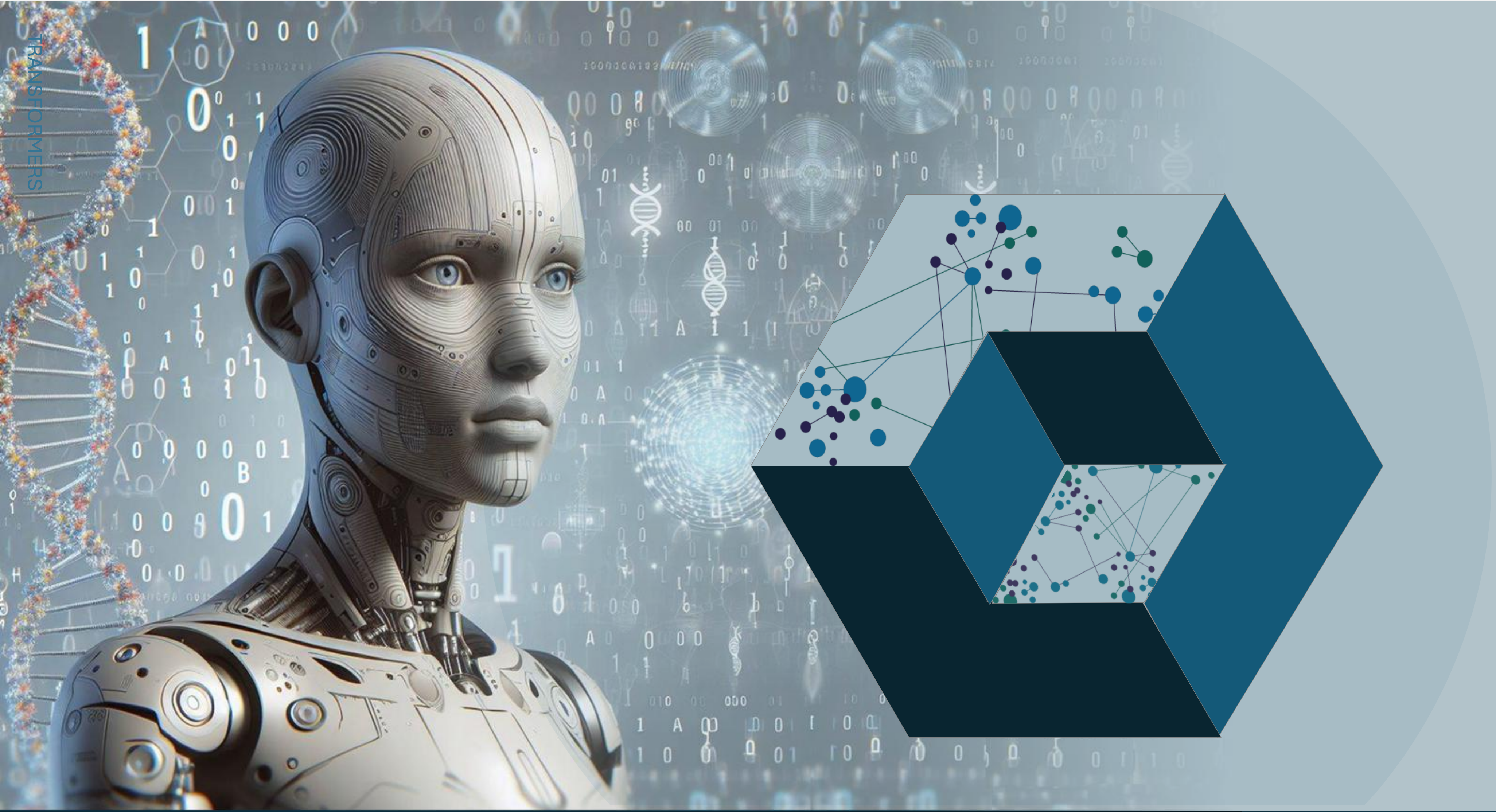


SYNTHETIC DATA

Generate synthetic data without sacrificing privacy

REALM

2025-02-15

REALLY BIG DATA

By 2025, the big data market will be worth a whopping

$229.4 BILLION

Gartner estimates that by 2030, synthetic data will completely overshadow real data in AI models. You can learn more about synthetic

There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days

WHAT IS SYNTHETIC DATA

Synthetic data is information that's artificially generated rather than produced by real-world events. Typically created using algorithms, synthetic data can be deployed to validate mathematical models and to train machine learning models.

BY DATA TYPE

Text

Media

Tabular data

Timelines and sequences

BY METHOD

1. **Rule-based systems**: These are systems that generate text based on predefined rules and templates, such as grammar, syntax, and vocabulary. Rule-based systems can produce consistent and coherent text, but they may lack diversity and creativity. They are often used for simple and structured tasks, such as filling forms, generating reports, or creating chatbot responses/

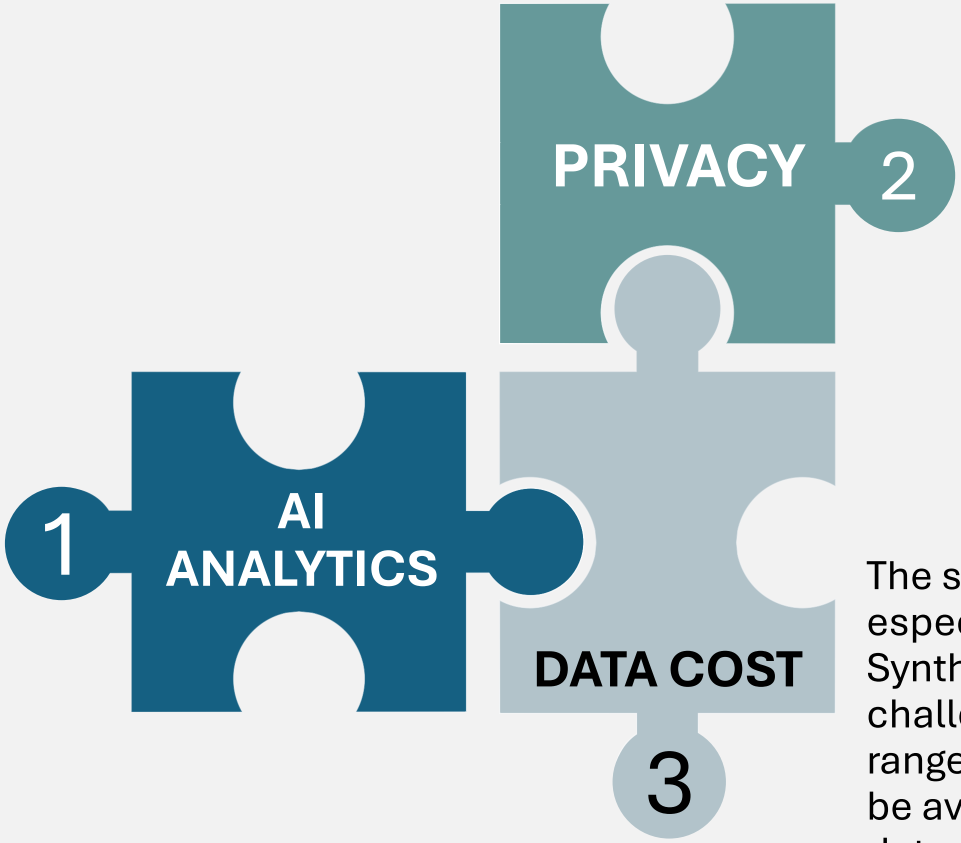
2. **Statistical models**: These are models that generate text based on statistical methods, such as n-grams, hidden Markov models, or Bayesian networks. Statistical models can learn the probabilities of words and phrases from large amounts of text data and use them to generate new text. Statistical models can produce more varied and natural text than rule-based systems, but they may also generate errors or nonsensical sentences. They are often used for complex and unstructured tasks, such as machine translation, text summarization, or text classification

3. **Neural networks**: These are models that generate text based on deep learning methods, such as recurrent neural networks, transformers, or generative adversarial networks. Neural networks can learn the semantic and syntactic features of text data and use them to generate new text. Neural networks can produce more realistic and fluent text than statistical models, but they may also require more computational resources and data. They are often used for advanced and creative tasks, such as text generation, text style transfer, or text augmentation1

WHY SYNTHETIC DATA

Synthetic data is a valuable resource for advancing artificial intelligence and data science, as it enables the generation of large, diverse, and high-quality datasets that overcome the limitations and challenges of real data.

The increasing demand for data-driven solutions and AI applications across various industries, such as healthcare, finance, manufacturing, transportation, and media. Synthetic data can help these industries to leverage the benefits of artificial intelligence and machine learning, such as improving efficiency, accuracy, and innovation



The rising concerns and regulations regarding data privacy and security, especially for sensitive and personal data. Synthetic data can help to address these issues by creating data that preserves the statistical properties and patterns of real data but does not contain any identifiable or confidential information. Synthetic data can also help to comply with data protection laws, such as PIPEDA, GDPR and CCPA

The scarcity and costliness of real data, especially for rare or complex scenarios. Synthetic data can help to overcome these challenges by creating data that covers a wide range of situations and conditions that may not be available or feasible in real data. Synthetic data can also help to reduce the time and effort required for data collection and labeling, which are often tedious and expensive tasks.

CREATED BY AI FOR AI

AI is a transformative technology that can bring significant benefits and opportunities to businesses, society, and humanity. AI needs more and more data to achieve its full potential and to solve the most challenging and impactful problems.

AI: Hungry for data, thirsty for knowledge

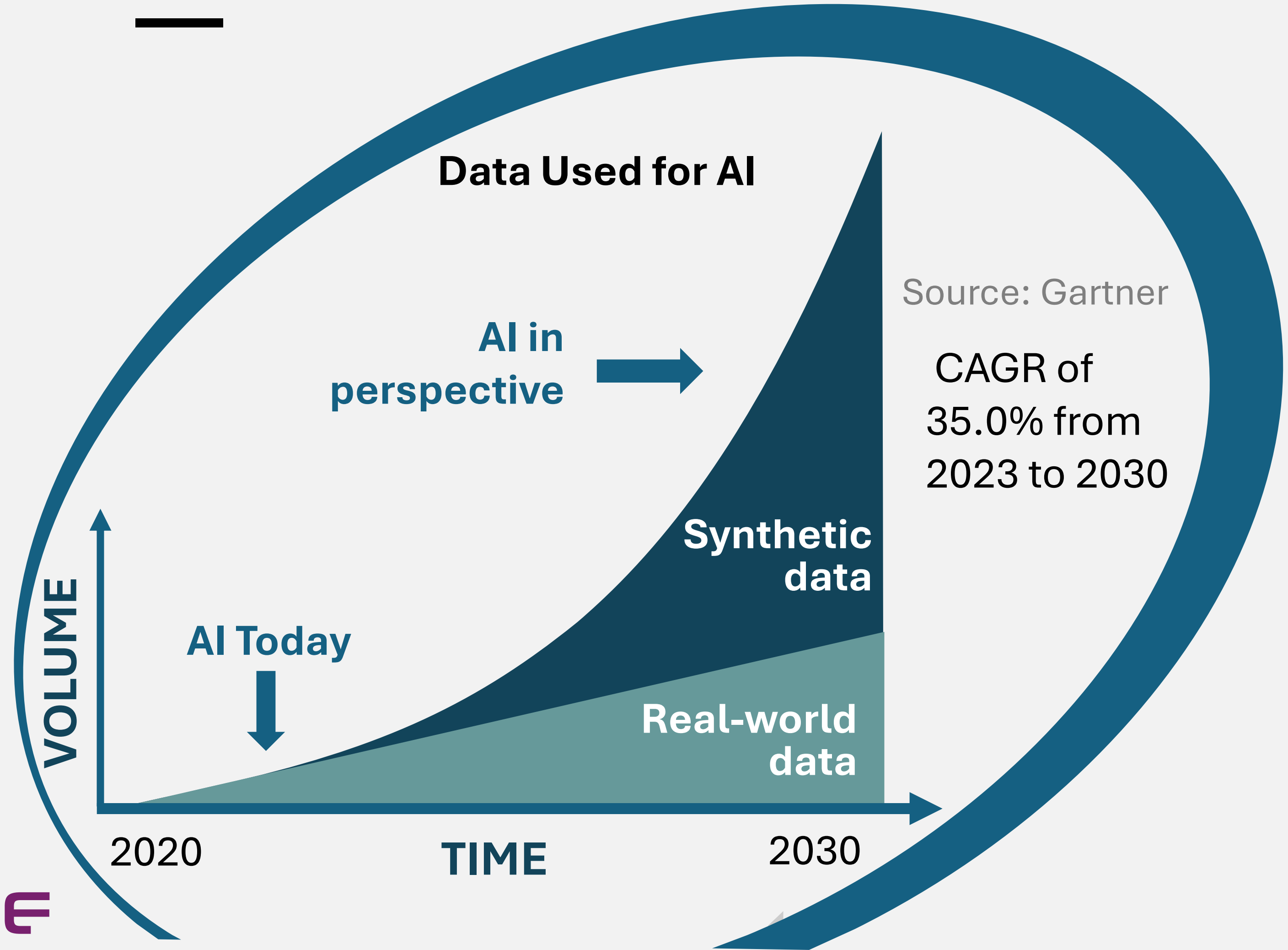
Synthetic data is a powerful tool for overcoming the challenges and limitations of real data, such as privacy, cost, scarcity, and diversity. Synthetic data can enable AI to learn from realistic and rich scenarios that are not feasible or available in real data, and thus improve the accuracy, reliability, and innovation of AI models.

