LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Data Science Without Borders

SYNTHETIC DATASETS TO OVERCOME DATA GOVERNANCE CHALLENGES

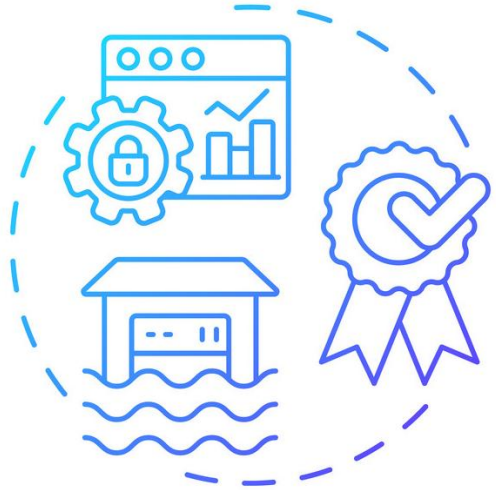
TATHAGTATA BHATTACHARJEE

AGENDA

1) INTRODUCTION
2) UNDERSTANDING DATA TYPES
3) DATA GOVERNANCE CHALLENGES
4) ROLE OF SYNTHETIC DATA IN OVERCOMING GOVERNANCE BARRIERS
5) KEY USE CASES IN POPULATION AND HEALTH RESEARCH
6) ETHICAL CONSIDERATIONS AND CHALLENGES
7) VALIDATION OF SYNTHETIC DATA
8) Q&A

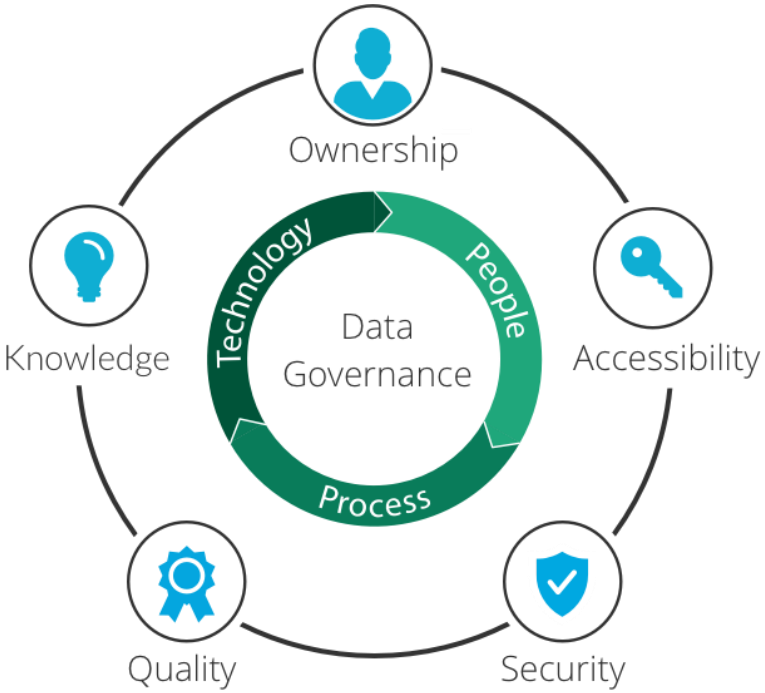
WHAT IS DATA GOVERNANCE

- A structured framework of **policies**, **processes**, and **controls** that ensures data is **accurate**, **secure**, and **ethically** used across the organization.
- It focuses on **quality**, **compliance**, **privacy**, and **accessibility**, enabling reliable data for decision-making while maintaining federated ownership.
- Key Features:
  - Strategic alignment – Balances decentralized control with enterprise-wide standards.
  - Risk & compliance – Enforces regulations (GDPR, CCPA, HIPPA) and internal policies.
  - Data integrity – Maintains consistency, accuracy, and usability.
  - Controlled access – Grants permissions based on roles and needs.
- Key to Success: Federated Data Governance
  - Collaborate with domain owners to create flexible, scalable policies.
  - Balance local control with enterprise-wide alignment.



