Generating Synthetic Data

// FLATIRON SCHOOL

Agenda

- What is synthetic data?
- Key Ideas
- Why use it? (and when not to)

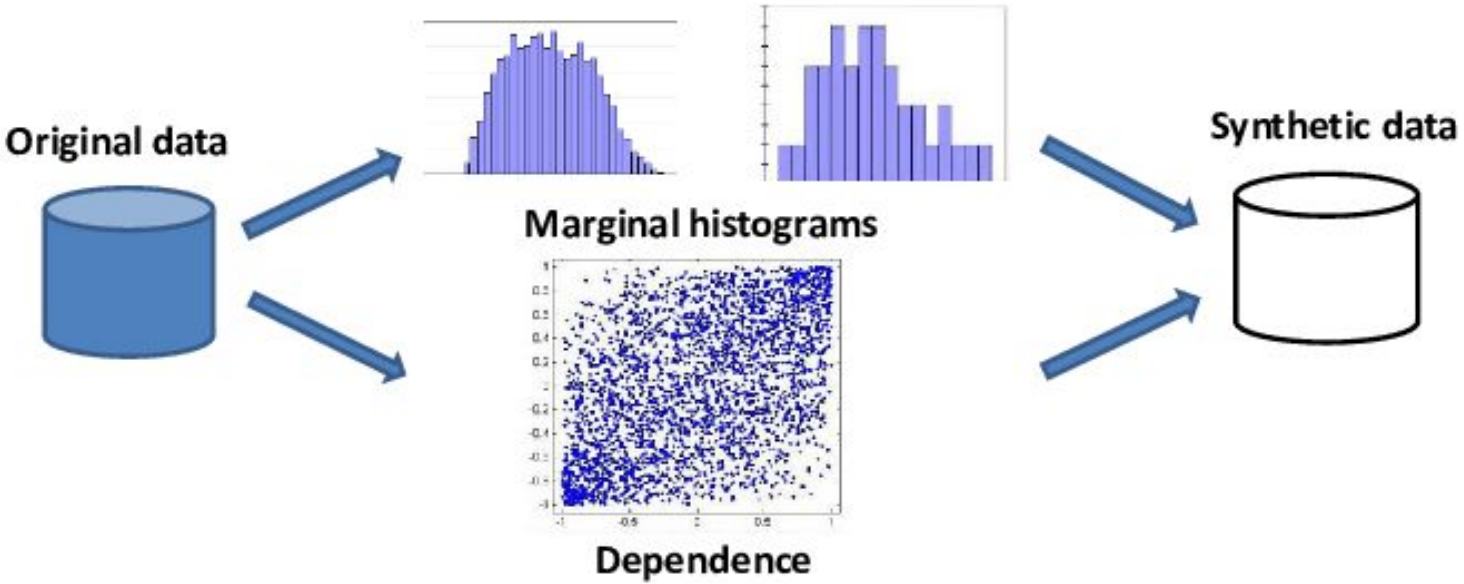
What is synthetic data?

// FLATIRON SCHOOL

Synthetic data...

- “Fake”data created to reflect some real world data or system
- Generated by a computer
- Similar characteristics to real world data
    - Univariate distributions
    - Structure (multivariate relationships)

Generation Process



Key Ideas

// FLATIRON SCHOOL

Terminology

- Data can be *structured* or *unstructured*
  - **Structured** - tabular format
  - **Unstructured** - images, video, text, etc
    - OpenAI GPT-3
    - OpenAI Dall-E 2
- ***Utility*** - how accurately simulated data reflects real data
  - Required utility depends on the use case

Generation Methods

- Real data
  - High utility
  - More resources required
- Existing model or knowledge
  - Lower utility
  - Fewer resources
- **Reproducibility** - generation process can be replicated

Synthetic Data Metrics

- Measure how well simulated data reflects real data
- Squared error
- Statistical Tests
  - Kolmogorov-Smirnov (KS) test
  - Chi-squared goodness of fit test
- Machine learning - *Distinguishability*

Why use it? (and when not to)

// FLATIRON SCHOOL

Efficient Data Access

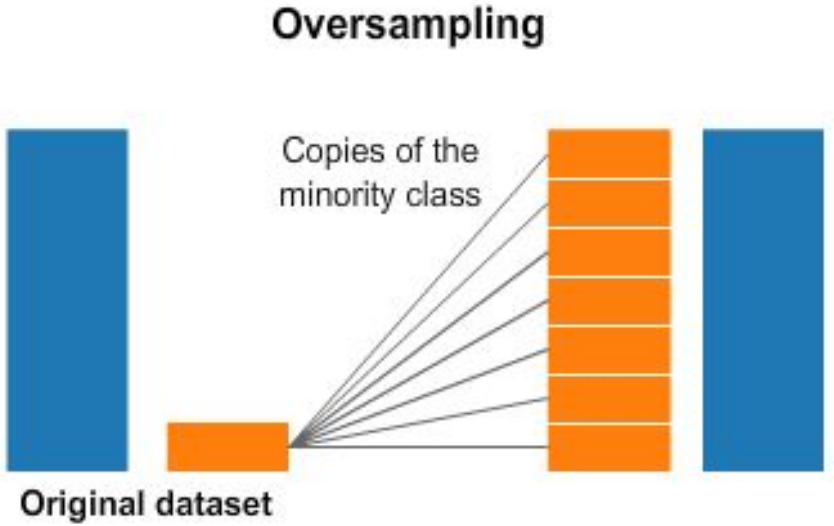
- Synthetic data is cheaper and faster
- Restrictions prevent access to data
- Solve privacy concerns
- Open data sources lead to:
    - Reproducibility
    - Innovation

