SECURE DATA PROCESSING

ANANONYMIZATION, SYNTHETIC DATA & FEDERATED LEARNING

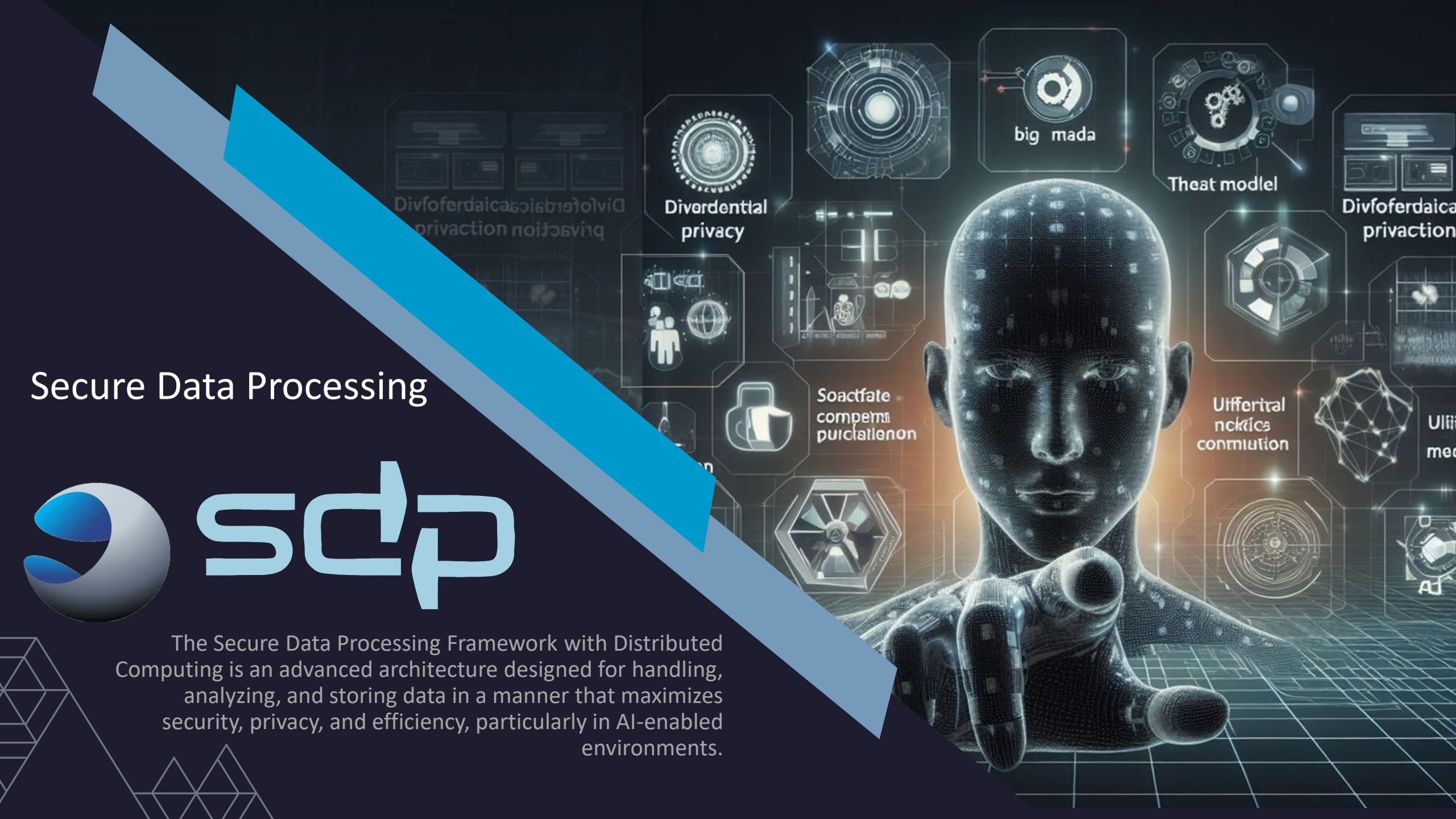
REALM

2025-02-15

Secure Data Processing

sdp

The Secure Data Processing Framework with Distributed Computing is an advanced architecture designed for handling, analyzing, and storing data in a manner that maximizes security, privacy, and efficiency, particularly in AI-enabled environments.



THE PROBLEMS

Data is growing exponentially and becoming more valuable and vulnerable due to integration and exchange across different domains and platforms.

Statista: the total amount of data created, captured, copied, and consumed worldwide reached 64.2 zettabytes in 2020, and is expected to grow to 180 zettabytes by 20252

Data breaches are security incidents that expose confidential and sensitive information to unauthorized parties. They can cause serious damage to individuals, organizations, and governments.

DataProt: there were 1,001 data breaches in 2020, exposing 155.8 million records. The average cost of a data breach was $3.86 million, and the average time to identify and contain a breach was 280 days

Machine learning and AI tools rely on data to provide insights and solutions, but they also pose challenges and risks for data protection and privacy . However, generating high-quality and secure synthetic data requires complex and protected models.

MIT News: synthetic data are simulated data that have many of the same properties as the original data, but with a much lower risk of revealing information about individuals.

Data-driven insight is essential for creating and capturing value from data, but it also requires balancing the trade-offs between data utility and data security.

IO Technologies: the data utilization rate is less than 10-15% in most companies, due to entropy, biases, and outdated tools).

THE SOLUTION

Flexibility Across Environments

Whether it's a centralized data center or a distributed computing setup, SDP adapts seamlessly, ensuring consistent data protection and processing efficiency.

Enhanced Data Privacy

With advanced anonymization techniques and synthetic data generation, including the innovative PATE-FL model, SDP offers robust privacy safeguards. This makes it an ideal tool for industries dealing with sensitive information.

Optimized for Leading AI Platforms

Fully compatible with PyTorch and TensorFlow, SDP integrates smoothly into existing AI and machine learning pipelines, enhancing their capability while safeguarding data privacy.

Balanced Risk Assessment

By evaluating both the confidentiality risks and the utility of data, SDP ensures that data-driven insights are derived without compromising security, aligning with regulatory compliance and ethical standards..

Tech-Savvy and Future-Ready

SDP is built to integrate with advanced platforms like DGT Data Guard and REALM, positioning it at the forefront of the next generation of secure data processing technologies.

TECHNOLOGIES

Privacy Enhancing Technologies (PETs)

- **k-Anonymity:** This is a method used to ensure that data cannot be re-identified to a specific individual. In a dataset that adheres to k-anonymity, each record is indistinguishable from at least (k−1) other records with respect to certain identifying attributes.
- **Differential Privacy**: Ensures that algorithms are designed in a way that the output doesn't significantly differ whether any single individual's data is included or not.
- **Secure Multi-party Computation (SMPC)**: Enables multiple parties to jointly compute a function over their data inputs while keeping those inputs private.
- **Synthetic Data**: Involves creating entirely new datasets that are statistically similar to the original data but do not contain any actual individual data points. This is particularly useful for training machine learning models were using real, sensitive data might pose privacy risks.

Synthetic Data Generation

- Generative Adversarial Networks (GANs): Used for generating synthetic datasets that mimic real data distributions without exposing sensitive information.
- Variational AutoEncoders (VAEs): Another approach for generating high-quality synthetic data, especially useful in scenarios where data is scarce or highly sensitive.
- PATE-FL (Private Aggregation of Teacher Ensembles-Federated Learning): This method combines the privacy benefits of the PATE architecture with the distributed nature of Federated Learning. In PATE-FL, multiple models (referred to as "teachers") are trained on disjoint subsets of data. These models then collectively contribute to training a student model, which generates synthetic data. The aggregation process ensures that the student model learns from the ensemble without compromising the privacy of the individual datasets. PATE-FL is particularly effective in decentralized environments where data cannot be pooled due to privacy or regulatory concerns.

Scalable Architectures

- Federated Learning: Collaborative machine learning without centralizing data, crucial for decentralized or distributed data processing.
- PATE (Private Aggregation of Teacher Ensembles): A framework combining the benefits of multiple models to enhance privacy and data utility, especially in federated learning scenarios.
- Blockchain Technologies: For secure and decentralized data logging and management, ensuring data integrity and traceability.

Metrics for Security and Data Utility

- Data Utility Metrics: Quantitative measures to assess how useful the processed data is for specific analytical or operational purposes.
- Security Metrics: Measures to evaluate the level of data protection against unauthorized access or breaches.

RISK MODELS

The risk model is the heart of the SDP. This involves identifying and understanding the potential threats and attack vectors that could compromise data privacy and security. It includes analyzing who the potential attackers are, what resources they might have, and what their motivations could be. The threat model helps in anticipating how and where a system might be attacked. **Key elements of a risk model**:

- **Threat Model**: This involves identifying and understanding the potential threats and attack vectors that could compromise data privacy and security.
- **Communication Scenario**: Determines the nature of data flow (data volume, one-way, multiple participants, workload, confidential category)
- **Risk Identification**: Recognizing potential threats to data security and privacy.
- **Risk Assessment**: Evaluating the likelihood and potential impact of these risks.
- **Risk Mitigation**: Developing strategies to reduce or manage identified risks.
- **Monitoring and Review**: Continuously monitoring risks and adapting strategies as needed.

Re-identification Model

This model aims to minimize the risk of re-identifying individuals within a dataset. It operates on the principle of k-anonymity, ensuring each individual's data is "hidden" among at least (k−1) other individuals, является обобщением модели k-anonymity, направлена на атаки вывода (allocation/inference)

$$R_{ID_k} = \frac{1}{\min\limits_{i \in D_S} k_i}$$

Divergent Model

Focuses on quantifying the confidentiality of synthetic or anonymized data by measuring the distance or discrepancy between this data and the original source data.

$$d_{KL}(D_S, D_O) = \sum_{x \in X} D_S(x) \log_2 \frac{D_S(x)}{D_O(x)}$$

Information Leakage Model

Considers the probability of three types of privacy attacks: singling out, linkage, and inference. The model is predicated on the notion that synthetic or anonymized data can be considered as a noisy version of the original dataset.

