High Fidelity Data

Balancing Privacy and Usage

Author: Zilong Tang and Terry Cheng

Version 1.3

2024-08-19

1. Introduction

The effective de-identification algorithms that balance data usage and privacy is critical. Industries like healthcare, finance and advertising, rely on accurate and secure data analysis. However, existing de-identification methods often compromise either the data usability or privacy protection, and limits the advanced applications like knowledge engineering and AI modeling.

To address these challenges, we introduce High Fidelity(HiFi) data, a novel approach to meet the dual objectives of data usability and privacy protection. High Fidelity data maintains the original data's usability while ensuring compliance with stringent privacy regulations.

Firstly, the de-identification approaches and their strengths and weaknesses are examined. Then four fundamental features of HiFi data: visual integrity, population integrity, statistical integrity, and ownership integrity are specified and rationalized. Lastly the balancing of data usage and privacy protection is discussed with examples.

2. Current Status of De-Identification

De-identification is the process of reducing the informative content in data to decrease the probability of discovering an individual’s identity. The growing use of personal information for extended purposes may introduces more risk of privacy leakage.

Various metrics and algorithms have been developed to de-identify data. HHS published a detailed guide at https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html), known as Safe Harbor, to measure de-identified patient health records. Common de-identification approaches:

- **Redaction and Suppression**: removing certain data elements from database records.
    - A common difficulty with these approaches is to define "done properly"
    - Removal of elements can significantly impact the effective use of data and possible loss of critical information for analysis.
- **Blurring**: reducing the data precision of by combining several data elements. Three main approaches are Aggregation: Combining individual data points into larger groups (e.g., summarizing data by region instead of individual address).
    - *Generalization*: Replacing specific data with broader categories (e.g., replacing age with age range).
    - *Pixelation*: Lowering the resolution of data (e.g. less precise geographic coordinates)
    - Blurring methods are used in various reports or statistical summaries to provide a level of anonymity without fully protecting - individual data rather than general-purpose de-identification.
- **Masking**: replacing data elements with either random or made-up value, or with another value in the dataset. It may decrease the accuracy of computations in many cases, affecting the validity and usability. Main variants in this category include:
- *Pseudonymization*: Assigning pseudonyms to data elements to mask their original values while maintaining consistency across the dataset.

- *Perturbation Randomization*: Adding random noise to data elements to mask their true values without completely distorting the overall dataset.
- *Swapping/Shuffling*: Exchanging values between records to mask identities while preserving the dataset's statistical properties.
- *Noise Differential Privacy*: Injecting statistical noise into the data to protect privacy while allowing for meaningful aggregate analysis.

> Each Masking method comes with limitations and challenges, one of them is to make sure the data remains useful while adequately protecting individual privacy.

3. What is High Fidelity Data

There are several key needs for HiFi Data, including but not limited to:

- **Privacy and Regulatory Compliance**: Ensuring data privacy and adhering to associated regulations.
- **Safe Data Utilization**: Discover business insight without risking privacy
- **AI Modeling**: Train AI models with real-world data for better and accurate behavior of model itself and agents.
- **Rapid Data Access for Production Issues**: Access to production quality data during issues or unexpected network traffic without compromising privacy.

Given these complex and multifaceted requirements, a breakthrough solution is necessary which ensures:

- **Privacy Protection**: Privacy and sensitive data is encoded to prevent privacy leaks.
- **Data Integrity**: The transformed data retains the same structure, size, and logical consistency as the original data.
- **Usage for Analysis and AI**: For analysis, projections and AI modeling, the transformation should preserve statistical characteristics and population properties ideally in lossless fashion.
- **Quick Accessible**: Transforming should be quick and on-demand based to ensure the transformed be accessible for production issues.

> High Fidelity Data provides a comprehensive solution to the challenges of data privacy, utilization, and rapid access.

4. High Fidelity Data Specification

High Fidelity Data refers to data which is faithfulness to original features after transformation and/or encoding, including

- **Visual Integrity**: The transformed data retains its original format, making it "look and feel" the same as the original ones (e.g., dates still appear as dates, phone numbers as phone numbers).
- **Population Integrity**: The transformed data preserves the population characteristics of the original dataset, ensuring that the distribution and relationships within the data remain intact.
- **Statistical Integrity**: The statistical properties are maintained, ensuring that analyses performed on the encoded data yield results same to those on the original data.
- **Ownership Integrity**: The data retains information about its origin, ensuring that the ownership and provenance of the data are preserved to avoid unnecessary extended use.