Synthetic data: AI’s way of learning from itself

Synthetic data is data that is artificially generated by AI algorithms, instead of being collected from real-world sources. Synthetic data can help AI to overcome the limitations and challenges of real data, such as privacy, cost, scarcity, and diversity. Synthetic data can enable AI to learn from realistic and rich scenarios that are not feasible or available in real data, and thus improve the accuracy, reliability, and innovation of AI models. Synthetic data is AI’s way of learning from itself, by creating its own data that suits its needs and goals.

SYNTHETIC DATA VS REAL-WORLD DATA

The demand for synthetic data will increase exponentially till 2030, as more industries and applications adopt artificial intelligence and machine learning, and face the challenges and limitations of real data, such as privacy, cost, scarcity, and diversity.

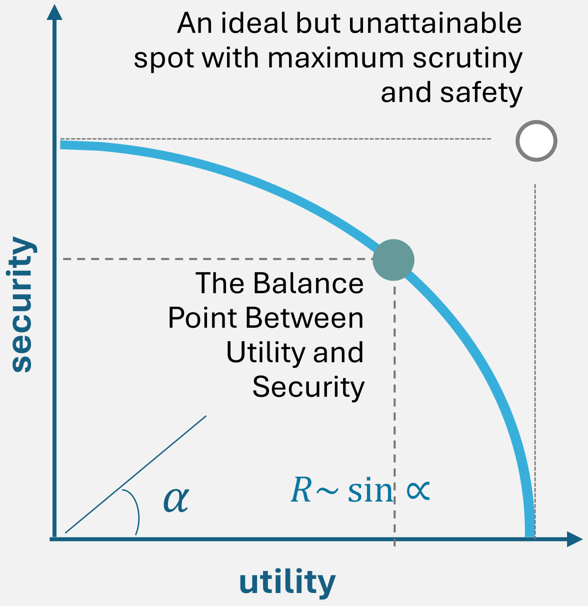


- **The benefits of synthetic data** include reducing the constraints associated with using regulated or sensitive data, customizing data to suit specific conditions or scenarios, and creating large and diverse datasets without manual labeling
- **Limitations associated with synthetic data** include ensuring the quality and reliability of the data generated, preserving the statistical properties and correlations of real data, and eliminating the ethical and legal implications of creating and using synthetic data.

SYNTHETIC DATA PRIVACY

Synthetic data, based on real data, may reveal private information due to its ‘fingerprint’. This requires careful methods to create and use synthetic data.

- **Quality Metrics for Synthetic Data:** The utility of synthetic data is largely gauged by its quality — specifically, how accurately it mirrors the real data's inherent characteristics, including correlations and dependencies. Assessing this quality necessitates the use of sophisticated metrics designed to quantify the fidelity and utility of the synthetic dataset.
- **Trade-off Between Quality and Privacy:** Similar to the challenges faced with real datasets, synthetic data is subject to a pivotal trade-off between quality and privacy. High-fidelity synthetic data might inadvertently lead to the risk of re-identification, wherein the synthetic dataset reveals identifiable links to real-world data, potentially exposing sensitive attributes.
- **Risk Assessment and Mitigation:** To navigate the delicate balance between data utility and privacy, it's crucial to quantify the level of risk. Employing specialized risk models enables stakeholders to assess potential privacy threats tailored to the data's nature and the contexts of its application. This risk-informed approach paves the way for crafting a harmonized balance between data quality and privacy safeguards.

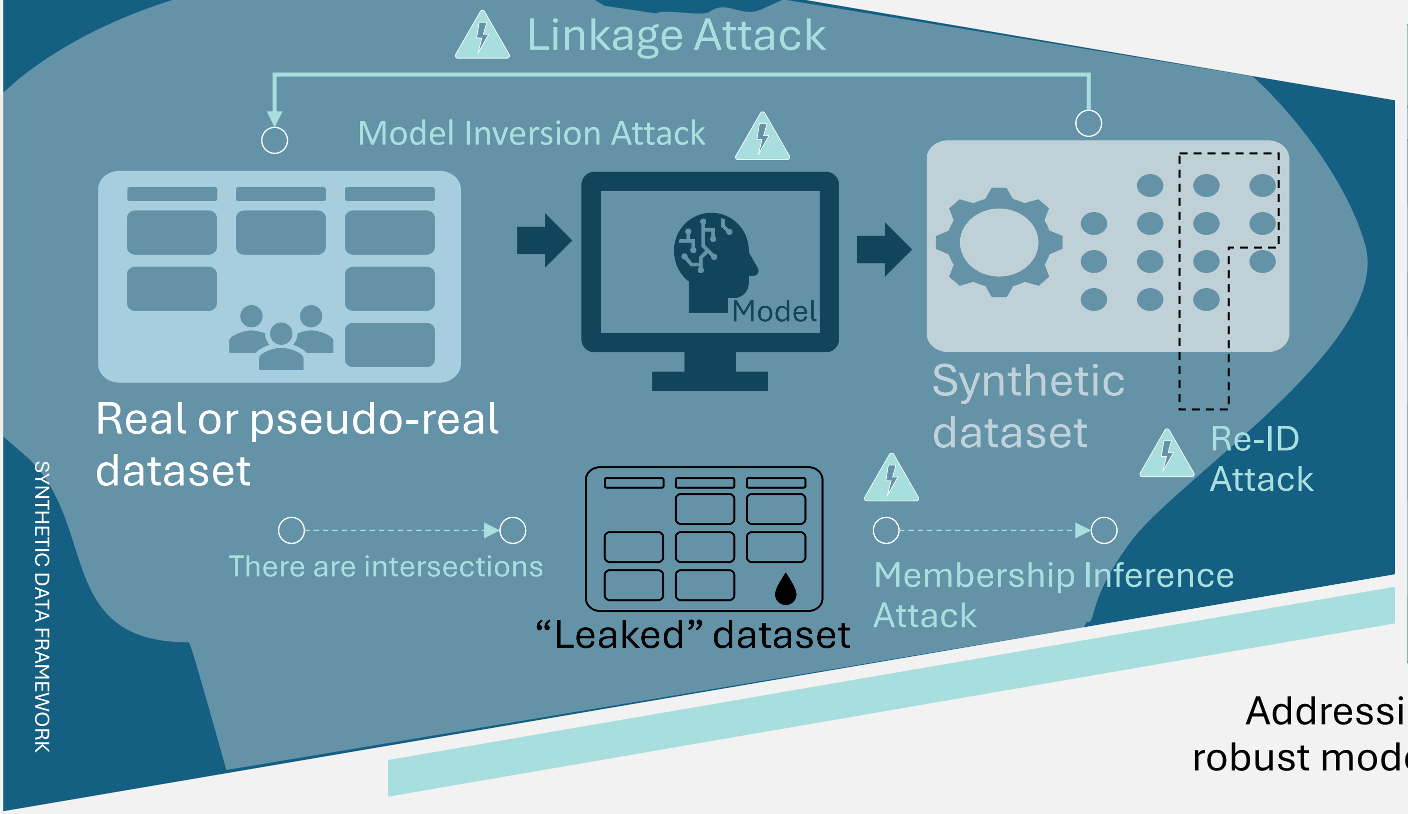


Merely generating synthetic data involving personal information is not an end in itself. It's imperative to constantly seek an equilibrium between data quality and security. In traditional data scenarios, techniques like k-anonymity are commonly employed to preserve privacy. For synthetic data engineered through AI methodologies, the principle of differential privacy stands out as a robust mechanism, offering a structured approach to managing privacy risks while retaining the utility of the synthetic datasets. This approach ensures that the synthetic data serves its intended purpose without compromising individual privacy.

THE DARK SIDE OF SYNTHETIC DATA

While synthetic data promises enhanced privacy and boundless analytical opportunities, it harbors hidden dangers. Understanding the potential for misuse and exposure is crucial.

By establishing a robust threat model and meticulously examining the attack surface, we can anticipate, recognize, and fortify against malicious exploits. This proactive stance is not just about defense; it's a commitment to the responsible and secure use of synthetic data.



SYNTHETIC DATA ATTACKS

Attack Type	Description	Consequence
Model Inversion Attack	Adversaries use the model's output to infer sensitive input data.	Exposure of sensitive information, compromising individual privacy.
Membership Inference Attack	Attackers determine if specific data was part of the model's training set.	Potential identification of individual data contributions, leading to privacy breaches.
Data Poisoning	Malicious data is introduced into the training set, affecting learning.	Compromised model integrity, leading to skewed or harmful outputs.
Adversarial Manipulation	Deceptive data input exploits model vulnerabilities, causing wrong outputs.	Eroded trust in model accuracy and potential manipulation for nefarious purposes.
Model Stealing/Extraction	Reverse-engineering a model to replicate its functionality and data.	Unauthorized access and potential misuse of proprietary algorithms and data insights.
Re-identification Attack	Cross-referencing anonymized data with external sources to identify individuals.	Violation of anonymity guarantees, leading to privacy invasions and potential legal ramifications.
Attribute Inference Attack	Using model outputs to infer sensitive attributes of individuals in the dataset.	Exposure of sensitive attributes, leading to privacy breaches and potential misuse of data.

Addressing these threats necessitates a layered approach, combining robust model design, thorough data sanitization, continuous monitoring, and a keen awareness of the evolving threat landscape.

PXP : THE PARAGON OF SYNTHETIC DATA SECURITY

PxP Framework is ultimate safeguard in the synthetic data universe. Merging cutting-edge technology with ironclad security measures

Foundations of Safety: PXP exists at the crossroads of innovation and security, utilizing the principles of differential privacy to generate synthetic data that upholds the highest standards of safety and confidentiality.

Architectural Mastery: PXP harnesses the power of PATE-GAN for robust synthetic tabular data generation, optionally integrating DP-auto-GAN with DT-GAN for enhanced versatility. For semi-structured data, DP-Former, a transformer fortified with differential privacy, stands as a sentinel of structure and meaning.

Precision and Correlation Guard: The framework isn't just about data generation; it's about perfection in replication. With an embedded quality model, PXP meticulously calibrates the precision and accuracy of synthetic data, preserving correlations with the precision of a master jeweler. For semi-structured data, BLEU & ROUGE-X metrics serve as the twin lenses of clarity and quality.

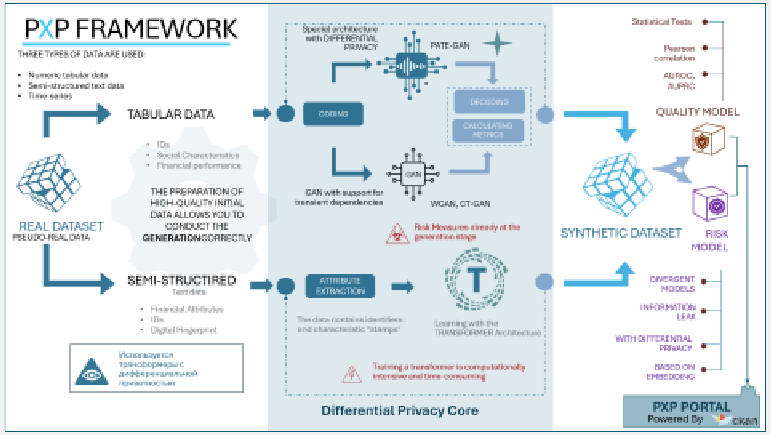
Privacy Permanence Matrix: PXP predicts. Employing a comprehensive suite of risk models — Information Leakage, Divergent Approach, and DP-Risk Model — PXP offers a panoramic view of data security, assuring that tabular synthetic data is shielded in a fortress of privacy. For semi-structured data, the Embedding Model with cosine similarity serves as the vigilant guardian, ensuring that every data narrative is both rich and secure.

Synthetic Data Sanctuary: PXP introduces an unparalleled portal, a central command for managing synthetic datasets and their metrics. It's not just a tool; it's a sanctuary where data is not only generated and assessed but revered and safeguarded.

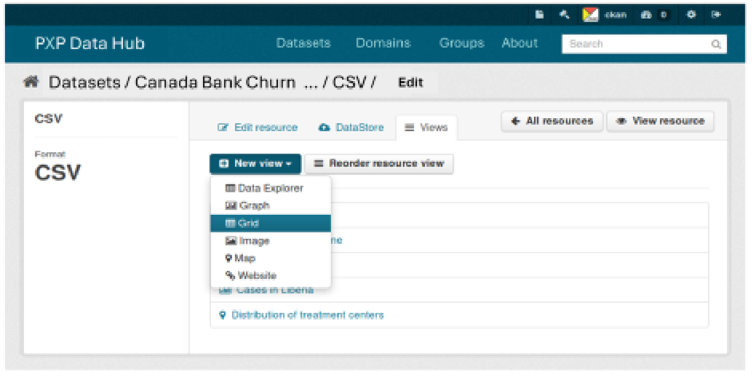
PXP ARCHITECTURE

The PXP framework is not just a structure; it's an ecosystem where every element plays a pivotal role in safeguarding the sanctity of synthetic data.

- **Differential Privacy Core**: Embeds Differential Privacy principles to ensure individual data anonymity while allowing aggregate data analysis.
- **PATE-GAN for Tabular Data**: Utilizes PATE-GAN to generate synthetic tabular data, balancing data utility with individual privacy.
- **DP-Former for Semi-Structured Data**: Implements DP-Former, integrating transformer architecture with differential privacy for generating semi-structured data.
- **Quality & Correlation**: Features an inbuilt quality model to assess the precision, accuracy, and preservation of correlations in synthetic data.
- **Multi-Faceted Risk Assessment Matrix**: Includes Information Leakage, Divergent Approach, and DP-Risk Model to provide comprehensive risk evaluation for synthetic data.
- **Embedding Model with Cosine Similarity**: Employs an Embedding Model coupled with cosine similarity measures to maintain semantic integrity in semi-structured data.
- **PXP Portal - Central Management Hub**: Offers a centralized portal for managing, evaluating, and monitoring synthetic datasets and related metrics.



