of privacy protection.
- Recognizes that privacy can be undermined even after data de-identification; e.g., "tallest person in room" is an identifier.
- Privatization mechanism m satisfies (ε, δ)-*differential privacy*:

$$\frac{p\big(m(\boldsymbol{z}) \in b | \boldsymbol{z} = \boldsymbol{z}\big)}{p\big(m(\boldsymbol{z}) \in b | \boldsymbol{z} = \boldsymbol{z}'\big)} \leq e^{\varepsilon} + \delta,$$

  For event b & adjacent $\boldsymbol{z}$, $\boldsymbol{z}'$ (substitute a single observation)

  - ε: privacy budget: δ: allowance. Small $\varepsilon \to$ strict privacy protection may reduce statistical accuracy of downstream analysis.
- Differentially private synthetic data: generated by a diffusion model with gaussian noise injection to gradient updates for stochastic gradient decent [Ghalebikesabi et al., 2023].

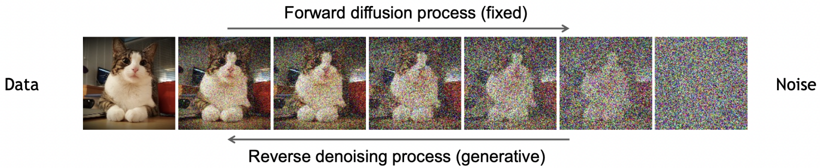
Outline

1. **Q1: Privacy.** **Can synthetic data satisfy data privacy standard?**
2. **Q2: Efficacy.** **Does a method gain accuracy on synthetic compared to raw data?**

Efficacy: Generational Effect

- **Raw sample**: $(\boldsymbol{z}_i)_{i=1}^n \sim$ cdf F.
- **Synthetic sample**: $(\tilde{\boldsymbol{z}}_i)_{i=1}^m \sim \tilde{F}$, produced from a generative model.
- **Method:** use synthetic $(\tilde{\boldsymbol{z}}_i)_{i=1}^m$ to perform any data analytics task.
- **Comparison:** accuracy of a method on $(\tilde{\boldsymbol{z}}_i)_{i=1}^m$ vs $(\boldsymbol{z}_i)_{i=1}^n$.
  - yes, $m = +\infty$ like simulations if no generation error ($\tilde{F} = F$).
    - Generation error: discrepancy between $\tilde{F}$ & F. high-fidelity: low error.
  - Generational effect: increasing m could diminish accuracy benefits or even a plateau due to generation error.
  - **Solution:** "syn" framework [Shen et al., 2023] — use empirical error measures to tune (Prediction error, Type-I error control) to choose optimum m.
  - Sample size expansion: $m >> n$.

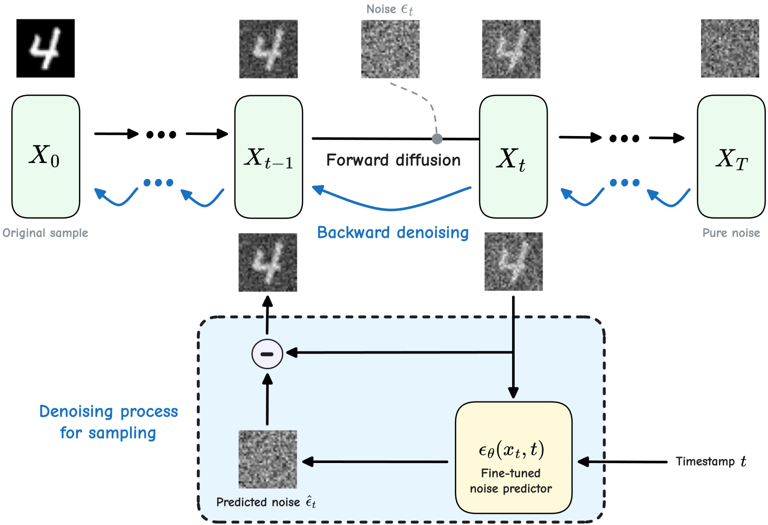
Generative Models: Diffusion



(image credit: https://cvpr2022-tutorial-diffusion-models.github.io/)