$$R_{IL} = w_1 \times R_{so} + w_2 \times R_L + w_3 \times R_I$$

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

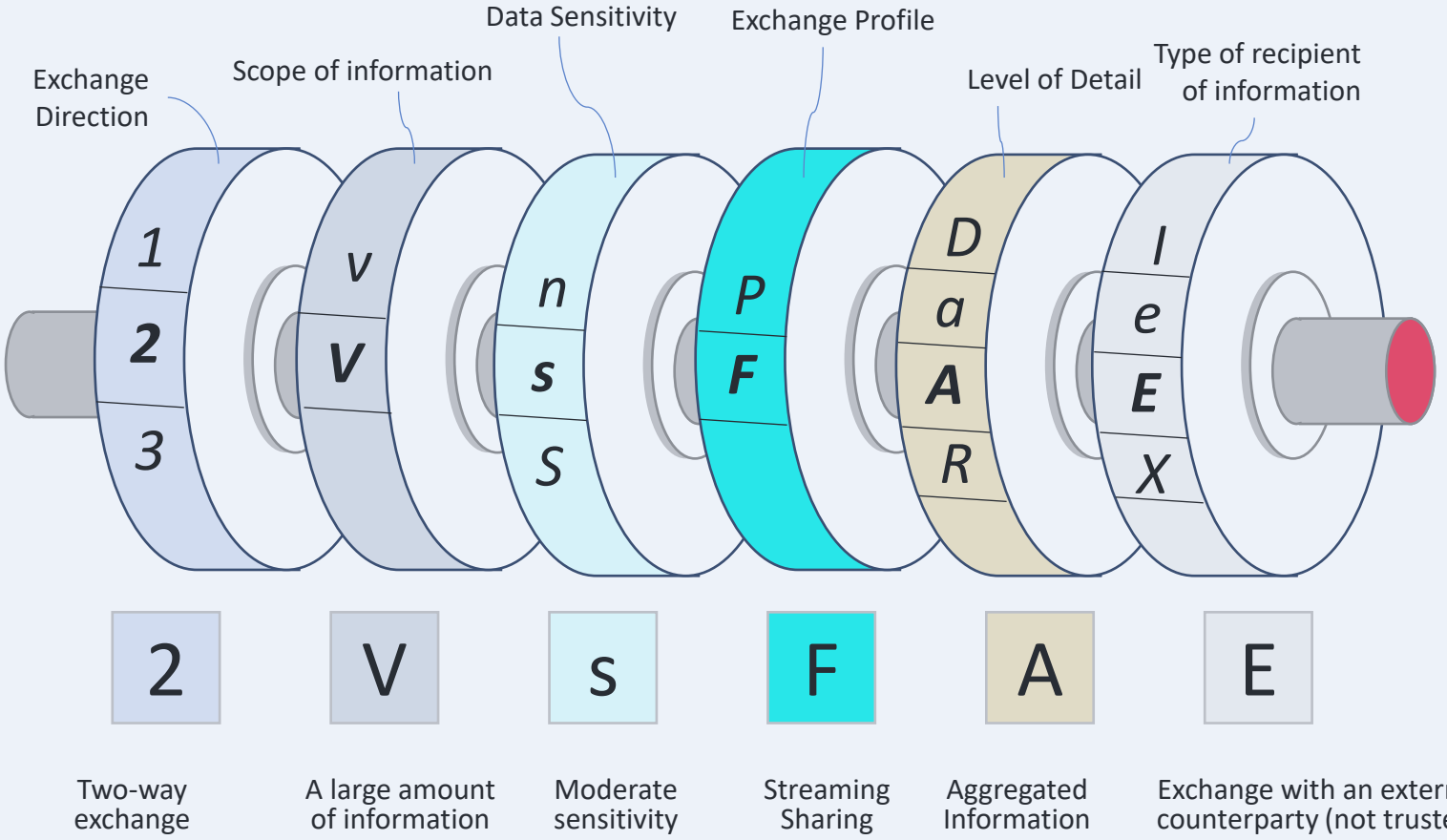
Differential Privacy

Operates within the differential privacy framework by integrating random noise into the data processing.

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

COMMUNICATION SCHEMES

In data privacy, understanding the communication scheme—whether it's one-way or multilateral data exchange—is key to implementing effective privacy controls. Adopting a unified coding approach is essential for maintaining data integrity, ensuring security, and achieving interoperability across different systems.



FULL RISK CALCULATION

1. In the world of big data, there is no absolute protection, a probabilistic approach is needed

2. Risk can be assessed on the basis of a comprehensive approach

RISK = (ATTACK PROBABILITY) × (DEPENDENCY FACTOR) ×

$P_a \rightarrow 1$ for data publication

mutual influence of F1 factors (simplification)

It is not considered in the main model, but can be considered for specific cases

(CONTEXTUAL RISK) ×

Environmental and Compliance Risk

(DATA RISK) ×

Risk associated with data protection

(DAMAGE)

the total loss of money from leaks and successful data attacks

3. Contextual Risk is evaluated using a scoring model that considers environmental risks, compliance requirements, and the implementation of proactive protection measures.

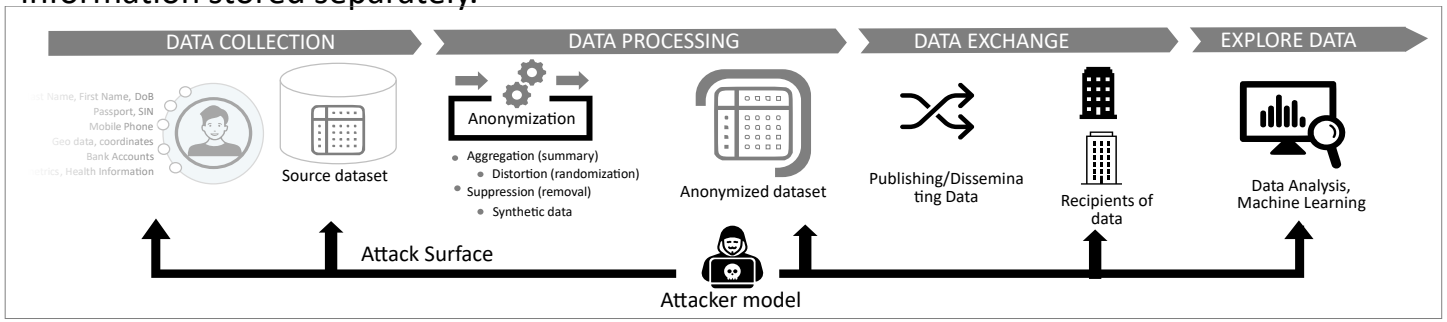
4. Data Risk is incorporated into the assessment of various data protection practices, including anonymization, the use of synthetic data, differential privacy techniques, and confidential computing.

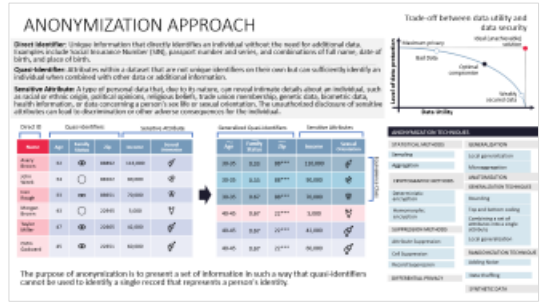
5. DAMAGE – financial losses related to reputation, fines and compensations, customer attrition, and other measurable indicators.

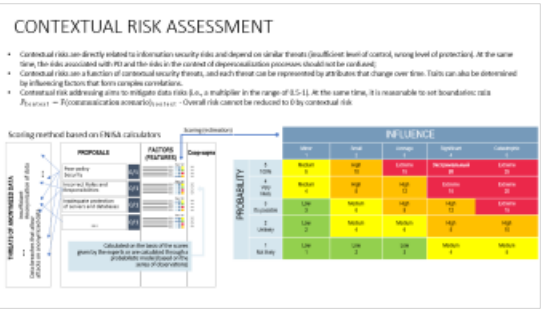
BASIC RISK MODEL

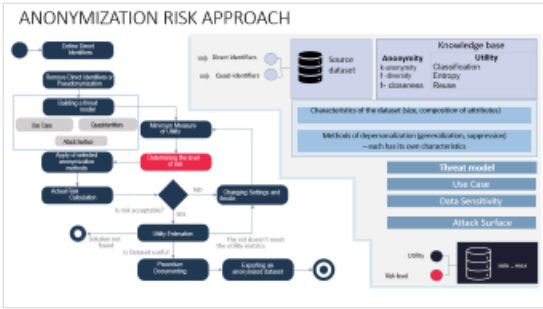
Personal Identifiable Information (PII) is any information that can be used on its own or with other data to identify, contact, or locate a single person, or to identify an individual in context. This may include information such as names, addresses, email addresses, telephone numbers, social insurance numbers, and may extend to more indirect forms of identification such as IP addresses or other digital identifiers when they can be linked to individuals.

Anonymization of information refers to the process of removing personally identifiable information from data sets, so that the individuals whom the data describe remain anonymous. This is achieved by eliminating or modifying personal identifiers, such as names or social security numbers, and other information that might allow someone to re-identify the data subjects indirectly. The goal of anonymization is to protect individuals' privacy by ensuring that personal data cannot be linked back to them, even by the data processor or holder, without additional information stored separately.









Re-identification Risk refers to the probability or likelihood that a specific individual can be correctly identified from a dataset that has been processed for anonymization.

$$P_{re\text{-}id} = P_{context} \times P_{data}$$

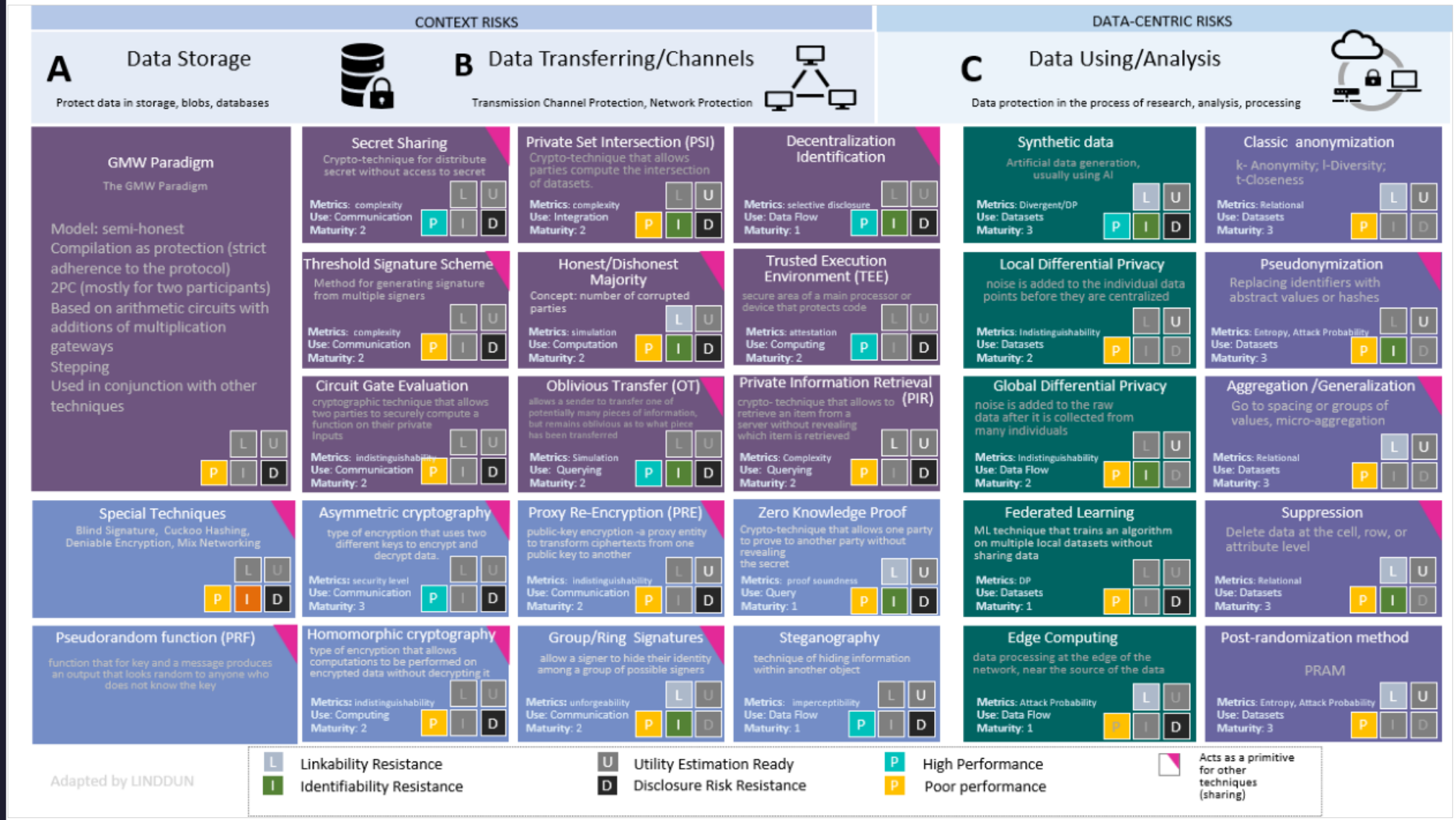
Data Storage Security Risks in the Corporate Environment (Scoring Methods)

