Synthetic Data as the Future of AI Privacy, Explainability, and Fairness: An Introduction for Data Scientists and Data Executives

with Alexandra Ebert

Synthetic Data for Data Scientists

Things to remember about synthetic data

1. Not all synthetic data is automatically privacy safe.
2. Consider "build versus buy" solutions for synthetic data. You can use a synthetic data generator included in the additional resources with this handout. Keep in mind: according to the European Union's Joint Research Centre, commercial "buy" solutions beat open-source options by a huge margin.
3. Use production-like, real-world data to assess the utility of synthetic data for your organization. Don't just rely on "toy data" or publicly available datasets.

To assess synthetic data quality, accuracy, and utility

1. Conduct a visual comparison.
2. Measure the data's realism using a Turing test assessed by humans.
3. Compare marginal distributions and deviations between training data, synthetic data, and holdout data.
4. Use machine learning to best assess synthetic data quality.

To ensure safety in synthetic data generators

1. Ensure there is no one-to-one relationship between synthetic data and real-world data.
2. Have noise added into the process.
3. Protect the entire value range of your customers.
4. Empirically evaluate and measure the level of anonymity in your data.

Hands-on, exploring, and generating synthetic data

Synthetic Data Python Tutorials	https://github.com/mostly-ai/mostly-tutorials Great not only to expand on what we covered in the hands-on sections during the course (for example, learn how to deal with multitable setups during synthetic data generation), but also to explore synthetic data beyond privacy, from fairness and explainable AI to data augmentation, smart imputation, and more.
Synthetic Data Generators	Synthetic Data Vault: the most widespread open-source SD generator library
	Mostly.ai's synthetic data generator
Synthetic Data Benchmarks	To benchmark different SD generators regarding privacy and accuracy, check out this blog post.
Sample Synthetic Data	Humana's Synthetic Data Exchange. You can sign up and get a sample of granular synthetic health insurance data. This is a great way to see how rich SD can be in contrast to the type of dummy data one can get from other sources.

Initiatives to be aware of and participate in:

The IEEE Synthetic Data IC Expert Group	The IEEE Synthetic Data IC Expert Group: the world's first multinational expert group bringing together industry practitioners, academics, and regulators to work on a standard for synthetic data privacy and accuracy. Everybody is welcome to join and contribute at https://standards.ieee.org/industry-connections/synthetic-data/
NIST's Collaborative Research Cycle on Synthetic Data	There's a lot going on at this project and it's well worth checking out. One interesting element is the leader board summarizing how different SD generation algorithms, both open and closed source, performed in terms of privacy and accuracy.

Further readings and resources:

Synthetic Data Definition	https://mostly.ai/blog/define-synthetic-data
The *Data Democratization* Podcast	Check out this podcast, available on most podcast platforms, to learn more about synthetic data, responsible AI, AI governance, and how to accelerate AI innovation across an enterprise organization.
	Episode 40: *Synthetic data beyond privacy: data augmentation powered by AI*
Synthetic Data beyond Privacy	Fair synthetic data: https://mostly.ai/blog/tackling-ai-bias-at-its-source-with-fair-synthetic-data-fairness-series-part-4
	Data augmentation: https://mostly.ai/blog/data-augmentation
	Data simulation: https://mostly.ai/blog/data-simulation
	Smart imputation: https://mostly.ai/blog/smart-imputation-with-synthetic-data