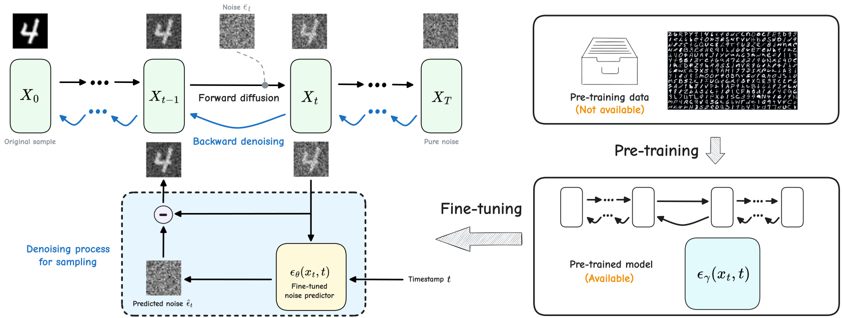
- Diffusion: Inject noises in forward process and denoise backwards.
- Forward: $\boldsymbol{x}_t = \sqrt{1-\beta_t} \cdot \boldsymbol{x}_{t-1} + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon_t}$; $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}_d, \boldsymbol{I}_d)$.
- Backward: $\boldsymbol{x}_{t-1} = \mu_\theta(\boldsymbol{x}_t, t) + \sqrt{\beta_t} \cdot \boldsymbol{\epsilon_t}, \quad \boldsymbol{\epsilon}_t \sim mcN(\mathbf{0}_d, \boldsymbol{I}_d)$,
  - $\mu_\theta(\boldsymbol{x}_t, t) = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\prod_{i=1}^t(1-\beta_i)}} \cdot \epsilon_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right)$.
  - $\beta_t \in (0,1)$ controls the amount of noise at step t.
  - $\epsilon_\theta(\boldsymbol{x}_t, t)$: a neural network parameterized by $\boldsymbol{\theta}$, predicting noise $\boldsymbol{\epsilon}_t$.
- Sampling is conducted by feeding noise into the backward process.

Denoising Network



Knowledge Transfer with Diffusion Models

Fine-tune pre-trained diffusion model on raw datasets.

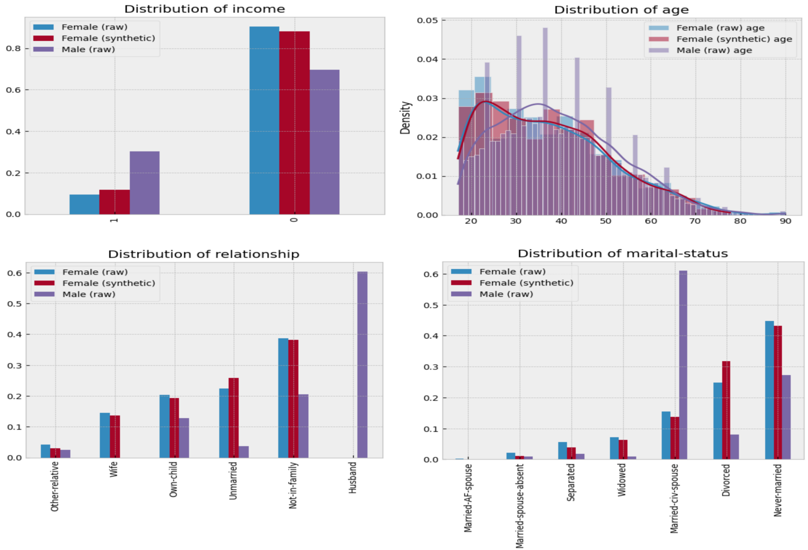


Q2: Efficacy. Does a method gain accuracy on synthetic compared to raw data?

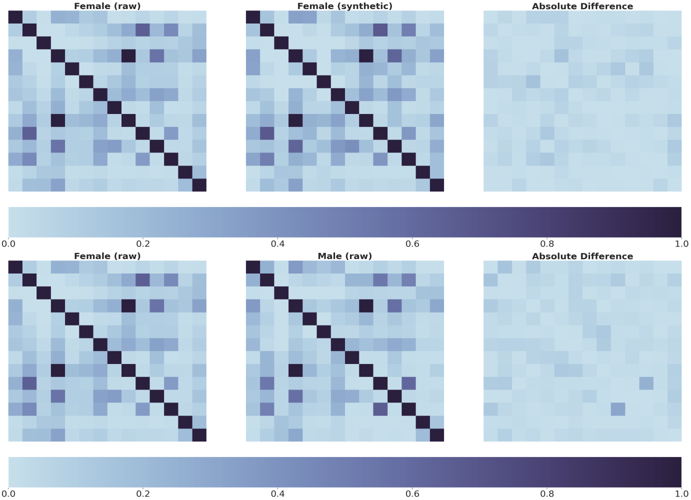
Classification on Adult-Female

- Adult dataset [Kohavi et al., 1996]:
  - Predict if annual income > 50k (classification) for adult-female data $(16,192)$ using 6 numerical & 8 nominal features: age, work class, final weight, # years in education, marital status, working hours per week, native country,.. .
- Boosting applies to syn-female: knowledge transfer from males
  - Pre-training (knowledge transfer): train tdm [Kotelnikov et al., 2023] on adult-male of size $32,650$, as our pre-trained generator.
  - raw: adult-female subset of size $n = 1,350$.
  - test: an independent adult-female subset of size $1,350$.
- Three prediction models:
  - Catboost: boosting on raw data, traditional.
  - Synboost: boosting on synthetic data with knowledge transfer.
  - FNN: feed-forward-network with knowledge transfer.
- Effect of synthetic size m: tune $m/n \in \{1, 2, \ldots, 30\}$ on misclassification error.
- Pre-training data are often unavailable in practice but Pre-trained models may be available $\rightarrow$ knowledge transfer via fine-tuning.