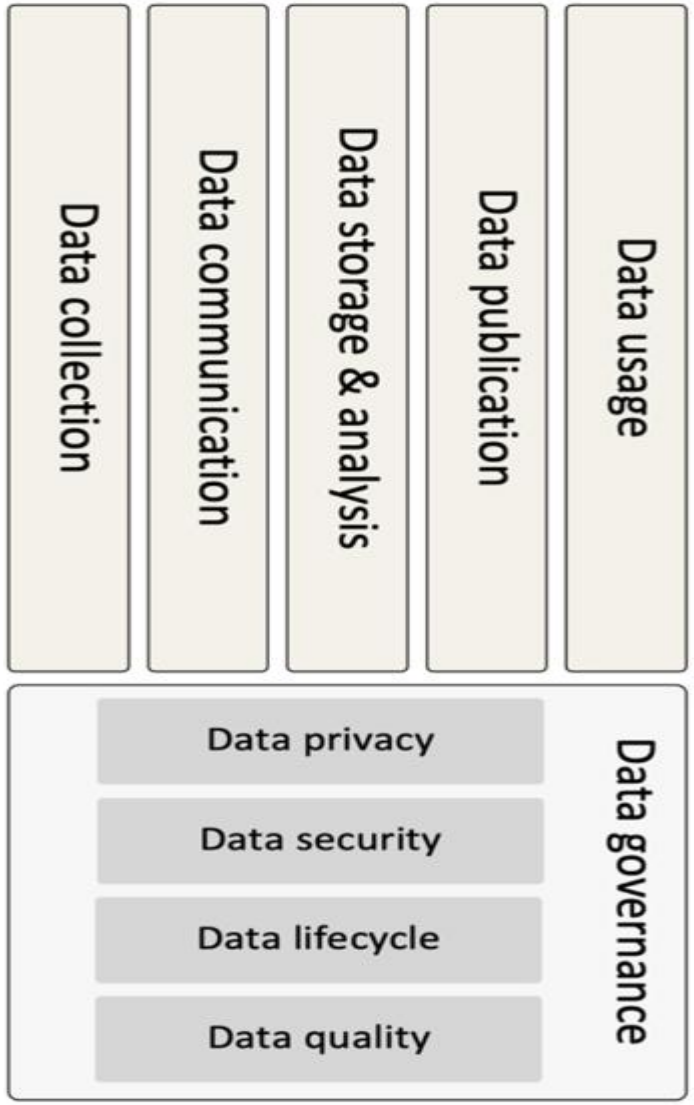
OVERVIEW OF DATA GOVERNANCE CHALLENGES IN POPULATION AND HEALTH RESEARCH

1) **Data Ownership and Sovereignty:** Countries enforce data protection laws to ensure that data generated within their borders is controlled locally. This restricts international research collaboration.
2) **Ethical Concerns:** Privacy risks, informed consent challenges, and participant protection limit population and health research data sharing.
3) **Institutional Policies:** Universities, research institutes, and government bodies impose internal restrictions on data usage, slowing collaborative efforts.
4) **Regulatory Frameworks:** Compliance with frameworks such as GDPR, HIPAA, or country-specific data laws adds complexity to data access processes.
5) **Cultural and Social Norms:** Local customs may influence data accessibility and usage policies.



IMPACT OF DATA ACCESS RESTRICTIONS ON SCIENTIFIC PROGRESS

- **Delayed Discoveries:** Restrictions hinder researchers from accessing datasets needed for timely insights and interventions.
- **Reduced Collaboration:** Strict governance limits data sharing across institutions and countries, reducing collective knowledge advancement.
- **Bias in Research Outcomes:** Limited access to diverse datasets may skew findings and reduce model generalizability.
- **Replication Challenges:** The inability to access original data makes validating and reproducing research findings difficult.
- **Innovation Barriers:** Data restrictions limit opportunities to train advanced machine learning models, particularly in population and health studies.



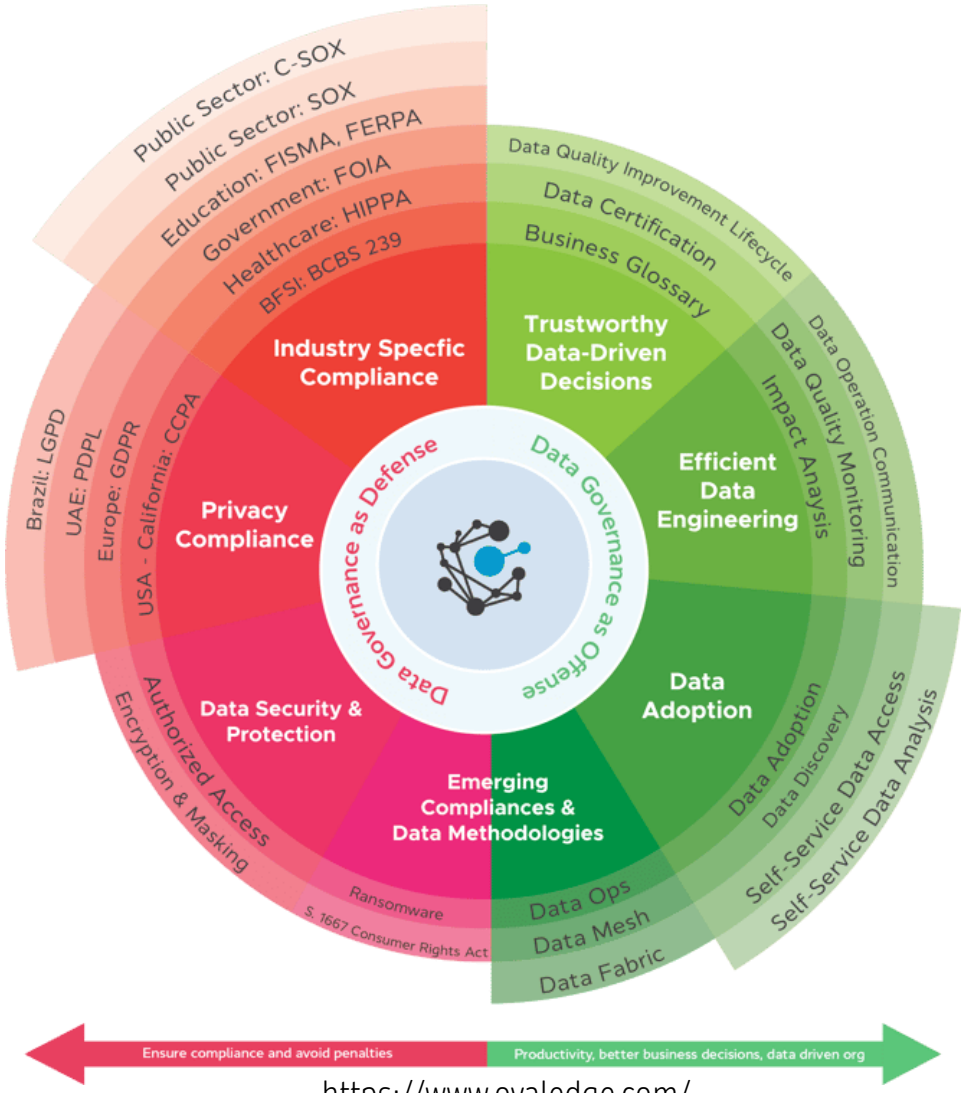
IMPORTANCE OF DATA GOVERNANCE

Defensive Data Governance

Defensive Data Governance is a risk-averse approach focused on protecting sensitive data, ensuring regulatory compliance, and minimizing exposure to breaches or legal penalties

Key Aspects:

1) **Industry Compliance:** Ensures adherence to regulations like HIPAA
2) **Privacy Laws:** Mandates GDPR, CCPA, and evolving global standards (e.g., Meta's €405M fine for mishandling minors' data).
3) **Security & Protection:** Requires encryption, access controls, and PII safeguards (e.g., can be fined for unencrypted data leaks).
4) **Emerging Threats:** Adapts to new risks (e.g., ransomware) and methodologies (e.g., Social Media Privacy Act).



https://www.ovaledge.com/

Offensive Data Governance

Offensive Data Governance proactively leverages high-quality, well-managed data to drive innovation, efficiency, and competitive advantage across the organization

Key Aspects:

1) Trusted Decisions: High-quality, standardized data enables better analysis
2) Efficient Engineering: Improved data quality accelerates model development
3) Wider Adoption: Self-service access fosters data-driven innovation
4) Future-Ready: Supports advanced methodologies like Data Mesh

Note: Data Mesh: A decentralized architecture where different teams own their data, using shared infrastructure under federated governance