Risks of the set itself based on record uniqueness metrics (equivalence classes)

PIPEDA

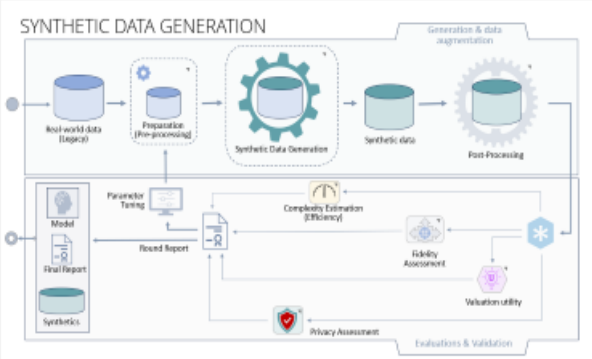
CCPA

PRIVACY ENHANSING TECHNOLOGIES

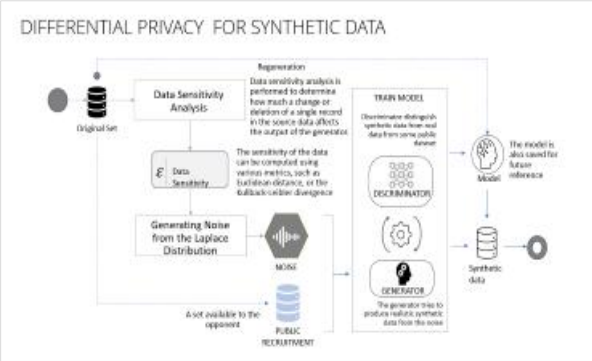


SYNTHETIC DATA

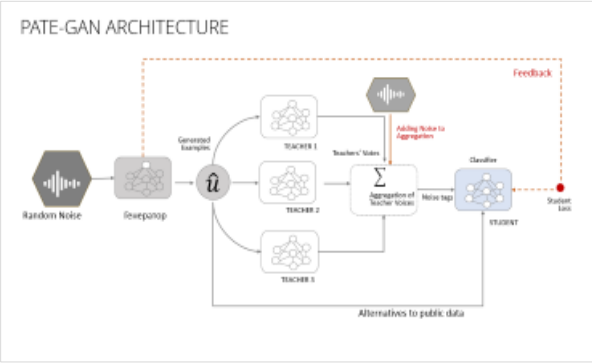
Synthetic data generation is important for privacy protection because it allows data owners to share data that mimics the characteristics and patterns of real data without revealing any sensitive or personally identifiable information. Synthetic data can enable data analysis, machine learning, and data sharing without compromising the privacy of individuals in the original data. Synthetic data can also prevent re-identification attacks that can link anonymized data to other identifiable datasets.



Differential privacy is a mathematical framework that quantifies the privacy loss of a data analysis algorithm. Differential privacy provides a formal guarantee that the output of the algorithm is not significantly affected by the presence or absence of any individual data point. Differential privacy can be used to enhance the privacy and security of synthetic data generation by adding calibrated noise to the data or the algorithm, and by bounding the influence of any individual data point on the output.



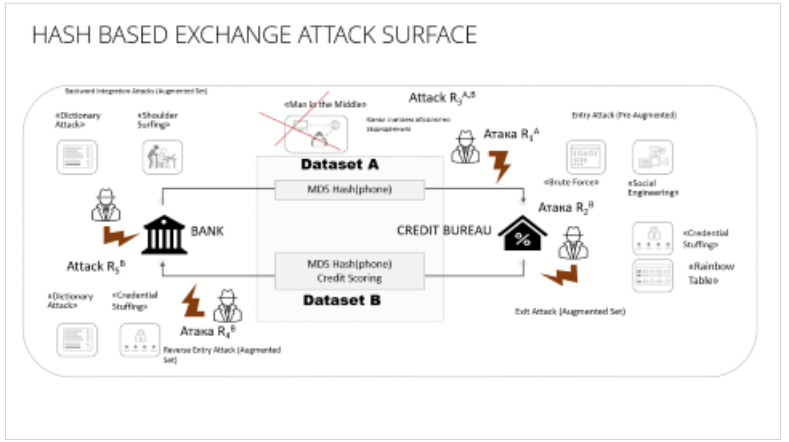
PATE-GAN is a differentially private synthetic data generation method that adapts the Private Aggregation of Teacher Ensembles (PATE) framework to the Generative Adversarial Networks (GAN) framework. PATE-GAN trains multiple teacher models on disjoint partitions of the original data and transfers their knowledge to a student model that generates synthetic data.



FL-PATE is a differentially private federated learning framework with knowledge transfer. FL-PATE combines the Federated Learning (FL) and the PATE frameworks to train teacher models on distributed and sensitive data, and to generate a student model that can be publicly accessed for prediction. FL-PATE uses a secure aggregation protocol to protect the communication between the participants and the server during the FL process, and a differential privacy mechanism to protect the data privacy during the PATE process.

DIRECT HASHING

In the context of the case, the BANK transfers the mobile phone numbers of its clients to scoring agencies and receives enriched data in return – the scoring of the client associated with the phone number. Bilateral information exchange is carried out once a quarter with a volume of 500,000 records. Phone numbers are protected by hashing with a salt.

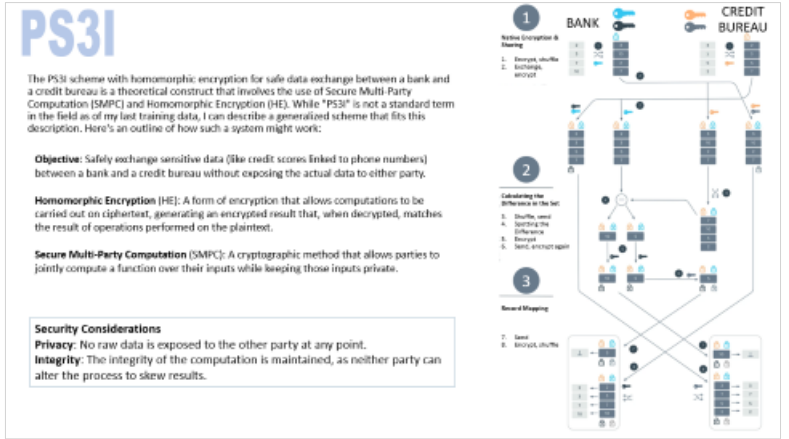


Direct exchange with static salt was found to be ineffective, and a new safety architecture was developed.

As a result, an optimal scheme has been developed:

- **Option 1**: Exchange with dynamic salt using
  SHA-3(SHA-3(SALT*PHONE))
  $$Risk = \frac{1}{\mu(\tau)\cdot\tau} 0.05$$

- **Option 2**: Using dynamic SMPC scheme PS3I,
  $$Risk = 10^{-8}.$$

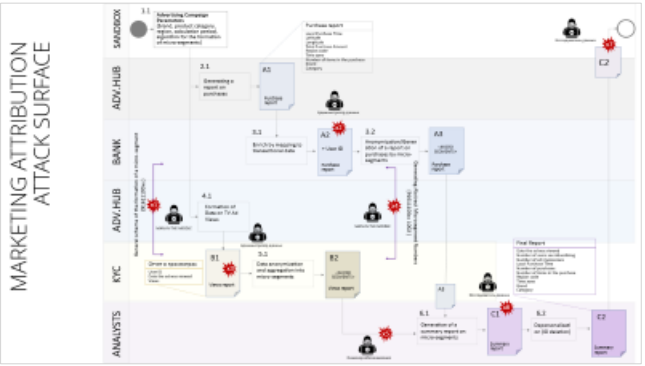


MARKETING ATTRIBUTION

This case involves three parties - a marketing platform (Side A), a financial institution (Party B), and a third party (Side C) - each playing a role in processing user data for targeted marketing:

- **Side A (Marketing Platform)**: Generates data about user ad views, where users are identified by phone numbers.
- **Party B (Financial Institution)**: Provides information on payments made for the advertised products, along with profiles of individuals who made these payments.
- **Side C**: Combines the data sets from Side A and Party B to build a machine learning model for marketing targeting.

The primary challenge is to protect the privacy and integrity of user data during this exchange and processing.