High Fidelity Data maintains privacy, usability, and integrities, making it suitable for data analysis, AI modeling, and reliable deployment by testing of production quality data.

4.1 Visual Integrity

Visual Integrity means the transformed data should comply with the original data in ways:

- **Length of Words and Phrases**: transformations should maintain the original length of the data. For instance, Base64 or AES encrypted names would make them 15-30% longer, which is undesirable.
- **Data Types**: data type should be preserved (e.g., phone numbers should remain as dashed digital characters). Last four digits extracting as integers would break or change validation pipeline.
- **Data Format**: remain consistent with the original.
- **Internal Structure of Composite Data**: complex data types, like addresses, should maintain their internal structure.

Although Visual Integrity might not seem significant at first glance, it profoundly impacts how analysts use the data and how trained LLMs predict outcomes.

As shown following HiFi Data Visual Integrity:

trait	original	transformed
SSN	372-46-1176	447-21-8841
DOB	1983-03-02	1970-10-11
email	frankj@icloud.com	ltgtoo@icloud.com
Phone	301-369-7653	042-347-7255

- Transformed birthdates still appear as dates.
- Transformed phone numbers or SSNs still resemble phone numbers or SSNs, rather than random strings.
- Transformed emails look like valid email addresses but cannot be looked up on a server. No need for popular domains like "Gmail" does encoding, but for less common domains, the domain is encoded as well.

Visual Integrity is critical in complex software ecosystems, especially production environments . Changes in data type and length could cause database schema changes, which is labor-intensive, time-consuming, error prone. Validation failures during QA could restart development sprints, and may even trigger configuration changes in firewalls and security monitoring systems. For instance, invalid email addresses or phone numbers might trigger security alerts.

> Preserving the "Look & Feel" of data is essential for data engineers and analysts, leading to less error-prone insights.

4.2 Population Integrity

Population Integrity ensures the consistency of report and summary statistics is maintained in lossless fashion before and after transformation

- **Population Distributions**: The transformed data should mirror the original data's population distribution (e.g., in healthcare, the percentage of patients from different states should remain consistent).
- **Correlations and Relations**: The internal relationships and correlations between data elements should be preserved which is crucial for analyses that rely on understanding the interplay between different variables. For example, if one "John" had 100 records in the database, after transforming, there would still be 100 records of "John", with each "John" represented only once.

Maintaining Population Integrity is essential to ensure the transformed data remains useful for statistical analysis and modeling for these reasons:

- **Accurate Analysis**: Analysts can rely on the transformed data to provide the same insights as the original data, ensuring that trends and patterns are correctly discovered.
- **Reliable Data Linkage**: Encoded data can still be linked across different datasets without loss of information, allowing for comprehensive analyses that require data integration.
- Consistent Results: Ensures that the results of data queries and analyses are consistent with what would be obtained from the original dataset.

In healthcare, maintaining population integrity ensures accurate tracking of patient records and health outcomes even after data de-identification. In finance, it enables precise analysis of transaction histories and customer behavior without compromising privacy. For example, in a region defined by a set of zip codes, the ratio of vaccine takers to non-takers should remain consistent before and after data de-identification.

> Preserved Population Integrity ensures that encoded datasets remain useful and reliable for all analytical purposes without the privacy risk.

4.3 Statistical Integrity

Statistical Integrity ensures that the statistical properties, like mean, standard deviation(STD), entropy and more, of the original dataset are preserved in the transformed data. This integrity allows for accurate and meaningful analysis, projection and deep mining of the insight and knowledge. It includes:

- **Preservation of Statistical Properties**: Mean, STD, and other statistical measures should be maintained. Ensures that statistical analyses yield consistent outcomes cross transformation.
- **Accuracy of Analysis and Modeling**: crucial for applications in machine learning and AI modeling, like user pharmacy visiting projection and visiting.

