deploy

by DigitalOcean

I Can’t Believe It’s Not Real Data!

An Introduction to Synthetic Data

Mason Egger

Lead Developer Advocate - Gretel.ai

Imagine

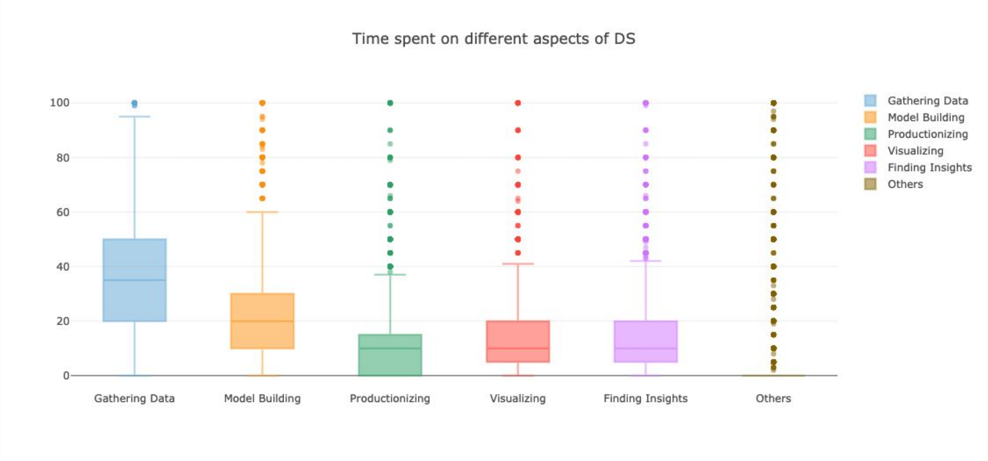
- You're a developer working on a web application (Django) at work that manages students in a classroom
  - Time to test!
  - Can't access production DB for security reasons
    - FERPA data is protected by law
  - Have to use a test DB with only a handful of records
  - **An edge case slips through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

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

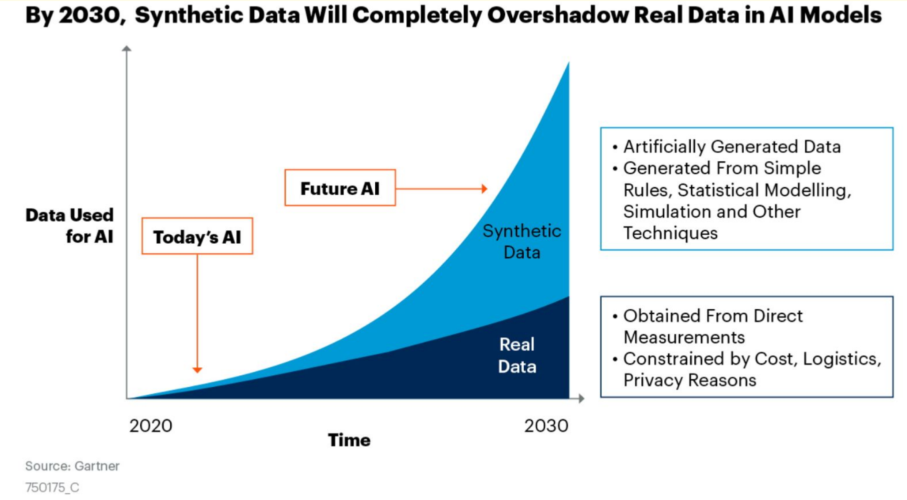
- Biased data is a big problem
  - Leads to inaccurate models, unfair results, and may even cause harm
- If you can identify the bias in your data, you can use Synthetic Data to balance your data set
  - Reducing AI Bias with Synthetic Data in heart disease prediction models
  - 68% male data, 32% female, 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!

Data is a Scarce Resource

- 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



By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models

- Artificially Generated Data
- Generated From Simple Rules, Statistical Modelling, Simulation and Other Techniques

- Obtained From Direct Measurements
- Constrained by Cost, Logistics, Privacy Reasons

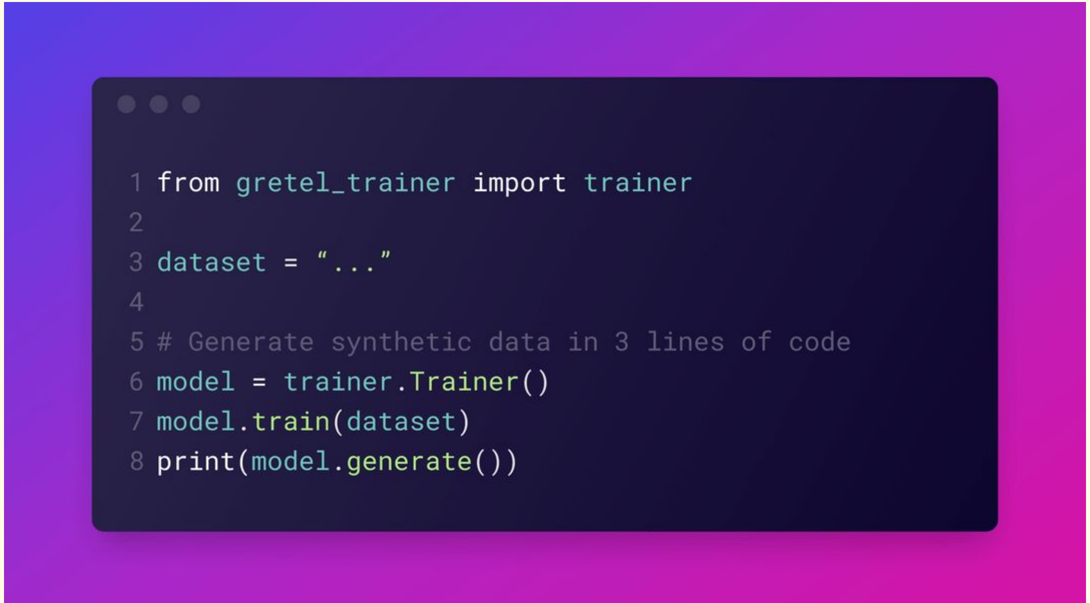
Source: Gartner
750175_C

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 Cloud

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



```python
from gretel_trainer import trainer

dataset = “...”

# Generate synthetic data in 3 lines of code
model = trainer.Trainer()
model.train(dataset)
print(model.generate())
```

Demo

by DigitalOcean

deploy

Free Swag!

- Fill out https://grtl.ai/deploy2022 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
- Have questions? Join the Synthetic Data Discord! https://grtl.ai/discord
- Get started with Gretel https://gretel.ai
- Slides can be found on my website https://grtl.ai/deploy2022-slides or scan the QR code

Thank you for attending deploy by DigitalOcean!

by DigitalOcean

deploy