Improved Scheme

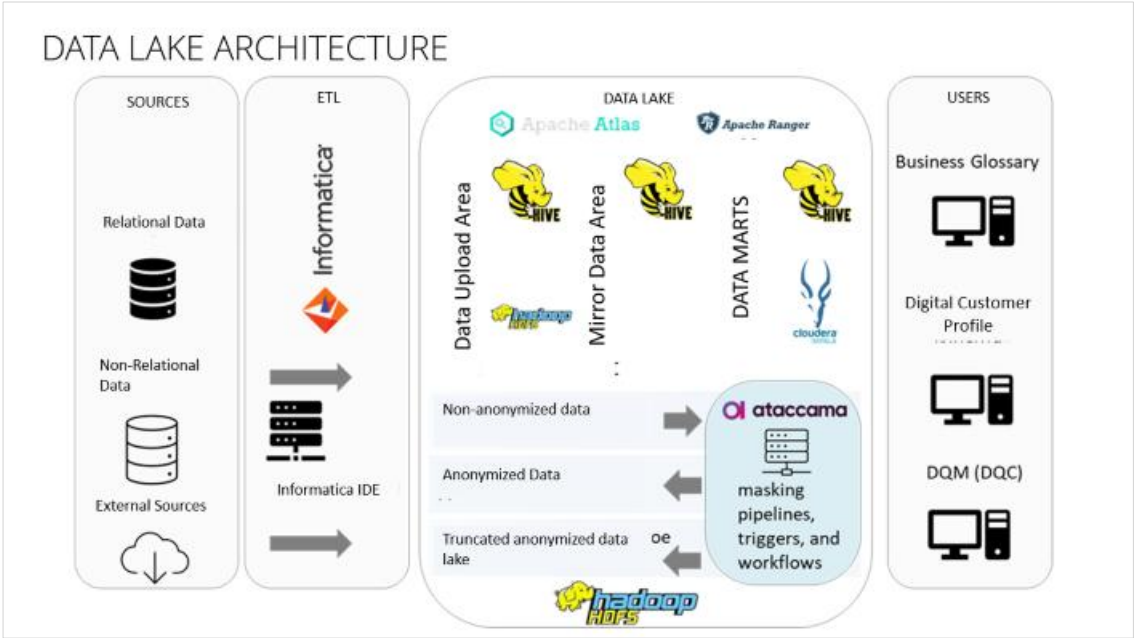
1. **PS3I and Secure Multi-Party Computation**: Implements a more complex and secure scheme utilizing the principles of Secure Multi-Party Computation (SMPC).
2. **Homomorphic Encryption and Oblivious Transfer**: The scheme uses multiple keys and homomorphic encryption to enable secure computations on encrypted data. Oblivious Transfer is used to protect the exchange of information, ensuring that parties only learn specific pieces of data they are entitled to, without accessing the entirety of the other party's information.
3. **Attack Success Probability**: The probability of a successful attack on this scheme is extremely low (0.00000001), indicating a high level of security.

DATA LAKE

The BANK is creating a data lake to develop digital client profiles, enhance process mining, and perform advanced analytics. This initiative aims to leverage extensive intra-bank information to gain deeper insights into client behavior and financial patterns.

Anonymization Techniques:

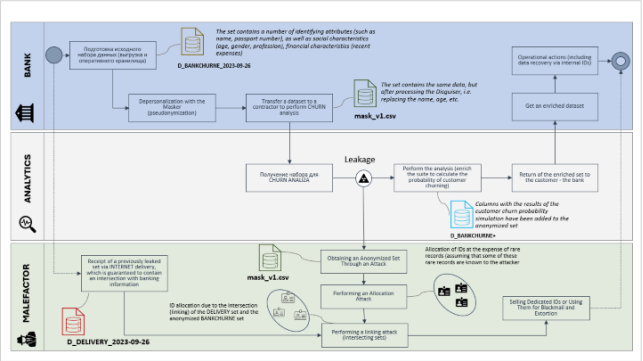
- **Automated Anonymization**: Investigating the use of the Mondrian algorithm (a greedy multidimensional algorithm for generalization) versus Datafly/Incognito for pre-processing before applying Ataccama's maskerization.
- **Goal:** Achieving specific threshold values for the risk of re-identification while maintaining data utility.
- **Challenge**: Ataccama's built-in pseudonymization methods are not meeting the required thresholds for risk reduction.



Conclusions:

- **Balancing Anonymization and Utility**: The key challenge is to balance the anonymization to protect client privacy (minimizing re-identification risk) with the preservation of data utility, especially for AI/ML applications.
- **Method Selection**: The selection between Mondrian, Datafly/Incognito, and Ataccama's methods will depend on their effectiveness in achieving k-anonymity and the specific requirements of different use cases within the bank.
- **Synthetic Data for AI/ML**: The use of synthetic data is a promising approach for AI/ML cases, as it can provide a balance between data utility and privacy.
- **Risk Assessment and Compliance**: Continuous assessment of re-identification risks and compliance with data protection regulations (like GDPR) is crucial.
- **Insider Threat Mitigation**: Given the insider as a potential infringer, robust access controls, monitoring, and audit trails are essential to safeguard the data lake.

CHURN DATASET: LINKING ATTACK



The bank is analyzing the likelihood of its clients switching to another bank (CHURN) to improve client retention. To do this, it creates and then anonymizes the CHURN dataset and then passes it on to external analysts. At this stage, an anonymized data set is leaked, on which the attacker carries out a number of attacks:

- **Singling Out Attack**: Isolating rare records in the hope that the attacker knows some of the clients.
- **Linkage Attack**: Crossing the two datasets to identify clients (A previously available set of online deliveries is used).

Security Testing and Analysis:

- Simulating attacks using the Fellegi-Sunter method. Simulating potential attacks and analyzing the security and utility of the data is a key element in risk assessment and choosing anonymization methods.
- Calculating security and utility metrics, including Pearson correlation analysis and information loss.

Results and Improvements:

- Development of an improved data anonymization procedure.
- Creation of a synthetic dataset to enhance security while maintaining the data's analytical utility.

APPLICATIONS

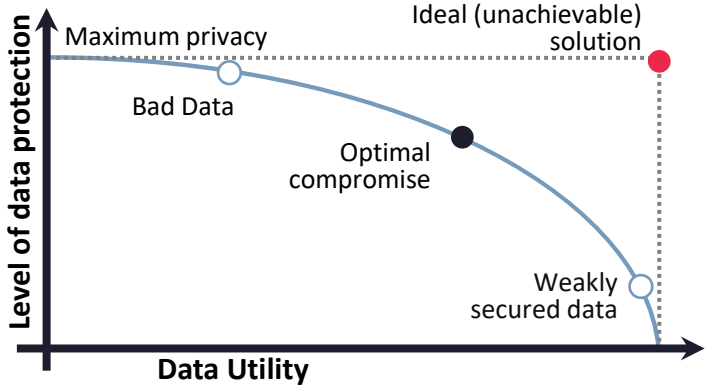
ANONYMIZATION APPROACH

Direct Identifier: Unique information that directly identifies an individual without the need for additional data. Examples include Social Insurance Number (SIN), passport number and series, and combinations of full name, date of birth, and place of birth.

Quasi-Identifier: Attributes within a dataset that are not unique identifiers on their own but can sufficiently identify an individual when combined with other data or additional information.

Sensitive Attribute: A type of personal data that, due to its nature, can reveal intimate details about an individual, such as racial or ethnic origin, political opinions, religious beliefs, trade union membership, genetic data, biometric data, health information, or data concerning a person's sex life or sexual orientation. The unauthorized disclosure of sensitive attributes can lead to discrimination or other adverse consequences for the individual.

Trade-off between data utility and data security



Direct ID	Quasi-Identifiers			Sensitive Attribute	
Name	Age	Family Status	Zip	Income	Sexual Orientation
Avery Brown	32	⚭	88892	116,000	⚥
John Week	34	◯	88262	90,000	⚥
Iren Rough	33	⚮	88651	70,000	⚢
Morgan Brown	43	◯	22945	5,000	⚧
Taylor Miller	47	⚭	22665	42,000	⚥
Ostin Godward	45	⚭	22651	60,000	⚥

Generalized Quasi-Identifiers			Sensitive Attributes		
~ Age	~ Family Status	~ Zip	Income	Sexual Orientation	
30-35	0.33	88***	116,000	⚥	Equivalence Class
30-35	0.33	88***	90,000	⚥	
30-35	0.67	88***	70,000	⚢	
40-45	0.67	22***	5,000	⚧	
40-45	0.67	22***	42,000	⚥	
40-45	0.67	22***	60,000	⚥	

The purpose of anonymization is to present a set of information in such a way that quasi-identifiers cannot be used to identify a single record that represents a person's identity.

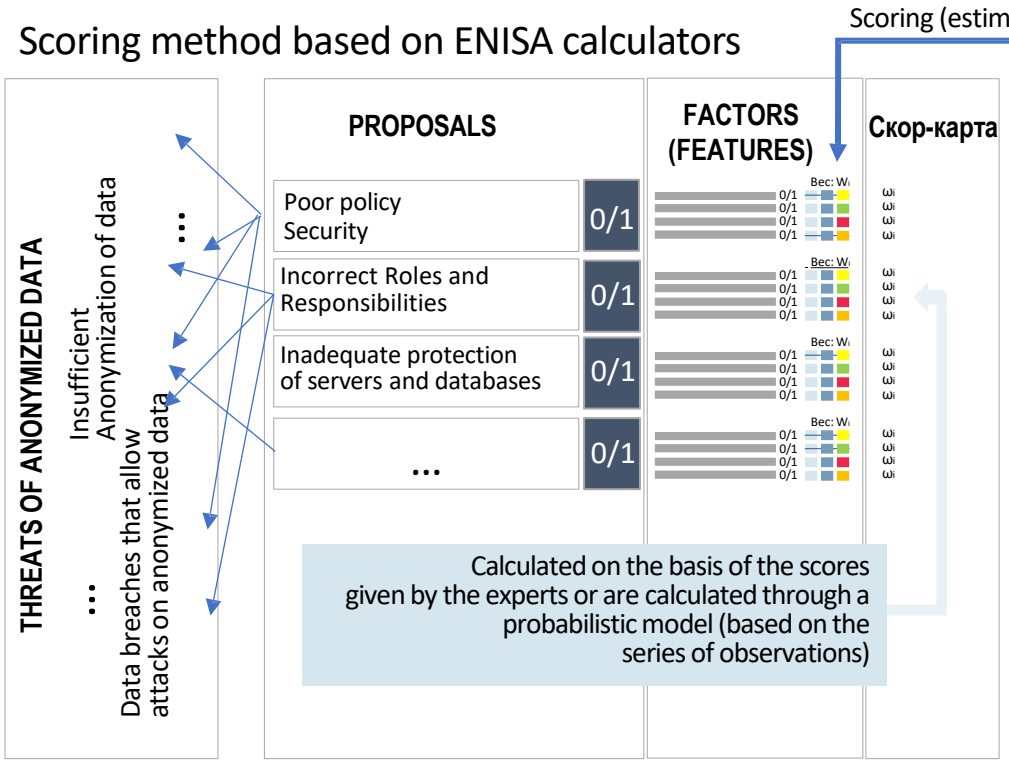
ANONYMIZATION TECHNIQUES

- STATISTICAL METHODS
  - Sempling
  - Aggregation
- CRYPTOGRAPHIC METHODS
  - Deterministic encryption
  - Homomorphic encryption
- SUPPRESSION METHODS
  - Attribute Suppression
  - Cell Suppression
  - Record Suppression
- DIFFERENTIAL PRIVACY
- GENERALIZATION
  - Local generalization
  - Microaggregation
- ANATOMIZATION
- GENERALIZATION TECHNIQUES
  - Rounding
  - Top and bottom coding
  - Combining a set of attributes into a single attribute
  - Local generalization
- RANDOMIZATION TECHNIQUE
  - Adding Noise
  - Data Shuffling
- SYNTHETIC DATA

