

# I Can't Believe It's Not Real Data!

## An Introduction into Synthetic Data

Mason Egger

[@masonegger](https://twitter.com/masonegger)

Lead Developer Advocate - Gretel



gretel™

# Imagine

- You're a developer working on a web application (Django) at work that manages students in a classroom
  - Time to test!
  - Can't access production DB for security reasons
    - FERPA data is protected by law
  - Have to use a test DB with only a handful of records
  - **An edge case slips through that wasn't represented in the test DB**

# Imagine

- You're a Data Scientist trying to build a model
  - Figured out what you want to do, you want to try to predict a rare disease
  - Start looking for relevant data sets, but find out you don't have enough of the data you need
  - Have to train the model with the limited data set
  - **The model is unsuccessful due to size**
  - But wait! Someone in another hospital has a similar data set you think will work!
  - **Can't get access to it due to PII (Personally Identifiable Information) in the dataset**

# Data is a Scarce Resource

- Lack of access to usable testing data
  - 35% of DS time is spent in the “data gathering” stage
  - Data is inaccessible due to PII

- Limited Data Sets

- Lack of quality data can affect model training results
  - Prohibitively expensive or even impossible to collect more

- Biased Data

- Data can be skewed towards representation of subjects in a data set



# Solution: Synthetic Data

- **Synthetic Data:** *Synthetic data is artificially annotated information that is generated by computer algorithms or simulations, commonly used as an alternative to real-world data.* – Alex Watson
- Synthetic data is created by training a generative machine learning model on your data.

# Isn't That Just Fake Data?

- Synthetic data is different from “fake” or “mock” data
  - You may be thinking of Faker
- Fake/mock data may not be representative. It is purely random
  - Fake/mock data can be “too clean”
- Synthetic Data is generated from existing data
  - It will look and behave like the initial dataset
- Synthetic data can be nearly as representative as the initial dataset

# How Accurate is Synthetic Data?

- Unlike “fake” data, Synthetic data can be nearly as accurate as the real data
  - In some cases, [accuracy is improved](#)
- Downstream data consumers can readily make use of Synthetic Data
  - Eg: A classifier trained on Synthetic Data can get the same accuracy as a classifier trained on the original dataset
    - Did the user buy pizza or not?

Model	Accuracy	Recall	Prec.	F1
Logistic Regression on Synthetic Data	0.9450	0.2545	0.9100	0.9249
Logistic Regression on Real Data	0.9390	0.2471	0.9029	0.9206

# What Can I Use Synthetic Data For?

- Synthetic Data acts as an alternative to real-world data
- Any task where you need data, you can use Synthetic Data
  - Training models
  - Testing applications
  - Creating sample data for demos
  - Anonymizing data
  - and more!



# How Do I Use Synthetic Data?

- Make private data accessible and safely shareable
- Generate more samples with limited data sets
- Reduce bias in machine learning datasets

# Make Private Data Accessible & Shareable

- Data often contains PII (Personally Identifiable Information) making it *very risky* or even *illegal* for developers to work with
  - Developers and Data Scientists often don't want access to PII, developers want access to data that is relevant to their problem
- Generating a Synthetic Dataset allows you to have statistically similar data while removing the PII
  - This allows you to share your data, not only within the company but externally as well
  - Eg: You can have your data in an S3 bucket and then automatically generate synthetic data on access

# Augment Small Data Sets

- Not having enough of the right data is a serious bottleneck
  - Data is often your most valuable asset and collecting data is expensive and hard
- Synthetic Data allows you generate an unlimited amount of data based on a relatively small data set
  - Eg: You have a Machine Learning model and a small amount of data, you can use Synthetic Data to regularize your model training
  - Eg: From a testing standpoint, you can load/stress test your application

# Reduce bias in Data Sets

- Biased data is a *big* problem
  - Leads to inaccurate models, unfair results, and may even cause harm
- If you can identify the bias in your data, you can use Synthetic Data to balance your data set
  - [Reducing AI Bias with Synthetic Data in heart disease prediction models](#)
  - 68% male data, 32% female, 2:1 ratio
  - Use Synthetic Data to generate more female patients to balance the data set
  - Increase in accuracy from 88.5% to 96.7%
  - 6.17% more females with heart disease can now be accurately diagnosed