PXP Portal Data Hub (powered by CKAN)



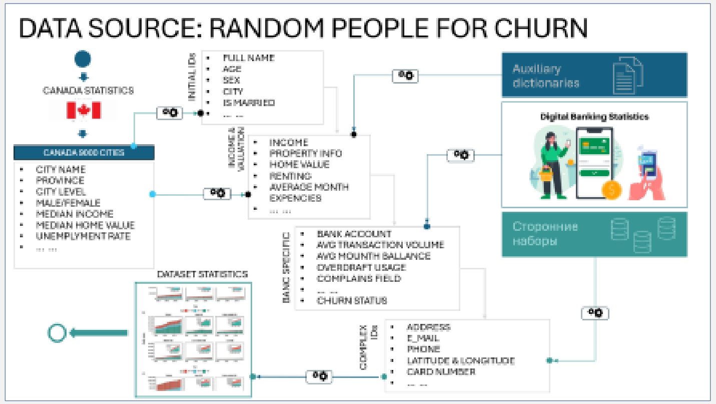
SOURCE QUALITY FOR SYNTHETIC DATA GENERATION

The genesis of every synthetic data ecosystem lies in the quality of its source data. Like a master artist requires pristine canvas and colors, high-quality synthetic data generation demands a well-curated, robust source dataset. It's not just about quantity

- **Richness in Source Data**: The bedrock of high-quality synthetic data isn't merely its volume but the wealth of attributes, intricate statistics, and inherent correlations. These elements are the raw materials for the synthetic data generation model, teaching it the subtle dance of relationships between different attributes and ensuring the preservation of underlying patterns without falling prey to Bayesian traps or biases.

- **Pseudo-Real as a Training Ground**: Synthetic models thrive on diversity. Training on pseudo-real data, constructed through rule-based algorithms or statistical methods, provides a robust playground for models. Within the PXP framework, this diversified training regimen is recognized under the term 'RANDOM DATASET,' distinguishing it from purely 'SYNTHETIC DATASET', and enriching the model's ability to navigate and replicate the complexities of real-world data.

- **Normalization and Anonymization**: Preparing the real and pseudo-real datasets for the generation journey involves more than just standardization. It's a meticulous process of normalization, assigning arbitrary identifiers and employing sophisticated tools within the framework. These tools skillfully replace sensitive information like addresses, names, documents, contact details, and coordinates with intelligently generated random values. This process ensures that the synthetic data mirrors the original's range and distribution, all while upholding the sanctity of privacy and data security.



QUALITY MODEL

The genesis of every synthetic data ecosystem lies in the quality of its source data. Like a master artist requires pristine canvas and colors, high-quality synthetic data generation demands a well-curated, robust source dataset. It's not just about quantity

1. Fidelity to Original Data:

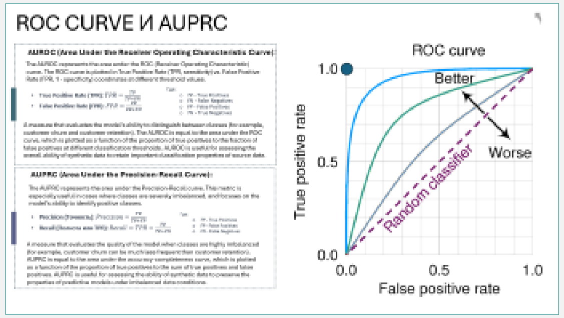
- **Statistical Similarity:** Measures how closely the synthetic data matches the original data in terms of statistical distributions (e.g., mean, variance) for individual attributes.
- **Correlation Preservation:** Ensures that the relationships and dependencies between different attributes in the original data are accurately reflected in the synthetic data.

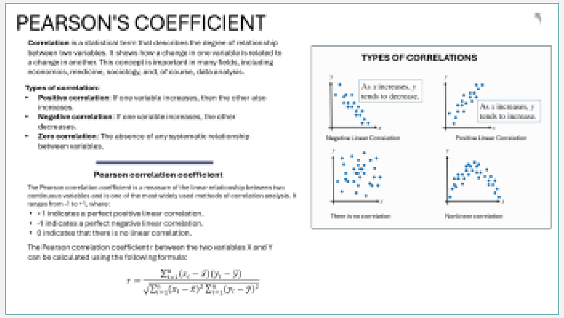
2. Utility and Usability:

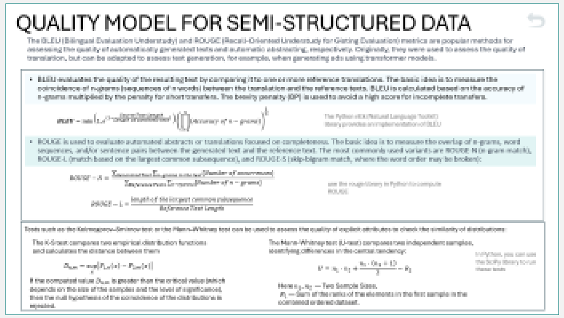
- **Task-Specific Quality:** Evaluates how well the synthetic data performs in specific downstream tasks, such as training machine learning models, compared to the original data.
- **Data Completeness:** Checks for missing values or data sparsity that could affect the utility of the synthetic data.

3. Data Diversity and Coverage:

- **Variability:** Ensures the synthetic data contains sufficient variability and does not overly replicate specific data points, leading to overfitting.
- **Representation of Minority Groups:** Checks that the synthetic data adequately represents all subgroups or demographics present in the original data, avoiding bias.







RISK MODEL

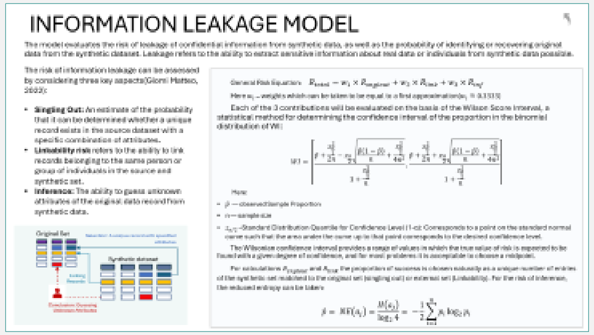
Risk in the context of synthetic data generation pertains to the potential for data to expose sensitive information, lead to incorrect conclusions, or otherwise harm individuals or organizations

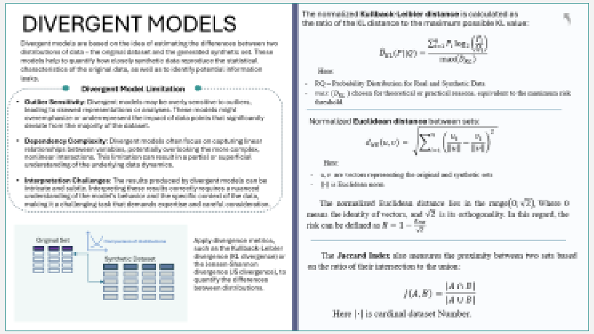
Security metrics are calculated using a risk model. **A risk model** is a framework or method for measuring, evaluating, and managing the privacy risk associated with synthetic data. A risk model typically consists of the following components:

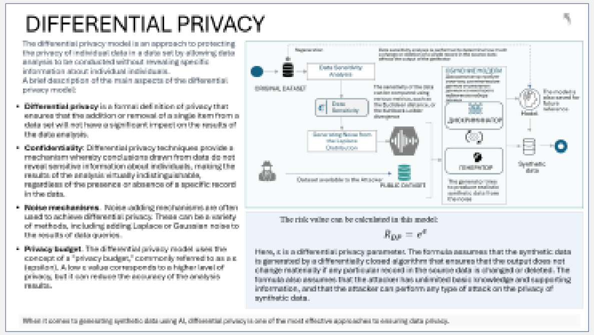
- **A threat** model that identifies an attacker's capabilities, goals, and strategies to execute attacks on the privacy of synthetic data.
- **A privacy criterion** that determines the desired level of protection or acceptable level of risk for synthetic data.
- **A risk metric** that quantifies the privacy risk of synthetic data by comparing synthetic data to raw data, or by assessing the likelihood or impact of attacks on the privacy of synthetic data.
- **Risk score,** which applies a risk metric to synthetic data and source data and calculates a risk score or risk level for synthetic data.
- **Risk mitigation,** in which some technique or mechanism is applied to reduce the privacy risk of synthetic data, such as adding noise, distorting features, or synthesizing new data.

SPECIFIC RISK MODELS

- INFORMATION LEAKAGE
- DIVERGENT MODEL
- DIFFERENTIAL PRIVACY
- EMBEDDING BASED

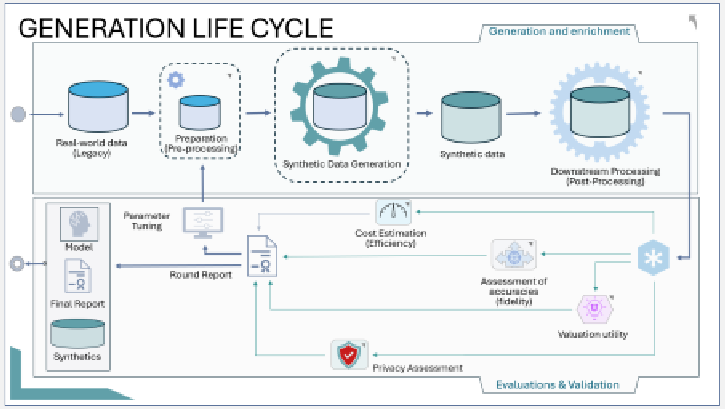






ORCHESTRATING THE SYNTHETIC DATA SYMPHONY

Embark on a journey through the synthetic data generation cycle, a meticulously orchestrated symphony of steps, each playing a crucial role in transforming raw data into a synthetic masterpiece..



Upstream Process:

- **Importance of Preparation:** The pre-processing phase is fundamental, as it ensures the data is clean, consistent, and structured. This step lays the groundwork, much like tuning instruments before a performance, to ensure the subsequent synthetic data generation is accurate and effective.

- **Necessity of Coding and Decoding :** Coding transforms complex, real-world data into a more abstract, manageable form, enabling sophisticated generative models to learn and mimic intricate data patterns. Decoding then reverses this process, converting the abstract representations back into tangible, meaningful data. This two-part act ensures that the essence of the original data is captured and reflected in the synthetic version.

Downstream Process:

- **Metric Calculation - Measuring Excellence:** Metric calculation involves evaluating the synthetic data against various performance indicators. This step is akin to a dress rehearsal, where each aspect of the synthetic data is scrutinized to ensure it meets the desired standards of quality, utility, and resemblance to the original data.

- **Dual-Stream Approach - A Balanced Ensemble:** The synthetic data generation cycle is a dual-stream process, encompassing both the upstream (preparation and creation) and downstream (assessment and refinement) phases. This balanced approach ensures a comprehensive and meticulous treatment of data, akin to how both composition and critique are vital in creating a musical masterpiece. The upstream stream focuses on generating the data, while the downstream stream ensures the data's quality, respects privacy, and aligns with the intended use cases.

GENERATION APPROACH

Step into the realm of Generative Adversarial Networks (GANs) and Private Aggregation of Teacher Ensembles (PATE-GAN), where innovation meets privacy in the generation of synthetic data.

TABULAR DATA WITH DP

GAN Core Mechanism: GANs consist of two neural networks, the generator and the discriminator, engaged in a continuous game. The generator creates synthetic data, while the discriminator evaluates its authenticity, leading to a dynamic training process where the generator strives to produce increasingly realistic data.

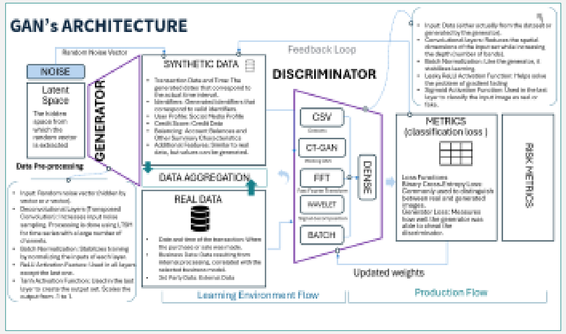
PATE-GAN : Privacy-Centric Approach: PATE-GAN extends the principles of GANs by integrating the PATE framework. It focuses on generating synthetic data that preserves privacy, making it particularly suitable for sensitive datasets where data utility and individual privacy must be carefully balanced.

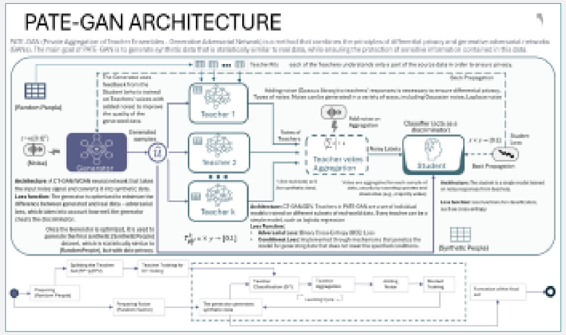
Mechanism: PATE-GAN leverages an ensemble of teacher models, each trained on disjoint subsets of the private data. It then aggregates their knowledge to guide the training of a student model (the generator) in a differentially private manner, ensuring that the synthetic data generated does not compromise individual privacy.

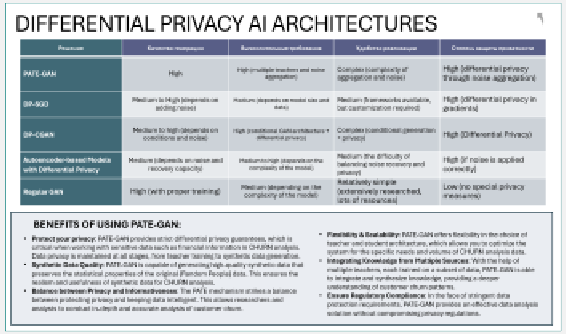
Benefits and Challenges: PATE-GAN provides strong privacy guarantees, making it an ideal choice for scenarios requiring compliance with stringent data protection standards.

