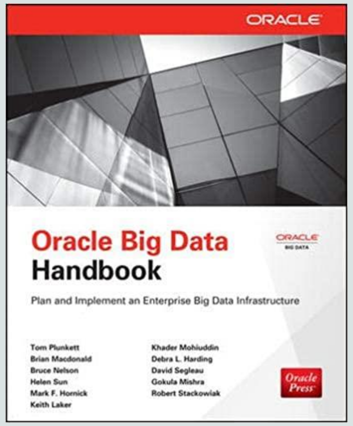
Beyond The Real

How Synthetic Data Fuels the Future of AI/GenAI

Gokula Mishra

August 2024

Gokula Mishra



Author

DATA JUICE
101 STORIES OF HOW ORGANIZATIONS ARE SQUEEZING VALUE FROM AVAILABLE DATA ASSETS
DOUGLAS B. LANEY
MAKE INSIGHTS ACTIONABLE WITH AI & BI
Collective advice from 20 + Data Leaders

Contributor

Roles

- VP, Data Science & AI/ML (Former), Direct Supply
- Global Leader Data Analytics & Supply Chain, McDonalds
- Founding Editor: The CDO Magazine
- Keynote speaker CDOIQ Symposium (MIT)
- Chair Illinois Chief Data & Analytics Officer Forum

- https://www.linkedin.com/in/gokulamishra/

Agenda

The Data Bottleneck

The Rise of Synthetic Data

Unlocking the Potential

Q&A

The Data Bottleneck

Challenges of Real-World Data for AI/GenAI

The Data Bottleneck
Challenges of Real-World Data for AI/GenAI

Inaccessability

“By 2024, 75% of the Global Population will have Its Personal Data Covered Under Privacy Regulations.” Gartner

Bias & Representation

Quality & Quantity

The Rise of Synthetic Data

Crafting Artificial Reality for AI/GenAI

Overcoming Data Challenges and unlocking new possibilities

What is Synthetic Data?

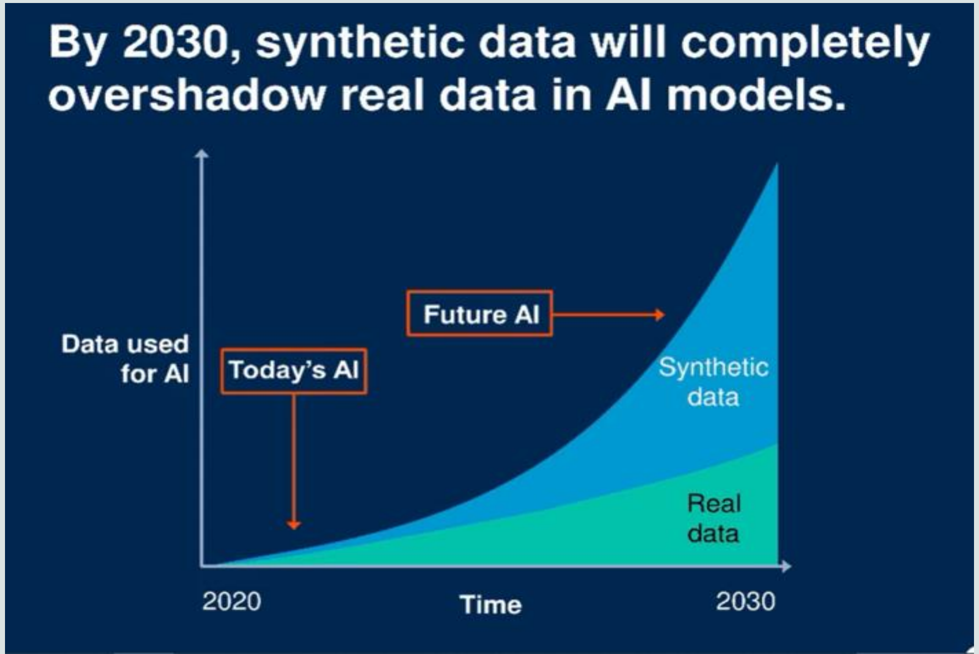
- Synthetic data is artificially generated, most of the time using real-world data collected from real-world events
- Now multi-modal synthetic data is created by advanced generative algorithms learning from real world multi-modal data samples

- Advantages:
  - Scalability: Easily scaled to create large datasets.
  - Privacy friendly alternative to real-world data
  - Offers flexibility in creating variations such as:
    - Lighting, poses in images
    - Filling missing pieces in photos and videos…
  - Flexibility: Allows for the creation of diverse and controlled datasets

According to Gartner

- Synthetic Data is data that is artificially generated rather than obtained by direct measurement produced through algorithms that model the statistical properties of real-world data, enabling the creation of new, realistic datasets without compromising privacy or security.

- Gartner predicts that over 60% of data used to train AI models will be synthetic, highlighting its growing importance in various sectors like healthcare, finance, and retail data by 2024.

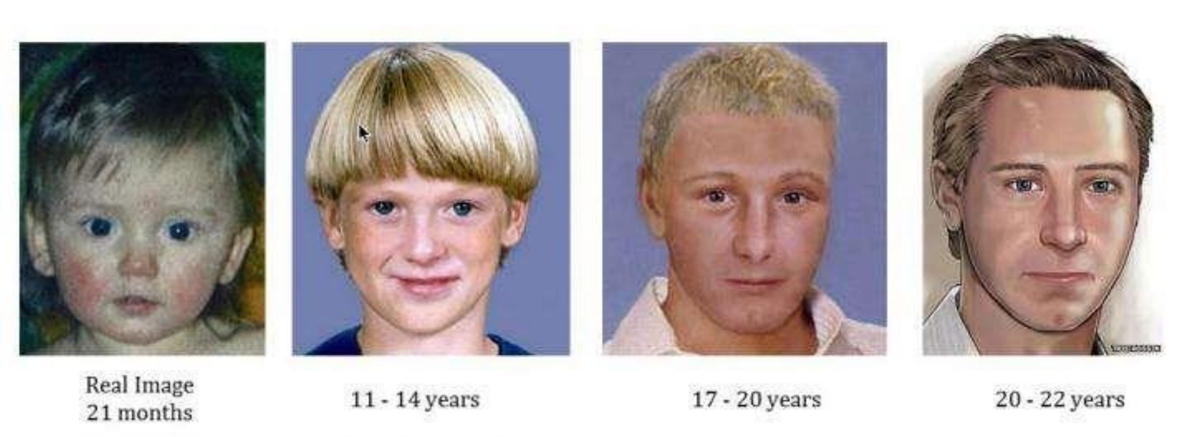


Unlocking the Potential

Applications of Synthetic Data in AI/GenAI Growth

Use Cases

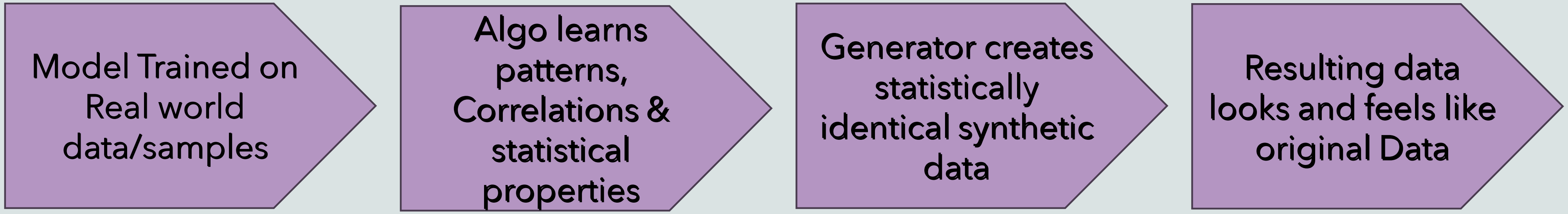
- Training AI Models in privacy-preserving environment
- Accelerating AI Development and Testing
- Improving Model Performance and accuracy
- Real World Data set augmentation
- Data Sharing and Collaboration
- De-identification
- Enhanced design and simulation in product development
- Training computer vision algorithm with synthetic images
- Help to find missing person



Training ML Models

- **Autonomous Vehicles:** Simulates various driving scenarios for safer testing.
  - Face Recognition
  - Hands-on-wheel Detection
  - Drowsy Driver/Eye gaze on the road
- **Financial Services**
  - Simulation of economic conditions/market fluctuations to enhance risk & decision models
  - Simulate various financial fraud scenarios
  - Enhance trading algorithm
- **Life Sciences & Healthcare**
  - Clinical Trial: Simulation of patient data for initial testing phases and protocol development for clinical trial
  - Drug Discovery: Accelerates the creation of diverse datasets for training models.
  - Medical Imaging: Generates diverse medical images for improved diagnosis models
- **Real-World Success**: MIT Study on Synthetic Data
  - Study used 150,000 synthetic video clips for AI training[3]
  - Models trained on synthetic data excelled, especially with simpler backgrounds[3]
  - Demonstrated potential in enhancing real-world machine learning applications[3]
  - Addressed ethical and privacy challenges in AI development[3]

Synthetic Data Generation Technology



Model Trained on Real world data/samples

Algo learns patterns, Correlations & statistical properties

Generator creates statistically identical synthetic data

Resulting data looks and feels like original Data

- **Random Sampling and Noise Addition**
- **Rule & Constraints-based generation**
- **Generative Models** (such as GANs and VAEs): Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), learn the underlying data distribution and generate new samples that resemble the original data.
- **Mixture Model and Interpolation:** Mixture models combine multiple distributions to generate data points. Interpolation techniques can also be used to create data points that lie between existing data points.
- **Data Transformation and re-sampling:** transforming or resampling existing data to create new instances. For example, in image data, you could apply rotations, translations, and other transformations to generate new images.
- **Markov Chain Monte Carlo (MCMC) Methods:** MCMC methods generate data sequences based on the transition probabilities of a Markov chain. These methods are useful for simulating sequential data or generating samples that follow a specific distribution.
- **Simulators and Simulations:** They generate synthetic data by modeling the underlying mechanisms. These methods are commonly used in scientific research and for simulating complex scenarios that are difficult to observe directly.

Vendors

- **Synthetic Data Generation Market:**
  - 2023 was $300 million
  - 2028 projected to be $2.1B according to a report from MarketsAndMarkets.
- **Factors driving the synthetic Data Market:**
  - especially opportunities in heavily regulated industries
  - address gaps for testing and training performant AI models
  - Rapid growth in AI/GenAI products demand in corporations
- **Vendors**:
  - Zuno Synth by Cognida, Subsalt, Ydata, Tonic, Mostly.ai, Gretel, Datomize, GenRocket, Betterdata, etc.
  - Opensource products such as Synner, Datagene, mirrorGen etc…

Challenges and Future Directions

- **Challenges**
  - Ethical Considerations: Using Synthetic Data responsibly
  - How synthetic data will be integrated into AI workflows & Real Data?
  - Lineage of synthetic data before and after it is integrated
  - Quality of synthetic data in multi-modal data landscape
  - Being Sober – willing to experiment but not drink the Kool-Aid blindly
  - Understanding the constraints & boundaries of technique used – especially multi-technique solutions inside a product
  - Model collapse: Training on generated data alone could lead to a degradation of AI models. And hallucinate more often, fail to answer questions and performance would falter.

- **Future Directions**
  - Innovations: Emerging technologies in synthetic data generation, and quality assessment
  - Contributing to explainable AI and algorithm bias elimination
  - Transforming how we approach AI research and application development
  - Enabling data-driven innovation while prioritizing privacy and ethics

Q&A