Improve Analytics

- Make data open
- Test hypotheses
- Increase data size for modeling
- Account for edge cases and rare events

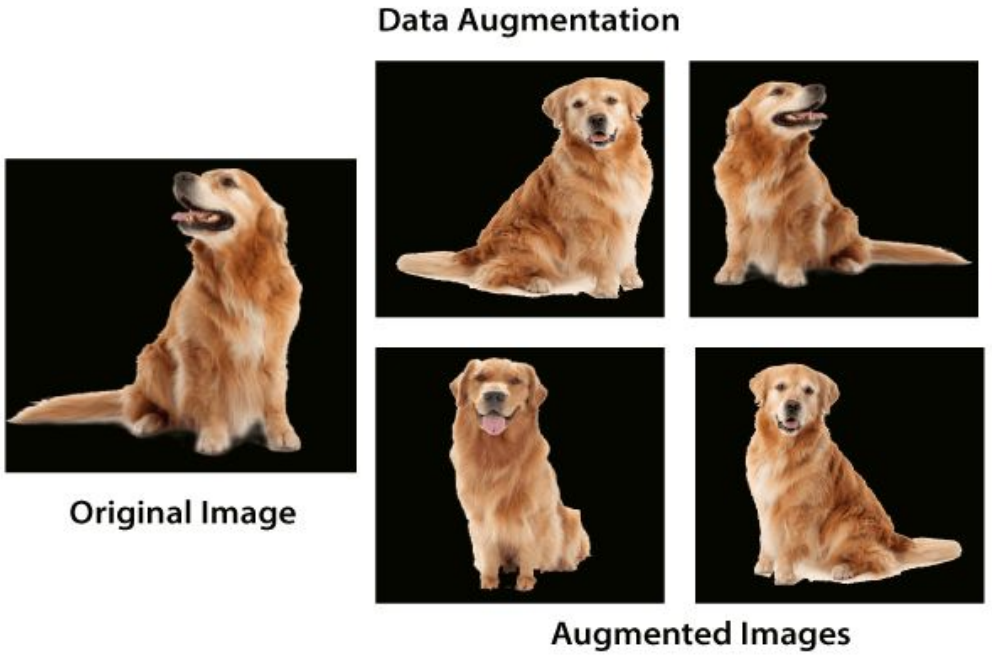
Machine Learning

- Synthetic data to solve for class imbalance
  - Fraud
  - Car crashes
  - Medical scans
- More data for training



Computer Vision

- Requires immense amount of data
- Impractical and expensive to collect
- Data augmentation creates synthetic samples for model

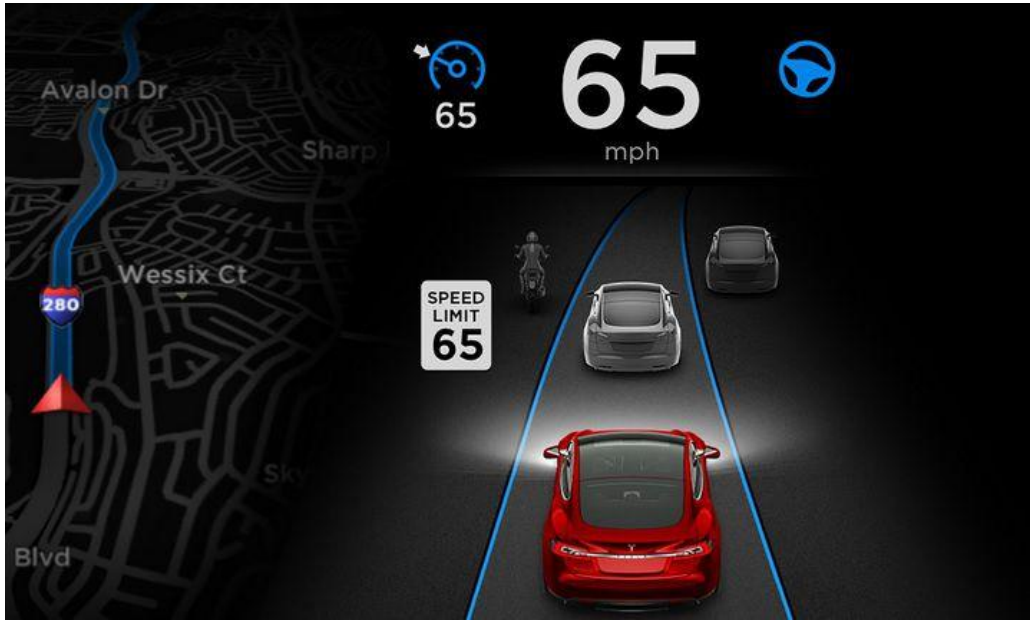


Healthcare

- Challenges
  - Health data is often private
  - Expensive to collect real data
- Synthetic data solution
  - Explore new digital health technologies
  - Sharing data open source

Autonomous Vehicles

- Limited data on edge cases
- Requires lots of training data
- Run simulations for rare events



Why NOT to use it

- Lack of resources
- Limited understanding of data and/or process
- Concern of privacy breach
- Misinterpreted as real data
    - Control narrative with misinformation
    - Always communicate data was simulated

References

Date, subhead or category

// FLATIRON SCHOOL

References

- Practical Synthetic Data Generation - OReilly
- Webinar: What is Synthetic Data?
- We need Synthetic Data - Medium Article