Crowdruption

BETTER DATA RE:IMAGINED.

Innovative & Disruptive Data Science Solutions For Fair and Responsible AI

Raj Mehta, Founder & CEO

rajcm365@gmail.com

Key Challenges Faced by the Data Science/AI/ML Industries

Inherent/Pre-Existing Biases in Real-World Data

- Researchers from University of Southern California (USC) found biases rampant in ≈38.6% of 'facts' used by AI.

Algorithmic Biases

- Gartner (2018 study) predicted that up to 85% of AI projects could potentially deliver erroneous outcomes due to biases in data, algorithms, or the teams responsible for managing them.

Data Privacy/Security

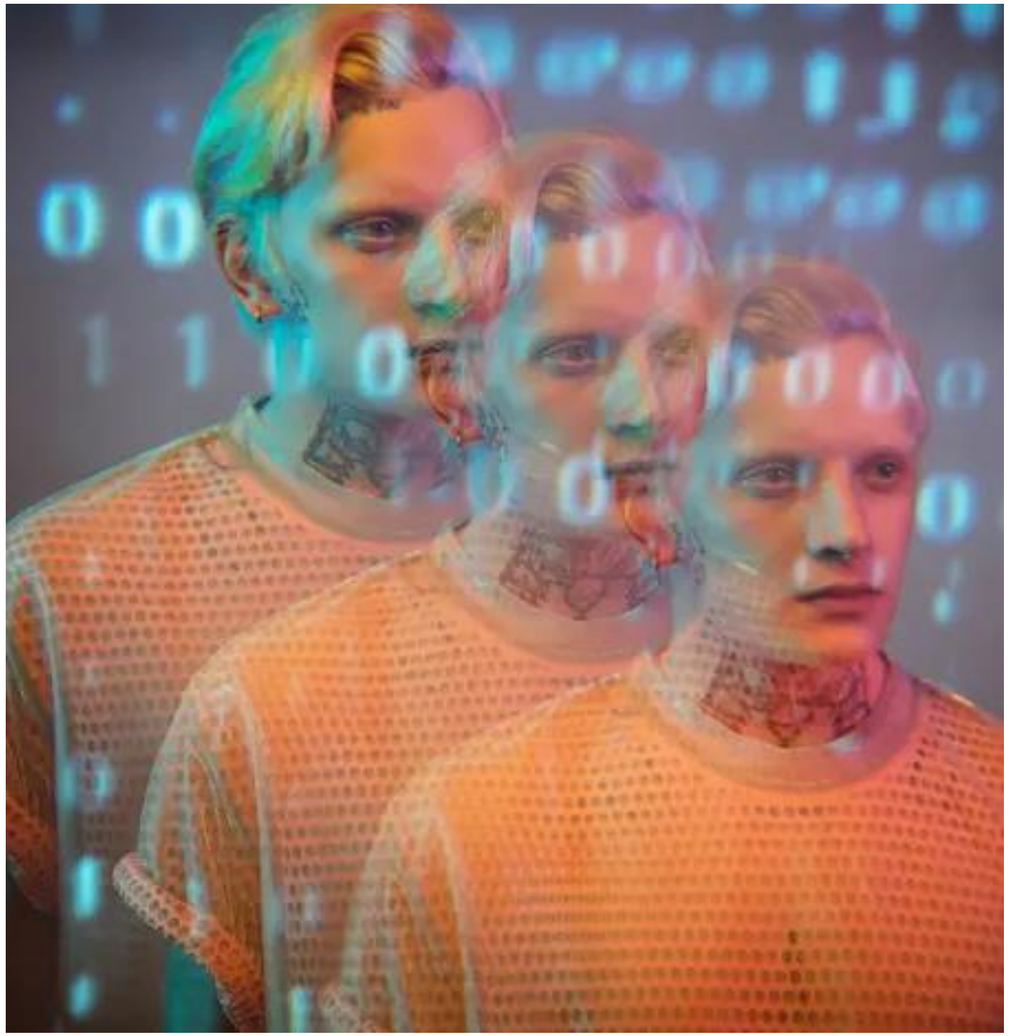
- Stringent regulatory compliance requirements such as HIPAA, GDPR, etc., restricting data sharing/collaboration especially in regulated industry verticals like BFSI, Healthcare/Life Sciences, etc.

Lack of Sufficient Data

- Where limited real-world data is existent, such as rare diseases in the field of medical science, as well as in the fields of computer vision & marketing, where collecting sufficient data is not as feasible due to resource & capital constraints.

What is Synthetic Data?

- Synthetic Data is a novel and nascent concept which mimics traditional data while improving AI models, enabling life-critical applications like Healthcare, and other industry applications in Finance, IoT, Robotics, Retail, etc.
- Synthetic data is essentially computer-generated data that mirrors the characteristics and patterns of real-world data, without containing any actual confidential or sensitive information.
- It can be likened to creating a digital twin of traditional real-world data.
- It can be synthesized in two primary ways: by either duplicating real data with algorithms or by simulating data using aggregated statistics and algorithms guided by a touch of randomness.



How Synthetic Data Helps (1 of 2)