Quality of Synthetic Data: While PATE-GAN aims to generate high-quality synthetic data, the balance between data utility and privacy is delicate and requires careful tuning of model parameters.

Computational Complexity: The intricate architecture of PATE-GAN, involving multiple teacher models and a student model, can lead to increased computational complexity and resource requirements.



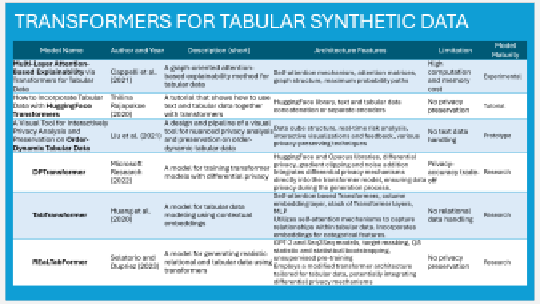


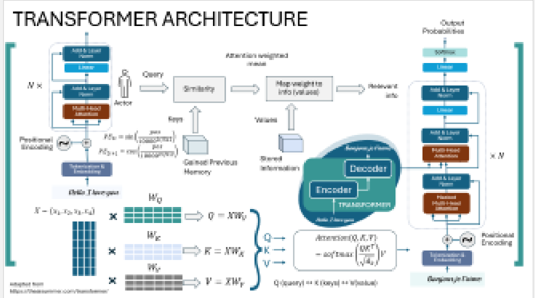


INTEGRATING TRANSFORMER APPROACH IN P²

From high computational demands to privacy concerns, and the intricate task of processing tabular data, the path is fraught with obstacles. Yet, the potential for transformative change in handling semi-structured data is undeniable.

- **Transformers at the AI Frontier**: Transformers are a breakthrough in AI, especially in NLP, yet they come with significant limitations. They are resource-intensive, often centralized, raising privacy concerns, and are not inherently designed for handling tabular numerical data where capturing statistical dependencies is crucial.
- **Harnessing Transformers for Semi-Structured Data**. In specific scenarios, such as processing semi-structured data where text intertwines with structured attributes, transformers can be effectively utilized. By forming a stream of text information transformation, transformers can augment other generation methods, enhancing the richness and accuracy of the generated content.
- **Adapting Transformers for Tabular Data and Privacy**. Initiatives to mold transformer models for tabular data processing and integrate differential privacy are in motion. Prominent examples in this realm are undergoing rigorous experimentation and research, with most not yet primed for commercial deployment. These efforts aim to preserve the intuitive data handling of transformers while fortifying privacy and adapting to the unique structure of tabular data.
- **Quality and Risk Model Adaptation**: To effectively incorporate transformer models in semi-structured data processing within P², both the quality and risk models require meticulous adaptation. This involves accommodating the nuances of word and attribute vectorization (embedding), ensuring that the generated synthetic data maintains high fidelity to the original, both in content and context, while upholding stringent standards of privacy and data security.

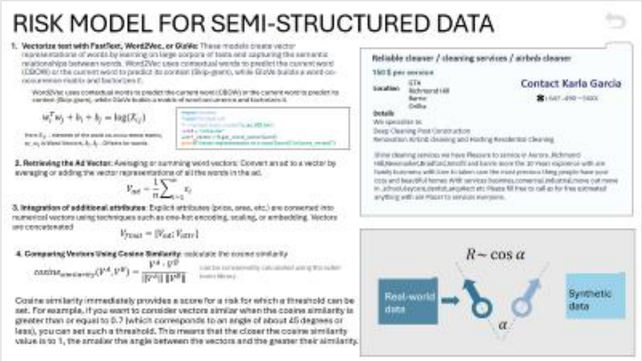


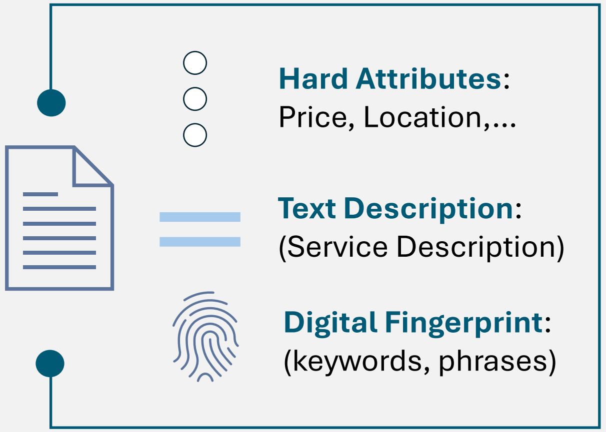


EMBEDDING-BASED RISK MODEL

In the intricate world of semi-structured synthetic data, safeguarding privacy while maintaining data utility is a fine art. Our Embedding-Based Risk Model is the guardian at the gates, leveraging advanced embedding techniques and cosine similarity measures to navigate the subtleties of data security.

- **Embedding as the Foundation**: Harness the power of embedding techniques to transform semi-structured text data into numerical vectors, laying a robust foundation for precise risk assessment and maintaining the richness of semantic information.
- **Attribute Integration**: Seamlessly concatenate embeddings of textual data with vectors representing additional attributes, ensuring a comprehensive representation of the dataset's multi-faceted nature and intricacies.
- **Cosine Similarity for Insightful Comparisons**: Employ cosine similarity to measure the closeness between vectors, enabling a nuanced understanding of the relationships and similarities within the data, crucial for identifying potential privacy risks and data linkages.
- **Preserving Privacy in Vector Space**: Meticulously monitor the distribution of vector representations to ensure that the transformation into the embedding space does not compromise privacy, keeping individual data points indistinguishable and secure.
- **Iterative Refinement for Enhanced Security**: Continuously refine the embedding techniques and similarity thresholds based on feedback loops and risk assessment outcomes, fostering a model that evolves and adapts to emerging threats and data landscapes.


RISK MODEL FOR SEMI-STRUCTURED DATA

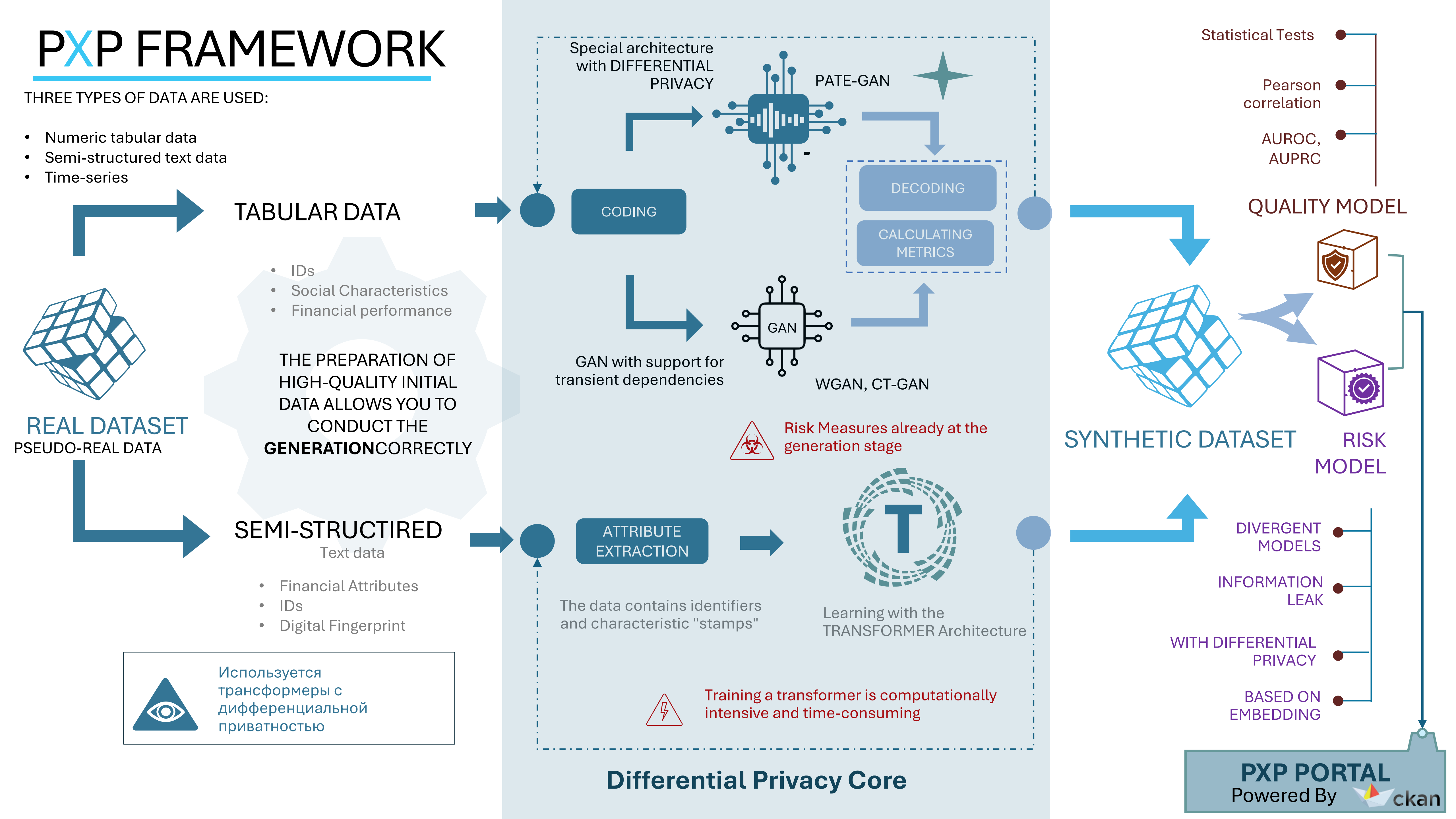


THANK YOU

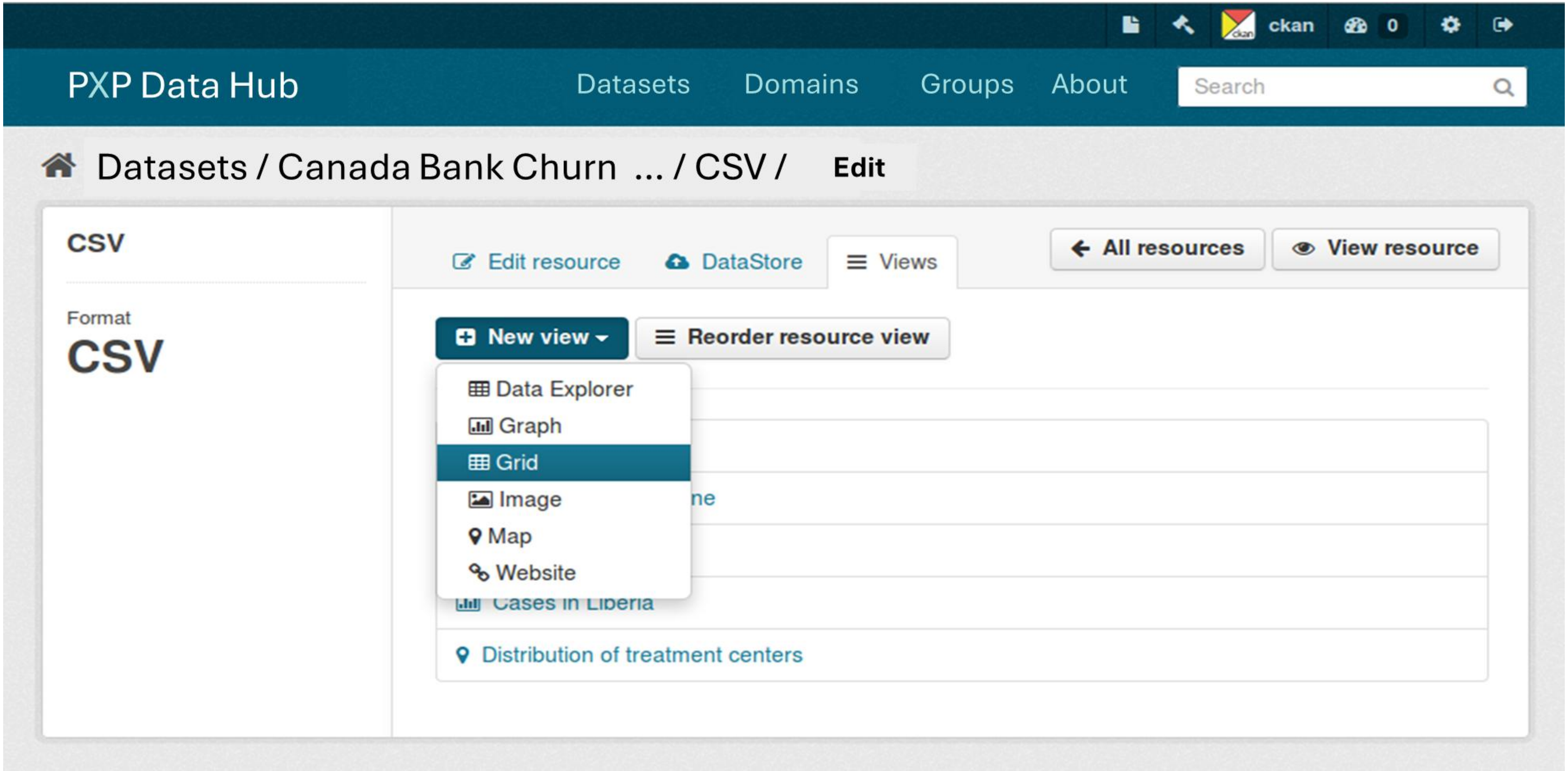
APPENDIX

SYNTHETIC DATA ATTACKS

Attack Type	Description	Consequence
Model Inversion Attack	Adversaries use the model's output to infer sensitive input data.	Exposure of sensitive information, compromising individual privacy.
Membership Inference Attack	Attackers determine if specific data was part of the model's training set.	Potential identification of individual data contributions, leading to privacy breaches.
Data Poisoning	Malicious data is introduced into the training set, affecting learning.	Compromised model integrity, leading to skewed or harmful outputs.
Adversarial Manipulation	Deceptive data input exploits model vulnerabilities, causing wrong outputs.	Eroded trust in model accuracy and potential manipulation for nefarious purposes.
Model Stealing/Extraction	Reverse-engineering a model to replicate its functionality and data.	Unauthorized access and potential misuse of proprietary algorithms and data insights.
Re-identification Attack	Cross-referencing anonymized data with external sources to identify individuals.	Violation of anonymity guarantees, leading to privacy invasions and potential legal ramifications.
Attribute Inference Attack	Using model outputs to infer sensitive attributes of individuals in the dataset.	Exposure of sensitive attributes, leading to privacy breaches and potential misuse of data.