THE GOVERNANCE PARADOX: DATA STANDARDS VS. RESEARCH REALITIES IN POPULATION HEALTH

- **Rigid Standards vs. Collaborative Agility**
  - Governance demands structured metadata & access controls
  - Research requires rapid, cross-institutional data sharing
- **Privacy Compliance vs. Scientific Utility**
  - Anonymization protocols reduce data granularity
  - Epidemiological research needs detailed demographic linkages
- **Centralized Control vs. Distributed Ownership**
  - Single source of truth requirements
  - Federated health systems maintain local data sovereignty
- **Documentation Burden vs. Research Velocity**
  - Complete provenance tracking slows studies
  - Public health crises demand real-time analysis

Bridge the Gap:

- ✔ Adaptive governance frameworks
- ✔ Tiered access models
- ✔ Federated learning approaches
- ✔ Automated compliance tools

The Bottom-Line
Perfect governance shouldn't be the enemy of research

UNDERSTANDING DATA TYPES

- **Real Data:**
  - Collected directly from the source (e.g., patient records, census data).
  - Reflects true observations and is governed by data privacy regulations.
- **Fake Data:**
  - Randomly generated data with no meaningful correlation to real-world data.
  - It is used for software testing and lacks analytical value.
- **Synthetic Data:**
  - Artificially generated data that mimics the statistical properties of real data while preserving privacy.
  - Created using machine learning techniques and is ideal for research and model training.
- **Simulated Data:**
  - Data produced from mathematical models or simulations to represent hypothetical scenarios.
  - It is used to test theories, predict outcomes, or model potential system behaviors.

SYNTHETIC VS SIMULATED DATA

- **Synthetic Data:**
  - Created by analyzing and replicating patterns from real data.
  - Preserves statistical properties while ensuring privacy.
  - Example: Generating synthetic patient profiles that maintain real-world correlations between age, disease status, and medication use, allowing researchers to develop risk prediction models without accessing sensitive medical records.
- **Simulated Data:**
  - Generated using theoretical models or system rules to create hypothetical scenarios.
  - Often lacks real-world variability but is useful for exploring system behaviors, testing models, or training in controlled environments.
  - Example: During the early months of the COVID-19 pandemic in 2020, researchers used simulation models to predict infection rates, hospitalization needs, and potential death tolls. These models incorporated assumptions about transmission rates, social distancing measures, and healthcare capacity to guide public health interventions and inform policy decisions.

Key Difference: Synthetic data mirrors real data characteristics, while simulated data is based on hypothetical models rather than true observational patterns.

SYNTHETIC DATA BREAKS GOVERNANCE BARRIERS

- Bypasses Privacy Restrictions
  - Generates artificial data that mirrors real patterns
  - Contains no actual personal/confidential information
  - Meets strict regulations (GDPR, HIPAA) by design
- Preserves Analytical Value
  - Maintains statistical relationships of original data
  - Works for ML training, testing, and analytics
  - Validated to deliver accurate business insights
- Enables Secure Collaboration
  - Share across teams/partners without governance risks
  - Combine datasets that normally couldn't be merged
  - Accelerate research while complying with policies