Facilitating Data Sharing and Collaboration

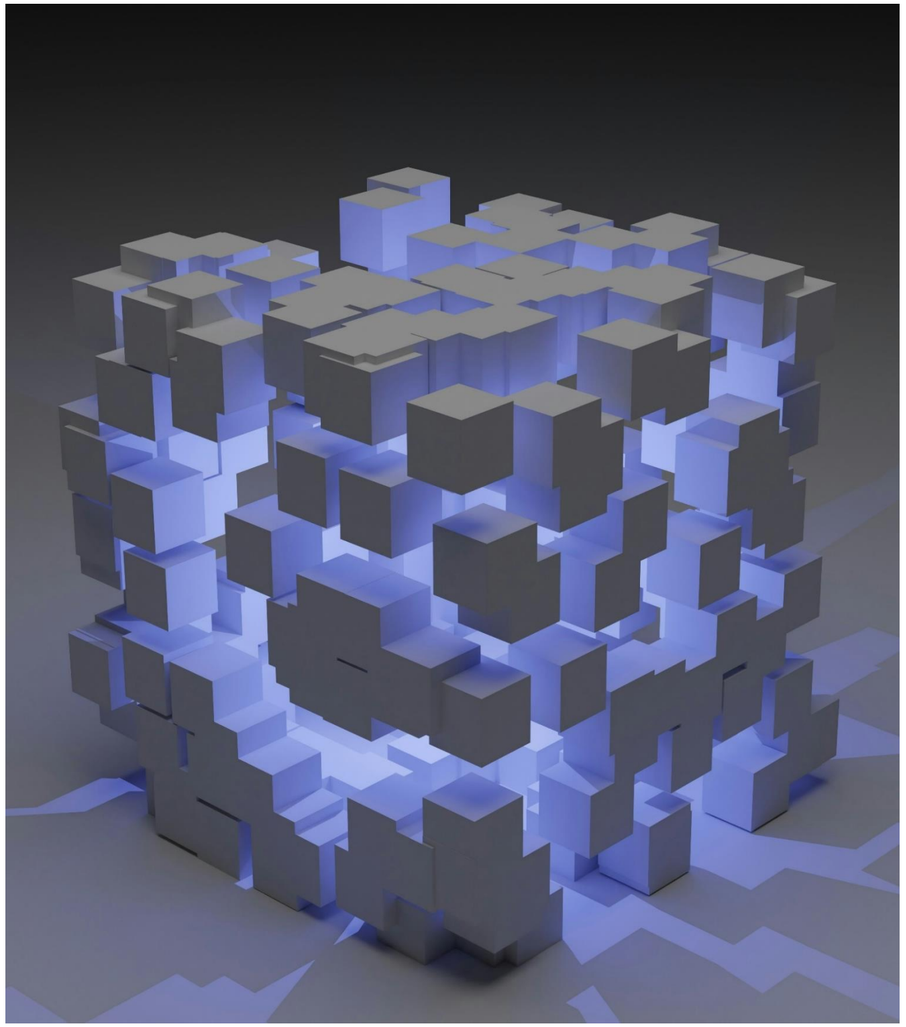
- Synthetic data provides a private & secure alternative where organizations can exchange insights, fostering collaboration in research projects and industries where data sharing is crucial but regulated (e.g. GDPR/HIPAA Compliance).

AI/ML Model Development

- It enhances data by upscaling rare patterns, mitigating biases, & boosting AI performance. It facilitates incorporating domain knowledge into models, forms the foundation for Explainable AI, & offers insights into model decisions.

Testing and Product Development

- In complex enterprise settings, obtaining realistic data is often hindered by anonymization tools, schema limitations, and prohibitions against using production data. Synthetic test data offers a practical solution, creating realistic, privacy-compliant replicas of customer data, expediting development and reducing costs.



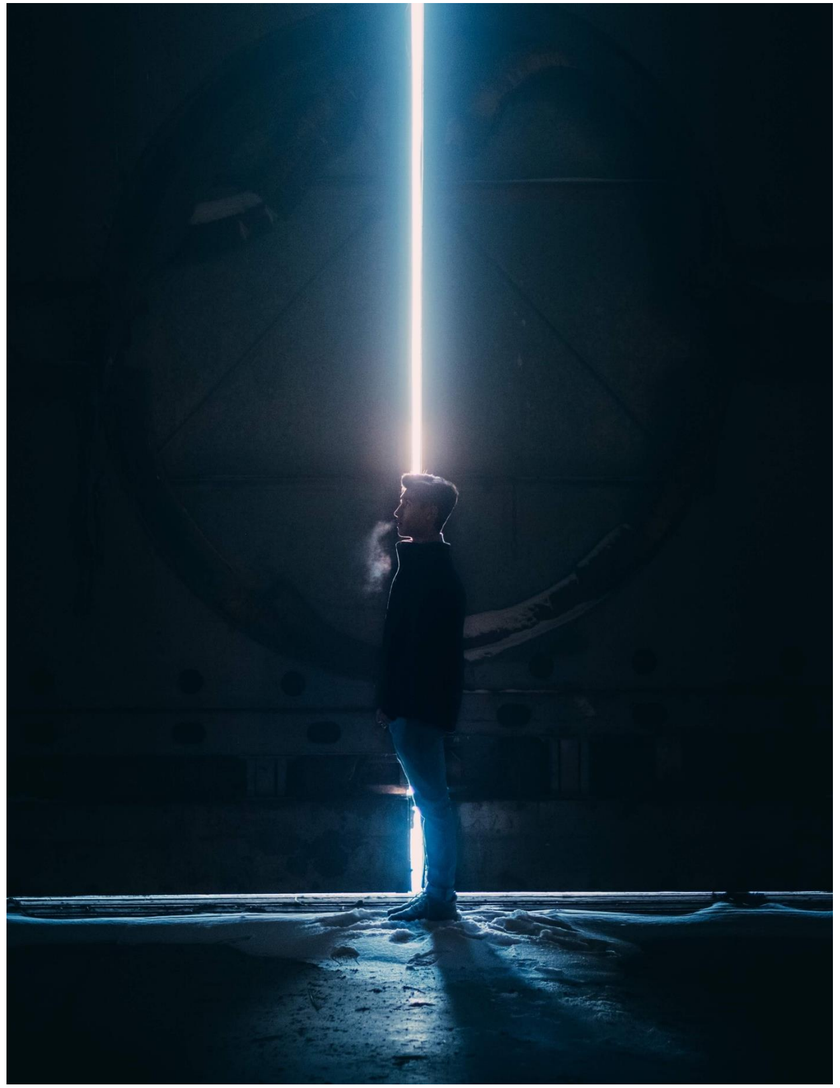
How Synthetic Data Helps (2 of 2)

Improving AI/ML Fairness and Explainability

- Global AI regulations, like the EU's initiative, underscore the need for fairness and transparency, yet organizations struggle to demonstrate compliance—implementing Fair and Explainable AI remains a complex work in progress. Synthetic data offers promise, effectively addressing bias in datasets, enhancing fairness, and providing an audit trail for transparent AI decisions, pivotal in gaining trust & meeting regulatory standards while maintaining privacy compliance.
- Synthetic data can allow data scientists to augment real data with underrepresented datasets or demographics, to create a much more balanced data set.

Data Availability

- Synthetic data can augment traditional data in use cases where only limited real-world data is available, especially in life-critical applications, such as rare diseases & medical science, as well as in the fields of computer vision and marketing, where collecting sufficient data is not as feasible due to resource & capital constraints.