Marginal Distributions: Females vs Males



Pairwise Correlations: Females vs Males



Knowledge transfer → information gain: synthetic resembles raw females

Adult-Female: Information gain

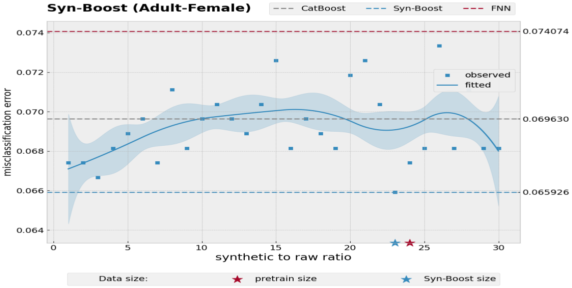
- Measure fidelity based on distributional distances:

	FID (gaussian)	Wasserstein-1	Wasserstein-2
Female (raw) vs male (raw)	1.971	1.968	2.125
Female (raw) vs male (pre-trained)	2.051	1.967	2.127
Female (raw) vs female (fine-tuned)	**0.249**	**1.170**	**1.399**

Table 1: FID-scores, Wasserstein-1, and -2 distances between the true female sample and other samples. "raw", "pre-trained", and "fine-tuned" denote raw data, synthetic data generated from a pre-trained model, and synthetic data generated from a fine-tuned model.

- Knowledge transfer via fine-tuning improves distribution closeness.

SynBoost: Sample Size Augmentation

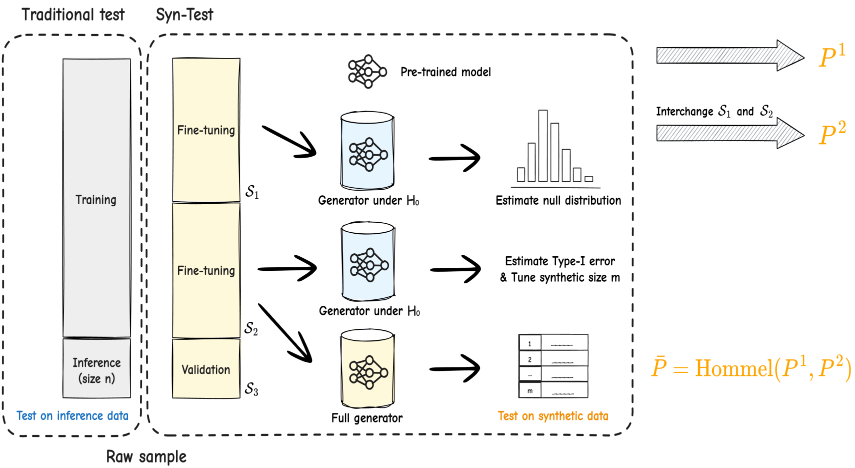


- Efficacy enhancement through synthetic size of larger size Through knowledge transfer.
- Generational effect: optimal $m \approx 23n$ (trade-off between generation error and accuracy).

Inference: Syn-Test

- Inference for complicated models, e.g., boosting, deep neural network (FNN): no asymptotic dist, lacks power, sample splitting [Dai et al., 2021, Wasserman et al., 2020] for black-box learners (dnn).
- Use synthetic data to boost power while controlling Type-I error.
- **Syn-Test:** training sample equally split to $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$.
  - Train or fine-tune generative models using $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ to estimate null distribution and Type-I error using MC approach.
  - Choose largest synthetic size m that Type-I error is controlled.
  - Testing with synthetic data of size m (usually $> n$).
  - Need a validation sample $\boldsymbol{S}_3$ for tuning to avoid overfitting.
- Trade-off between generation effect and estimation error.

