Welcome To

Synthetix.ai

•••

Unlocking Insights, Preserving Privacy

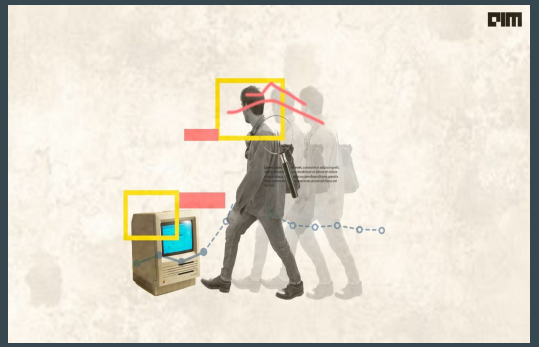
github.com/synthetic-ai/franklin-templeton

synthetixai@gmail.com

Data is #1 bottleneck in AI

Data Collection

1. Limited Avenues
2. Privacy Concerns
3. Biased Data



Data Labelling

1. Human Errors
2. Inconsistent
3. Labelling Bias

Data Entry/Processing

1. Human Errors
2. Slow
3. No Standardization

Solution:
Synthetic Data

Artificial Data which retains the properties of the real data while protecting data privacy and security

Real data is typically limited in size, difficult to access, and may not reflect the full range of possible values or behaviors, making it difficult to manage and analyze.

Main data categories in Finance:

- Tabular Data
- Time series Data

Benefits:

- Greater control over quality and format
- Increased privacy and security
- Lower costs
- Better performance in models
- Faster turnaround time

Tackling the problem on its chin

Privacy Regulations

Protect individuals' privacy and protection against re-identification attacks and promote a safe data access. Synthetic data enables data access compliant with GDPR and CCPA.

Limited Availability

Financial institutions often lack sufficient data on rare events like market crashes or fraud scenarios, which are crucial for developing robust risk management models.

Unlock Data Sharing

Data silos and lack of data sharing across organizations within the finance industry hinder access to diverse and comprehensive datasets.

Understanding the market

Key Market Trends and Opportunities

- According to PwC, **global financial institutions** lose over $1 trillion annually due to **fraud** and other economic crimes.
    - Synthetic data can be used to generate realistic fraud scenarios for training and testing fraud detection models, improving their accuracy and effectiveness.
- The **global consumer lending market** is expected to reach $24.8 trillion by 2025 (Source: Wintergreen Research).
    - Synthetic data can create diverse synthetic customer financial profiles, enabling the development of fair and unbiased credit scoring models.
- Regulatory bodies like the Federal Reserve and the European Banking Authority mandate **regular stress testing** for financial institutions.
    - Synthetic data can simulate various economic scenarios and market conditions, enabling comprehensive stress testing and risk management strategies.
- The **global wealth management** market is projected to reach $1.1 trillion by 2027 (Source: Allied Market Research).
    - Synthetic data can generate diverse synthetic investment profiles, enabling the development of personalized financial planning and advisory services.
- The **global insurance industry** is worth over $6 trillion, with underwriting and pricing being critical functions (Source: IBISWorld).
    - Synthetic data can create realistic synthetic policyholder data, enabling the development of accurate risk models and pricing algorithms.Key Market Trends and Opportunities

Target audience

Overview of the Finance and Insurance Industry

- The finance and insurance industry is a critical component of the global economy, with a combined market size of over $22 trillion as of 2022 (Source: IBISWorld)
- Data plays a crucial role in making informed decisions, managing risks, and providing personalized services in this industry.
- However, using real customer data poses significant challenges due to data privacy regulations (e.g., GDPR, CCPA) and the risk of data breaches, which can lead to hefty fines and reputational damage.

- Insurance Companies
  - Expand risk factors and demographic variables for robust risk assessment
  - Ability to account for extreme and rare events
  - Safeguard customer privacy
  - Data for software testing
  - Combat bias in data driven algorithms
- Banking and Finance Companies
  - Train Anti Money Laundering models
  - Simulate risk scenarios for fraud detection
  - Evaluate portfolio performance over

Other Markets

Our target audience includes all companies which use data for analytics/training but face problems in procurement of required data

- Health Care
  - Payer Claim data is protected by HIPAA which makes it difficult to develop models in the healthcare space
  - Virtual clinical trials, drug development, disease development and prediction, medical imaging and radiology
- Cybersecurity and network security
  - Sharing network traffic and attack data can compromise network security and expose vulnerabilities
  - Intrusion detection, threat modelling, vulnerability testing and penetration testing
- Autonomous Vehicle and Robotics
  - Collecting and annotating data can be expensive, time consuming and potentially dangerous
  - Create various simulated environments for testing autonomous vehicles, drones, and robots in a controlled and safe environment
- E-commerce and Marketing
  - Purchase histories, browsing patterns and personal preferences are subject to privacy regulations and customer consent
  - Virtual clinical trials, drug development, disease development and prediction, medical imaging and radiology

Our approach: GANs

CTGAN (Conditional Tabular GAN)- Credit Card Fraud Detection:

FEATURES

- Specialized GAN for generating high-quality synthetic tabular data.
- Handles imbalanced data and mixed data types (continuous and categorical).

MODEL TRAINING:

- **Generator-** Learns to create synthetic samples; mimics real data distribution.
- **Discriminator-** Distinguishes between real and synthetic samples.
- **Training Process:** Generator and Discriminator improve through adversarial training; Uses conditional vectors to ensure balanced representation of minority classes

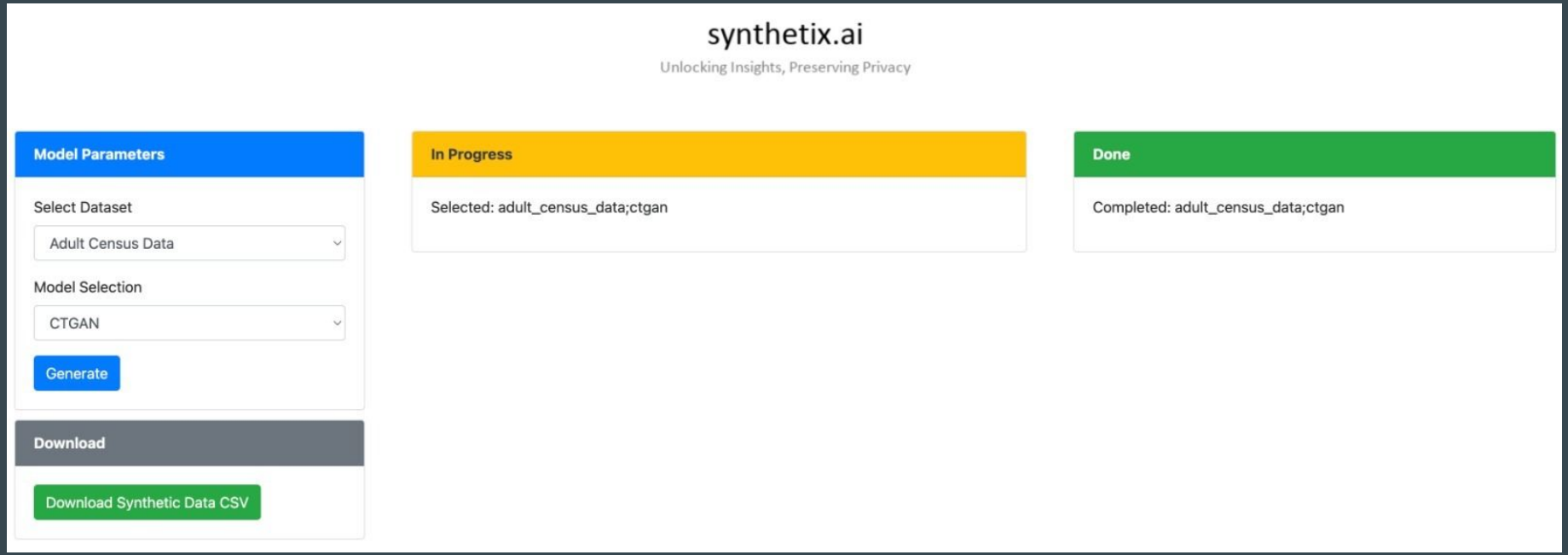
TimeGAN (Time-series GANs)- Stock Price Data:

FEATURES

- Combines GANs with recurrent neural networks (RNNs) for sequential data.
- Preserves temporal dynamics and correlations in time series data.

MODEL TRAINING:

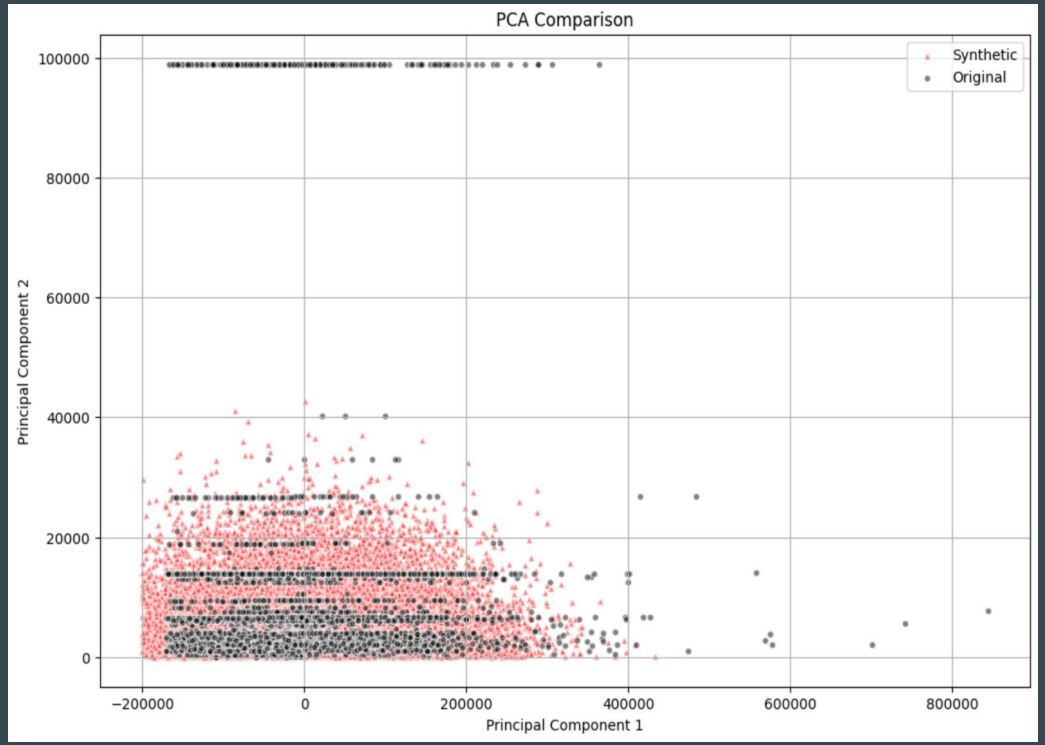
- **Embedding Network-** Encodes time series into a latent space.
- **Generator and Discriminator-** Generate and distinguish between real and synthetic sequences.
- **Adversarial and Supervised Losses:** Combines adversarial loss with supervised loss to maintain temporal dynamics.