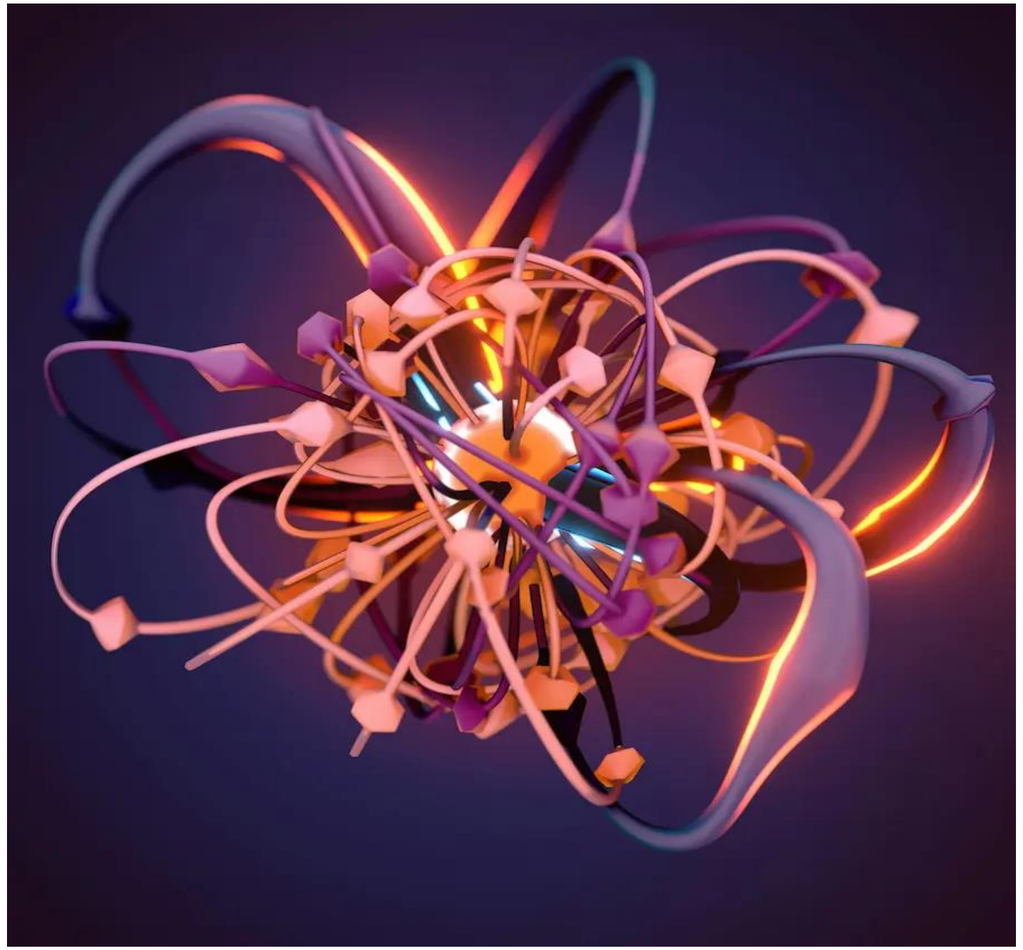
Synthetic Data Industry Trends

- Verticals like BFSI, Healthcare and Life Sciences, Retail, IoT, Robotics, etc., are expected to both drive demand and benefit from the emergence of synthetic data. This market growth is particularly driven by key factors like the increasing industry demand for data security/privacy and rising investment in advanced/emerging "neo-generation" technologies, leading to increased demand for simulated data for privacy-preservation solutions.
- According to a report by Gartner, only 1% of all AI training data was synthetic in 2021, but it is projected to grow up to 60% in the coming years, substantially outweighing real data in AI models by 2030.
- Per Allied Market Research, the synthetic data industry is expected to rapidly grow at a CAGR of more than 35%, from ≈$169 million in 2021 to ≈$3.5 billion by 2031, of which the North America region is expected to hold the largest share.

Synthetic Data Generation Methodologies

Generative Adversarial Networks (GANs)

GANs are at the heart of our synthetic data generation process. GANs consist of two neural networks – a generator and a discriminator – continuously engaged in a competition. The generator strives to produce data that is indistinguishable from real-world data, while the discriminator diligently assesses its authenticity. This adversarial training results in a dynamic feedback loop that pushes the generator to continually refine its creations. The outcome? Synthetic data that not only mimics the statistical properties of real data but also exhibits the nuances and complexities essential for training AI models effectively.