PXP FRAMEWORK
THREE TYPES OF DATA ARE USED:
• Numeric tabular data
• Semi-structured text data
• Time-series
TABULAR DATA
• IDs
• Social Characteristics
• Financial performance
THE PREPARATION OF HIGH-QUALITY INITIAL DATA ALLOWS YOU TO CONDUCT THE GENERATIONCORRECTLY
REAL DATASET
PSEUDO-REAL DATA
SEMI-STRUCTIRED
Text data
• Financial Attributes
• IDs
• Digital Fingerprint
Используется трансформеры с дифференциальной приватностью
Special architecture with DIFFERENTIAL PRIVACY
PATE-GAN
CODING
DECODING
CALCULATING METRICS
GAN
GAN with support for transient dependencies
WGAN, CT-GAN
Risk Measures already at the generation stage
ATTRIBUTE EXTRACTION
T
The data contains identifiers and characteristic "stamps"
Learning with the TRANSFORMER Architecture
Training a transformer is computationally intensive and time-consuming
Differential Privacy Core
Statistical Tests
Pearson correlation
AUROC, AUPRC
QUALITY MODEL
SYNTHETIC DATASET
RISK MODEL
DIVERGENT MODELS
INFORMATION LEAK
WITH DIFFERENTIAL PRIVACY
BASED ON EMBEDDING
PXP PORTAL
Powered By
ckan

PXP Portal Data Hub (powered by CKAN)



DATA SOURCE: RANDOM PEOPLE FOR CHURN

CANADA STATISTICS

CANADA 9000 CITIES

- CITY NAME
- PROVINCE
- CITY LEVEL
- MALE/FEMALE
- MEDIAN INCOME
- MEDIAN HOME VALUE
- UNEMPLYMENT RATE
- … …

INITIAL IDs

- FULL NAME
- AGE
- SEX
- CITY
- IS MARRIED
- … …

INCOME & VALUATION

- INCOME
- PROPERTY INFO
- HOME VALUE
- RENTING
- AVERAGE MONTH EXPENCIES
- … …

BANC SPECIFIC

- BANK ACCOUNT
- AVG TRANSACTION VOLUME
- AVG MOUNTH BALLANCE
- OVERDRAFT USAGE
- COMPLAINS FIELD
- … …
- CHURN STATUS

COMPLEX IDs

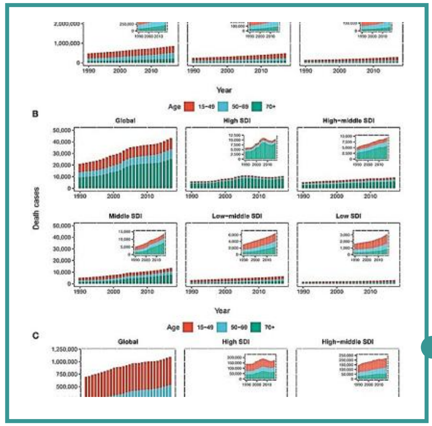
- ADDRESS
- E_MAIL
- PHONE
- LATITUDE & LONGITUDE
- CARD NUMBER
- … …

Auxiliary dictionaries

Digital Banking Statistics

Сторонние наборы

DATASET STATISTICS



{CHURN RANDOM People} Attributes

#	Attribute	Type	Description	Churn
1	ID	Long	On-premises client ID	id
2	FULLNAME	String	Client's full name (last name and first name)	q-id
3	SEX	String	The sex of the individual. Examples of values: 'M', 'F'	score
4	AGE	Integer	The age of the individual. Range: 20 to 80	score
5	IsMarried	Boolean	Marital status. True if married, False otherwise	score
6	FAMILYID	String	Family ID. If not empty, points to family members	score
7	RACE	String	Nationality and other characteristics	score
8	Income	Float	The level of income of an individual.	score
9	IsHomeOwner	Boolean	Property ownership status. True if owned, False otherwise	score
10	HOMEVALUE	Float	The value of the property.	score
11	RENTVALUE	Float	Estimating the amount of rent if he rents a house.	score
12	IsEduBachelors	Boolean	Level of education. True if the level of education is not lower than higher education, False otherwise	score
13	IsUnemployed	Boolean	Employment status. True if unemployed, False otherwise	score
14	AME	Float	Average spending per month (estimate).	score
15	AMB	Float	Average monthly account balance.	score
16	credit_score	Integer	Credit rating of an individual.	score
17	AccountType	String	Account type. Examples of values: Debit, Joint, Credit, Investment	score
18	account_age	Integer	The age of the account in years.	score
19	avg12_tx	Integer	The average monthly number of transactions over the past 12 months.	score
20	BankIsPrimary	Boolean	Whether the bank is the main bank. True if yes, False otherwise	score
21	avg12_tx_volume	Float	Average monthly transaction volume over the last 12 months.	score
22	loan_status	String	Credit status of an individual. Examples of values include "Mortgage", "Personal loan", or blank if there is no credit	score
23	credit_card_status	Boolean	Credit card status. True if the customer is using a credit card, False otherwise	score
24	overdraft_usage	Boolean	Use of an overdraft. True if used, False otherwise	score
25	branch_visits	Integer	Number of physical visits to the bank in 12 months. Range: 0-12+	score
26	digital_usage_level	String	The extent to which digital channels are used. Examples of values: Low, Medium, High	score
27	customer_service	Integer	Number of support calls. Range: 0-20+	score
28	complaints_filed	Integer	Number of complaints filed during the period. Range: 0-10+	score
29	satisfaction_level	Integer	Rating of satisfaction with the bank's services. Range: 0-10	score
30	BankPresenceRating	Float	Penetration rate (current market share) in the city. For example, 0.15	score
31	Address	String	The physical address of an individual.	q-id
32	Postal Code	String	The postal code associated with the address.	q-id
33	Latitude	Float	The geographic latitude of the address.	q-id
34	Longitude	Float	Geographic longitude of the address.	q-id
35	Distance to Metropolis	Float	Distance to the nearest metropolis from the address.	score
36	Mobile Phone	String	An individual's mobile phone number.	Id
37	Card Number	String	A bank card number linked to an individual.	Id
38	Card Type	String	Type of bank card. For example, 'Credit', 'Debit'	q-id
39	E-MAIL	String	An individual's email address.	q-id
40	ChurnProbability	Float	The estimated probability that a person will leave the bank. (0-1)	Target

ROC CURVE И AUPRC

AUROC (Area Under the Receiver Operating Characteristic Curve):

The AUROC represents the area under the ROC (Receiver Operating Characteristic) curve. The ROC curve is plotted in True Positive Rate (TPR, sensitivity) vs. False Positive Rate (FPR, 1 - specificity) coordinates at different threshold values.

- **True Positive Rate (TPR)**: $TPR = \frac{TP}{TP+FN}$
- **False Positive Rate (FPR)**: $FPR = \frac{FP}{FP+TN}$

Где:

- *TP* - True Positives
- *FN* - False Negatives
- *FP* - False Positives
- *TN* - True Negatives

A measure that evaluates the model's ability to distinguish between classes (for example, customer churn and customer retention). The AUROC is equal to the area under the ROC curve, which is plotted as a function of the proportion of true positives to the fraction of false positives at different classification thresholds. AUROC is useful for assessing the overall ability of synthetic data to retain important classification properties of source data.

AUPRC (Area Under the Precision-Recall Curve):

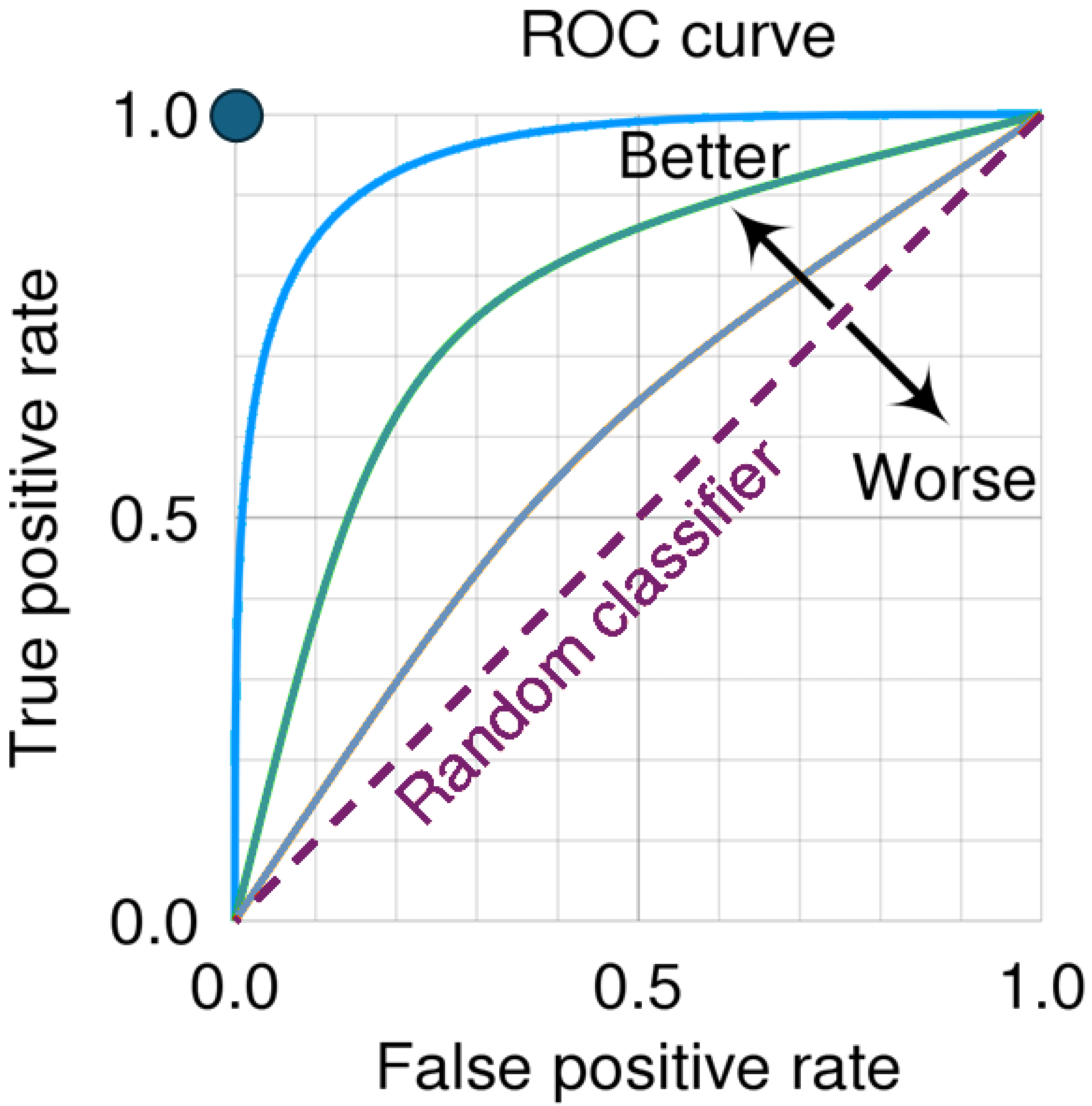
The AUPRC represents the area under the Precision-Recall curve. This metric is especially useful in cases where classes are severely imbalanced, and focuses on the model's ability to identify positive classes.

- **Precision (Точность)**: $Precision = \frac{TP}{TP+FP}$
- **Recall (Полнота или TPR)**: $Recall = TPR = \frac{TP}{TP+FN}$

Где:

- *TP* - True Positives
- *FP* - False Positives
- *FN* - False Negatives

A measure that evaluates the quality of the model when classes are highly imbalanced (for example, customer churn can be much less frequent than customer retention). AUPRC is equal to the area under the accuracy-completeness curve, which is plotted as a function of the proportion of true positives to the sum of true positives and false positives. AUPRC is useful for assessing the ability of synthetic data to preserve the properties of predictive models under imbalanced data conditions.

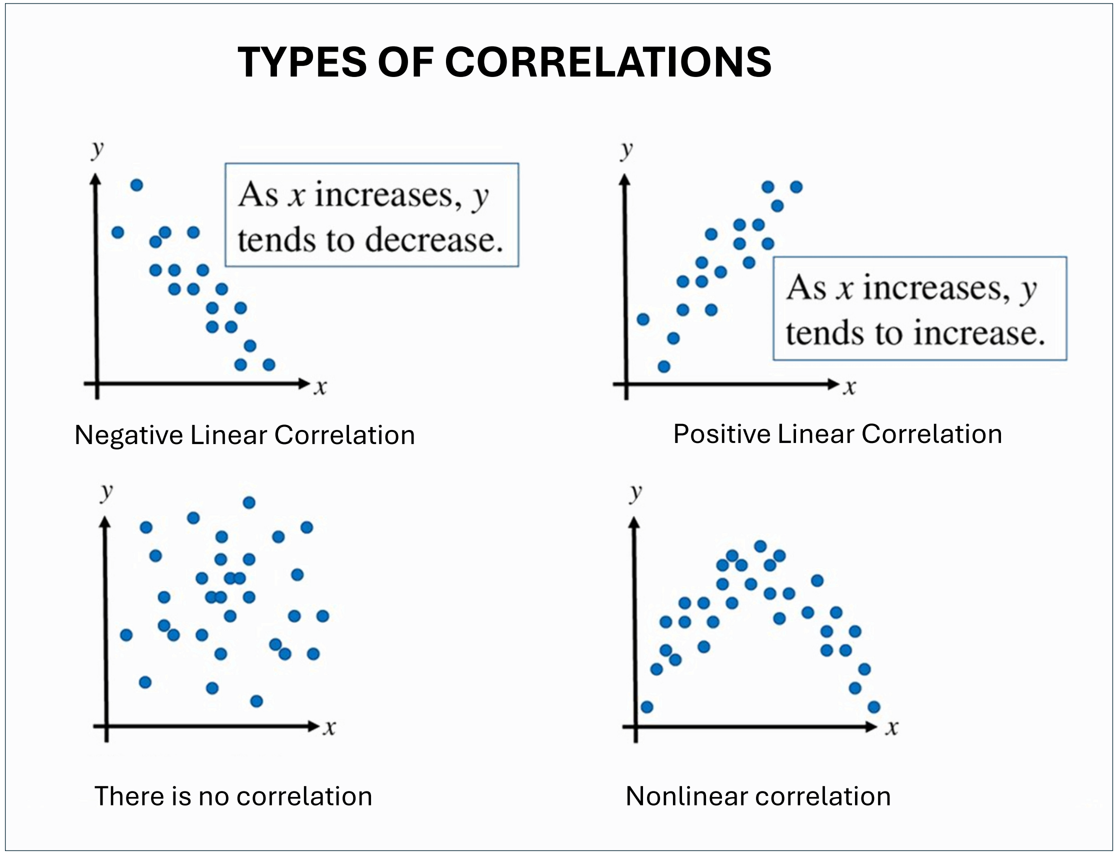


PEARSON'S COEFFICIENT

Correlation is a statistical term that describes the degree of relationship between two variables. It shows how a change in one variable is related to a change in another. This concept is important in many fields, including economics, medicine, sociology, and, of course, data analysis.

Types of correlation:

- **Positive correlation**: If one variable increases, then the other also increases.
- **Negative correlation**: If one variable increases, the other decreases.
- **Zero correlation**: The absence of any systematic relationship between variables.



Pearson correlation coefficient

The Pearson correlation coefficient is a measure of the linear relationship between two continuous variables and is one of the most widely used methods of correlation analysis. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear correlation.
- -1 indicates a perfect negative linear correlation.
- 0 indicates that there is no linear correlation.

The Pearson correlation coefficient r between the two variables X and Y can be calculated using the following formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

QUALITY MODEL FOR SEMI-STRUCTURED DATA

The BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are popular methods for assessing the quality of automatically generated texts and automatic abstracting, respectively. Originally, they were used to assess the quality of translation, but can be adapted to assess text generation, for example, when generating ads using transformer models.

- BLEU evaluates the quality of the resulting text by comparing it to one or more reference translations. The basic idea is to measure the coincidence of n-grams (sequences of n words) between the translation and the reference texts. BLEU is calculated based on the accuracy of n-grams multiplied by the penalty for short transfers. The brevity penalty (BP) is used to avoid a high score for incomplete transfers.

$$\boldsymbol{BLEU} = \min\left(1, e^{\left(1-\frac{Source\ Text\ Length}{Length\ of\ synthetic\ text}\right)}\right)\left(\prod_{i=1}^{n}(Accuracy\ of\ n-grams)\right)^{\frac{1}{n}}$$

The Python nltk (Natural Language Toolkit) library provides an implementation of BLEU

- ROUGE is used to evaluate automated abstracts or translations focused on completeness. The basic idea is to measure the overlap of n-grams, word sequences, and/or sentence pairs between the generated text and the reference text. The most commonly used variants are ROUGE-N (n-gram match), ROUGE-L (match based on the largest common subsequence), and ROUGE-S (skip-bigram match, where the word order may be broken):

$$ROUGE-N = \frac{\sum_{Generated\ Text}\sum_{n-grams\ in\ the\ text}\{Number\ of\ occurrences\}}{\sum_{Reference\ Texts}\sum_{n-\text{граммы}}\{Number\ of\ n-grams\}}$$

$$ROUGE-L = \frac{length\ of\ the\ largest\ common\ subsequence}{Reference\ Text\ Length}$$

use the rouge library in Python to compute ROUGE

Tests such as the Kolmogorov–Smirnov test or the Mann–Whitney test can be used to assess the quality of explicit attributes to check the similarity of distributions:

The K-S test compares two empirical distribution functions and calculates the distance between them

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

If the computed value $D_{n,m}$ is greater than the critical value (which depends on the size of the samples and the level of significance), then the null hypothesis of the coincidence of the distributions is rejected.

The Mann-Whitney test (U-test) compares two independent samples, identifying differences in the central tendency:

$$U = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

Here n_1, n_2 — Two Sample Sizes,
R_1 — Sum of the ranks of the elements in the first sample in the combined ordered dataset.

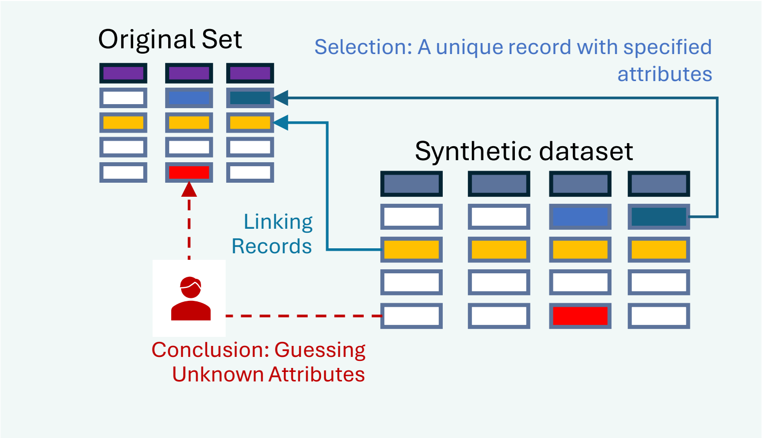
In Python, you can use the SciPy library to run these tests

INFORMATION LEAKAGE MODEL

The model evaluates the risk of leakage of confidential information from synthetic data, as well as the probability of identifying or recovering original data from the synthetic dataset. Leakage refers to the ability to extract sensitive information about real data or individuals from synthetic data possible.

The risk of information leakage can be assessed by considering three key aspects(Giomi Matteo, 2022):

- **Singling Out:** An estimate of the probability that it can be determined whether a unique record exists in the source dataset with a specific combination of attributes**.**
- **Linkability risk** refers to the ability to link records belonging to the same person or group of individuals in the source and synthetic set.
- **Inference:** The ability to guess unknown attributes of the original data record from synthetic data.



General Risk Equation: $R_{total} = w_1 \times R_{snglout} + w_2 \times R_{link} + w_3 \times R_{inf}$

Here w_i – weights which can be taken to be equal to a first approximation($w_i \approx 0.3333$)

Each of the 3 contributions will be evaluated on the basis of the Wilson Score Interval, a statistical method for determining the confidence interval of the proportion in the binomial distribution of WI:

$$WI = \left[\frac{\hat{p} + \frac{z_{\frac{\alpha}{2}}^2}{2n} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4n^2}}}{1 + \frac{z_{\frac{\alpha}{2}}^2}{n}}, \frac{\hat{p} + \frac{z_{\frac{\alpha}{2}}^2}{2n} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4n^2}}}{1 + \frac{z_{\frac{\alpha}{2}}^2}{n}} \right]$$

Here:

- $\hat{p}$ — observed Sample Proportion
- n — sample size
- $z_{\alpha/2}$ –Standard Distribution Quantile for Confidence Level (1-ɑ): Corresponds to a point on the standard normal curve such that the area under the curve up to that point corresponds to the desired confidence level.

The Wilsonian confidence interval provides a range of values in which the true value of risk is expected to be found with a given degree of confidence, and for most problems it is acceptable to choose a midpoint.

For calculations $R_{snglout}$ and R_{link} the proportion of success is chosen naturally as a unique number of entries of the synthetic set matched to the original set (singling out) or external set (Linkability). For the risk of inference, the reduced entropy can be taken:

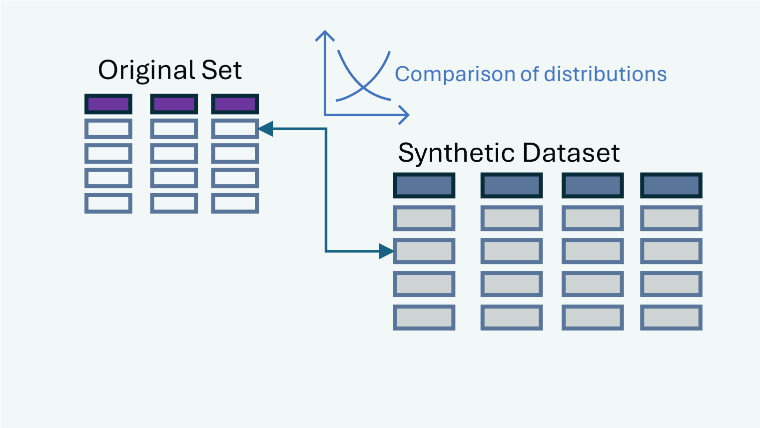
$$\hat{p} = NE(a_j) = \frac{H(a_j)}{\log_2 4} = -\frac{1}{2}\sum_{i=1}^{n} p_i \log_2 p_i$$

DIVERGENT MODELS

Divergent models are based on the idea of estimating the differences between two distributions of data – the original dataset and the generated synthetic set. These models help to quantify how closely synthetic data reproduce the statistical characteristics of the original data, as well as to identify potential information leaks.

Divergent Model Limitation

- **Outlier Sensitivity**: Divergent models may be overly sensitive to outliers, leading to skewed representations or analyses. These models might overemphasize or underrepresent the impact of data points that significantly deviate from the majority of the dataset.
- **Dependency Complexity**: Divergent models often focus on capturing linear relationships between variables, potentially overlooking the more complex, nonlinear interactions. This limitation can result in a partial or superficial understanding of the underlying data dynamics.
- **Interpretation Challenges**: The results produced by divergent models can be intricate and subtle. Interpreting these results correctly requires a nuanced understanding of the model's behavior and the specific context of the data, making it a challenging task that demands expertise and careful consideration.



Apply divergence metrics, such as the Kullback-Leibler divergence (KL divergence) or the Jensen-Shannon divergence (JS divergence), to quantify the differences between distributions.

The normalized **Kullback-Leibler distance** is calculated as the ratio of the KL distance to the maximum possible KL value:

$$\widehat{D}_{KL}(P||Q) = \frac{\sum_{i=1}^{n} P_i \log_2\left(\frac{P_i}{Q_i}\right)}{\max(D_{KL})}$$

Here:

- P,Q – Probability Distribution for Real and Synthetic Data
- $max\ (D_{KL}\ )$ chosen for theoretical or practical reasons, equivalent to the maximum risk threshold.

Normalized **Euclidean distance** between sets:

$$d_{NE}(u,v) = \sqrt{\sum_{i=1}^{n}\left(\frac{u_i}{\|u\|} - \frac{v_i}{\|v\|}\right)^2}$$

Here:

- u, v are vectors representing the original and synthetic sets
- $\|\cdot\|$ is Euclidean norm

The normalized Euclidean distance lies in the range$\left(0; \sqrt{2}\right)$, Where 0 means the identity of vectors, and $\sqrt{2}$ is its orthogonality. In this regard, the risk can be defined as $R = 1 - \frac{d_{NE}}{\sqrt{2}}$

The **Jaccard Index** also measures the proximity between two sets based on the ratio of their intersection to the union:

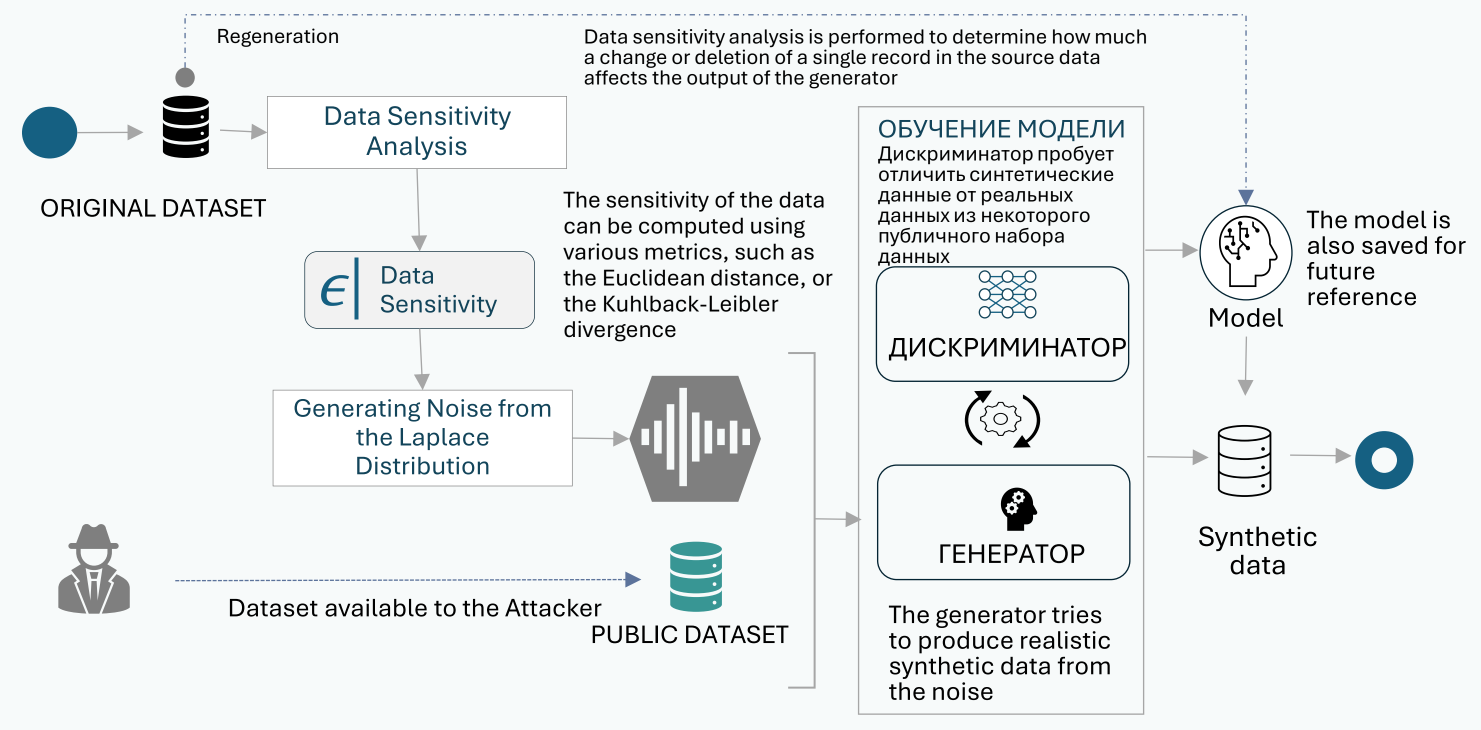
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Here $|\cdot|$ is cardinal dataset Number.

DIFFERENTIAL PRIVACY

The differential privacy model is an approach to protecting the privacy of individual data in a data set by allowing data analysis to be conducted without revealing specific information about individual individuals.
A brief description of the main aspects of the differential privacy model:

- **Differential privacy** is a formal definition of privacy that ensures that the addition or removal of a single item from a data set will not have a significant impact on the results of the data analysis.
- **Confidentiality**: Differential privacy techniques provide a mechanism whereby conclusions drawn from data do not reveal sensitive information about individuals, making the results of the analysis virtually indistinguishable, regardless of the presence or absence of a specific record in the data.
- **Noise mechanisms**. Noise-adding mechanisms are often used to achieve differential privacy. These can be a variety of methods, including adding Laplace or Gaussian noise to the results of data queries.
- **Privacy budget**. The differential privacy model uses the concept of a "privacy budget," commonly referred to as a ε (epsilon). A low ε value corresponds to a higher level of privacy, but it can reduce the accuracy of the analysis results.



The risk value can be calculated in this model:

$$R_{DP} = e^{\epsilon}$$

Here, ε is a differential privacy parameter. The formula assumes that the synthetic data is generated by a differentially closed algorithm that ensures that the output does not change materially if any particular record in the source data is changed or deleted. The formula also assumes that the attacker has unlimited basic knowledge and supporting information, and that the attacker can perform any type of attack on the privacy of synthetic data.

When it comes to generating synthetic data using AI, differential privacy is one of the most effective approaches to ensuring data privacy.

RISK MODEL FOR SEMI-STRUCTURED DATA

1. **Vectorize text with FastText, Word2Vec, or GloVe:** These models create vector representations of words by learning on large corpora of texts and capturing the semantic relationships between words. Word2Vec uses contextual words to predict the current word (CBOW) or the current word to predict its context (Skip-gram), while GloVe builds a word co-occurrence matrix and factorizes it.

Word2Vec uses contextual words to predict the current word (CBOW) or the current word to predict its context (Skip-gram), while GloVe builds a matrix of word occurrence and factorizes it.

$$w_i^T w_j + b_i + b_j = \log(X_{ij})$$

Here X_{ij} - element of the word co-occurrence matrix, w_i, w_j is Word Vectors, b_i, b_j - Offsets for words.

```
import fasttext
import fasttext.util
ft = fasttext.load_model('cc.en.300.bin')
word = 'computer'
word_vector = ft.get_word_vector(word)
print(f"Vector representation of a word'{word}':\n{word_vector}")
```

2. Retrieving the Ad Vector: Averaging or summing word vectors: Convert an ad to a vector by averaging or adding the vector representations of all the words in the ad.

$$V_{ad} = \frac{1}{n}\sum_{i=1}^{n} v_i$$

3. Integration of additional attributes: Explicit attributes (price, area, etc.) are converted into numerical vectors using techniques such as one-hot encoding, scaling, or embedding. Vectors are concatenated

$$V_{final} = [V_{ad}; V_{attr}]$$

4. Comparing Vectors Using Cosine Similarity: calculate the cosine similarity

$$cosine_{similarity}(V^A, V^B) = \frac{V^A \cdot V^B}{\|V^A\| \, \|V^B\|}$$

can be conveniently calculated using the scikit-learn library

Cosine similarity immediately provides a score for a risk for which a threshold can be set. For example, if you want to consider vectors similar when the cosine similarity is greater than or equal to 0.7 (which corresponds to an angle of about 45 degrees or less), you can set such a threshold. This means that the closer the cosine similarity value is to 1, the smaller the angle between the vectors and the greater their similarity.

Reliable cleaner / cleaning services / airbnb cleaner

150 $ per service

Location GTA
Richmond Hill
Barrie
Orillia.

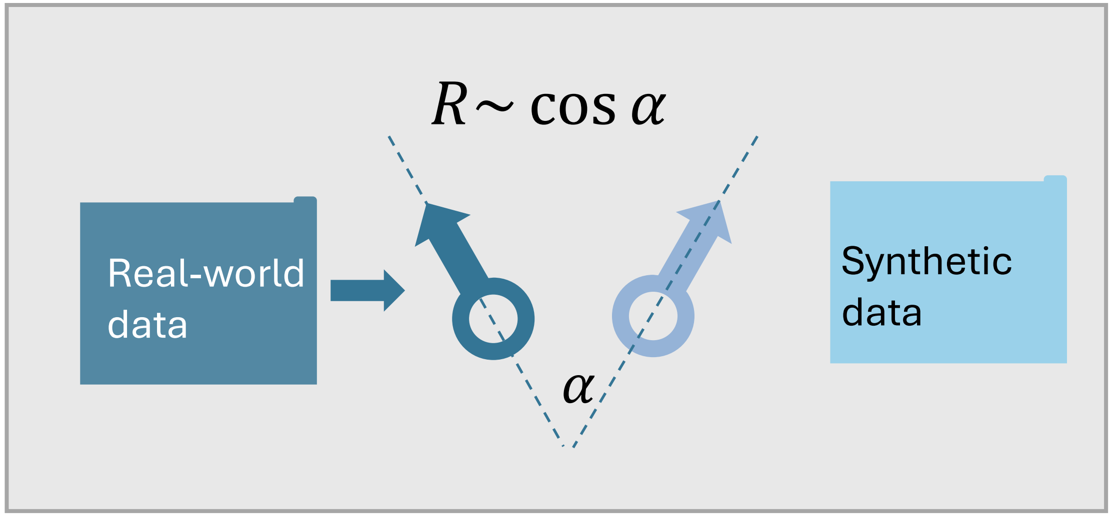
Contact Karla Garcia

☎: 647--490—5XXX

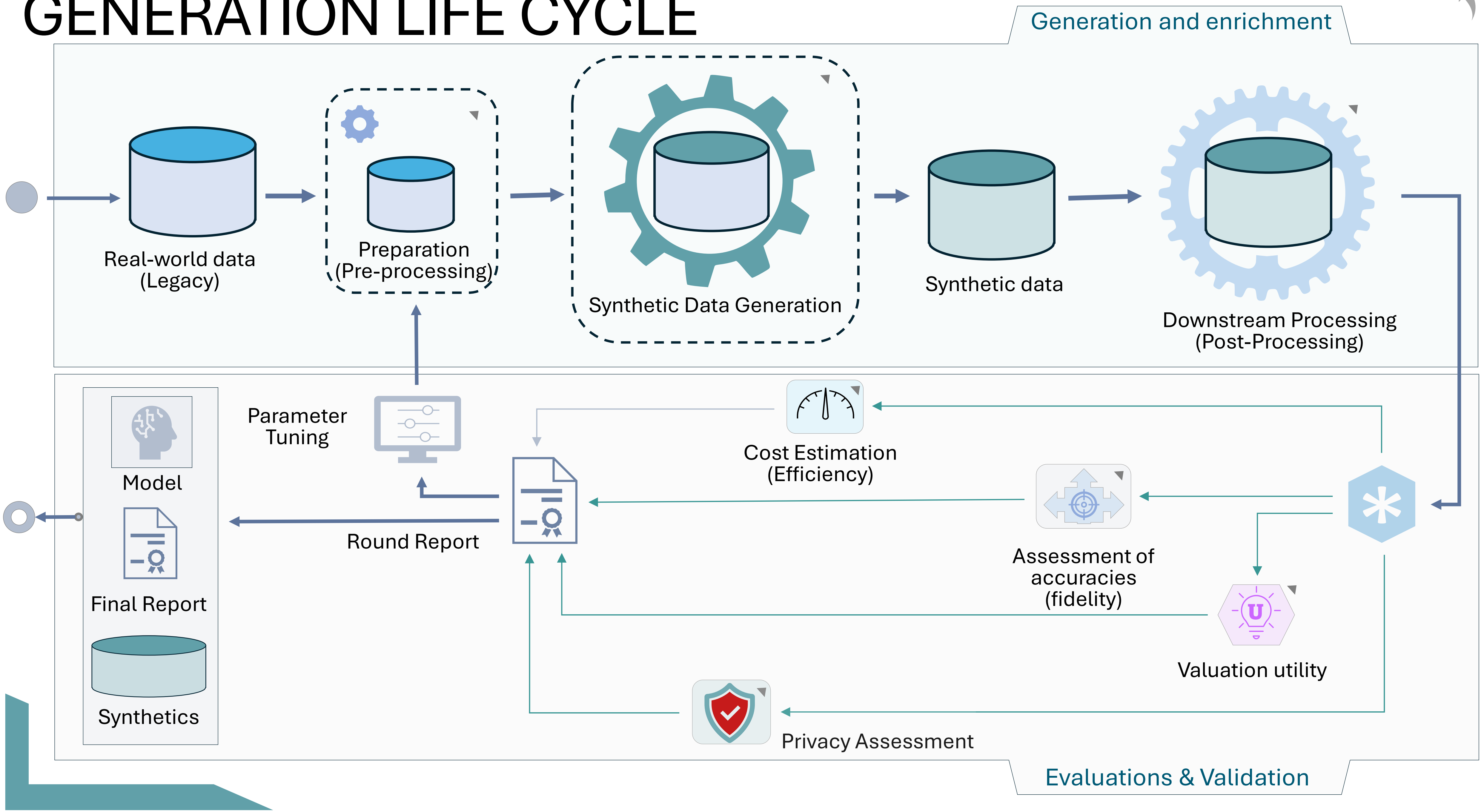
Details

We specialize in:
Deep Cleaning Post Construction
Renovation Airbnb cleaning and Hosting Residential Cleaning

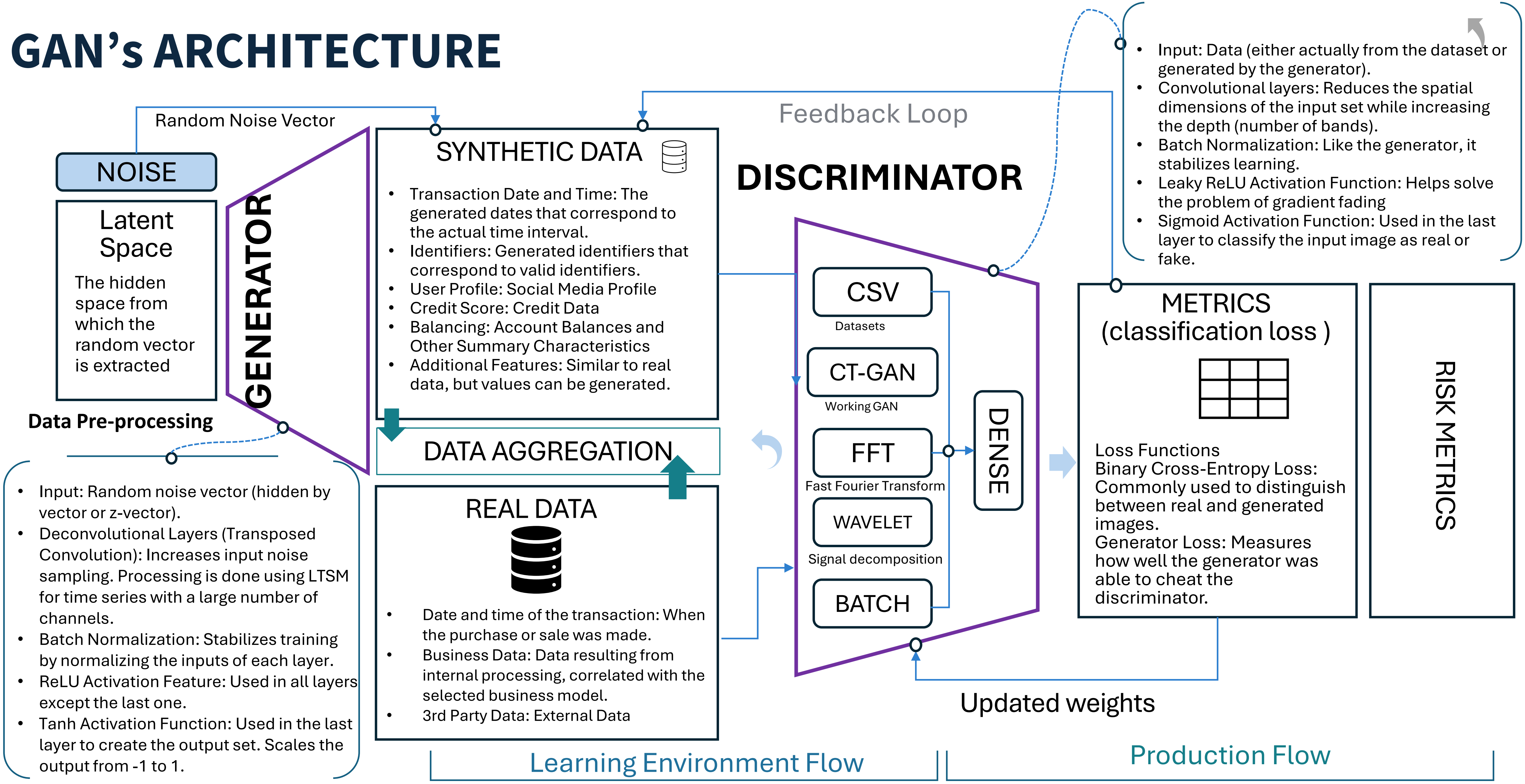
Shine cleaning services we have Pleasure to service in Aurora ,Richmond Hill,Newmarket,Bradford,Innisfil and barrie more the 10 Years expirence with are family businness with love to taken care the most precious thing people have your cozy and beautiful homes With services businnes,comercial,industrial,move out move in ,school,daycare,dentist,arquitect etc Please fill free to call as for free estimated anything with are Placer to services everyone.



GENERATION LIFE CYCLE

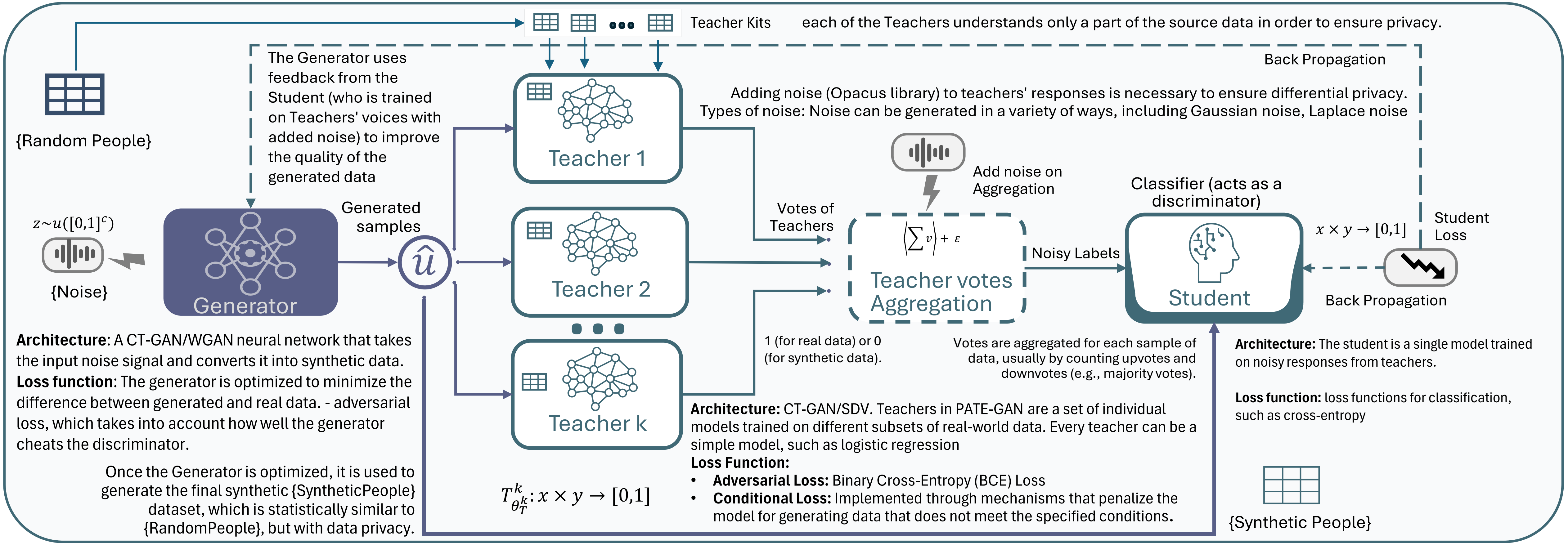


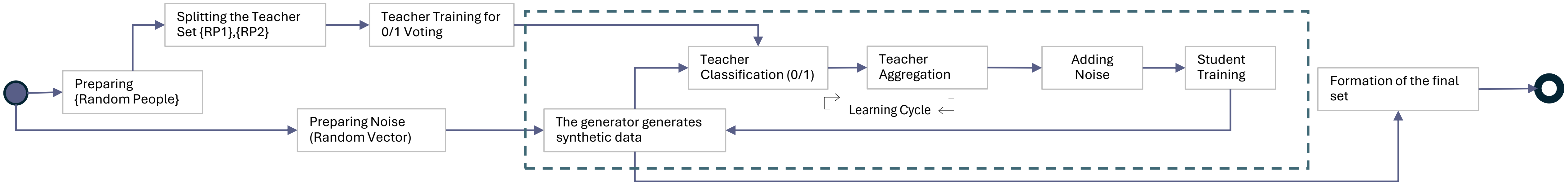
GAN's ARCHITECTURE



PATE-GAN ARCHITECTURE

PATE-GAN (Private Aggregation of Teacher Ensembles - Generative Adversarial Network) is a method that combines the principles of differential privacy and generative adversarial networks (GANs). The main goal of PATE-GAN is to generate synthetic data that is statistically similar to real data, while ensuring the protection of sensitive information contained in this data.





DIFFERENTIAL PRIVACY AI ARCHITECTURES

Решение	Качество генерации	Вычислительные требования	Удобство реализации	Степень защиты приватности
PATE-GAN	High	High (multiple teachers and noise aggregation)	Complex (complexity of aggregation and noise)	High (differential privacy through noise aggregation)
DP-SGD	Medium to High (depends on adding noise)	Medium (depends on model size and data)	Medium (frameworks available, but customization required)	High (differential privacy in gradients)
DP-CGAN	Medium to high (depends on conditions and noise)	High (conditional GAN architecture + differential privacy)	Complex (conditional generation + privacy)	High (Differential Privacy)
Autoencoder-based Models with Differential Privacy	Medium (depends on noise and recovery capacity)	Medium to high (depends on the complexity of the model)	Medium (the difficulty of balancing noise recovery and privacy)	High (if noise is applied correctly)
Regular GAN	High (with proper training)	Medium (depending on the complexity of the model)	Relatively simple (extensively researched, lots of resources)	Low (no special privacy measures)

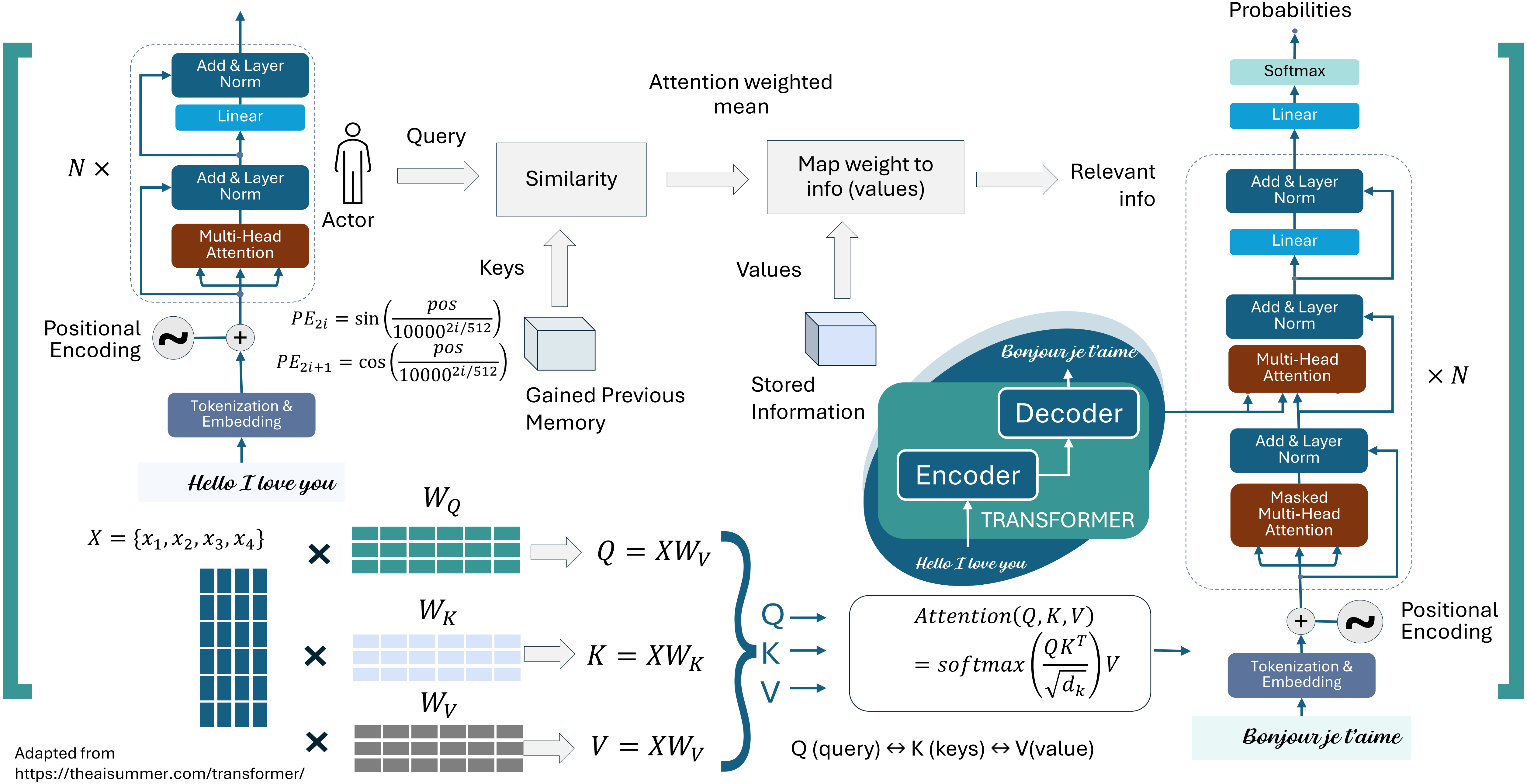
BENEFITS OF USING PATE-GAN:

- **Protect your privacy**: PATE-GAN provides strict differential privacy guarantees, which is critical when working with sensitive data such as financial information in CHURN analysis. Data privacy is maintained at all stages, from teacher training to synthetic data generation.
- **Synthetic Data Quality**: PATE-GAN is capable of generating high-quality synthetic data that preserves the statistical properties of the original {Random People} data. This ensures the realism and usefulness of synthetic data for CHURN analysis.
- **Balance between Privacy and Informativeness**: The PATE mechanism strikes a balance between protecting privacy and keeping data intelligent. This allows researchers and analysts to conduct in-depth and accurate analysis of customer churn.
- **Flexibility & Scalability**: PATE-GAN offers flexibility in the choice of teacher and student architecture, which allows you to optimize the system for the specific needs and volume of CHURN analysis data.
- **Integrating Knowledge from Multiple Sources**: With the help of multiple teachers, each trained on a subset of data, PATE-GAN is able to integrate and synthesize knowledge, providing a deeper understanding of customer churn patterns.
- **Ensure Regulatory Compliance**: In the face of stringent data protection requirements, PATE-GAN provides an effective data analysis solution without compromising privacy regulations.

TRANSFORMERS FOR TABULAR SYNTHETIC DATA

Model Name	Author and Year	Description (short)	Architecture Features	Limitation	Model Maturity
Multi-Layer Attention-Based Explainability via Transformers for Tabular Data	Cappelli et al. (2021)	A graph-oriented attention-based explainability method for tabular data	Self-attention mechanism, attention matrices, graph structure, maximum probability paths	High computation and memory cost	Experimental
How to Incorporate Tabular Data with **HuggingFace Transformers**	Thilina Rajapakse (2020)	A tutorial that shows how to use text and tabular data together with transformers	HuggingFace library, text and tabular data concatenation or separate encoders	No privacy preservation	Tutorial
A Visual Tool for Interactively Privacy Analysis and Preservation on **Order-Dynamic Tabular Data**	Liu et al. (2021)	A design and pipeline of a visual tool for nuanced privacy analysis and preservation on order-dynamic tabular data	Data cube structure, real-time risk analysis, interactive visualizations and feedback, various privacy-preserving techniques	No text data handling	Prototype
DPTransformer	Microsoft Research (2022)	A model for training transformer models with differential privacy	HuggingFace and Opacus libraries, differential privacy, gradient clipping and noise addition Integrates differential privacy mechanisms directly into the transformer model, ensuring data privacy during the generation process.	Privacy-accuracy trade-off	Research
TabTransformer	Huang et al. (2020)	A model for tabular data modeling using contextual embeddings	Self-attention based Transformers, column embedding layer, stack of Transformer layers, MLP Utilizes self-attention mechanisms to capture relationships within tabular data. Incorporates embeddings for categorical features.	No relational data handling	Research
REaLTabFormer	Solatorio and Dupriez (2023)	A model for generating realistic relational and tabular data using transformers	GPT-2 and Seq2Seq models, target masking, Qδ statistic and statistical bootstrapping, unsupervised pre-training Employs a modified transformer architecture tailored for tabular data, potentially integrating differential privacy mechanisms	No privacy preservation	Research

TRANSFORMER ARCHITECTURE



Adapted from
https://theaisummer.com/transformer/