Maintaining Statistical integrity is essential for several reasons:

- **Accurate Statistical Analysis**: Analysts can perform statistical tests and derive insights from the transformed data with confidence, knowing that the results will be reflective of the original data.
- **Valid Predictive Modeling**: Machine learning models and other predictive analytics can be trained on the transformed data without losing the accuracy and reliability of the predictions.
- **Consistency Across Studies**: Ensures that findings from different studies or analyses are consistent, facilitating reliable comparisons and meta-analyses.

For example, in the healthcare industry, preserving statistical integrity allows researchers to accurately assess the prevalence of diseases, the effectiveness of treatments, and the distribution of health outcomes. In finance, it enables the precise evaluation of risk, performance metrics, and market trends.

> By ensuring consistent statistical properties, Statistical Integrity supports robust and reliable data analysis, enabling stakeholders to make informed decisions based on accurate and trustworthy insights.

4.4 Ownership Integrity

Owner means an entity who has full controls of original data set. **Entity** usually refers to a person, but it can also mean a company, an application, or a system.

Ownership Integrity ensures that the provenance and ownership information of the data is preserved throughout the transformation process. The data owner can perform additional new transformations as needed in case of the scope/requirement is changed.

- **Data Ownership**: retaining ownership is crucial for maintaining data governance and regulation compliance.
- **Provenance**: reserving the data source origination plays important role for traceable and accountable of the transformed data.

Maintaining ownership integrity is crucial for several reasons:

- **Regulation Compliance**: helps organizations comply with legal and regulatory requirements by maintaining clear records of data provenance and ownership.
- **Data Accountability**: Since the transformation is project based, it can be designed to be reusable or not reusable. For example, different purposes for data analysis and/or model training may transform data accordingly with different data subsets of its origin without cross reference.
- **Data Governance**: supports robust data governance through its lifecycle to avoid unnecessary or unintentional reuse. Trust and Transparency: builds trust with stakeholders by demonstrating that the organization maintains high standards of data integrity and accountability. Users of the transformed data can be assured that it comes from the original owner.

In healthcare, ownership integrity allows the tracking of patient records back to the original healthcare provider. In finance, it ensures that transaction data can be traced back to the original financial institution, supporting regulatory compliance and auditability.

> Preserved Ownership Integrity ensures that encoded datasets remain transparent, accountable, and compliant with regulations, providing confidence to all stakeholders involved.

5. Summary of High Fidelity Data

High Fidelity Data offers a balanced approach to data transformation, combining privacy protection with the preservation of data usability, making it a valuable asset across various industries.

Specification

High Fidelity Data (HiFi Data) specification aims to maintain the original data's usability while ensuring privacy and compliance with regulations. HiFi Data should offer the features:

- **Visual Integrity**: The encoded data retains its original format, ensuring it looks and feels same as the raw data .

- **Population Integrity**: The transformed data preserves the population characteristics of the original dataset, like distribution and frequency.
- **Statistical Integrity**: The preserved statistical properties ensure accurate analysis and projection.
- **Ownership Integrity**: The ownership and provenance is preserved through the transformation which prevents unauthorized re-used.

Benefits

Regulatory Compliance: Helps organizations comply with legal and regulatory requirements by maintaining data ownership and provenance.

- **Data Usability**: Encoded data retains its usability for analysis, reporting, and machine learning, without compromising privacy and re-architect the complicated process management.
- **Data Accountability**: Population and statistical and ownership integrity make the data governance be consistency and accountability.
- **Enhanced Security**: Makes re-identification extremely difficult.
- **Consistency**: Supports consistent encoding across different data sources and projects, promoting uniformity in data handling.

Usage

- **Healthcare**: Ensuring compliance with HIPAA. HiFi Data can be used for population health research and health services research without risking patient privacy.
- **Finance**: Financial models and analyses can be conducted accurately without exposing sensitive information.
- **Advertising**: Enables the use of detailed customer data for targeted advertising while protecting individual identities.
- **Data Analysis and AI modeling**: Provides high-quality data for training models, ensuring they reflect real-world scenarios without compromising privacy sensitive information.

References

1. Health Information Privacy, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," U.S. Department of Health & Human Services, HHS.gov.
2. Gartner, "How to Reduce Bias in AI," Gartner, gartner.com.
3. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule