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

An Introduction into Synthetic Data

Mason Egger

[@masonegger]

Lead Developer Advocate - 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**

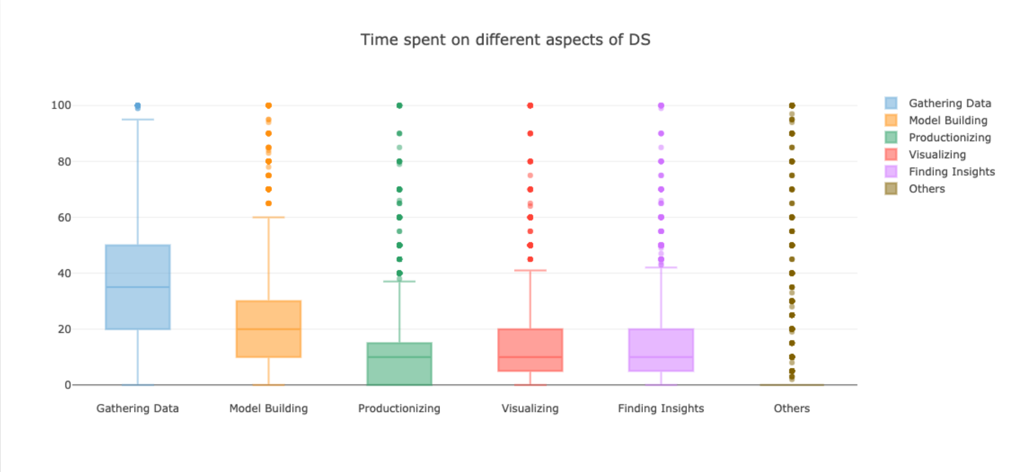
gretel™

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**

Common Data Challenges

- 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 sets 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.*

gretel™

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

gretel™

The Benefits of Synthetic Data

1. Make private data accessible and safely shareable
2. Generate more samples with limited data sets
3. Reduce bias in machine learning datasets

1. 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

gretel™

2. 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
  - Especially prevalent in the public sector, where poor data practices (such as storing data in "unreadable formats") causes for an abundance of inaccessible data

gretel™

3. 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

Is Synthetic Data Accurate?

- Unlike “fake” data, Synthetic data is nearly as accurate as the real data
  - In some cases, [accuracy is improved]
- Augmenting a training set with Synthetic Data had a mean ML accuracy less than 1% from their real-world equivalents

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!

gretel™

Getting Started Using Synthetic Data

- Many resources available
  - https://www.opensourceagenda.com/tags/synthetic-data
- Gretel makes it easy
  - All models are open source
  - No code options
  - Run in cloud or on-prem

gretel-synthetics

- Open Source
- Multiple models
  - LSTM
  - GPT-3
  - 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
- Many ways to run
  - Dashboard (No Code)
  - CLI
  - Python SDK
  - REST API

gretel™

Additional Resources

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

gretel™

Free Swag!

- Fill out https://grtl.ai/pyohio2022 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
- Follow Gretel on Twitter @gretel_ai to keep up with all things Synthetic Data
- Get started with Gretel https://gretel.ai

- Slides on my website, https://mason.dev

gretel™