CONTEXTUAL RISK ASSESSMENT

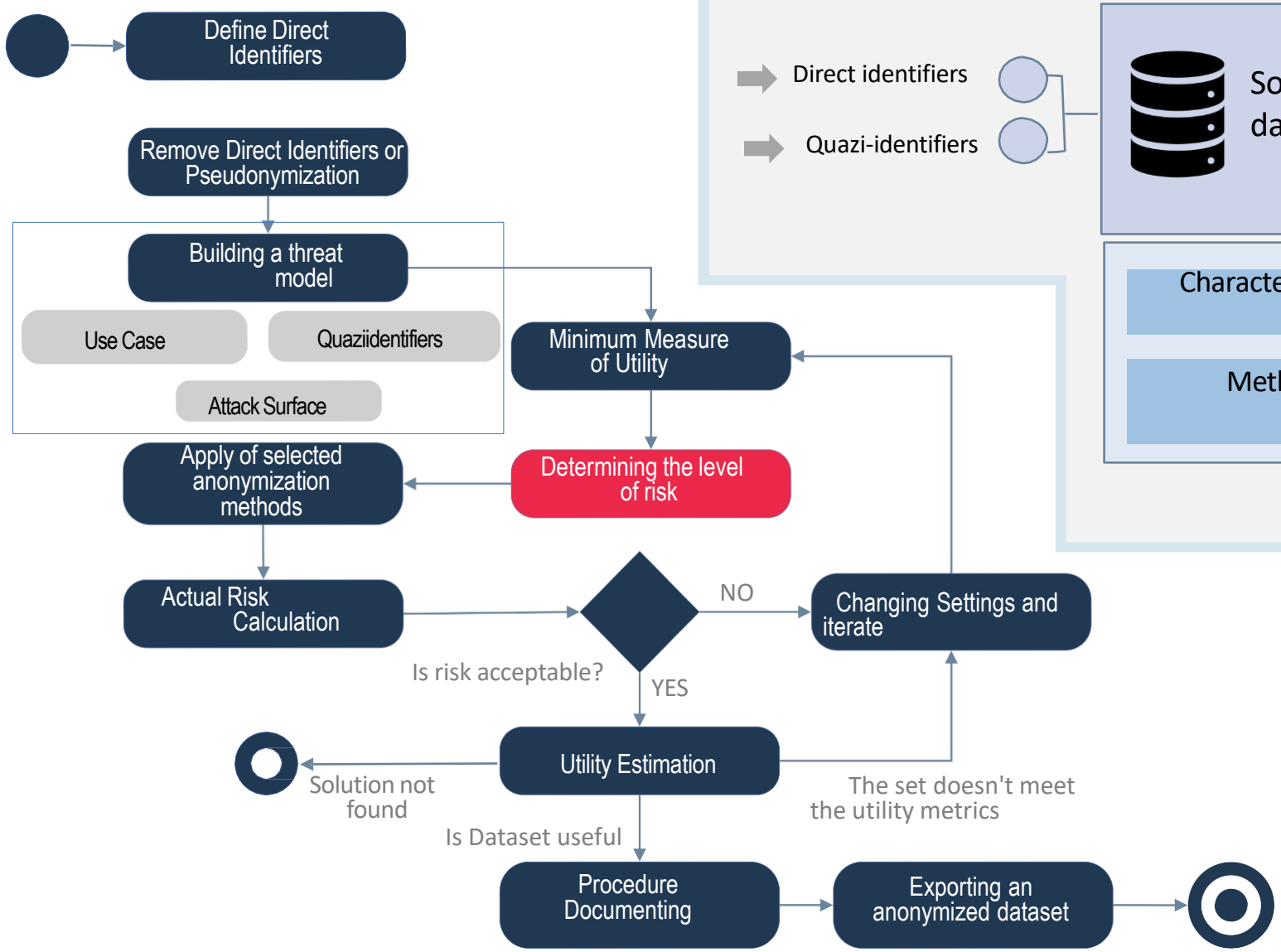
- Contextual risks are directly related to information security risks and depend on similar threats (insufficient level of control, wrong level of protection). At the same time, the risks associated with PD and the risks in the context of depersonalization processes should not be confused;
- Contextual risks are a function of contextual security threats, and each threat can be represented by attributes that change over time. Traits can also be determined by influencing factors that form complex correlations.
- Contextual risk addressing aims to mitigate data risks (i.e., a multiplier in the range of 0.5-1). At the same time, it is reasonable to set boundaries: $\min P_{\text{Context}} = \text{F}(\text{communication scenario})_{\text{context}}$ - Overall risk cannot be reduced to 0 by contextual risk

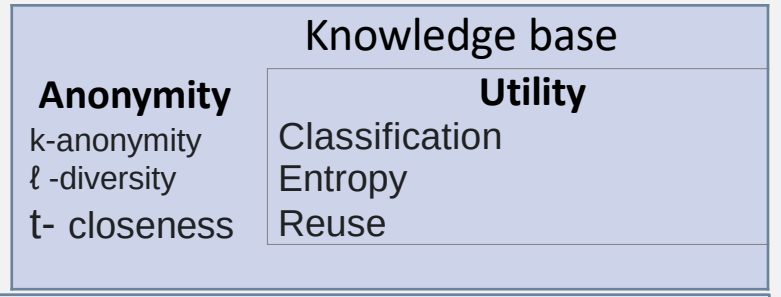
Scoring method based on ENISA calculators



PROBABILITY	INFLUENCE: Minor 1	Small 2	Average 3	Significant 4	Catastrophic 5
5 100%	Medium 5	High 10	Extreme 15	**Экстремальный 20**	Extreme 25
4 Very likely	Medium 4	High 8	High 12	Extreme 16	Extreme 20
3 It's possible	Low 3	Medium 6	High 9	High 12	Extreme 15
2 Unlikely	Low 2	Medium 4	Medium 6	High 8	High 10
1 Not likely	Low 1	Low 2	Low 3	Medium 4	Medium 5

ANONYMIZATION RISK APPROACH





Characteristics of the dataset (size, composition of attributes)

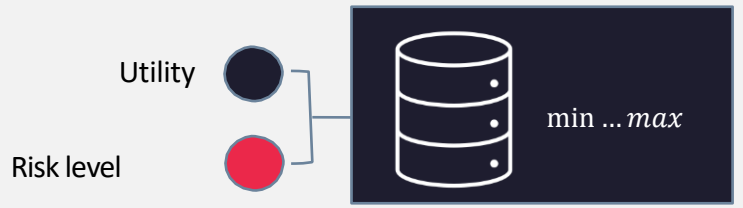
Methods of depersonalization (generalization, suppression)
– each has its own characteristics

Threat model

Use Case

Data Sensitivity

Attack Surface



CONTEXT RISKS

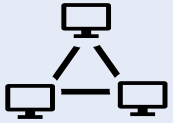
DATA-CENTRIC RISKS

A Data Storage

Protect data in storage, blobs, databases

B Data Transferring/Channels

Transmission Channel Protection, Network Protection



C Data Using/Analysis

Data protection in the process of research, analysis, processing

GMW Paradigm

The GMW Paradigm

Model: semi-honest
Compilation as protection (strict adherence to the protocol)
2PC (mostly for two participants)
Based on arithmetic circuits with additions of multiplication gateways
Stepping
Used in conjunction with other techniques

P L U I D

Secret Sharing

Crypto-technique for distribute secret without access to secret

Metrics: complexity
Use: Communication
Maturity: 2

P L U I D

Private Set Intersection (PSI)

Crypto-technique that allows parties compute the intersection of datasets.

Metrics: complexity
Use: Integration
Maturity: 2

P L U I D

Decentralization Identification

Metrics: selective disclosure
Use: Data Flow
Maturity: 1

P L U I D

Synthetic data

Artificial data generation, usually using AI

Metrics: Divergent/DP
Use: Datasets
Maturity: 3

P L U I D

Classic anonymization

k- Anonymity; l-Diversity; t-Closeness

Metrics: Relational
Use: Datasets
Maturity: 3

P L U I D

Threshold Signature Scheme

Method for generating signature from multiple signers

Metrics: complexity
Use: Communication
Maturity: 2

P L U I D

Honest/Dishonest Majority

Concept: number of corrupted parties

Metrics: simulation
Use: Computation
Maturity: 2

P L U I D

Trusted Execution Environment (TEE)

secure area of a main processor or device that protects code

Metrics: attestation
Use: Computing
Maturity: 2

P L U I D

Local Differential Privacy

noise is added to the individual data points before they are centralized

Metrics: Indistinguishability
Use: Datasets
Maturity: 2

P L U I D

Pseudonymization

Replacing identifiers with abstract values or hashes

Metrics: Entropy, Attack Probability
Use: Datasets
Maturity: 3

P L U I D

Circuit Gate Evaluation

cryptographic technique that allows two parties to securely compute a function on their private Inputs

Metrics: indistinguishability
Use: Communication
Maturity: 2

P L U I D

Oblivious Transfer (OT)

allows a sender to transfer one of potentially many pieces of information, but remains oblivious as to what piece has been transferred

Metrics: Simulation
Use: Querying
Maturity: 2

P L U I D

Private Information Retrieval (PIR)

crypto- technique that allows to retrieve an item from a server without revealing which item is retrieved

Metrics: Complexity
Use: Querying
Maturity: 2

P L U I D

Global Differential Privacy

noise is added to the raw data after it is collected from many individuals

Metrics: Indistinguishability
Use: Data Flow
Maturity: 2

P L U I D

Aggregation /Generalization

Go to spacing or groups of values, micro-aggregation

Metrics: Relational
Use: Datasets
Maturity: 3

P L U I D

Special Techniques

Blind Signature, Cuckoo Hashing, Deniable Encryption, Mix Networking

P L U I D

Asymmetric cryptography

type of encryption that uses two different keys to encrypt and decrypt data.

