Data Protection and Privacy

University of Genoa

Lesson 7: Privacy Preserving Test Data Manufacturing

Gaspare Ferraro <ferraro@gaspa.re>

Fundamental of Privacy Preserving Testing

- Testing is an important part of the systems development life cycle (SDLC)
- The quality of software application depends on the quality of testing
- Software testing is a process, or a series of processes, designed to ensure that computer code does what it was designed to do and that it does not do anything unintended
- High-quality testing requires high-quality data
- Testing activity are nowadays mostly outsourced
- Testing data should be anonymized according to specific regulations → it could impact utility

Privacy of Test Data

- A fundamental problem in test outsourcing is how to allow a database-centric application owner to release its private data with guarantees that the entities in these data (e.g., people, organizations) are protected at a certain level while retaining testing efficacy
- Anonymizing Multidimensional Data for Testing:
    - Non-Perturbative anonymization techniques (e.g., generalization) that are suitable for data release are not useful in anonymizing testing data
    - Perturbative data anonymization techniques are preferred, like transformation, rotation, and noise addition have been proposed

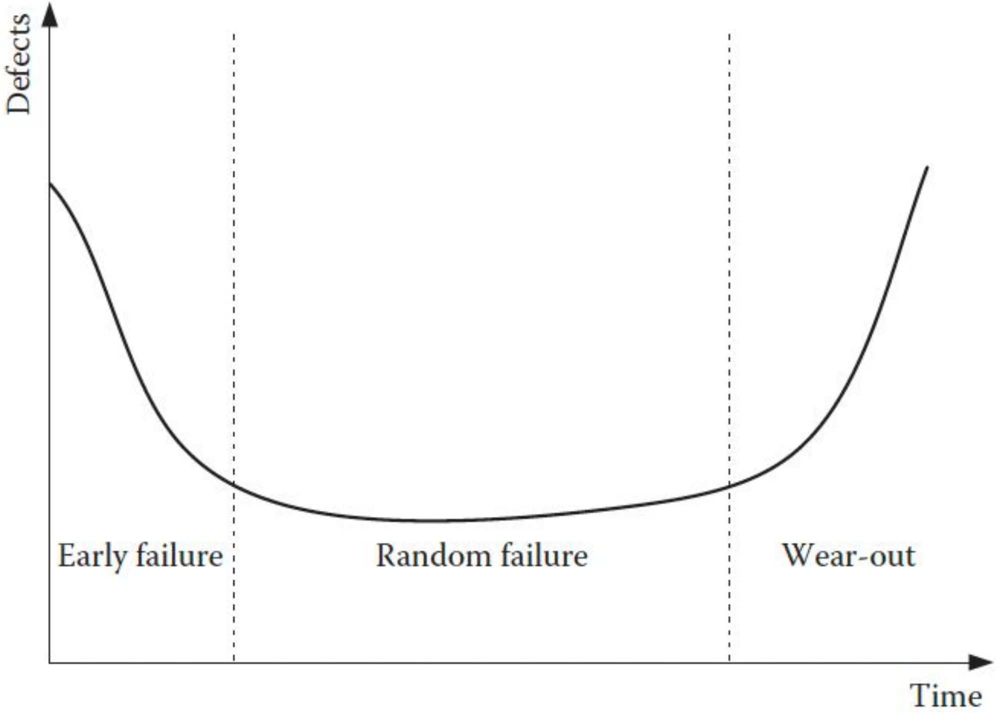
Testing Data Fundamentals I

- Functional vs. Non-Functional Testing
    - **Functional Testing**: System testing is aimed at evaluating specific parts of the system w.r.t. specific use cases. After system testing is completed, integration testing evaluates how the different (tested) parts integrate.
    - **Non-Functional testing**: Stress or load, scalability, responsiveness, reliability, security, etc., look at various nonfunctional characteristics of the software system and assess it against the agreed benchmarks.
- Test Data
    - Good test case requires testing all possible inputs: impossible.
    - Realistically, test data are limited to the predicates laid out by the testing team with the aim to maximize the test and code coverage.
    - The quality of test data determines the success or the failure of a test case.
    - The best source of data is production data, but they must be anonymized as they carry personal information (especially in financial and health fields).

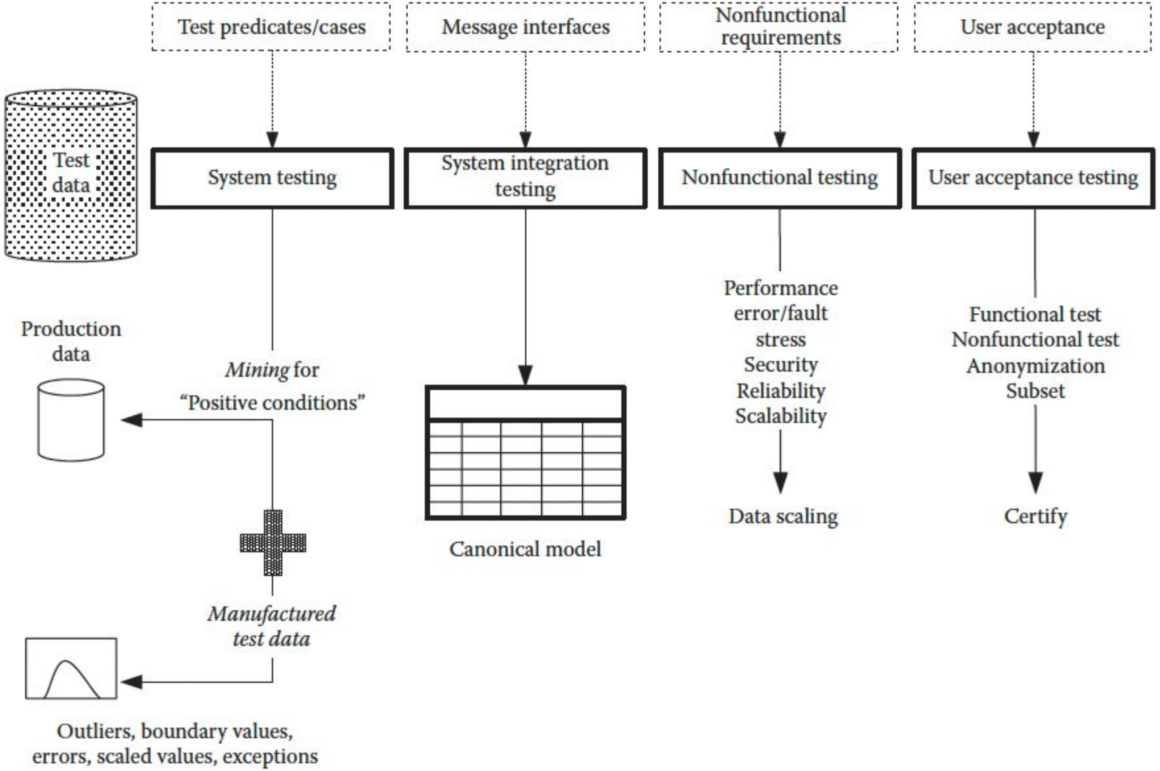
Test Data and Reliability

- Any device or software has three phases of reliability:
  - **Early failure**: Software exhibits many defects that are uncovered by initial rounds of testing. Defects are high at this time.
  - **Random failure**: Past the initial phases, the software stabilizes and exhibits a steady state with random defects that get discovered intermittently, while minor amendments are made to the software system.
  - **Wear-out**: Software system is becoming old and needs updating to keep up with changes in policy/business/technology that require numerous amendments. During this phase, again the number of defects begins to increase and continues to do so until the software system is unable to adapt any further and is abandoned

Bathtub curve of reliability



Test Data Design



Test Data and Reliability II

- As test environments do not get the same kinds of resources as production, there is the need to reduce the amount of data that could be brought into test → samples should be defined.
- Subsets are samples of data picked from original data in a way that they represent the entire data in syntax, semantics and statistics → the goal of creating a sample is to ensure that the responses derived from the sample are identical to those derived by using the original data itself.

Utility of Test Data: Coverage I

abbiamo bisogno di un numero che ci indichi quanto del nostro codice è testato. è il numero di righe testate

- Coverage is a quality assurance metric that determines how thoroughly a test suite exercises a given program → The loss of test coverage can be measured using the lines of code that were covered using original and anonymized data.
- Problem: how much does anonymization impact test coverage? How much utility is lost?
- The utility loss of test coverage can be measured in terms of lines of codes that can be tested using the original and the anonymized data.

gli outliners sono utili per il testing

Utility of Test Data: Coverage II

- Let’s suppose we have a software S to test with data D
- For each test phase t the test coverage (with N the total number of test cases) is given by the formula:

$$TC(t) = \sum^{N}_{n=1} tc(t_i)$$

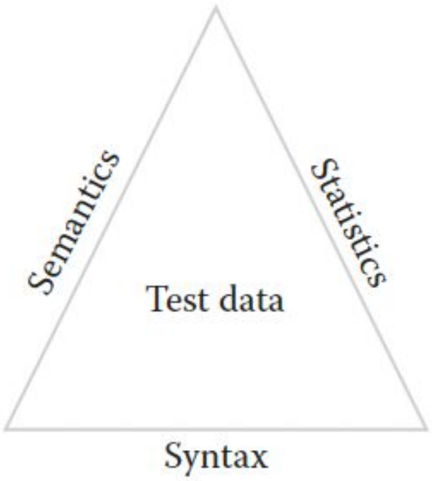
- Where the test case tc_n:

$$tc_n = f(S_n, d_n)$$

- depends on the system being tested (S_n) and the data provisioned (d_n).

Utility of Test Data: Coverage III

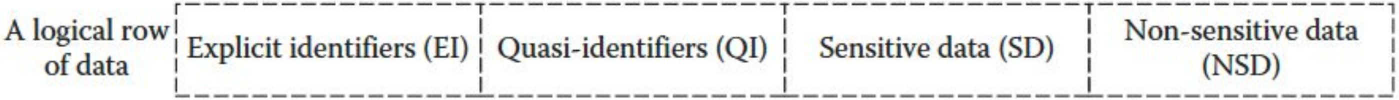
- Utility metrics: The utility of test data is given measured according to three characteristics, namely Syntax, Semantics and Statistics.



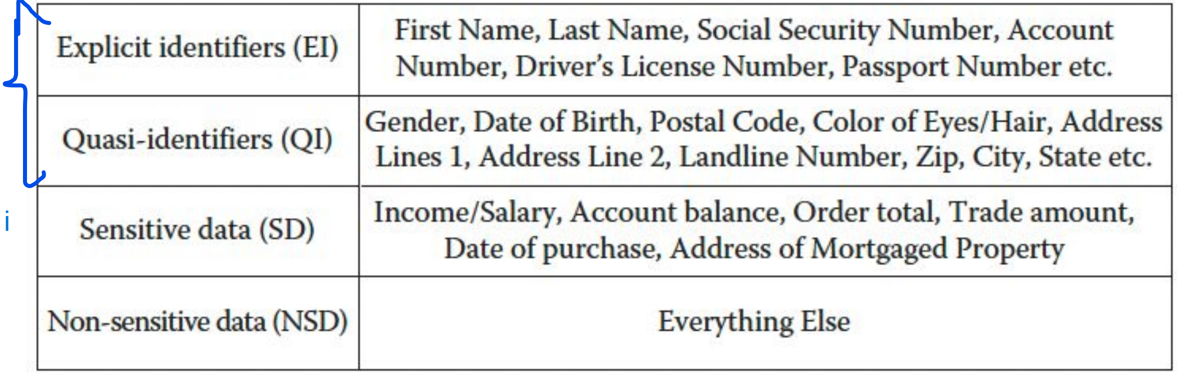
Utility of Test Data: Coverage IV - Privacy vs Utility

- Test data anonymization starts from defining EI, QI and SD (as expected).
  - EI should be anonymized ! semantics or statistics of EI are irrelevant for the utility of test data.
  - QI depends on the application context.
  - SD represent the facts in a software system. It projects the current state of the system, whereas EI and QI are details around the current state.
  - Outliers: Outliers have significance in test data as they are important since they tend to invoke parts of code that do not execute often.
- Outliers dilemma:
  - the best way to deal with outliers from a privacy perspective is to suppress them as they have potential to reveal identities on their own.
  - For a test manager, removal of outliers does not help test coverage.

Utility of Test Data: Coverage V



A logical row of data	Explicit identifiers (EI)	Quasi-identifiers (QI)	Sensitive data (SD)	Non-sensitive data (NSD)



Explicit identifiers (EI)	First Name, Last Name, Social Security Number, Account Number, Driver's License Number, Passport Number etc.
Quasi-identifiers (QI)	Gender, Date of Birth, Postal Code, Color of Eyes/Hair, Address Lines 1, Address Line 2, Landline Number, Zip, City, State etc.
Sensitive data (SD)	Income/Salary, Account balance, Order total, Trade amount, Date of purchase, Address of Mortgaged Property
Non-sensitive data (NSD)	Everything Else

nella registrazione

collezionando i dati nel tempo

Utility of Test Data: Coverage VI

- Test coverage and privacy is given by:
  - D: Production dataset
  - T: Optimal subset of D
  - f : Anonymization function applied on T
  - L: LOC (line of code) tested
  - T’: Anonymized test data subset
  - L’: LOC tested with T’
  - T = Subset of D
  - T’ = f (T)
- The utility is given by:

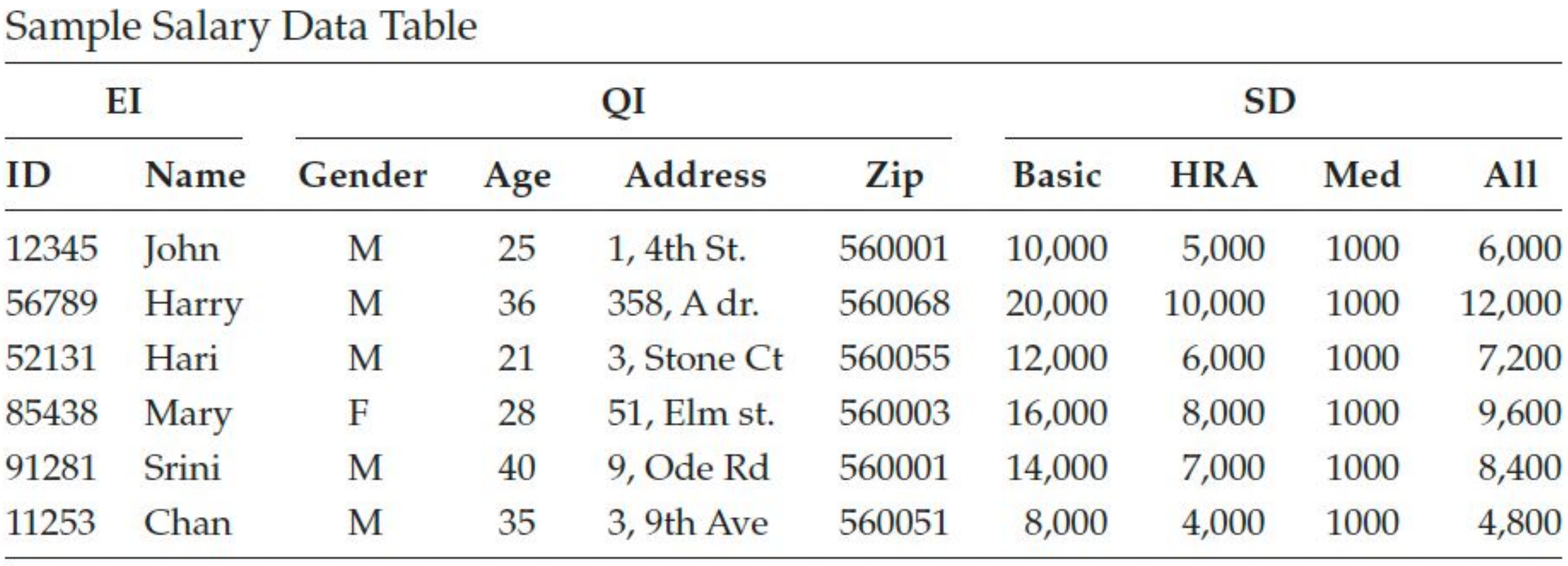
$$U = 1 - (L - L') / L$$

- An anonymization technique that leads to L - L’ = 0 is optimal for utility preservation

Privacy Preservation of Test Data: EI I

- EI must be masked, however, they cannot be randomly removed when dealing with test data. Instead, the masking algorithms should grant, on the anonymized test data:
  - Referential Integrity
  - Consistency
- Referential Integrity: There are cases where these primary keys appear as foreign keys in other tables. Hence, while masking any key field, care should be taken to propagate the same masked value to all respective rows of tables where this field appears. This is essential to preserve data integrity.

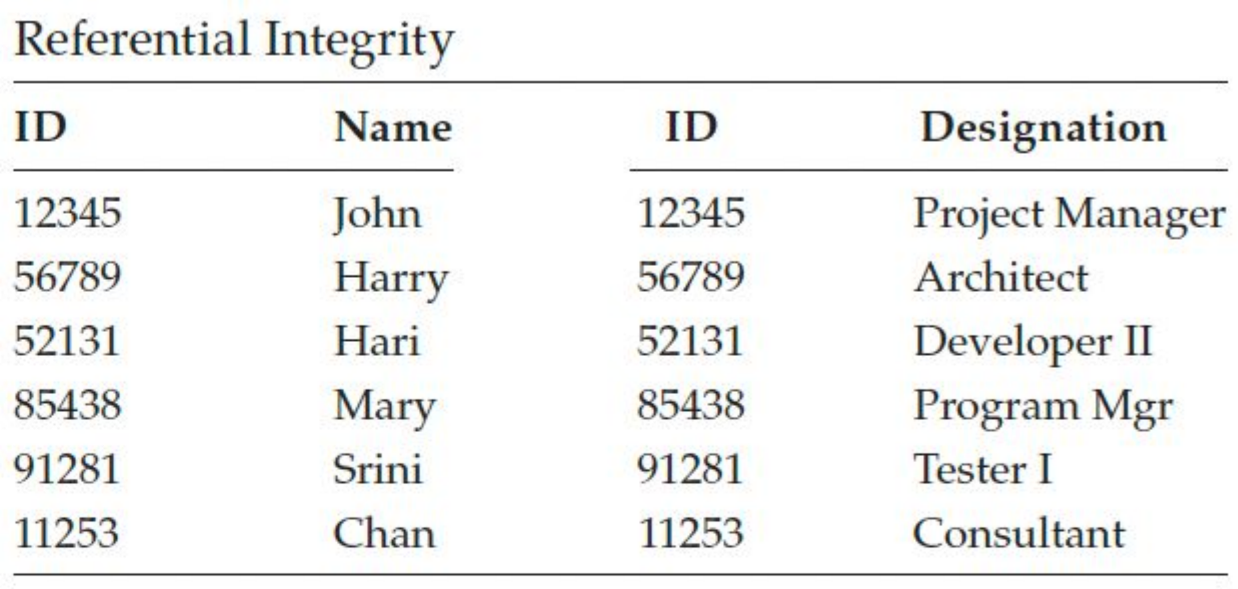
Privacy Preservation of Test Data: EI II



Sample Salary Data Table

EI		QI				SD			
ID	Name	Gender	Age	Address	Zip	Basic	HRA	Med	All
12345	John	M	25	1, 4th St.	560001	10,000	5,000	1000	6,000
56789	Harry	M	36	358, A dr.	560068	20,000	10,000	1000	12,000
52131	Hari	M	21	3, Stone Ct	560055	12,000	6,000	1000	7,200
85438	Mary	F	28	51, Elm st.	560003	16,000	8,000	1000	9,600
91281	Srini	M	40	9, Ode Rd	560001	14,000	7,000	1000	8,400
11253	Chan	M	35	3, 9th Ave	560051	8,000	4,000	1000	4,800

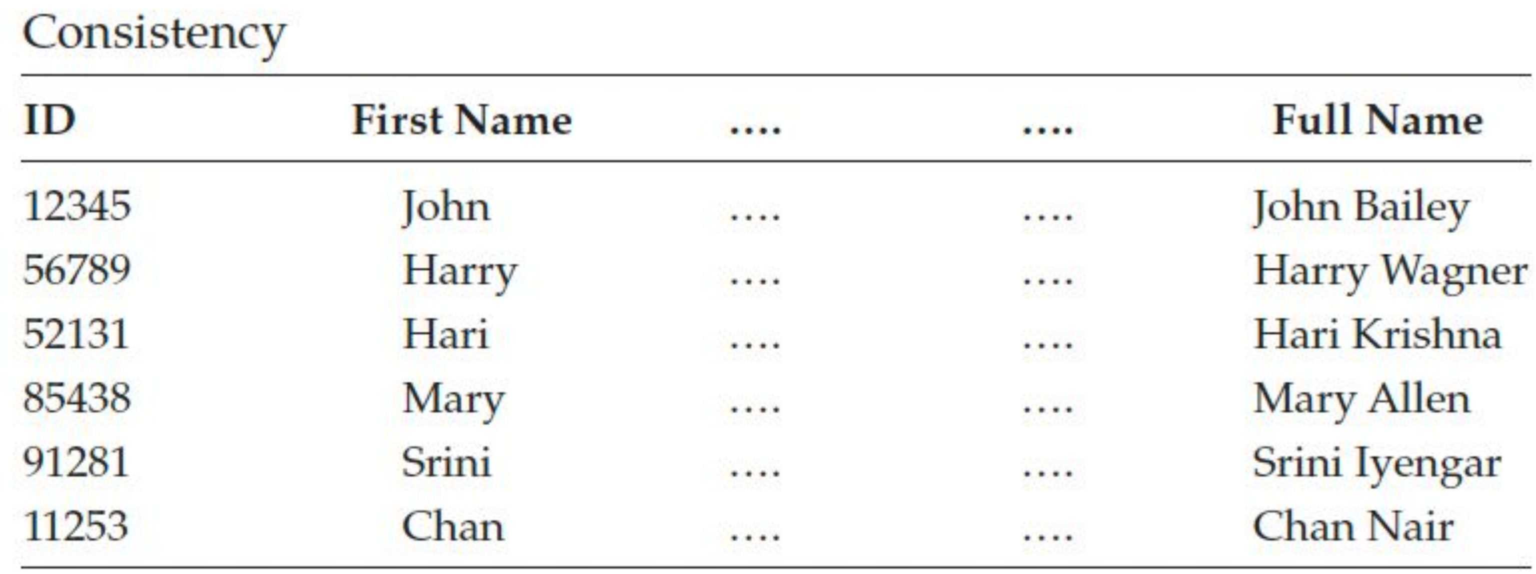
Privacy Preservation of Test Data: EI III



Referential Integrity

ID	Name	ID	Designation
12345	John	12345	Project Manager
56789	Harry	56789	Architect
52131	Hari	52131	Developer II
85438	Mary	85438	Program Mgr
91281	Srini	91281	Tester I
11253	Chan	11253	Consultant

Privacy Preservation of Test Data: EI IV



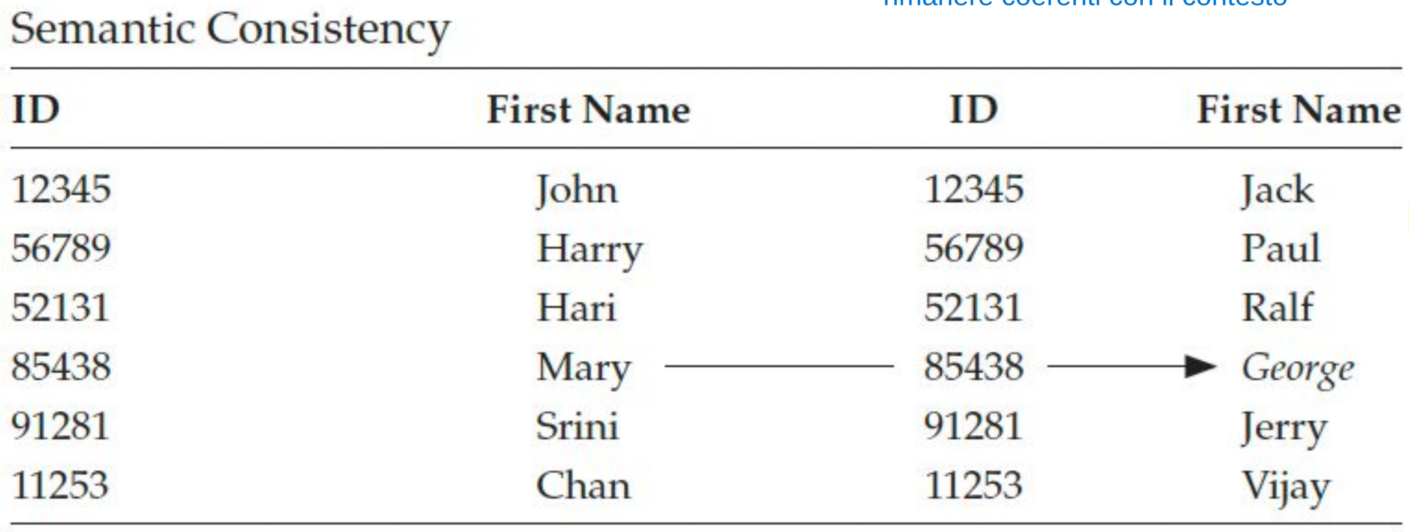
Consistency

ID	First Name	….	….	Full Name
12345	John	….	….	John Bailey
56789	Harry	….	….	Harry Wagner
52131	Hari	….	….	Hari Krishna
85438	Mary	….	….	Mary Allen
91281	Srini	….	….	Srini Iyengar
11253	Chan	….	….	Chan Nair

Privacy Preservation of Test Data: EI V

- Consistency: If there are semantic relationships between EI and QI, such relationship these must be maintained

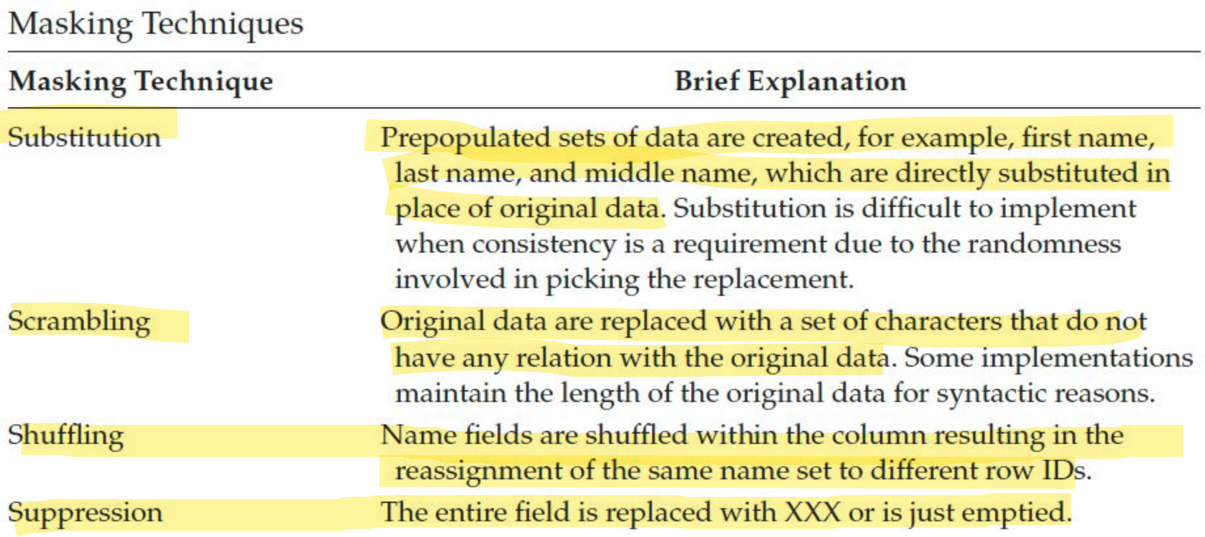
se vogliamo mischiarli comunque dobbiamo
rimanere coerenti con il contesto



Semantic Consistency

ID	First Name	ID	First Name
12345	John	12345	Jack
56789	Harry	56789	Paul
52131	Hari	52131	Ralf
85438	Mary →	85438 →	*George*
91281	Srini	91281	Jerry
11253	Chan	11253	Vijay

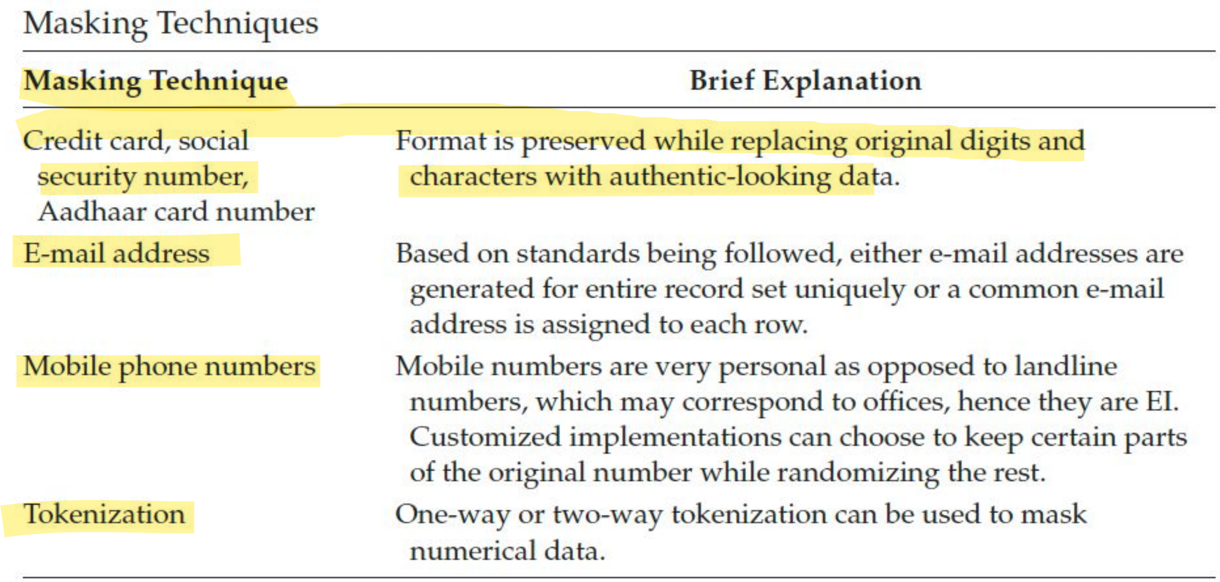
Privacy Preservation of Test Data: EI VI



Masking Techniques

Masking Technique	Brief Explanation
Substitution	Prepopulated sets of data are created, for example, first name, last name, and middle name, which are directly substituted in place of original data. Substitution is difficult to implement when consistency is a requirement due to the randomness involved in picking the replacement.
Scrambling	Original data are replaced with a set of characters that do not have any relation with the original data. Some implementations maintain the length of the original data for syntactic reasons.
Shuffling	Name fields are shuffled within the column resulting in the reassignment of the same name set to different row IDs.
Suppression	The entire field is replaced with XXX or is just emptied.

Privacy Preservation of Test Data: EI VII

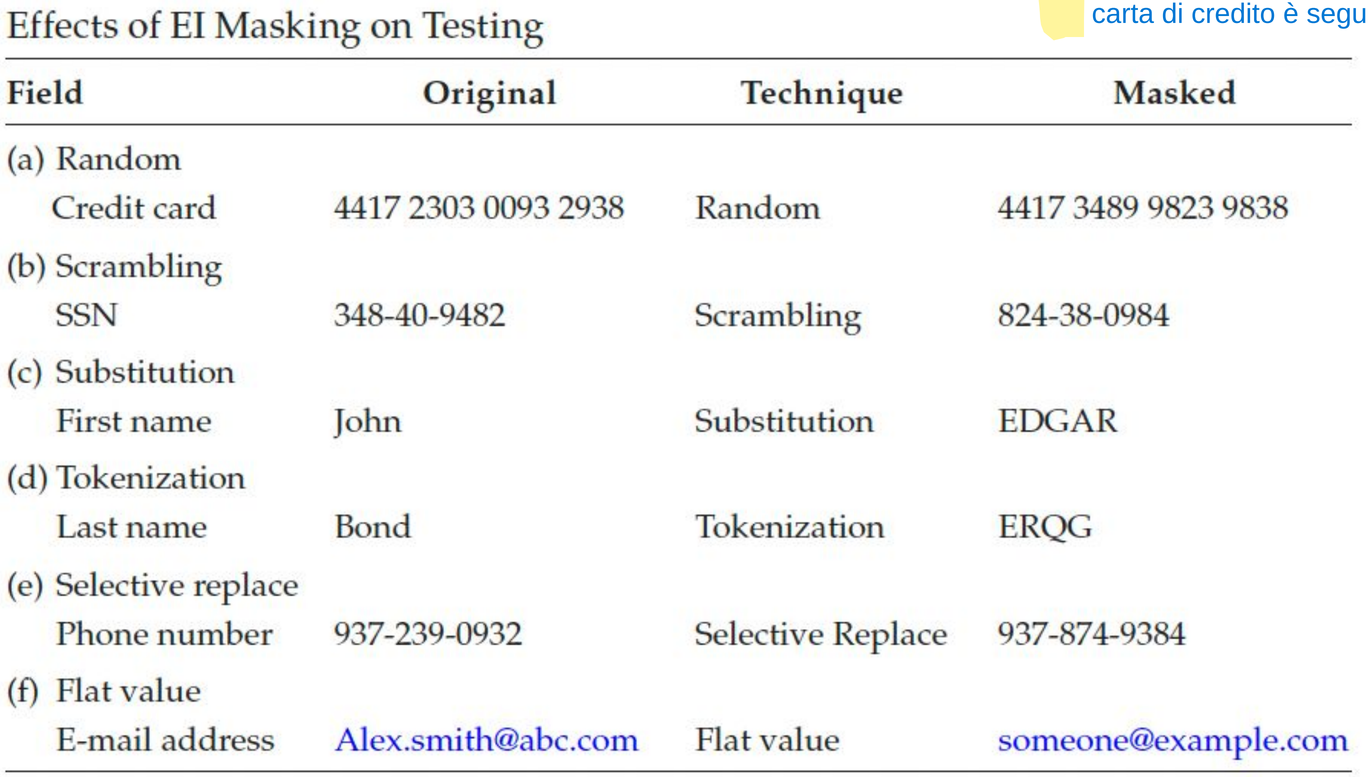


Masking Techniques

Masking Technique	Brief Explanation
Credit card, social security number, Aadhaar card number	Format is preserved while replacing original digits and characters with authentic-looking data.
E-mail address	Based on standards being followed, either e-mail addresses are generated for entire record set uniquely or a common e-mail address is assigned to each row.
Mobile phone numbers	Mobile numbers are very personal as opposed to landline numbers, which may correspond to offices, hence they are EI. Customized implementations can choose to keep certain parts of the original number while randomizing the rest.
Tokenization	One-way or two-way tokenization can be used to mask numerical data.

Privacy Preservation of Test Data: EI VIII

qui ad esempio anche se cambi i codici delle carte di credit devi essere sicuro che i nuovi numeri seguano le regole su come un numero della carta di credito è seguito.



Effects of EI Masking on Testing

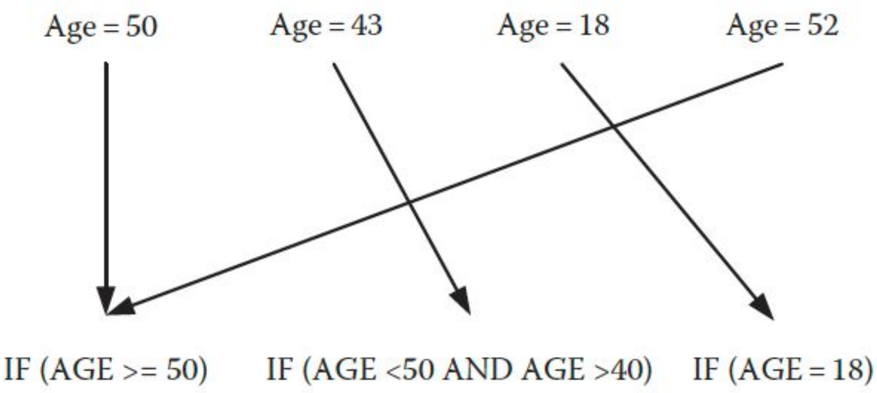
Field	Original	Technique	Masked
(a) Random			
Credit card	4417 2303 0093 2938	Random	4417 3489 9823 9838
(b) Scrambling			
SSN	348-40-9482	Scrambling	824-38-0984
(c) Substitution			
First name	John	Substitution	EDGAR
(d) Tokenization			
Last name	Bond	Tokenization	ERQG
(e) Selective replace			
Phone number	937-239-0932	Selective Replace	937-874-9384
(f) Flat value			
E-mail address	Alex.smith@abc.com	Flat value	someone@example.com

Privacy Preservation of Test Data: QI I

- In test data, QI play a vital role in business logic, driving the transactional data.
- A good example is demographics of individuals, where locations contribute to establishing parameters that either allow or disallow grant of loans.
- QI: Utility of Test Data vs. Data Release
  - Test data utility lies in the spread of values in the QI and their respective SD columns
  - As we discussed in previous lessons, anonymized data for release should maintain the correlation between QI and SD as much as possible.
  - In test data QI, each data value is unique and others a different flow that the program could take in its execution → granting correlation between QI and SD is not sufficient.

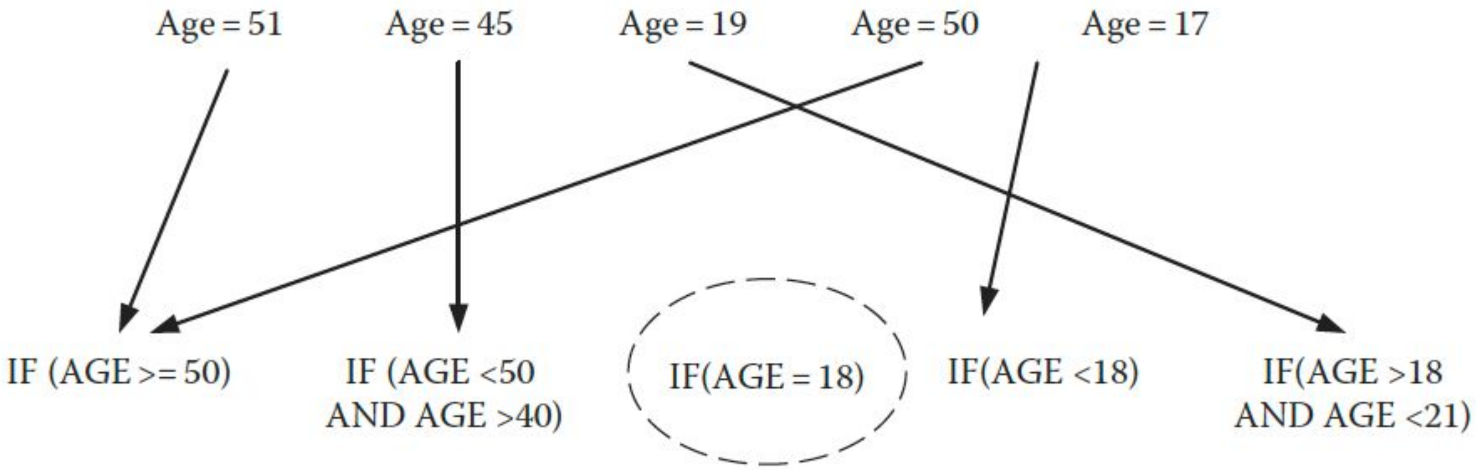
QI in Test Data: an example I

- Let’s suppose that the QI Age in the program is contained in statements which handle multiple conditions (i.e., potential flows of executions).



- If the QI is generalized and used for testing, the generalization would impact which part of the program will be tested → the reliability of the testing.

QI in Test Data: an example II

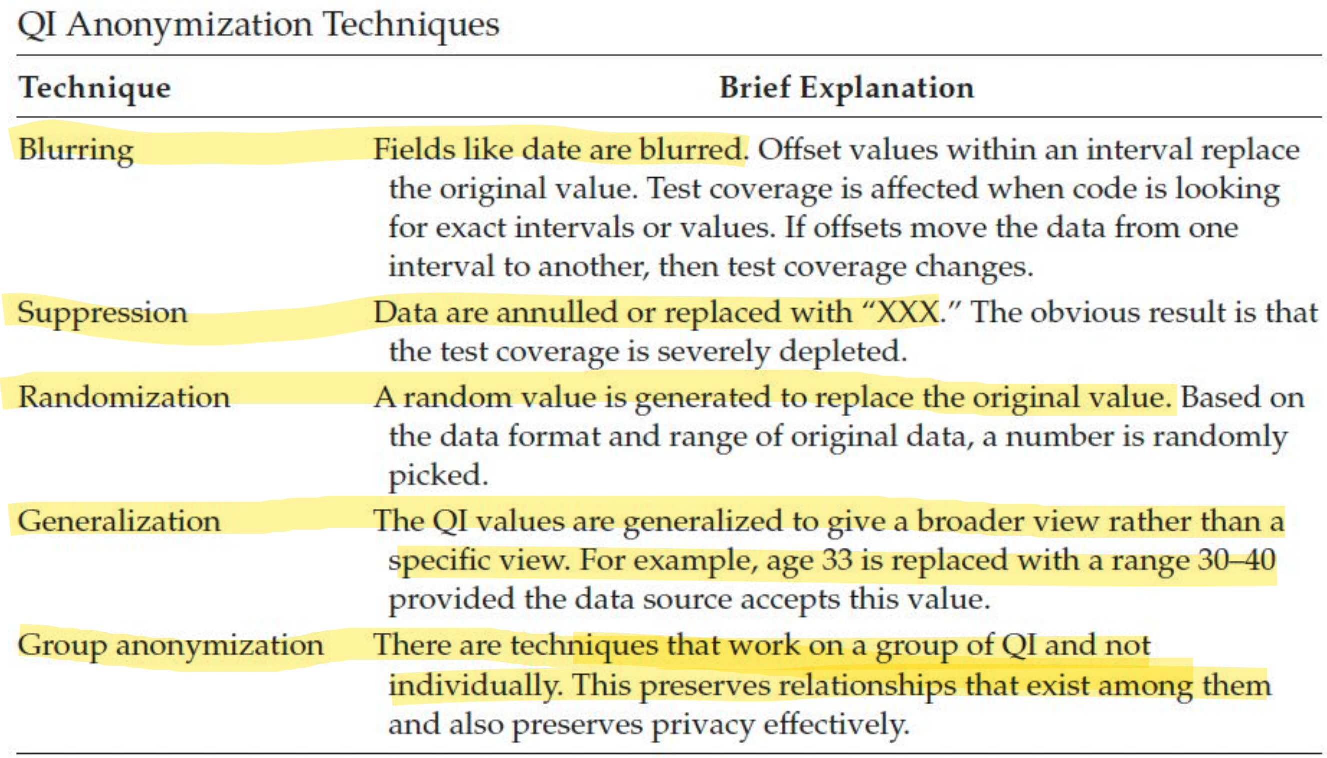


Privacy Preservation of Test Data: QI II

QI in Test Data:

- Perturbative techniques are not suitable for anonymizing test data QI.
- Individual values have high significance when it comes to anonymizing QI in test data as opposed to other areas like privacy preserving data mining, where the distribution as a whole is important and not its constituent data points

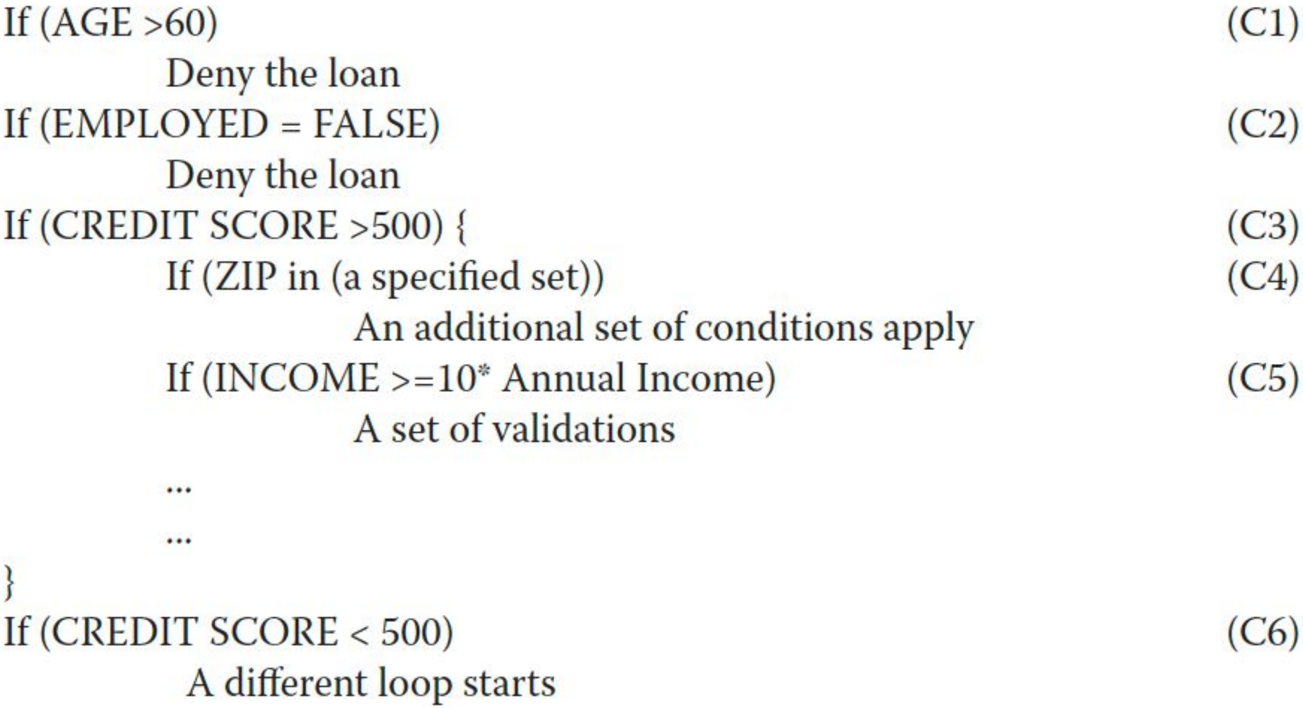
Privacy Preservation of Test Data: QI III



QI Anonymization Techniques

Technique	Brief Explanation
Blurring	Fields like date are blurred. Offset values within an interval replace the original value. Test coverage is affected when code is looking for exact intervals or values. If offsets move the data from one interval to another, then test coverage changes.
Suppression	Data are annulled or replaced with "XXX." The obvious result is that the test coverage is severely depleted.
Randomization	A random value is generated to replace the original value. Based on the data format and range of original data, a number is randomly picked.
Generalization	The QI values are generalized to give a broader view rather than a specific view. For example, age 33 is replaced with a range 30–40 provided the data source accepts this value.
Group anonymization	There are techniques that work on a group of QI and not individually. This preserves relationships that exist among them and also preserves privacy effectively.

Privacy Preservation of Test Data: QI IV



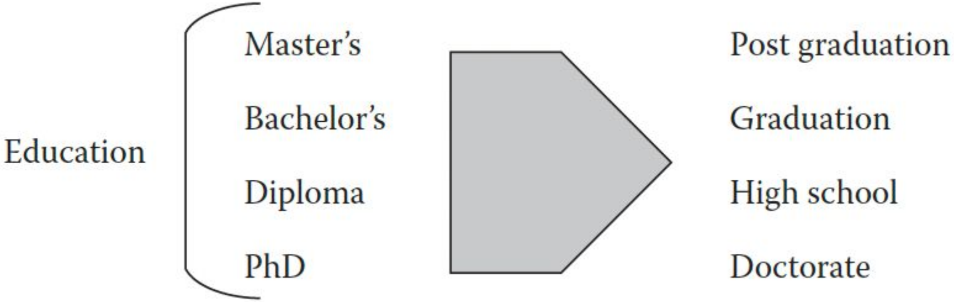
```
If (AGE >60)                                              (C1)
        Deny the loan
If (EMPLOYED = FALSE)                                     (C2)
        Deny the loan
If (CREDIT SCORE >500) {                                  (C3)
        If (ZIP in (a specified set))                     (C4)
                An additional set of conditions apply
        If (INCOME >=10* Annual Income)                   (C5)
                A set of validations
        ...
        ...
}
If (CREDIT SCORE < 500)                                   (C6)
         A different loop starts
```

Privacy Preservation of Test Data: QI V

Impact of QI Anonymization Techniques on Testing:

- *Blurring*: blurring moves the original DATE OF BIRTH on either side of the original value, C1 will be satisfied for applicants whose original age was below 60. The opposite may happen too: some people with ages above 60 may still be eligible for the loan.
- *Suppression*: A suppressed value does not satisfy any condition in the code. For this, if the EMPLOYMENT STATUS of the applicant is suppressed, C2 will never be checked.
- *Randomization* - Random values: A random value does not have the contours expected from the original attribute data. For example, a DATE OF BIRTH field with the original value of year as 1952 could get 2001 as the resultant value. This can definitely disrupt the way an application for loan gets processed and thus leading ot errand flows.
- *Generalization*: both domain and value generalization are unhelpful in test data. For instance, consider the example in the next slide where the EDUCATION attribute is generalized. The corresponding snippet of codes will behave oddly.

Privacy Preservation of Test Data: QI VI



Education

Master's

Bachelor's

Diploma

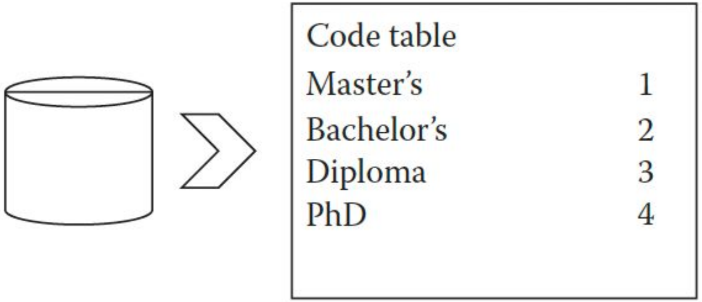
PhD

Post graduation

Graduation

High school

Doctorate



Code table

Master's	1
Bachelor's	2
Diploma	3
PhD	4

```
If (EDUCATION_LEVEL = 1)
  do this...
Else if (EDUCTATION_LEVEL = 2)
  do this...
...
...
```

Privacy Preservation of Test Data: QI VII

Impact of QI Anonymization Techniques on Testing (continue):

- *Group Anonymization*: k-anonymization makes QI indistinguishable and applies suppression, thereby leading to loss of test coverage. Besides, a high value of k means that many rows of important diverse data are being made homogeneous. This affects test coverage adversely. Conversely, a smaller value of k is not effective in protecting privacy.

We stated that the value of k depends on several factors:

$$k = f(P_R, U_R, C_R, G_L, C)$$

In this case, the utility requirement U_R is equivalent to the test coverage T_C

Quality of Test Data I

- The suitability and efficacy of test data depends on four aspects:
  1. Lines of coverage
  2. Query ability
  3. Time for testing
  4. Defect detection
- Lines of coverage:
  - Code coverage is a suitability measure. In this context, data quality is not an absolute metric, but one relative to the test cases.

Quality of Test Data II

- LOC Coverage

$$LCC = 1 - [\ (L_O - L_T) / L_O\ ]$$

- Where L_O is the LOC covered by original data and L_T by the test data

- Query Ability: Queries on anonymized data may yield a slight different result than what the original data would have. Nevertheless, that is the price paid to preserve privacy.

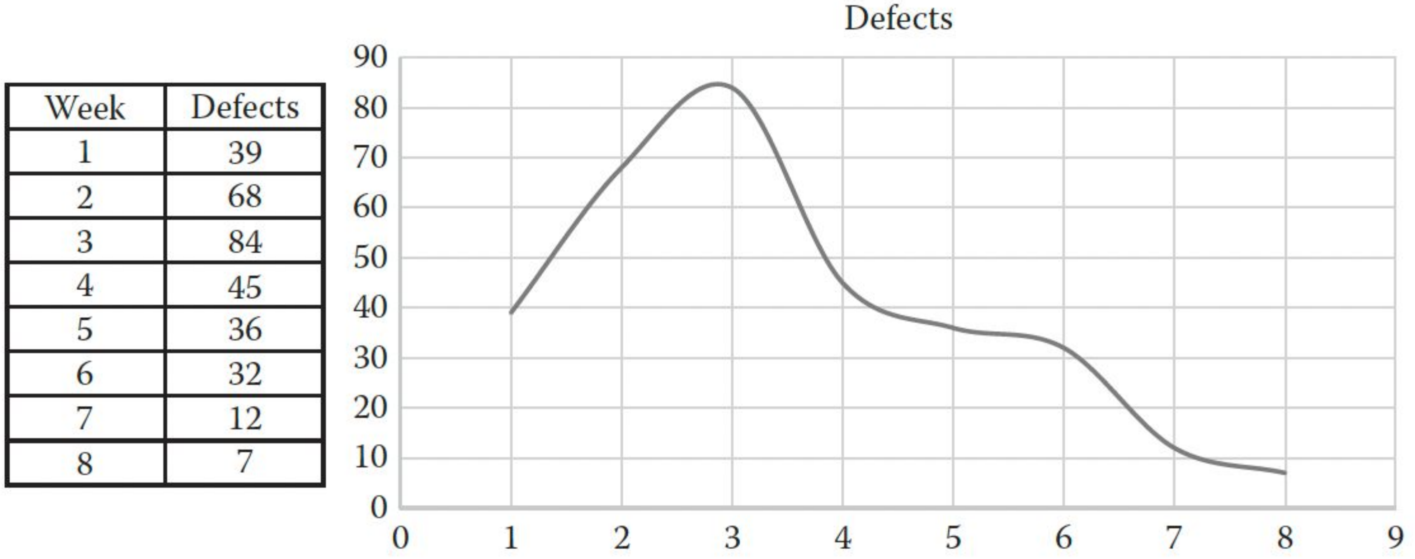
$$Q_A = 1 - [\ (Q_O - Q_T) / Q_O]$$

- being Q_O the number of queries that can be run on original data, Q_T the number of queries that can be run on test data, and Q_A the resulting query ability of the test data.

Quality of Test Data III

- Time for Testing: Time for testing refers to the total time required by testers to uncover the target number of defects from a data set.

piu vai avanti piu i bug sono difficili da trovare , e costa molto trovarli



Week	Defects
1	39
2	68
3	84
4	45
5	36
6	32
7	12
8	7

Quality of Test Data IV

- Time for Testing: The time for testing can be evaluated in terms of duration:

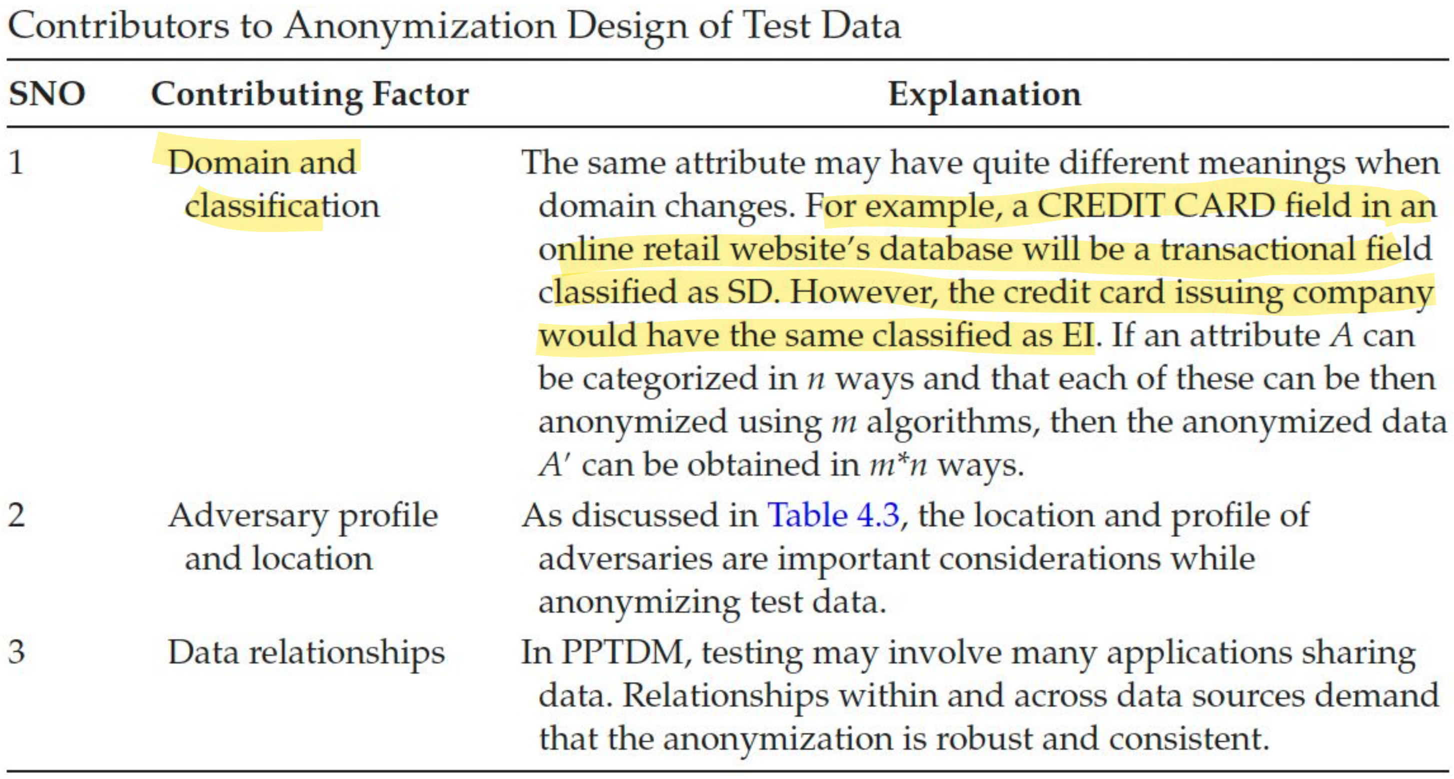
$$D_T = f(T, Q_{TD})$$

- where T are testing parameters, Q_{TD} the quality of test data, and D_T the test duration or number of defects to be detected over a fixed period of time.

Anonymization Design for PPTDM I

- Privacy regulations and data design:
- Privacy regulations impose restrictions on what can be borrowed from original data and what cannot.
- Therefore, most inadequacies encountered in test data are because no thought is given to fill the holes that privacy protection leaves in the test data.
- Inadequacies affect all the measures of test data quality discussed in the previous section.
- A good anonymization design takes into account privacy restrictions and utility needs to produce high-quality test data.
- In particular, you should pay special attention to the following contributors

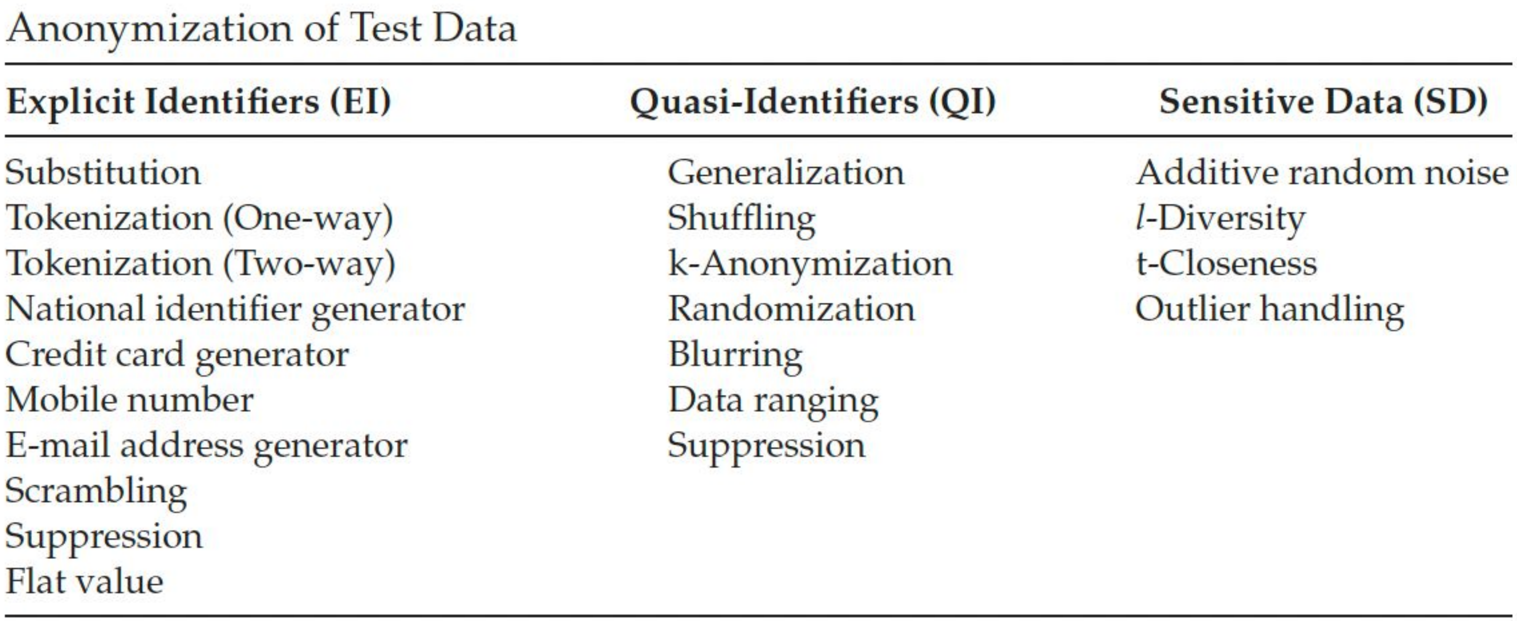
Anonymization Design for PPTDM II



Contributors to Anonymization Design of Test Data

SNO	Contributing Factor	Explanation
1	Domain and classification	The same attribute may have quite different meanings when domain changes. For example, a CREDIT CARD field in an online retail website's database will be a transactional field classified as SD. However, the credit card issuing company would have the same classified as EI. If an attribute A can be categorized in n ways and that each of these can be then anonymized using m algorithms, then the anonymized data A' can be obtained in m*n ways.
2	Adversary profile and location	As discussed in Table 4.3, the location and profile of adversaries are important considerations while anonymizing test data.
3	Data relationships	In PPTDM, testing may involve many applications sharing data. Relationships within and across data sources demand that the anonymization is robust and consistent.

Anonymization Design for PPTDM III



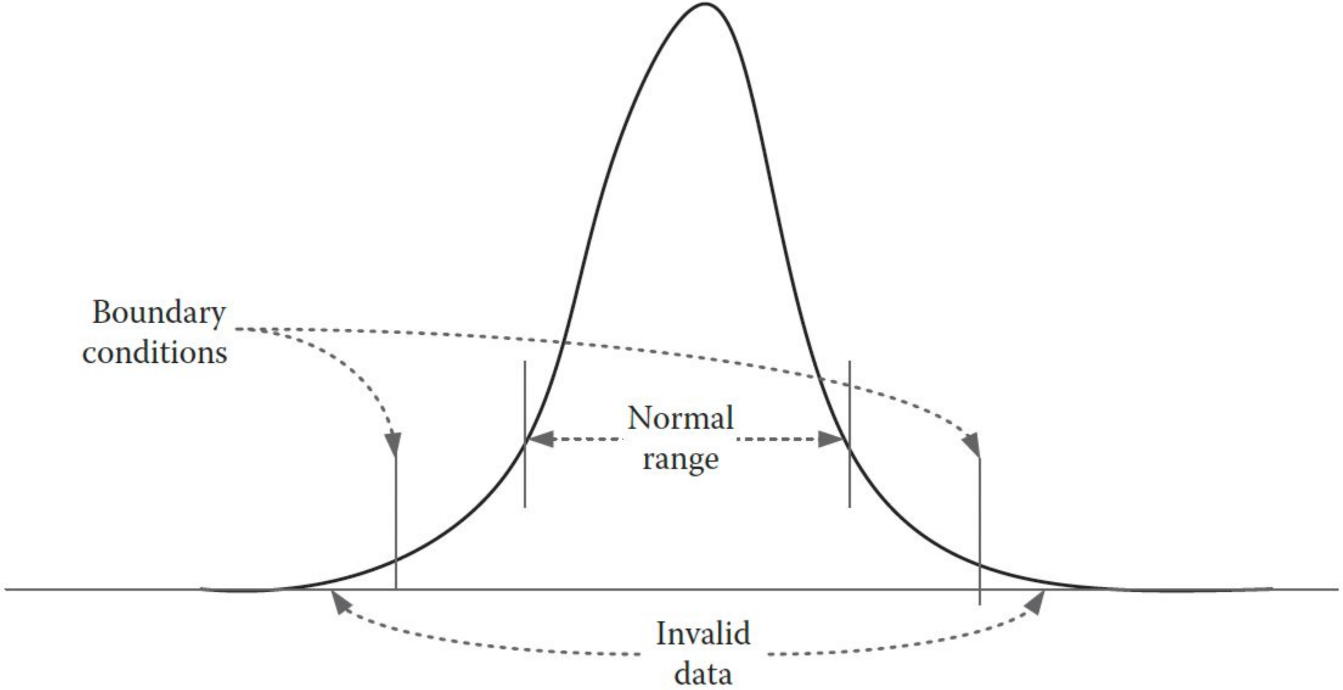
Anonymization of Test Data

Explicit Identifiers (EI)	Quasi-Identifiers (QI)	Sensitive Data (SD)
Substitution	Generalization	Additive random noise
Tokenization (One-way)	Shuffling	*l*-Diversity
Tokenization (Two-way)	k-Anonymization	t-Closeness
National identifier generator	Randomization	Outlier handling
Credit card generator	Blurring	
Mobile number	Data ranging	
E-mail address generator	Suppression	
Scrambling		
Suppression		
Flat value		

Insufficiencies of Anonymized Test Data I

- Anonymization of test data is a good approach to remain compliant with privacy regulation. However, certain circumstances do exist where anonymization is evidenced to be inadequate:
    - **Negative inputs**: System testing requires data satisfying both positive and negative input data to test the expected and alternate results, respectively.
    - **Sensitive domains**: highly sensitive domains are very confidential and rely on data security mechanisms rather than anonymization to protect their data. Access privileges combined with roles govern access to data in such domains.
    - **Regression Testing**: The random failure phase in the bathub curve is when the software system undergoes regression testing. These tests are conducted to test minor functional changes made to the software system periodically. During regression testing, multiple data refresh need to be done. As anonymization depends on production data, this process can turn out to be slow and also involve permissions for access to production data.

Insufficiencies of Anonymized Test Data III


Boundary conditions
Normal range
Invalid data