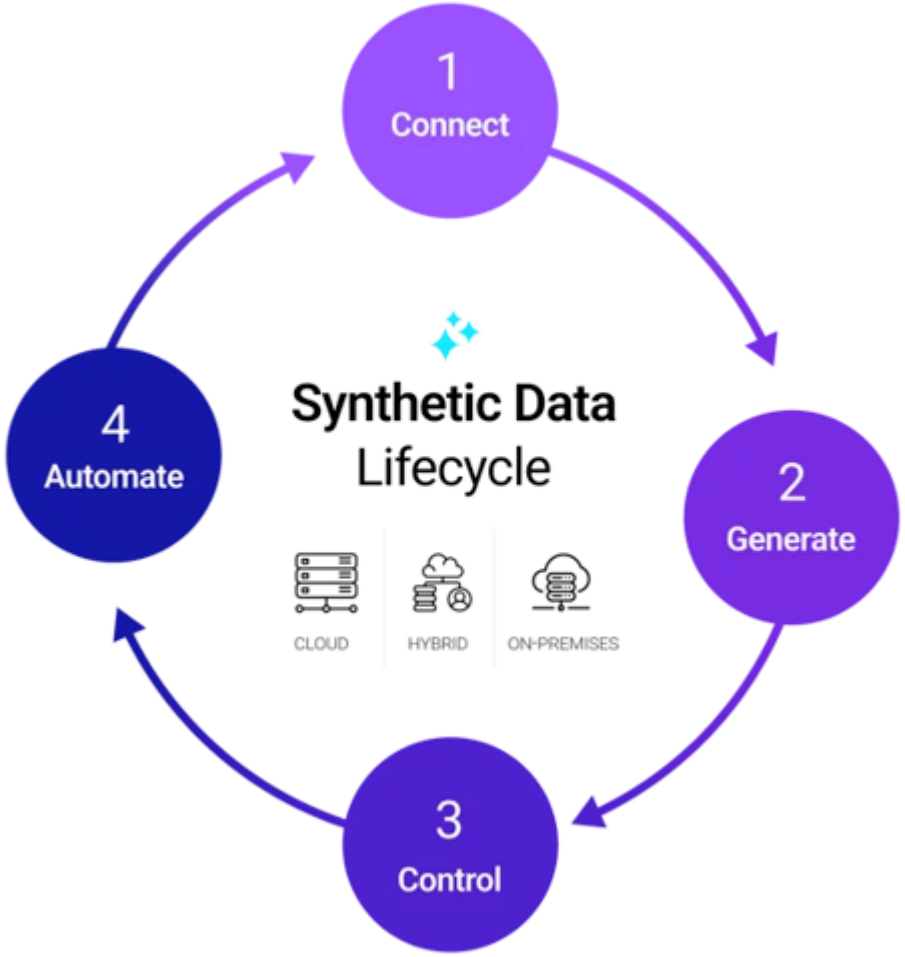
https://syntheticus.ai/blog/lifting-data-barriers-exploring-synthetic-data-in-healthcare-research

https://neptune.ai/blog/the-advantages-of-synthetic-data-over-real-data

Harnessing the power of synthetic data in healthcare: innovation, application, and privacy https://doi.org/10.1038/s41746-023-00927-3

GENERATING SYNTHETIC DATA

- Rule-Based Methods
  - Uses predefined rules and logic to generate synthetic data.
  - Example: Random sampling from distributions, shuffling real data.
- Statistical Methods
  - Generates synthetic data based on statistical properties of real data.
  - Example: Gaussian Mixture Models (GMM),
- Machine Learning-Based Methods
  - Uses ML models to learn patterns and generate synthetic data.
  - Example: Decision trees, regression models.
- Deep Learning-Based Methods
  - Uses neural networks to generate high-fidelity synthetic data.
  - Example: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs).