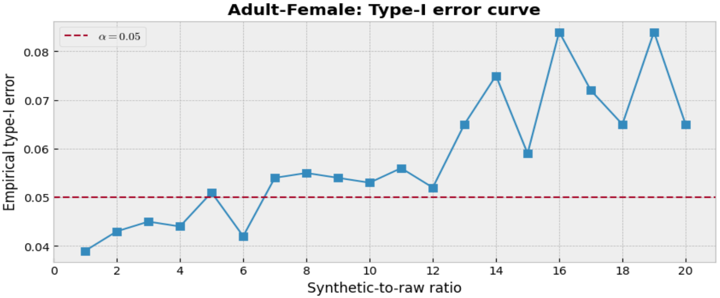
Syn-Test Illustration

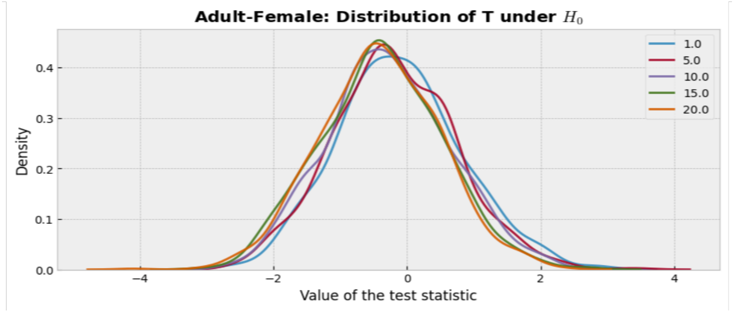


Inference: Syn-Test on Adult-Female

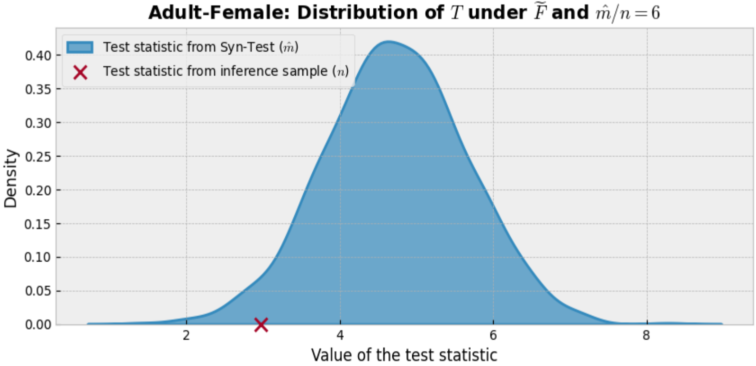
- **Dataset:** adult-female dataset [Kohavi et al., 1996] for predicting if their annual income exceeds 50k (binary classification).
- **Inference:** significance test for features age, education years, & working hours per week. black-box statistic [Dai et al., 2021] is applied here.
- Raw sample: training $(2,700)$ and inference size $(n = 300)$.
- Knowledge transfer: pre-train TDM on adult-male (larger size with distinct distributions) and fine-tune it on adult-female dataset.
- Tune $m/n \in \{1, 2, \ldots, 20\}$ with $\alpha = 0.05$.

Estimated Type-I Error and Null Distribution





Estimated Distribution of Test Statistic



Knowledge transfer $\rightarrow$ increase power through volume expansion.

Conclusion

- Impact of generative AI in data analytics is profound, presenting two primary advantages: diminished privacy concerns and enhanced statistical accuracy via sample size expansion through knowledge transfer.
- Statistical accuracy: Recognize the "generational effect" present in synthetic data.
- Development of large pre-trained models: Such advancements are crucial for furthering scientific research.
- This is just the start, with more advancements anticipated.

Thank you!

References: I

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Dai, B., Shen, X., and Pan, W. (2021). Significance tests of feature relevance for a blackbox learner. *arXiv preprint arXiv:2103.04985*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer.

Eastwood, B. (2023). What is synthetic data — and how can it help you competitively? *MIT Sloan School*.

Gao, C., Killeen, B. D., Hu, Y., Grupp, R. B., Taylor, R. H., Armand, M., and Unberath, M. (2023). Synthetic data accelerates the development of generalizable learning-based algorithms for x-ray image analysis. *Nature Machine Intelligence*, 5(3):294–308.

Gartner (2022). Is synthetic data the future of ai? *Gartner Newsroom Q&A*.

References: II

Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S. L., Wiles, O., and Balle, B. (2023). Differentially private diffusion models generate useful synthetic images. *arXiv preprint arXiv:2302.13861*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Kohavi, R. et al. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207.

Kotelnikov, A., Baranchuk, D., Rubachev, I., and Babenko, A. (2023). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR.

References: III

Liu, Q., Xu, J., Jiang, R., and Wong, W. H. (2020). Roundtrip: A deep generative neural density estimator. *arXiv preprint arXiv:2004.09017*.

OpenAI (2023). Gpt-4 technical report.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Shen, X., Liu, Y., and Shen, R. (2023). Boosting data analytics with synthetic volume expansion. *arXiv preprint arXiv:2310.17848*.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C., Rangwala, H., and Karypis, G. (2023). Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv preprint arXiv:2310.09656*.