- ➔ Self Service
- ➔ Granular Control
- ➔ Large Scale Data-Integration
- ➔ Domain Specific

Evaluation Metrics

	mean_diff(%)	std_diff(%)
age	0.000121	0.003302
workclass	0.006623	0.085787
fnlwgt	0.000913	0.000116
education	0.086580	0.182109
education-num	0.024354	0.249180
marital-status	0.041328	0.005342
occupation	0.037125	0.022192
relationship	0.000057	0.000357
race	0.024593	0.199508
sex	0.005574	0.002868
capital-gain	0.042823	0.000064
capital-loss	0.004981	0.000813
hours-per-week	0.000187	0.002652
native-country	0.050303	0.549395
target	0.014668	0.016460



Column Correlations

	age	workclass	fnlwgt	education	education num	marital status	occupation	relationship	race	sex	capital gain	capital loss	hours per week	native country
age	0.000	2.329	1.030	2.035	1.855	1.340	0.345	1.211	1.700	0.971	0.950	0.958	0.839	0.373
workclass	2.329	0.000	0.511	0.510	0.302	0.784	0.173	1.066	0.666	0.869	0.480	1.697	0.432	0.269
fnlwgt	1.030	0.511	0.000	0.356	0.201	0.968	6.439	1.582	1.968	0.745	0.351	0.030	1.666	0.473
education	2.035	0.510	0.356	0.000	0.226	0.015	1.101	0.507	0.717	0.226	0.383	0.177	0.190	0.174
education-num	1.855	0.302	0.201	0.226	0.000	0.463	0.398	1.008	1.038	4.578	0.375	0.414	0.341	0.133
marital-status	1.340	0.784	0.968	0.015	0.463	0.000	0.478	0.695	1.008	0.831	0.869	0.878	0.936	0.038
occupation	0.345	0.173	6.439	1.101	0.398	0.478	0.000	0.687	1.529	0.519	0.596	0.736	0.296	0.137
relationship	1.211	1.066	1.582	0.507	1.008	0.695	0.687	0.000	1.167	1.508	1.098	0.969	1.101	1.313
race	1.700	0.666	1.968	0.717	1.038	1.008	1.529	1.167	0.000	1.131	1.293	1.290	1.140	0.105
sex	0.971	0.869	0.745	0.226	4.578	0.831	0.519	1.508	1.131	0.000	0.986	0.805	0.607	0.227
capital-gain	0.950	0.480	0.351	0.383	0.375	0.869	0.596	1.098	1.293	0.986	0.000	1.071	1.026	0.526
capital-loss	0.958	1.697	0.030	0.177	0.414	0.878	0.736	0.969	1.290	0.805	1.071	0.000	0.921	0.781
hours-per-week	0.839	0.432	1.666	0.190	0.341	0.936	0.296	1.101	1.140	0.607	1.026	0.921	0.000	4.293
native-country	0.373	0.269	0.473	0.174	0.133	0.038	0.137	1.313	0.105	0.227	0.526	0.781	4.293	0.000

Evolution of Space

Early Days (2010-2015)

- The concept of synthetic data generation emerged from academic research in the early 2010s
- Initial startups like Synthetic Data Corp (now Syntelinc) and Tonic.ai (acquired by OpendataBot in 2021) were established to commercialize synthetic data solutions for industries like healthcare and finance.
- Adoption was slow due to skepticism about the accuracy and reliability

Growth Phase (2016-2020)

- Increasing concerns over data privacy, fueled by regulations like GDPR and CCPA, drove the demand
- Early acquisitions and consolidation began, with Delphix acquiring Synthetic Data Corp in 2018 and OpendataBot acquiring Tonic.ai in 2021 (Source: Crunchbase).

Recent Developments (2021- present)

- Major tech companies like Microsoft, Google, and Amazon have also invested in synthetic data generation capabilities.
- High-profile acquisitions have taken place, such as Microsoft's acquisition of Synthetic Data Corp (from Delphix) in 2022.
- Specialized solutions have emerged for industries like autonomous vehicles, manufacturing, and retail, indicating the broadening adoption of synthetic data across various sectors.