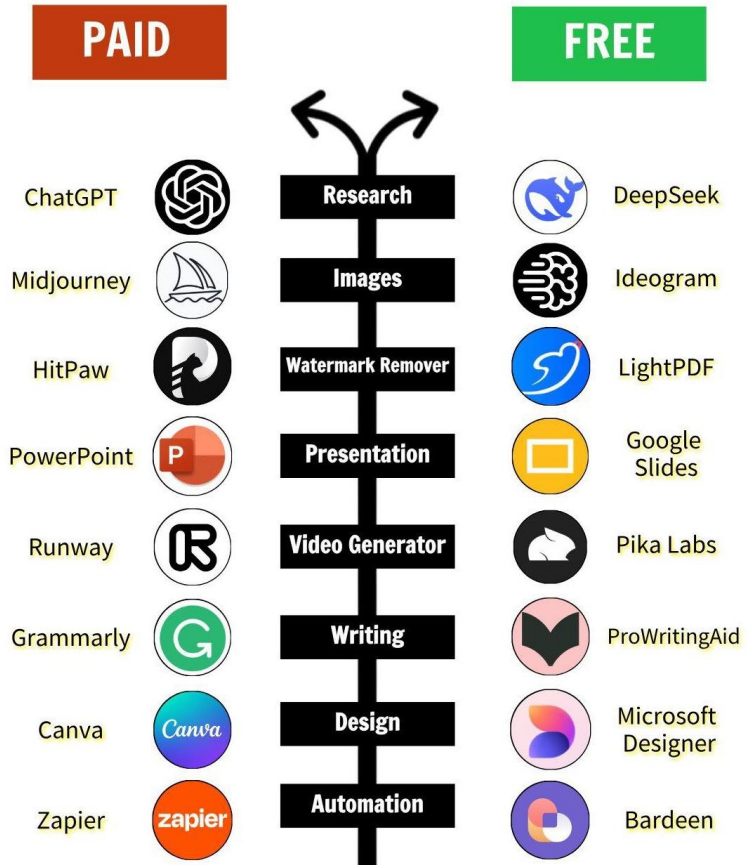


GENERATIVE AI

- A subset of artificial intelligence (AI)
- Focuses on creating new content, like text, images, audio, video, or even code, based on patterns and data it has been trained on
- Unlike traditional AI systems that are designed for specific tasks (e.g., classification or prediction), Generative AI models are capable of producing original outputs that mimic human creativity
- Characteristics of Generative AI
  - Creates New Content: It generates new data (e.g., text, images, music) rather than just analyzing or classifying existing data.
  - Learns from Data: It is trained on large datasets to understand patterns, structures, and relationships within the data
  - Uses Advanced Models:
    - Generative Adversarial Networks (GANs): Two neural networks (a generator and a discriminator) work together to create realistic outputs.
    - Variational Autoencoders (VAEs): Models that learn to encode and decode data to generate new samples.
    - Transformer-based Models: Models like GPT (Generative Pre-trained Transformer) that excel in generating text data.

EXAMPLES OF GENERATIVE AI

- **Text Generation:**
  - Tools like ChatGPT, GPT-4, and Bard, Copilot that generate human-like text for conversations, essays, or code.
  - Example: Writing a story, summarizing an article, or answering questions.
- **Image Generation:**
  - Tools like DALL·E, MidJourney, and Stable Diffusion create images from text descriptions.
  - Example: Generating a picture of "a futuristic city on Mars.“
- **Audio Generation:**
  - Tools like ElevenLabs or VALL-E generate realistic speech or music.
  - Example: Creating a voiceover or composing a song.
- **Video Generation:**
  - Tools like Synthesia or Runway ML generate videos from text or images.
  - Example: Creating a marketing video with AI-generated actors.
- **Code Generation:**
  - Tools like GitHub Copilot or Codex generate code based on natural language prompts.
  - Example: Writing a Python script for data analysis.



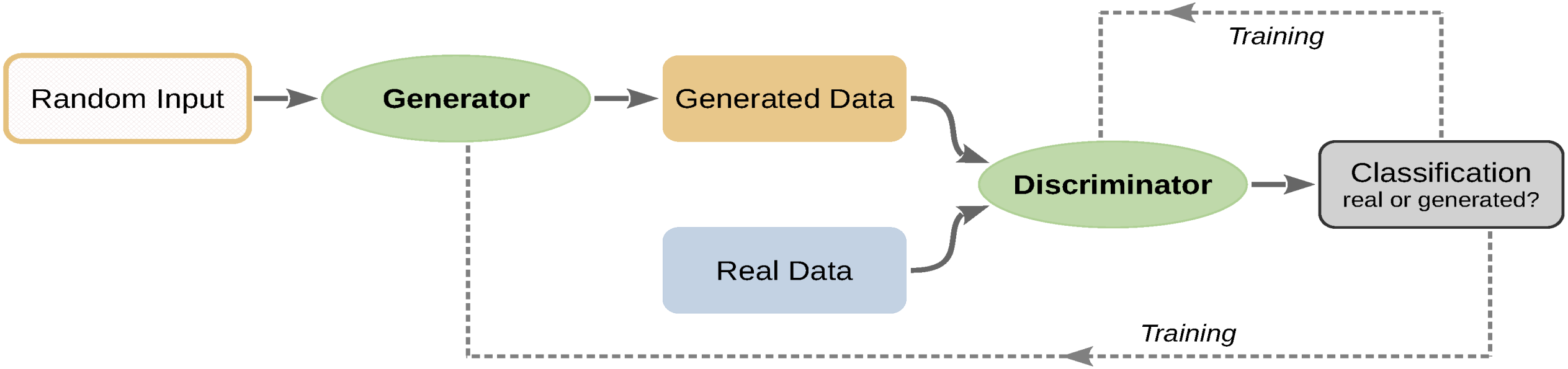
HOW GENERATIVE AI WORKS?

- **Training:**
  - The model is trained on a large dataset (e.g., text, images, or audio).
  - It learns patterns, relationships, and structures in the data.
- **Generation:**
  - Once trained, the model can generate new content by predicting the next word, pixel, or note based on the input it receives.
- **Fine-Tuning:**
  - Models can be fine-tuned for specific tasks or domains to improve their performance.

GENERATIVE ADVERSARIAL NETWORK (GAN)

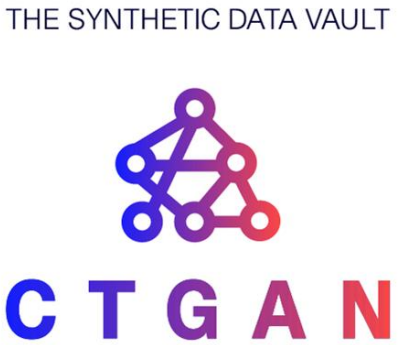
- A deep learning model designed to generate synthetic data that resembles real data.
- An algorithms that use two neural networks competing against each other (thus the “adversarial”) in order to generate synthetic data
- Ian Goodfellow introduced it in 2014, and it is widely used in fields such as image generation, data augmentation, and synthetic data creation.
- GAN consists of two competing neural networks:
  - Generator (G) – Creates fake data from random noise.
  - Discriminator (D) – Evaluates and distinguishes between real and fake data.
- Adversarial – meaning: involving or characterized by conflict or opposition.

The discriminator's goal is to be able to tell apart real and synthetic data. Meanwhile, the generator's goal is to create high-quality synthetic data that fools the discriminator.



CTGAN (CONDITIONAL TABULAR GENERATIVE ADVERSARIAL NETWORK)

- CTGAN is an advanced type of GAN designed specifically for generating realistic tabular data, including both continuous and categorical variables.
- The MIT Data introduced it to AI Lab and is part of the Synthetic Data Vault (SDV) project.
- Unlike traditional GANs, which struggle with tabular data due to complex distributions, CTGAN handles imbalanced categorical variables and multimodal distributions effectively.
- Why Use CTGAN for Synthetic Data Generation?
  - Handles Mixed Data Types – Works well with numerical and categorical data.
  - Solves Data Imbalance Issues – Resamples minority classes effectively.
  - Captures Complex Data Distributions – Uses a mode-specific normalization technique.
  - Preserves Relationships Between Features – Generates realistic synthetic data.



Reference: CTGAN-driven synthetic data generation: A multidisciplinary, expert-guided approach (TIMA)

https://doi.org/10.1016/j.cmpb.2024.108523

PREPARING FOR THE HANDS-ON

- **Database**
  - **PostgreSQL 17 (lower versions are also okay if already installed)**
  - **https://www.postgresql.org/download/windows/**
  - **Installation Demo: https://www.youtube.com/watch?v=GpqJzWCcQXY (Don't install Stack Builder when selecting components)**
- **Python** (You can use any IDE like PyCharm, Jupyter Notebook or any other)
  - **PyCharm Community Edition IDE**
    - **For Windows: https://www.jetbrains.com/pycharm/download/?section=windows**
    - **For Linux: https://www.jetbrains.com/pycharm/download/?section=linux**
    - **For Mac: https://www.jetbrains.com/pycharm/download/?section=mac**
    - **Sroll down to the community edition**
    - **Installation Demo: https://www.youtube.com/watch?v=Tmu_fkFwlvw**
  - **Jupyter Notebook IDE**

    https://github.com/tathagatabhattacharjee/Record-Linkage-Using-Machine-Learning-Techniques/blob/main/02%20Generate%20Synthetic%20Datasets/Installing%20Jupyter%20Notebook.pdf

QUIZ

- Is MIMIC-III a synthetic dataset? Give Reasons to justify your response.
  - https://physionet.org/content/mimiciii/1.4/
  - DOI (latest version): https://doi.org/10.13026/cd7z-wg25

SYNTHETIC DATA VALIDATION

Why Validate Synthetic Data?

- Utility: Ensures synthetic data maintains the statistical and analytical properties of the original data.
- Privacy: Prevents the synthetic data from leaking sensitive information about individuals.
- Fidelity: Ensures the structure, constraints, and patterns of the original dataset are preserved.

Reliable synthetic data enables accurate downstream analysis and ensures compliance with data privacy regulations.

UTILITY VALIDATION

1) Descriptive Statistics

Compare:

a. Mean, Median, Mode: Ensures the central tendencies are consistent.
b. Standard Deviation, Variance: Confirms variability is maintained.
c. Distribution checks: Visualize data patterns using histograms and boxplots.
d. Tool Suggestions: Pandas, NumPy, and Seaborn for visualization.

2) Correlation Analysis

Perform correlation tests:

a. Pearson Correlation: For linear relationships.
b. Spearman Correlation: For ranked variables.
c. Kendall Correlation: For ordinal data.
d. Ensures variable relationships are preserved to support meaningful insights.

3) Feature Importance

a. Train a model on the original dataset and synthetic data separately.
b. Compare feature importance scores to ensure key variables retain predictive power.

4) Model Performance

c. Train machine learning models on both datasets, then compare:
    a. Accuracy
    b. Precision / Recall
    c. F1 Score
d. Ensures synthetic data maintains predictive capabilities

5) Data Imbalance Analysis

a. Verify class distributions in categorical variables to ensure the balance matches the original dataset.

PRIVACY VALIDATION

1) Membership Inference Attack:
   a. This test checks if an attacker can tell whether someone's data was included in the dataset used to create the synthetic data.
   b. It's like asking: "Was this person's data part of the original training set?“
   c. Running this test helps assess privacy risks and ensures sensitive information isn’t exposed.
2) Attribute Inference Attack:
   a. This test checks if someone could guess missing details about individuals based on the synthetic data.
   b. For example, if someone knows a person’s age and gender, could they correctly guess their income?
   c. This helps identify weak points in privacy protection.
3) Distance-based Measures:
   a. These methods measure how closely synthetic data resembles real data without being too similar:
      i. k-Anonymity: Ensures that each combination of key attributes appears at least k times, making it harder to identify individuals.
      ii. l-Diversity: Ensures each group with similar attributes contains enough variety in sensitive values to prevent guessing.
      iii. t-Closeness: Ensures the distribution of sensitive values in each group closely matches the overall dataset.
4) Identifiability Tests:
   a. These tests check how easily someone could link synthetic data back to real individuals.
      i. Disclosure Risk Analysis: Assesses how likely it is for a person’s identity to be revealed.
      ii. Linkage Attack Simulations: Simulates attackers combining different datasets to uncover identities.

FIDELITY VALIDATION (ACCURACY AND COMPLETENESS OF DATA)

1) Schema Conformity:
   a. Column names and data types match the original dataset.
   b. Constraints and value ranges remain consistent.
2) Integrity Checks:
   a. Primary keys are unique.
   b. Foreign keys correctly reference linked tables.
   c. Composite keys maintain valid combinations.
3) Missing Data Patterns:
   a. Ensure null or missing value patterns in synthetic data resemble the original dataset.

OTHER VALIDATIONS

1) Visual Inspection

Recommended Visualization Tools:

a. Histograms: Show variable distributions.
b. Boxplots: Highlight outliers and quartile ranges.
c. Heatmaps: Illustrate correlation patterns.
d. PCA (Principal Component Analysis): Visualizes data clusters and patterns in reduced dimensions.

2) Domain-Specific Validation

Examples of Healthcare and Demographic Data:

a. Age Distributions: Align synthetic data with expected demographic patterns.
b. Event Timelines: Ensure birth dates precede death dates, vaccination dates follow birth dates, etc.
c. Geographical Distributions: Verify population distribution aligns with known regional data, if available.

SUMMARY: RECOMMENDED PROCESS FOR SYNTHETIC DATA VALIDATION

1) **Initial Inspection**: Confirm data structure, types, and value ranges.
2) **Statistical Tests**: Compare descriptive statistics and correlations.
3) **Privacy Assessment**: Perform privacy risk evaluations using attack models.
4) **Domain-Specific Testing**: Ensure meaningful trends and patterns remain intact.
5) **Model Testing**: Assess performance consistency in predictive models.
6) **Final Report**: Document strengths, weaknesses, and recommendations with clear visual evidence..

Synthetic Data:

- Trustworthiness in research outcomes
- Compliance with privacy standards
- Retention of key insights for downstream analysis

Thank you

tathagata.bhattacharjee@lshtm.ac.uk | https://orcid.org/0000-0001-9437-0894 | www.linkedin.com/in/tathagatabhattacharjee