Metrics: security level
Use: Communication
Maturity: 3

P L U I D

Proxy Re-Encryption (PRE)

public-key encryption -a proxy entity to transform ciphertexts from one public key to another

Metrics: indistinguishability
Use: Communication
Maturity: 2

P L U I D

Zero Knowledge Proof

Crypto-technique that allows one party to prove to another party without revealing the secret

Metrics: proof soundness
Use: Query
Maturity: 1

P L U I D

Federated Learning

ML technique that trains an algorithm on multiple local datasets without sharing data

Metrics: DP
Use: Datasets
Maturity: 1

P L U I D

Suppression

Delete data at the cell, row, or attribute level

Metrics: Relational
Use: Datasets
Maturity: 3

P L U I D

Pseudorandom function (PRF)

function that for key and a message produces an output that looks random to anyone who does not know the key

Homomorphic cryptography

type of encryption that allows computations to be performed on encrypted data without decrypting it

Metrics: indistinguishability
Use: Computing
Maturity: 2

P L U I D

Group/Ring Signatures

allow a signer to hide their identity among a group of possible signers

Metrics: unforgeability
Use: Communication
Maturity: 2

P L U I D

Steganography

technique of hiding information within another object

Metrics: imperceptibility
Use: Data Flow
Maturity: 1

P L U I D

Edge Computing

data processing at the edge of the network, near the source of the data

Metrics: Attack Probability
Use: Data Flow
Maturity: 1

P L U I D

Post-randomization method

PRAM

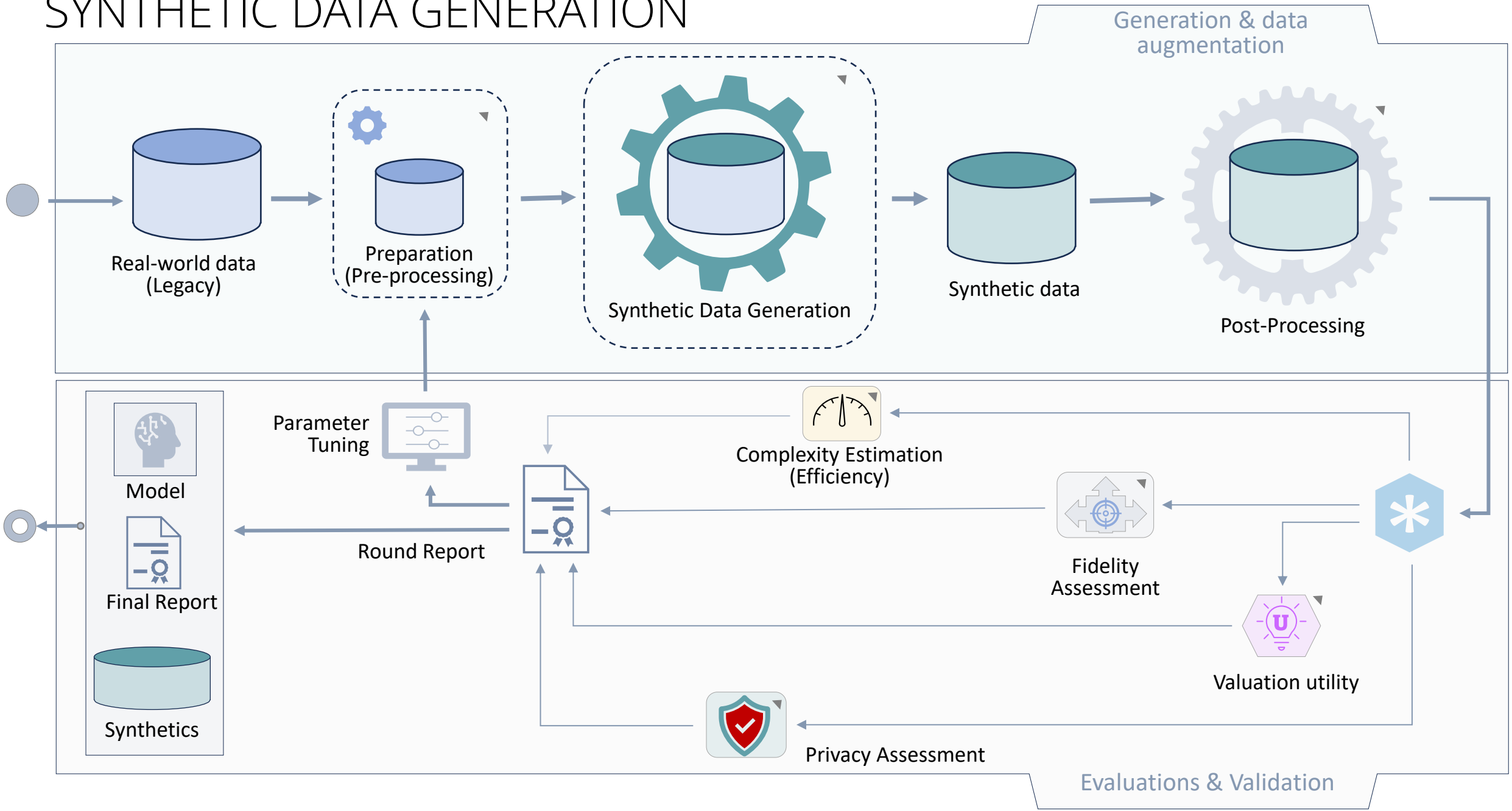
Metrics: Entropy, Attack Probability
Use: Datasets
Maturity: 3

P L U I D

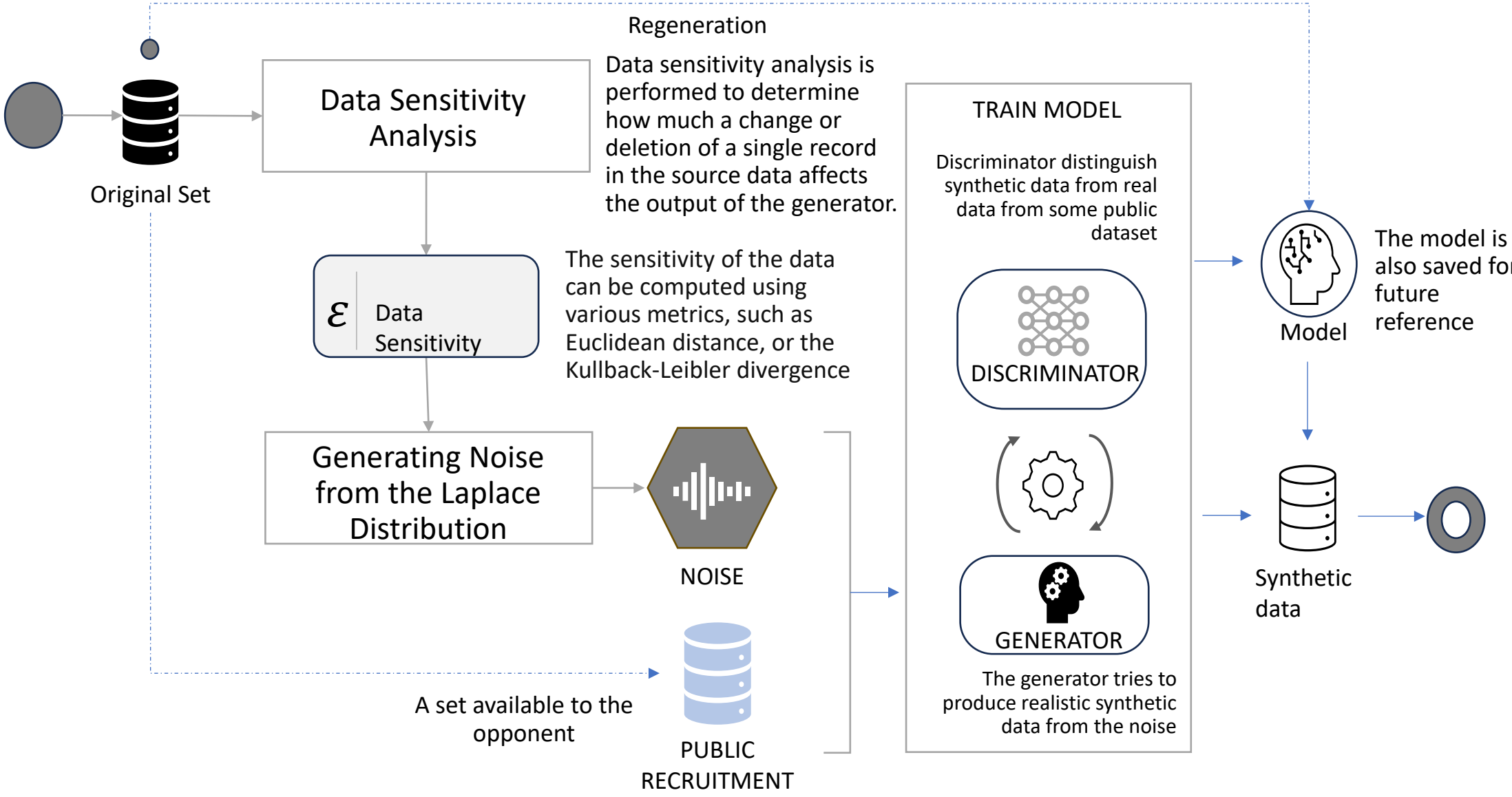
L Linkability Resistance
I Identifiability Resistance
U Utility Estimation Ready
D Disclosure Risk Resistance
P High Performance
P Poor performance
Acts as a primitive for other techniques (sharing)