Competitor analysis

	Mostly.ai	Open Source Libraries	Synthetix.ai
Cost	Follows a usage-based pricing model, with costs depending on the volume of data generated and the complexity of the use case	Typically free to use However, cost of integrating and maintaining, and the expertise required, should be considered.	We plan to offer both usage-based pricing and subscription-based pricing models. Pricing competitive to market
Ease of use	Provides a user-friendly web interface and APIs However, their platform may have a steeper learning curve for users without prior experience in data science or machine learning	Documentation and community support can vary, making it challenging for non-experts to use these libraries effectively.	We will prioritize user-friendly interfaces, comprehensive documentation, and intuitive workflows Our platform will be designed with a low barrier to entry, enabling easy integration and adoption within existing data pipelines
Features	Supports a wide range of data types, including structured, unstructured, and time-series data. Has industry-specific solutions tailored for verticals like finance, healthcare, and retail	Offer a wide range of synthetic data generation techniques May lack industry-specific features or pre-built solutions tailored to specific use cases May also have limitations in terms of scalability, performance, and support	We will offer industry-specific solutions tailored to finance, insurance, healthcare, and other sectors, with pre-built features and configurations for common use cases. We will continually invest in R&D to stay ahead of the curve and introduce innovative features to our platform.

Conclusion

- Synthetic data presents a compelling solution for the finance and insurance industry, enabling data-driven decision-making while addressing data privacy and regulatory concerns.
- Its ability to drive innovation, compliance, and risk management makes it a valuable asset in an increasingly data-driven and regulated industry.
- With the growing demand for synthetic data and its diverse applications, a startup in this space has significant growth potential and the opportunity to shape the future of data-driven finance and insurance.

Market Challenges and Considerations

- **Ensuring data quality and representativeness** is crucial, as synthetic data must accurately reflect the underlying patterns and distributions of real data.
- **Gaining trust and acceptance** from industry stakeholders, regulators, and customers may require education and demonstration of the benefits and safeguards of synthetic data.
- **Regulatory compliance** and evolving data privacy landscape must be closely monitored, as regulations may impact the generation and use of synthetic data.
- **Competition from established players** and alternative solutions, such as data anonymization and federated learning, may pose challenges for startups in this space.

Focus Areas

- **Investing in research and development** to develop cutting-edge techniques for generating high-fidelity synthetic data, ensuring data quality and representativeness.
- Identifying and specializing in high-value use cases, such as insurance underwriting, personalized finance, or trading applications, to differentiate from competitors and **build domain expertise.**
- **Implementing robust data quality measures** , security protocols, and regulatory compliance frameworks to **build trust** and address customer concerns.
- **Partnering with financial institutions** , insurers, technology providers, and industry associations to gain access to domain expertise, data sources, and distribution channels.
- **Allocating resources to educating customers** , demonstrating the benefits of synthetic data, and addressing concerns through proof-of-concept projects and case studies.
- Staying ahead of market trends, emerging technologies, and regulatory changes by continuously **monitoring the landscape and adapting strategies** accordingly.

Roadmap


May
July
Sep
Nov
Jan
July
Sep
Nov
Develop core synthetic data generation platform
Launch targeted solutions - fraud detection, credit scoring, trading
Expand to adjacent industries - insurance, risk management
Explore emerging technologies (blockchain, DeFi) and global expansion

$20 mm

Expected funding our Series A funding round

Allocation of Funds

40%
Research and Development

25%
Product Development and Commercialisation

15%
Talent Acquisition and Team Expansion

10%
Strategic partnerships and Acquisitions

5%
Marketing and Sales

5%
Infrastructure and Operations

The Team

Shashwat Mishra, CEO

Berkeley MFE, Mathematics and Computing at IIT Guwahati with minor in Robotics and AI, Ex-Mastercard, Hilabs

Sahil Gupta, CFO

Berkeley MFE,
Ex-Quant@Goldman Sachs

Rashi Mohta, CTO

Berkeley MFE,
Ex-Quant@JP Morgan,
Mathematics and Computing at IIT Guwahati

Harit Gupta, COO

Berkeley MFE,
Ex-Quant@Goldman Sachs,
Mathematics and Computing at IIT Guwahati