# Synthetic Data in Action

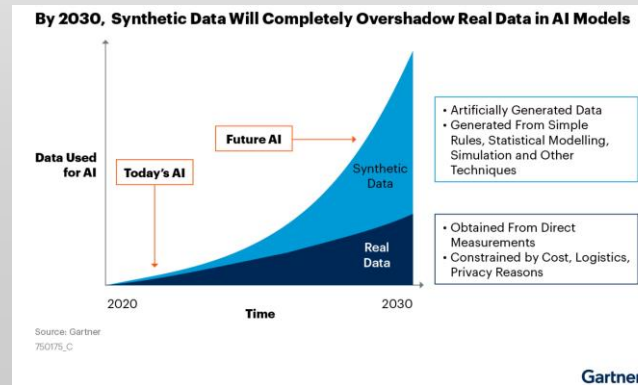
- **Automotive and Robotics** — leveraging synthetic data to create simulated environments for training robots, self-driving car software, and even testing safety and crash prevention technologies.
- **Financial Services** — creating synthetic time-series data to enable data sharing that doesn't compromise their customers' privacy
- **Cybersecurity and Infosec** — using synthetic data to train machine learning models to better detect rare events including fraud and cyber attacks
- **Healthcare and Life Sciences** — creating synthetic genomic data to fuel medical breakthroughs and encourage better medical care
- **Manufacturing** — using synthetic data to simulate complex supply chain operations and predict where failures may occur.
- And More!

# Current Challenges in Synthetic Data

- Highly dimensional datasets with hundreds or thousands of columns can be compute-intensive.
- Synthesizing relational datasets can require some manual configuration.
- Some privacy-preserving technologies such as differential privacy, require large amounts of data to provide strong privacy guarantees without degrading accuracy, and thus may not be appropriate for all datasets.
- Synthetic data generation requires time and effort.

# What does the future hold for Synthetic Data?

- By 2030 Gartner predicts that synthetic data will overshadow real data in AI models
  - Already happening today
  - Allows for easier compliance under data protections laws such as GDPR & CCPA
  - Reduces attack vector on data
- Eventually will solve the “cold start” problem
  - You have no data to start with



# Getting Started Using Synthetic Data

- Many resources available
  - <https://www.opensourceagenda.com/tags/synthetic-data>
  - <https://github.com/gretelai/awesome-synthetic-data>
- Open Source options available
  - gretel-synthetics
  - Synthetic Data Vault
  - Stable Diffusion - Images





# gretel-synthetics

- Open Source
- Multiple models
  - LSTM
  - GPT
  - CTGAN
  - More to come
- Train the synthetic data models yourself
  - You'll need a GPU
- <https://github.com/gretelai/gretel-synthetics>
- <https://synthetics.docs.gretel.ai/en/stable/>

# Gretel Cloud

- Don't have a GPU? Want to just try it out?
  - Try the [free tier](https://gretel.ai) at <https://gretel.ai>
- Train Synthetic Data in 3 lines of code

```
1 from gretel_trainer import trainer
2
3 dataset = "...
4
5 # Generate synthetic data in 3 lines of code
6 model = trainer.Trainer()
7 model.train(dataset)
8 print(model.generate())
```

# Additional Resources

- <https://docs.gretel.ai/>
- <https://github.com/gretelai/gretel-blueprints>
- <https://github.com/gretelai/fun-with-synthetic-data>

# Free Swag!



- Fill out <https://gretel.ai/dcus2022> and we'll mail you some stickers!
- Form closes a week after the premiere of this talk



# That's all for this time!

- Follow me on Twitter [@masonegger](https://twitter.com/masonegger)
- Follow Gretel on Twitter [@gretel\\_ai](https://twitter.com/gretel_ai) to keep up with all things Synthetic Data
- Have questions? Ask in our Slack! <https://grtl.ai/slack>
- Get started with Gretel <https://gretel.ai>
- Slides on my website, <https://mason.dev>