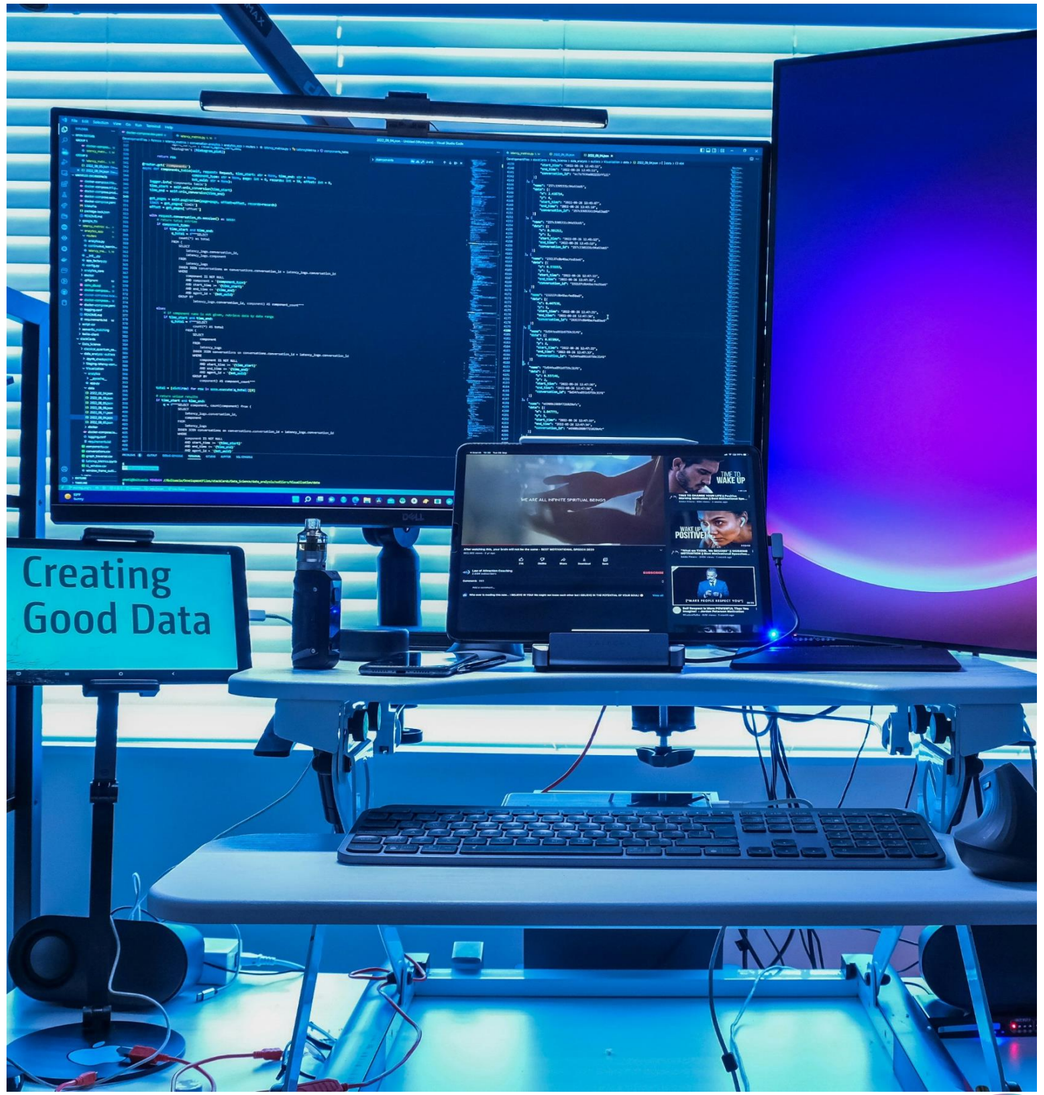
Synthetic Data Generation Methodologies

Variational Autoencoders (VAEs)

Complementing GANs, Variational Autoencoders (VAEs) play a crucial role in our synthetic data generation pipeline. VAEs excel at capturing the underlying structure and distribution of data. They do so by learning a compact representation, or latent space, where data can be manipulated and generated with precision. VAEs allow us to generate synthetic data that adheres not only to the statistical patterns but also the finer-grained features and relationships present in real data. This level of fidelity is essential for training AI models that generalize well and excel in real-world scenarios.

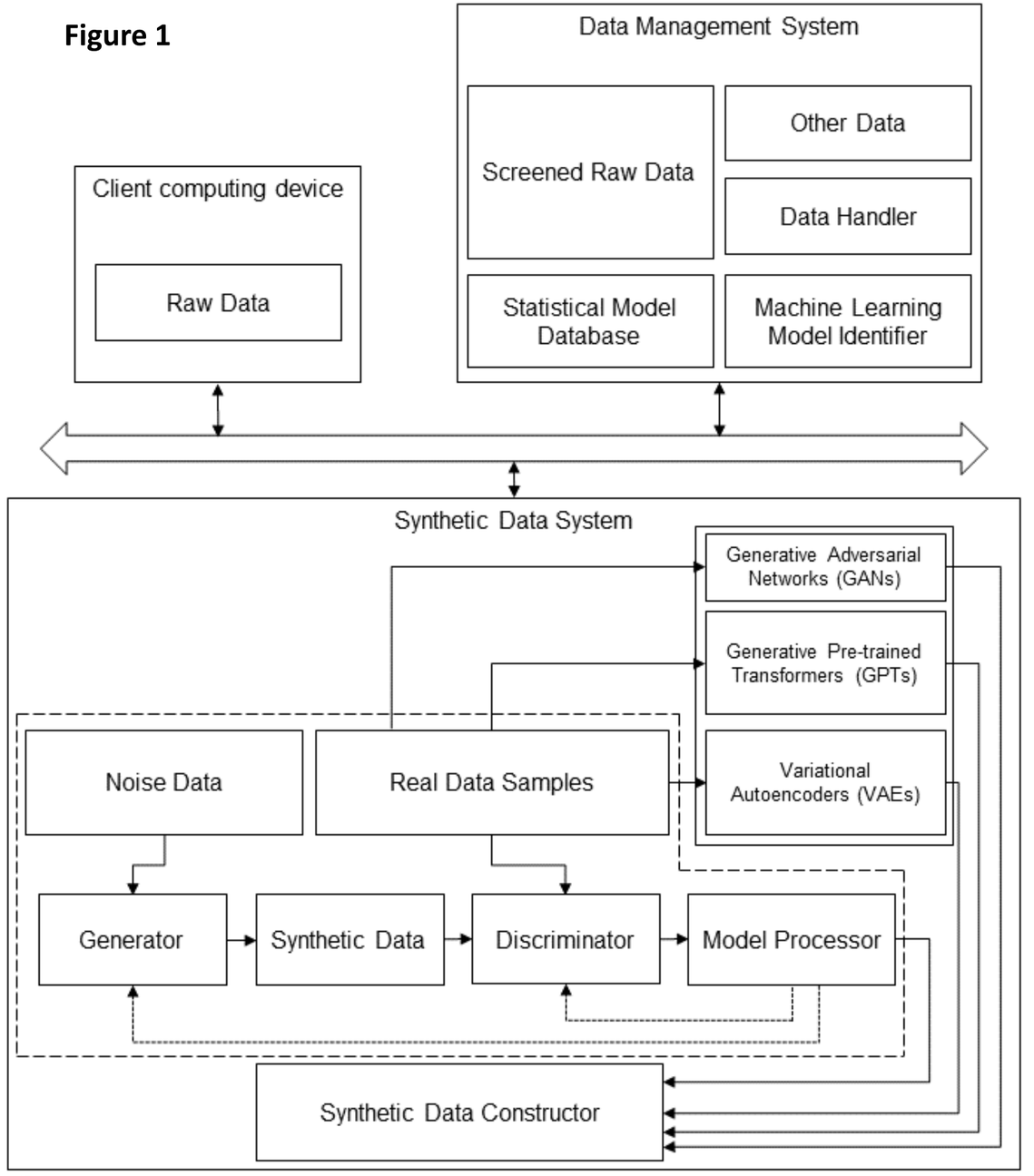
Crowdruption’s Unique Solution Approach

Crowdruption takes a differentiated approach to synthetic data generation models and methodologies. There are several different approaches to generating synthetic data, including Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), that till now, have predominantly been used in isolation rather than in combination. Through hypothesis-led simulations, we have developed hybrid statistical model approaches that synergistically integrate both the realistic data generation capabilities of GANs with the probabilistic frameworks of VAEs. This unique approach aims to improve and enhance data accuracy and availability, while also retaining the mathematical and statistical properties of traditional real-world data and preserving privacy, security, and transparency. I have already filed for a Patent for this innovative synergistic approach, and I am developing my Synthetic Data platform/toolkit.



