ENCRYPT Hackathon

Anonymisation Techniques

Stelios Erotokritou(8BELLS)

encrypt

Introduction

- Data Privacy & Anonymization
  - ✓**Data Collection**: Increasing amounts of personal data are gathered.
  - ✓**Anonymization Role**: Removes or conceals sensitive identifiers.
  - ✓**Regulatory Need**: Essential for GDPR compliance (anonymized data isn't personal data).
- Real-World Applications
  - ✓**Financial Services**: Anonymize customer data for PCI DSS compliance.
  - ✓**Healthcare**: Remove names/dates to enable HIPAA-compliant research.
  - ✓**Telecom & Media**: Drop personal identifiers to analyze usage patterns.

  - ✓**Key Insight**: Balances data utility with privacy protection.

Something to think about!

- **Reflect:** Where have you observed data privacy concerns in your field?

- **Consider:** Which types of personal data might need anonymization in your work?

Core Anonymization Techniques

- **Generalization**
  - ✓Replace specific values with broader categories.
- **Suppression**
  - ✓Remove sensitive data or entire records.
- **Masking**
  - ✓Hide parts of data values while preserving format.
- **Perturbation**
  - ✓Introduce controlled noise to modify data.
- **Pseudonymization**
  - ✓Replace real identifiers with artificial codes.
- **Synthetic Data**
  - ✓Generate new records mirroring real data's statistics.

Generalization

- **Definition**: Reduces data precision by grouping similar values.
- **How it works**: Replace exact values with broader categories.
- **Example**: “29” → “20–30” or full address → City/ZIP code.
- **Benefits**: Retains analytic insight; lowers re-identification risk.
- **Trade-offs**: Loss of exact detail; risk of over-generalization.

Suppression

- **Definition**: Removes or blanks out highly sensitive or identifying data.
- **How it works**: Omit individual fields or entire records.
- **Example**: Redact Social Security Numbers; remove unique records.
- **Benefits**: Completely eliminates risk from the suppressed data.
- **Trade-offs**: Reduces dataset completeness; may impact analysis.

Masking

- **Definition**: Obscures parts of a data value while preserving its format.
- **How it works**: Substitute characters in sensitive parts.
- **Example**: “1234-5678-9012-3456” → “XXXX-XXXX-XXXX-3456”; “John Doe” → “Jn De.”
- **Benefits**: Maintains recognizable structure; reduces disclosure risk.
- **Trade-offs**: Can be reversible if context is known; partial data remains visible.

Perturbation

- **Definition**: Alters data by adding controlled noise or inaccuracy.
- **How it works**: Modify values via random offsets or rounding.
- **Example**: Add ±5% noise to salaries; round birth dates to the 1st.
- **Benefits**: Preserves overall statistical trends; obscures individual specifics.
- **Trade-offs**: Reduces precision; may distort subtle data patterns.

Pseudonymization

- **Definition**: Replaces direct identifiers with artificial pseudonyms.
- **How it works**: Substitute names/IDs with consistent, *random codes*.
- **Example**: “John Doe” → “Patient 1001.”
- **Benefits**: Retains data linkages for analysis while concealing identities.
- **Trade-offs**: A re-identification risk exists if the key mapping is compromised.

Synthetic Data

- **Definition**: Generates entirely artificial records that mimic real data.
- **How it works**: Use statistical models/algorithms to create new data.
- **Example**: Synthetic customer dataset replicating purchase patterns.
- **Benefits**: No real personal data is used; safe for public sharing.
- **Trade-offs**: May miss subtle nuances of the original data.

Something to think about!

- **Scenario**: Table with Name, Email, Age, and Purchase Amount.

Example Dataset: Name, Email, Age, and Purchase Amount

Name	Email	Age	Purchase Amount ($)
Alice Smith	alice.smith@email.com	35	120.50
Bob Jones	bob.jones@email.com	45	89.99
Charlie Nguyen	charlie.nguyen@email.com	28	45.00
Diana Lee	diana.lee@email.com	40	150.00

- Which of the described techniques are best to use?

Something to think about!

- **Direct Identifiers**: Remove names/emails (suppression) or replace with codes (pseudonymization).

- **Numeric & Categorical Data**
  - ✓Generalize Age into ranges (e.g., 20–30, 30–40).
  - ✓Add noise to Purchase Amount (perturbation) to preserve trends.

- **Consider**
  - ✓How do these choices balance privacy with data usefulness?
  - ✓What trade-offs would you consider?

Something to think about!

Name (Suppressed)	Email (Suppressed)	Age Range	Noisy Purchase Amount ($)
Person 1	person1@email.com	30–39	121.75
Person 2	person2@email.com	40–49	90.50
Person 3	person3@email.com	20–29	44.20
Person 4	person4@email.com	40–49	149.20

- Suppression (Names, Emails): Remove direct identifiers or replace with pseudonyms (e.g., Person 1, Person 2).
- Generalization (Age): Group ages into ranges to anonymize individual ages.
- Perturbation (Purchase Amount): Add random noise to purchase amounts to protect exact transaction values.

Case Studies & Examples

- Let’s look at how these techniques appear in practice, and when each is appropriate

- **Purpose**: Examples of applying anonymization techniques.

- **Focus**: How each method suits different data types and risks.

- **Later on today**: Identify the best approach for a sample dataset.

Case Studies & Examples - Generalization & Suppression in Practice

- Healthcare Dataset Example
  - ✓Generalization: Band ages into ranges (e.g., 20–29, 30–39).
  - ✓Suppression: Remove outlier records (e.g., one 87-year-old with a rare disease).
- Benefit: Achieves anonymity by making individuals seem similar.
- Trade-off: Some loss of precise information.

Example Healthcare Dataset – Original Data

Patient ID	Age	Gender	Postal Code	Diagnosis
P001	27	F	02139	Asthma
P002	45	M	02138	Hypertension
P003	62	F	02140	Type 2 Diabetes
P004	35	M	02141	Migraine
P005	53	F	02139	Coronary Artery Disease

Age Range	Gender	Region	Diagnosis
20–29	F	021	Asthma
40–49	M	021	Hypertension
60–69	F	021	Type 2 Diabetes
30–39	M	021	Migraine
50–59	F	021	Coronary Artery Disease

- **Generalization**: Convert exact age to age ranges (e.g., 27 → 20–29).
- **Suppression/Truncation**: Remove patient IDs and reduce postal code precision (e.g., 02139 → 021).

Case Studies & Examples - Masking Example

- HR Database Example
  - ✓Masking: Hide parts of names, emails, phone numbers.
    - E.g., “John Doe” → “Jn De”, phone “415-xxx-xx89.”
  - ✓Pseudonymization: Optionally replace IDs to keep record linkages.
- Benefit: Maintains format for analysis while protecting details.
- Trade-off: Some information is still partially visible.

Original HR Dataset

EmpID	Name	Email	Phone	Department	Salary
1001	John Doe	john.doe@company.com	415-123-4567	Sales	$65,000
1002	Jane Smith	jane.smith@company.com	415-234-5678	Marketing	$70,000
1003	Alice Johnson	alice.j@company.com	415-345-6789	Engineering	$85,000
1004	Bob Brown	bob.brown@company.com	415-456-7890	HR	$60,000

Emp Code	Masked Name	Masked Email	Phone	Department	Salary
E1001	J** D**	j***@company.com	415-xxx-xx67	Sales	$65,000
E1002	J** S**	j***@company.com	415-xxx-xx78	Marketing	$70,000
E1003	A**** J**	a****@company.com	415-xxx-xx89	Engineering	$85,000
E1004	B** B****	b****@company.com	415-xxx-xx90	HR	$60,000

- **Pseudonymization**: Replace employee IDs with coded labels.
- **Masking**: Partially mask names and emails.
- **Phone Numbers**: Display only the area code and last two digits.

encrypt

Case Studies & Examples - Perturbation Example

- City Taxi Data Example:
  - ✓Perturbation: Add ±2-3% noise to trip distances and fares.
  - ✓Purpose: Retain overall trends, prevent exact matching.
- **Benefit**: Suitable for aggregate analysis and ML models.
- **Trade-off**: Reduces precision in individual data points.

Original City Taxi Dataset

Trip ID	Distance (km)	Fare ($)
T001	5.2	12.50
T002	3.8	9.80
T003	7.0	15.30
T004	4.5	10.20
T005	6.3	13.75

Trip ID	Distance (km)	Fare ($)
T001	5.3	12.7
T002	3.9	10.0
T003	7.1	15.5
T004	4.4	10.1
T005	6.2	13.6

- **Perturbation applied**: Add a small random variation (±2-3%) to each numeric value.

Case Studies & Examples - Pseudonymization Example

- Clinical Trial Data Example:
  - ✓Pseudonymization: Replace patient names with codes (e.g., “Subject A001”).
  - ✓Purpose: Allow follow-up if needed while hiding identities.
- **Benefit**: Enables analysis without direct identifiers.
- **Trade-off**: Risk if the key linking codes to identities is compromised.

Original Clinical Trial Dataset

Patient ID	Name	Age	Treatment Group	Lab Result
CT001	Alice Smith	52	Control	130/85
CT002	Bob Jones	47	Treatment	120/80
CT003	Carol Lee	63	Treatment	125/82
CT004	David Wong	55	Control	135/88

Subject Code	Age	Treatment Group	Lab Result
A001	52	Control	130/85
A002	47	Treatment	120/80
A003	63	Treatment	125/82
A004	55	Control	135/88

- **Pseudonymization**: Replace direct identifiers with subject codes, removing names and patient IDs; assign codes (e.g., “Subject A001”). Enables analysis while concealing direct identifiers. Maintains data utility for research and outcome analysis.

Case Studies & Examples - Synthetic Data Example

- Bank & Fintech Collaboration:
  - ✓Synthetic Data: Generate artificial transactions resembling real patterns.
  - ✓Purpose: Share data safely without exposing actual customer info.
- **Benefit**: Eliminates privacy risk entirely.
- **Trade-off**: Ensuring synthetic data faithfully represents real trends.

Original Bank Transaction Dataset

Transaction ID	Customer ID	Amount ($)	Transaction Date	Merchant Category
TX1001	C001	120.50	2023-04-15	Electronics
TX1002	C002	45.00	2023-04-16	Groceries
TX1003	C003	89.99	2023-04-17	Clothing
TX1004	C004	150.00	2023-04-18	Dining

Transaction ID	Synthetic Customer ID	Amount ($)	Transaction Date	Merchant Category
TXS001	SC001	118.75	2023-04-14	Electronics
TXS002	SC002	47.20	2023-04-16	Groceries
TXS003	SC003	91.00	2023-04-17	Clothing
TXS004	SC004	149.50	2023-04-18	Dining

- **Synthetic Data Generation**: New records are created to mimic real patterns without using actual customer data.
  Share realistic transaction patterns without exposing sensitive customer information.

encrypt

When to Use What

- When to Use What
  - ✓**Direct Identifiers**: Remove or pseudonymize immediately.
  - ✓**Quasi-Identifiers**: Generalize or perturb to prevent re-identification.
  - ✓**Precision Needs**: For high fidelity, consider generalization plus selective suppression.
  - ✓**Combination Approach**: Often multiple methods are used together for optimal balance.

Conclusion & Key Takeaways – Overview

- Recap: Why anonymization matters in protecting privacy
- Explored various techniques for anonymizing data
- Emphasis on balancing data utility with risk management

Know Your Data & Risk

- Identify Direct vs. Indirect Identifiers:
  - ✓Direct (e.g., names, SSNs) → remove or transform
  - ✓Indirect (quasi-identifiers) → require careful treatment
- Risk Awareness:
  - ✓Anonymized data can still be re-identified through cross-referencing
  - ✓Learn from high-profile cases (e.g., Netflix Prize de-anonymization)

Choose the Right Techniques (Often, Combine Them)

- Technique Combination:
  - ✓Pseudonymization, masking, generalization, suppression, etc.
- Tailored Approach:
  - ✓Select methods based on data type and analysis needs
- Balance:
  - ✓Ensure privacy without overly compromising data utility

Best Practices

- Early Integration:
  - ✓Incorporate anonymization in the data pipeline from the start
- Access Control:
  - ✓Limit access to raw personal data
- Documentation:
  - ✓Record anonymization methods for transparency and reproducibility
- Continuous Improvement:
  - ✓Monitor new techniques (e.g., differential privacy) and test for re-identification risks

Open Questions & Reflection

- **Reflect**:
  - ✓What challenges do you anticipate in your projects?
  - ✓How will you measure if anonymization is sufficient (e.g., k-anonymity targets)?
- **Consider**:
  - ✓Impact of emerging privacy laws and technologies on your approach
- **Question**: What key insight are you taking away, and how will it shape your data handling?

Thank you!

https://encrypt-project.eu/

encrypt-project

@encrypt_project

encrypt