Adapted by LINDDUN

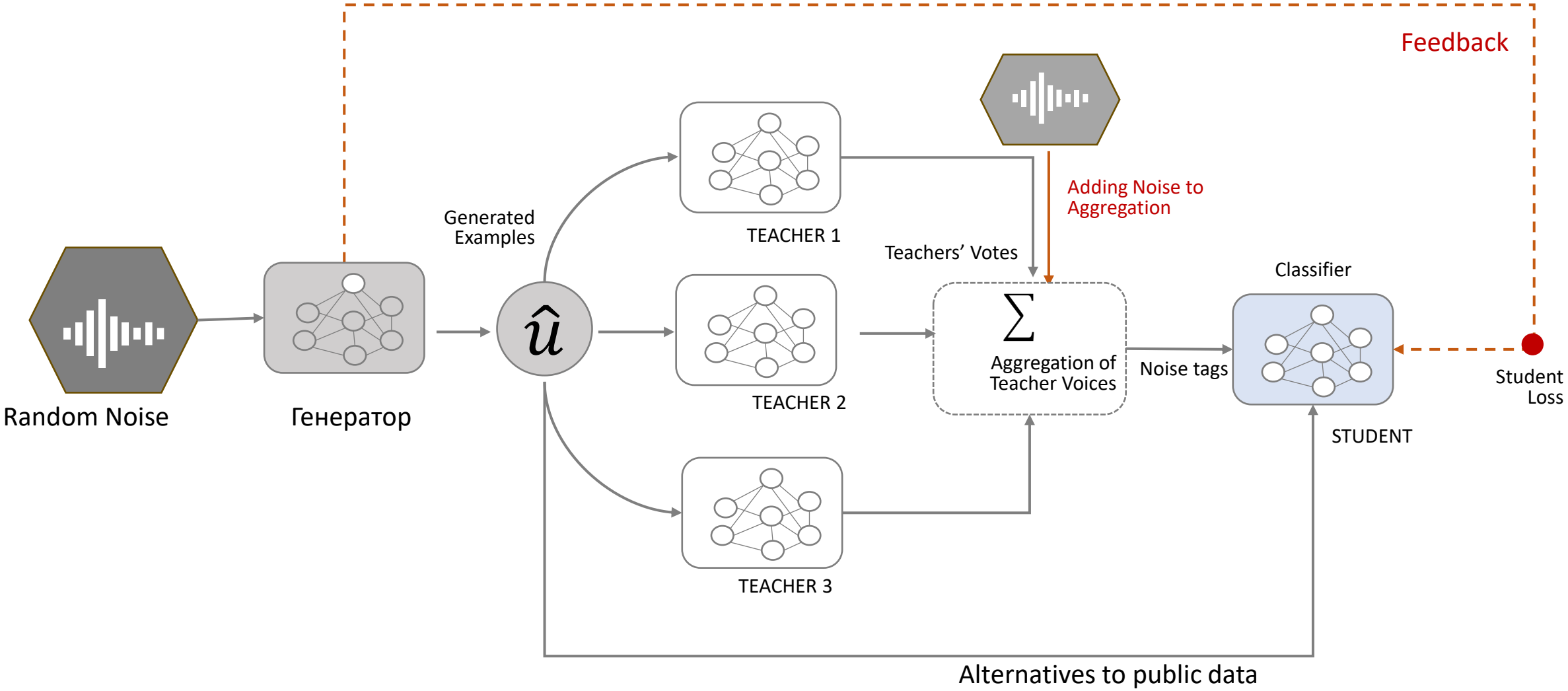
SYNTHETIC DATA GENERATION



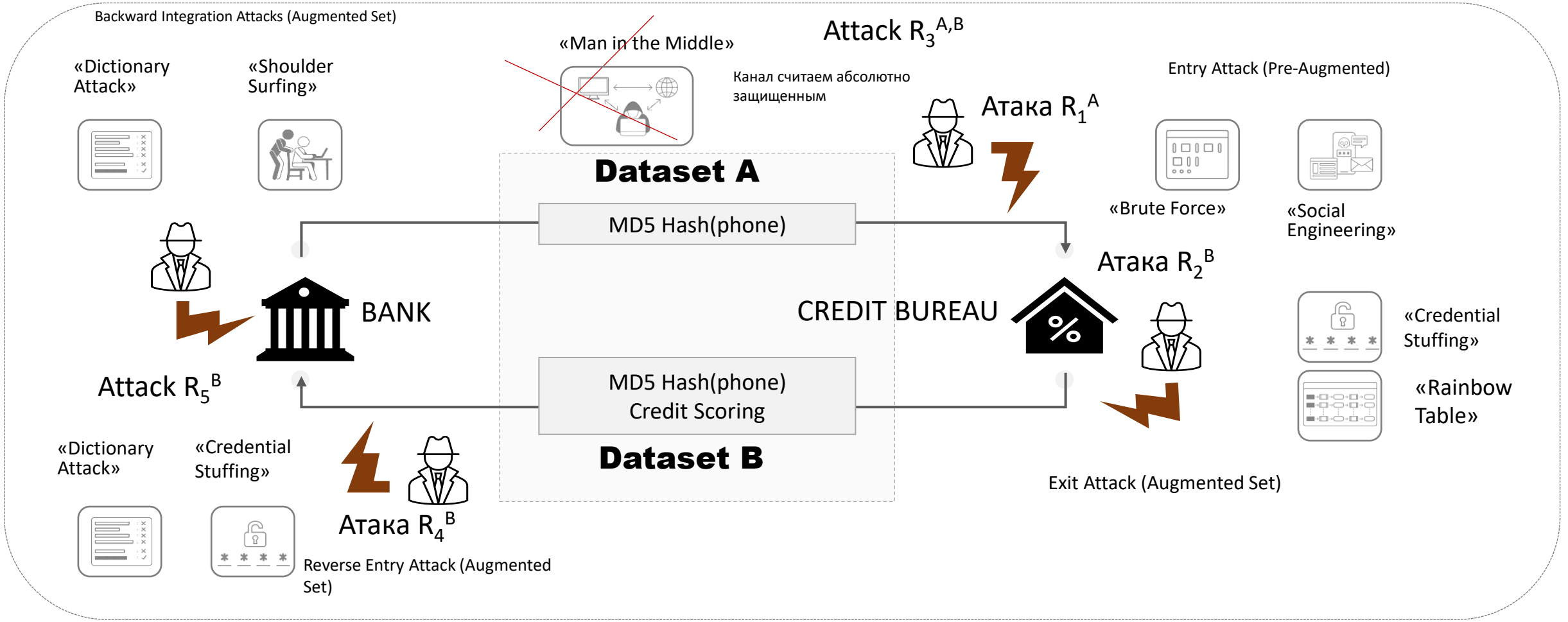
DIFFERENTIAL PRIVACY FOR SYNTHETIC DATA



PATE-GAN ARCHITECTURE



HASH BASED EXCHANGE ATTACK SURFACE



PS3I

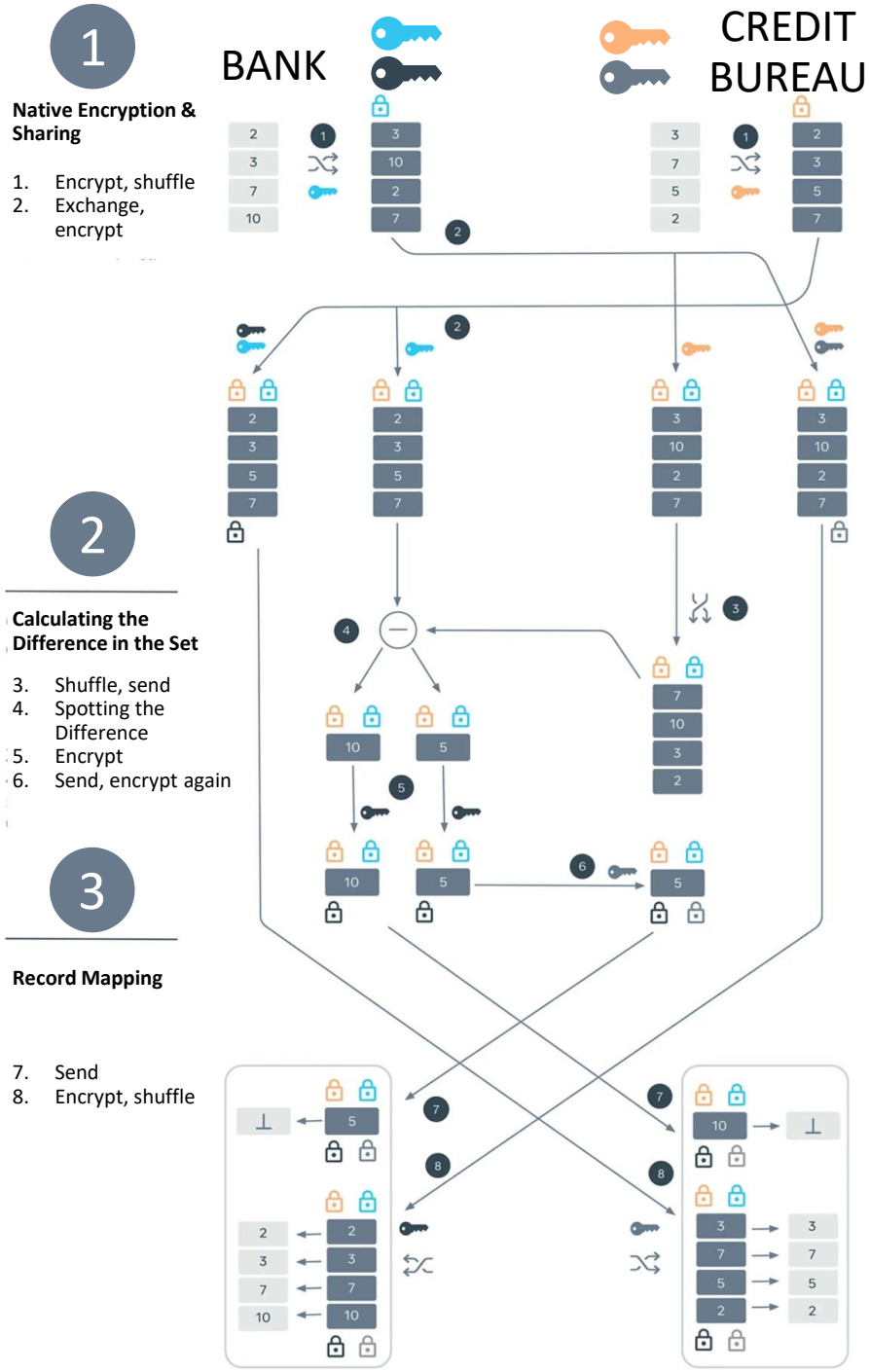
The PS3I scheme with homomorphic encryption for safe data exchange between a bank and a credit bureau is a theoretical construct that involves the use of Secure Multi-Party Computation (SMPC) and Homomorphic Encryption (HE). While "PS3I" is not a standard term in the field as of my last training data, I can describe a generalized scheme that fits this description. Here's an outline of how such a system might work:

Objective: Safely exchange sensitive data (like credit scores linked to phone numbers) between a bank and a credit bureau without exposing the actual data to either party.

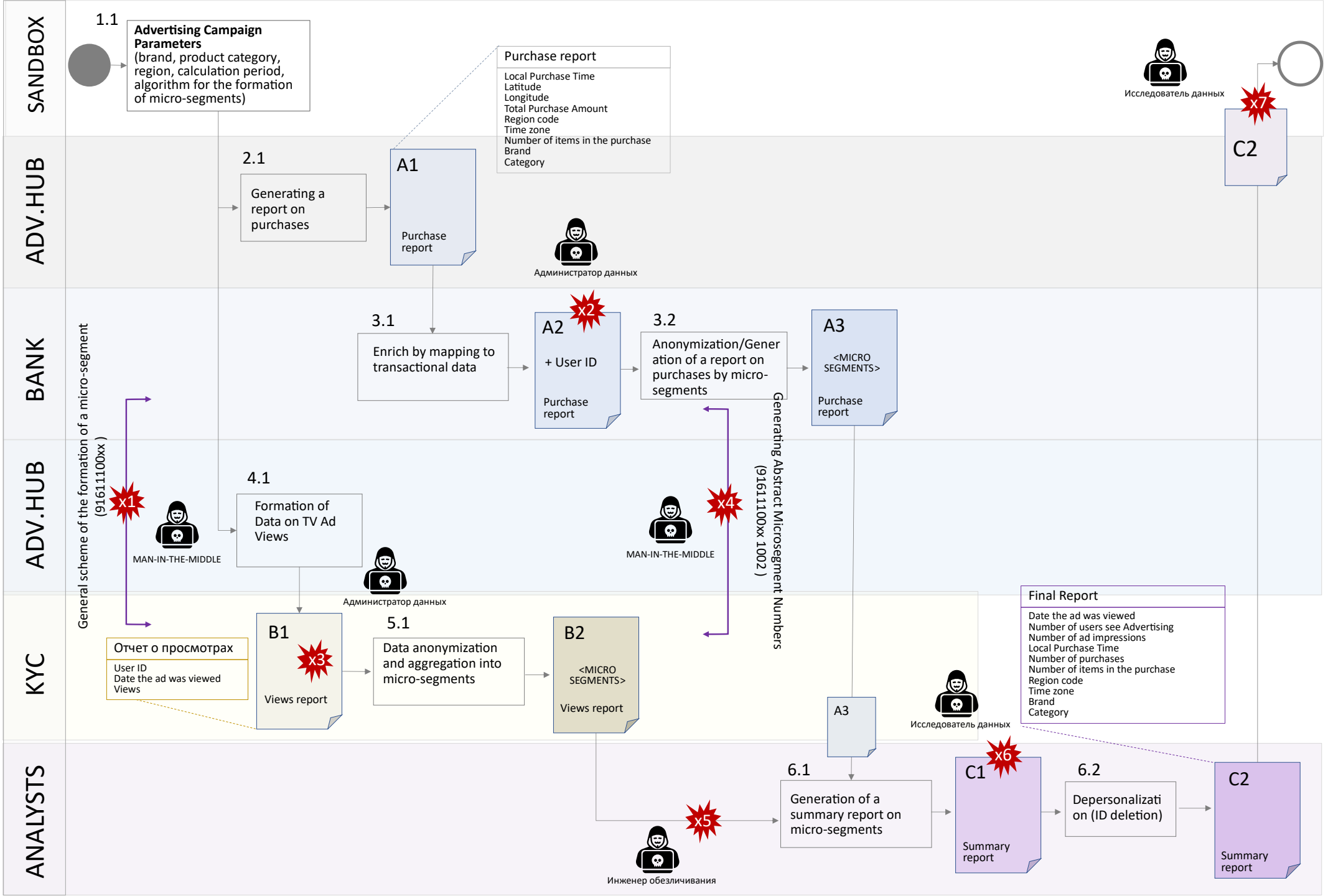
Homomorphic Encryption (HE): A form of encryption that allows computations to be carried out on ciphertext, generating an encrypted result that, when decrypted, matches the result of operations performed on the plaintext.