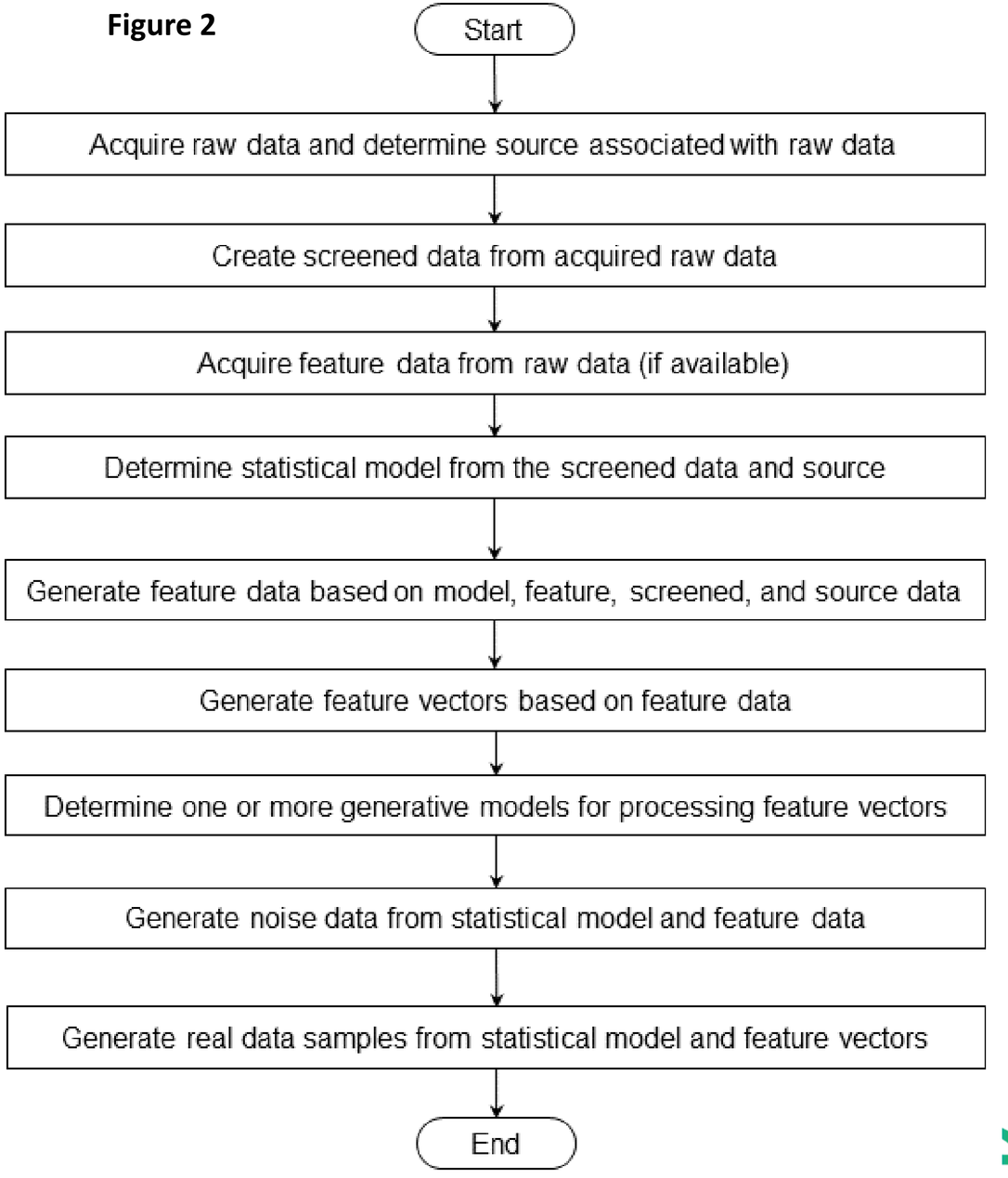
Concept Diagram of Synthetic Data Generation Methodologies

Figure 1



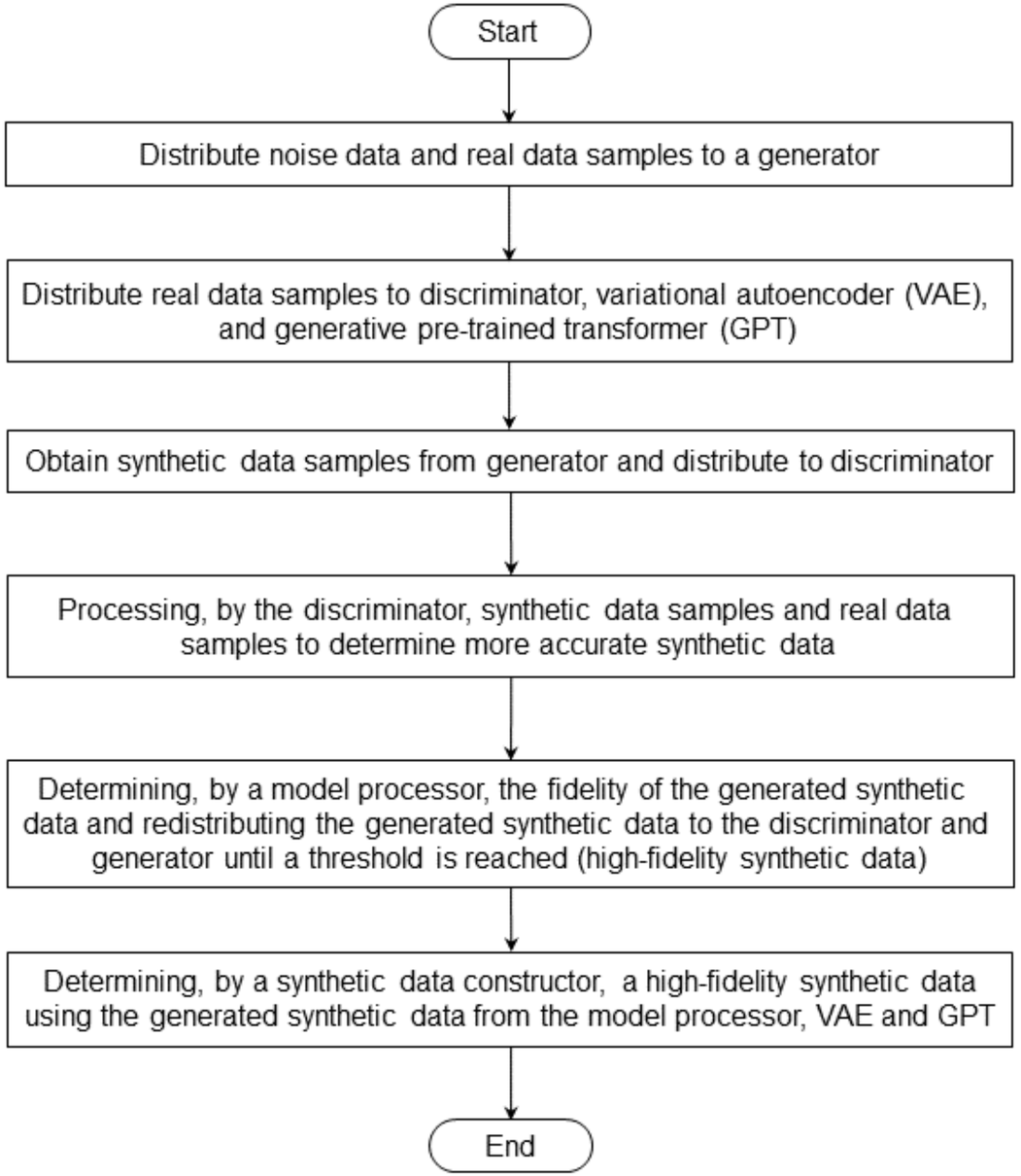
Concept Flowchart for Processing and Preparation of Raw Data

Figure 2



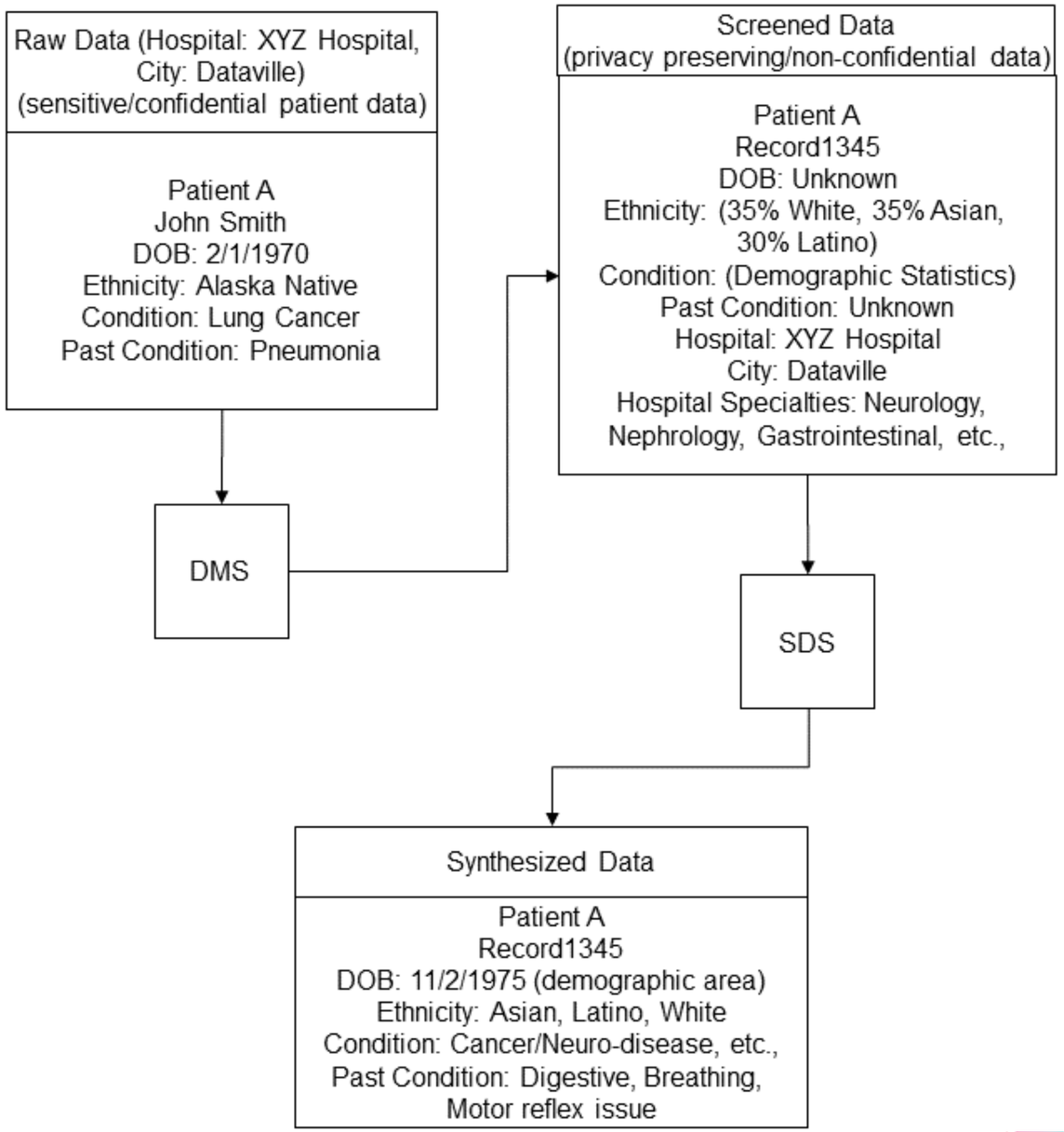
Concept Flowchart for Generating High-Fidelity Synthetic Data

Figure 3



Concept Example/Use Case of an Industry Application of Synthetic Data

Figure 4



Industry Use Cases – Financial Services

Synthetic Data for Risk Assessment

- Synthetic data empowers financial institutions to perform comprehensive risk assessments without compromising customer privacy. By generating synthetic profiles, financial institutions can analyze creditworthiness, detect anomalies, and predict risk factors with confidence. This not only ensures data privacy but also bolsters the accuracy of financial decisions.

Enhancing Fraud Prevention

- The financial sector is susceptible to various types of fraud. Synthetic data plays a crucial role in developing robust fraud prevention systems. Through the creation of synthetic transaction data and user behaviors, financial institutions can simulate countless fraud scenarios, train machine learning models effectively, and fortify their defenses against fraudulent activities.

Industry Use Cases – Healthcare & Life Sciences

Precision Medicine

- Synthetic data facilitates the development of precision medicine by creating datasets that emulate diverse patient profiles. Researchers and healthcare providers can develop targeted treatment plans with enhanced accuracy.

Drug Discovery

- In pharmaceutical research, synthetic data accelerates drug discovery processes by simulating patient responses to new medications. This aids in identifying potential drug candidates more efficiently.

Patient Privacy and Security

- Healthcare institutions are tasked with safeguarding patient data. Synthetic data supports the development and testing of electronic health records (EHR) systems, ensuring patient privacy while delivering high-quality care.

Healthcare Analytics

- Synthetic data enhances healthcare analytics by enabling the creation of large, diverse datasets. This facilitates disease outbreak prediction, resource allocation, and the identification of trends for evidence-based decision-making.

Industry Use Cases – Retail

Personalized Shopping Experiences

- Synthetic data allows retailers to create personalized shopping experiences. By simulating customer behavior and preferences, retailers can tailor recommendations, promotions, & store layouts, providing customers with a unique & engaging experience.

Inventory and Supply Chain Optimization

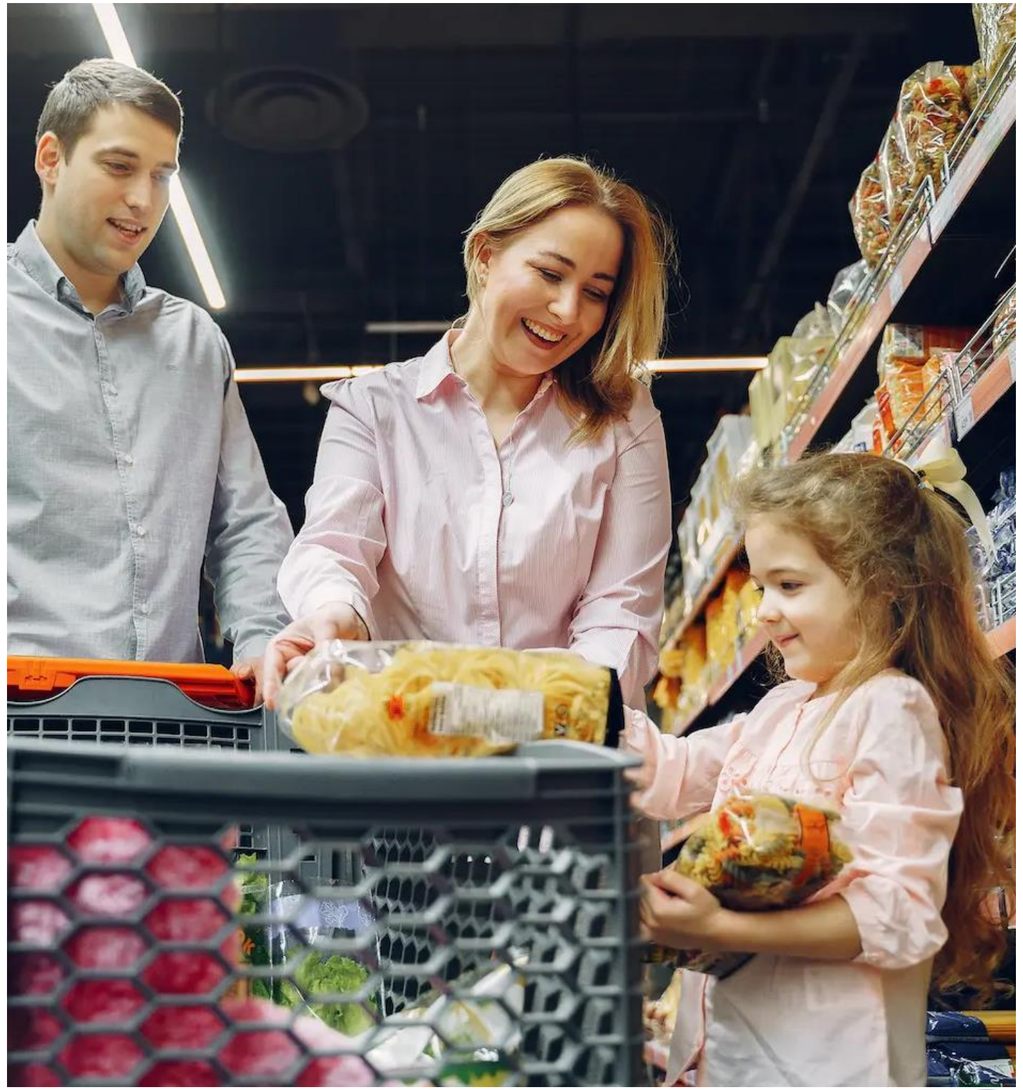
- Retailers rely on data for inventory management and supply chain optimization. Synthetic data streamlines inventory forecasting, demand planning, and logistics, helping retailers reduce costs and enhance efficiency.

Visual Merchandising and Store Layouts

- Synthetic data is a valuable tool for visual merchandising and store layout design. It enables retailers to experiment with various configurations, product placements, and designs to maximize in-store traffic and sales.

Loss Prevention and Security

- The retail industry faces challenges related to loss prevention and security. Synthetic data aids in developing and testing surveillance systems, ensuring that stores remain secure while safeguarding customer privacy.



Industry Use Cases – IoT & Robotics

Enabling Robust IoT & Robotics Solutions

- IoT devices collect vast amounts of data, enabling real-time monitoring, predictive maintenance, and improved user experiences. Synthetic data serves as a critical element in testing, fine-tuning, and optimizing IoT solutions. It allows developers to simulate diverse scenarios, test system reliability, and ensure seamless performance.

Data Privacy and Security

- In the IoT ecosystem, data privacy and security are paramount. Synthetic data mitigates privacy risks by creating data that closely resembles real-world data but without exposing sensitive information. This ensures IoT users' privacy while maintaining the integrity of data-driven applications.

Accelerating Innovation

- The ability to generate synthetic IoT data accelerates innovation by facilitating the development and testing of IoT applications. Researchers and developers can experiment with a wide range of data scenarios, fostering creativity and expanding the possibilities of IoT.

Potential Revenue Models

Offer a D2C Front-End and API Platform

- Targeted to individual developers, data enthusiasts, and personal users.
- Charge using a combination of monthly flat subscription fees ($100 to $300/month) and per-credit models ($2.00 to $10.00/credit).
- Credits could be defined as based on number of API calls, duration of API calls (5-10 minutes), number of synthetic data points generated, etc.
- Different price-point plans could have imposed limits/restrictions, such as on number of concurrent jobs, runtime limits, maximum input or output data set size, API availability SLAs, customer support, etc.
- Potentially offer a lite-version and a limited free plan to attract initial customers.

Offer an Enterprise/Business License

- Depending on the scale of the project and/or needs of the client/enterprise, bespoke pricing to be determined.
- Unlimited use of the synthetic data generation API and platform.
- Directly working with the Client/Enterprise to integrate synthetic data into their workflow and business applications.

Technology Stack We Work With…

Founder & CEO – Raj Mehta

Raj Mehta
Founder & CEO
https://www.linkedin.com/in/raj-c-mehta/

Raj Mehta brings a wealth of diverse experiences and a deep commitment to making a difference through cutting-edge technology, data science, and social impact initiatives.

Raj has work experience with leading neo-generation technology companies/startups, delivering turnkey Data Science/AI/ML projects for Fortune 500 companies.

Raj has deep quantitative skills and an analytical bent-of-mind, having won several prestigious math competitions (Qualified to Mathcounts Nationals, representing the State of Georgia, Qualified 4-times to the American Invitational Mathematics Examination (AIME), Nationally Top Ranked in AMC, etc.). Raj is an advanced professional chess player, having previously won the Georgia State Chess Championship.

Raj possesses advanced skills in AI and Data Science, having completed Google/Kaggle courses and FinTech course series by Wharton (Foundations & Applications of Financial Technology).

Raj is also the Founder & CEO of Fluid Ice Foundation (www.fluidice.org), a 501(c)(3) nonprofit organization dedicated to promoting financial inclusion and global social equity, reaching 10,000+ individuals in 50+ countries. Raj was interviewed by Emmy Award-winning journalist Ms. Jasmina Alston for Atlanta News First (CBS) for his community engagement/social impact (http://tinyurl.com/RajTVPressANF).
Additionally, covered in Forsyth County Newspaper (http://tinyurl.com/RajNewspaper).

Raj is also an accomplished Author ("SKope: The Social Kaleidoscope – Economics of Global Social Equity" – https://www.amazon.com/dp/B0CL3C65GH).

Raj has coauthored/presented published research papers, and was invited to be the first and only high-school student speaker at an annual leading economics conference (American Council on Consumer Interests).

Raj also received the honor to be the only high-school student to be invited to join the Outreach and Membership Committee of the Association for Social Economics (ASE), a research organization and learned society, alongside renowned professors globally. ASE is part of the member network of the prestigious American Economic Association (AEA), the publisher of the world-renowned American Economic Review journal.

In his spare time, Raj enjoys pursuing his passion for Jazz music and the saxophone (having qualified twice to the prestigious & highly-selective Georgia Allstate Band). Raj is also an avid golfer and book reader.

COllaborative INnovation Network (COINN)

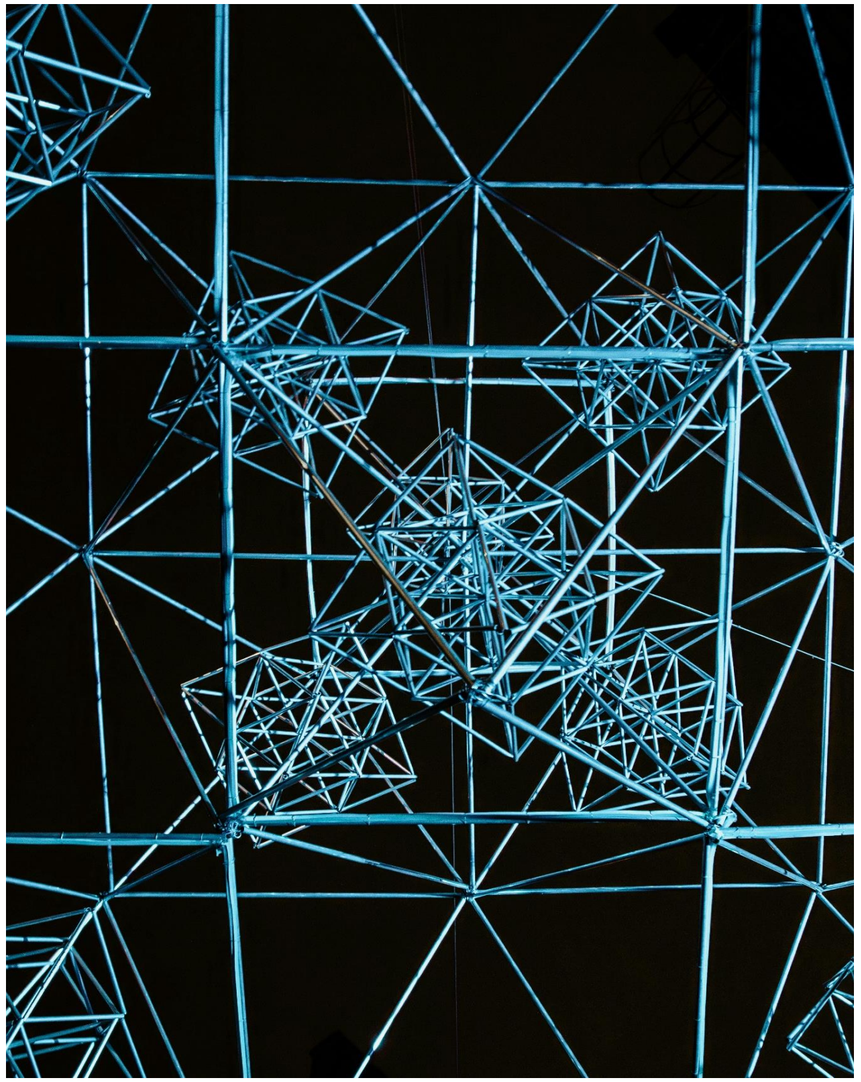
- Launched a crowdsourcing platform COllaborative INnovation Network (COINN) with an aim to build and nurture a vibrant online community of fellow passionate data enthusiasts and future data leaders.
- Through COINN, we foster continuous learning and exploration, by hosting hackathons, workshops, and inviting thought-leaders from industry and academia.

<u>Hackathons/Workshops</u>

- Facilitate advanced innovation and raise awareness for the viability of synthetic data in a wide range of industry applications, ultimately leading to the development of an open-source toolkit.

<u>Guest Speakers</u>

- Invite industry and academia leaders to share thought-provoking insights on the future potential of synthetic data and data science/AI/ML industry as a whole.




Thank you!
Website: https://www.crowdruption.com/
Contact: rajcm365@gmail.com
Crowdruption
BETTER DATA RE:IMAGINED.