Secure Multi-Party Computation (SMPC): A cryptographic method that allows parties to jointly compute a function over their inputs while keeping those inputs private.

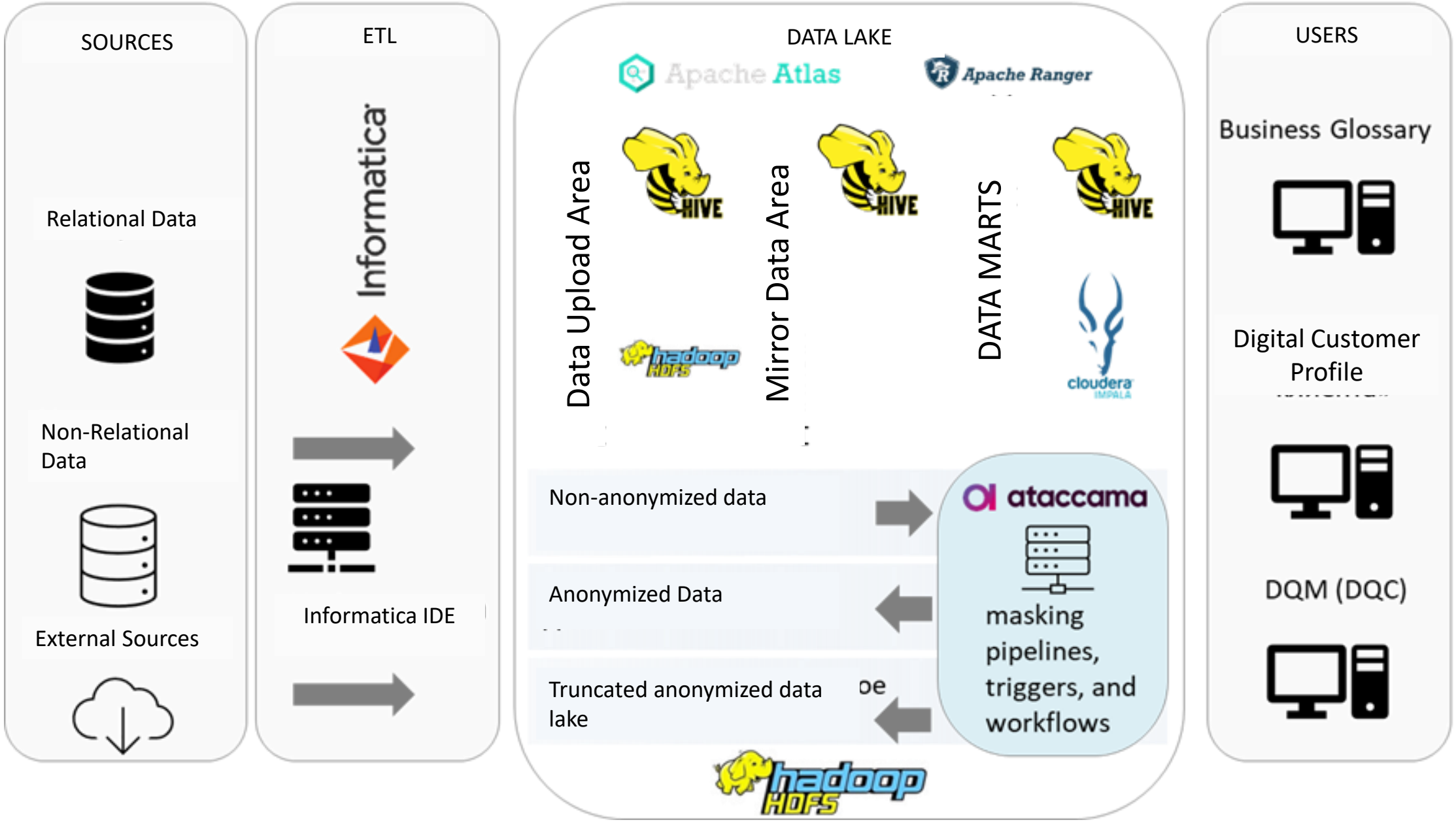
Security Considerations
Privacy: No raw data is exposed to the other party at any point.
Integrity: The integrity of the computation is maintained, as neither party can alter the process to skew results.

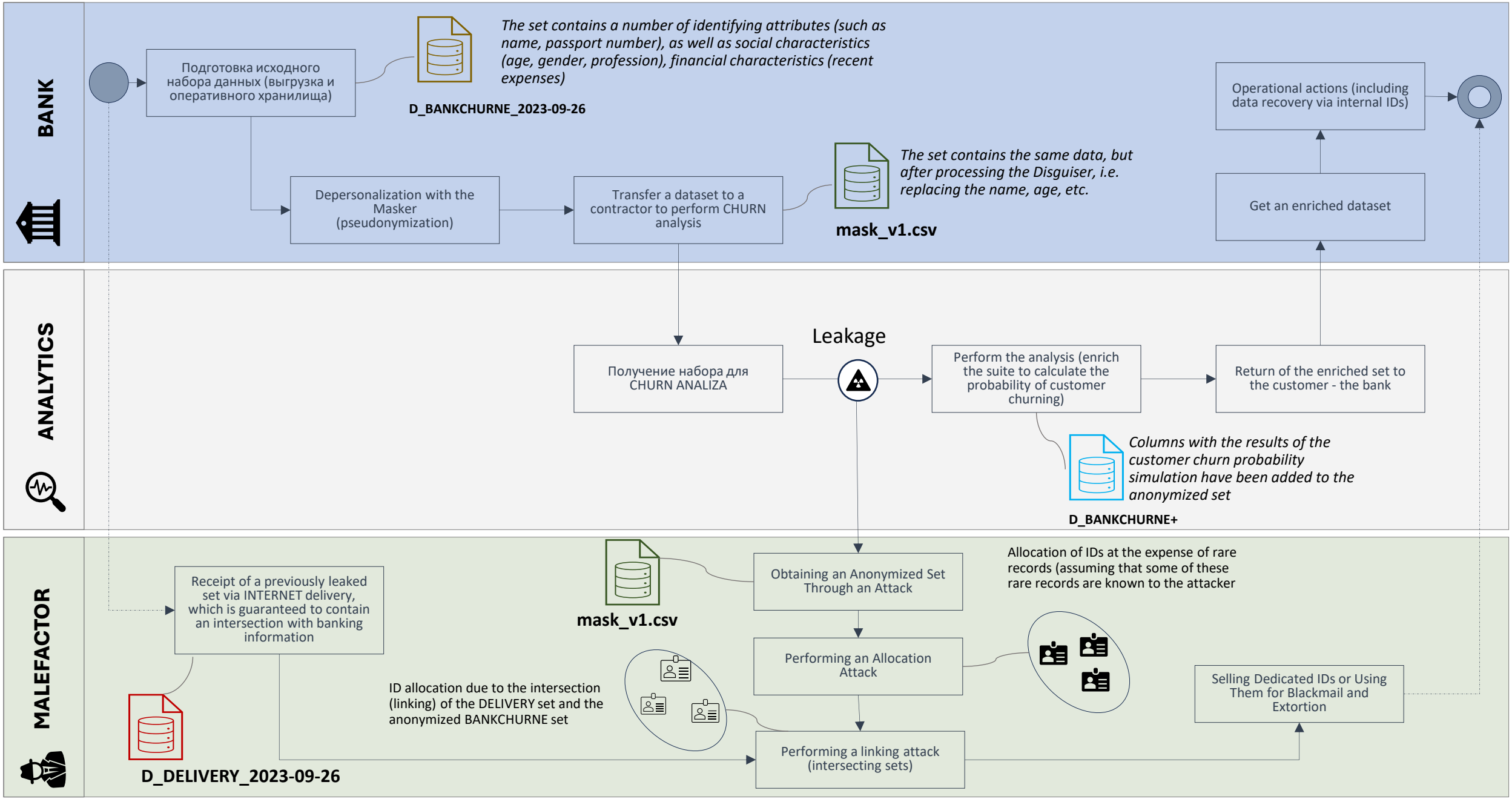


MARKETING ATTRIBUTION ATTACK SURFACE



DATA LAKE ARCHITECTURE




BANK
Подготовка исходного набора данных (выгрузка и оперативного хранилища)
The set contains a number of identifying attributes (such as name, passport number), as well as social characteristics (age, gender, profession), financial characteristics (recent expenses)
D_BANKCHURNE_2023-09-26
Depersonalization with the Masker (pseudonymization)
Transfer a dataset to a contractor to perform CHURN analysis
The set contains the same data, but after processing the Disguiser, i.e. replacing the name, age, etc.
mask_v1.csv
Operational actions (including data recovery via internal IDs)
Get an enriched dataset
ANALYTICS
Leakage
Получение набора для CHURN ANALIZA
Perform the analysis (enrich the suite to calculate the probability of customer churning)
Return of the enriched set to the customer - the bank
Columns with the results of the customer churn probability simulation have been added to the anonymized set
D_BANKCHURNE+
MALEFACTOR
Receipt of a previously leaked set via INTERNET delivery, which is guaranteed to contain an intersection with banking information
mask_v1.csv
Obtaining an Anonymized Set Through an Attack
Allocation of IDs at the expense of rare records (assuming that some of these rare records are known to the attacker
Performing an Allocation Attack
ID allocation due to the intersection (linking) of the DELIVERY set and the anonymized BANKCHURNE set
Performing a linking attack (intersecting sets)
Selling Dedicated IDs or Using Them for Blackmail and Extortion
D_DELIVERY_2023-09-26