Chapter 3:

Basics on Data Anonymization in Information Systems

Privacy-Preservation Technologies in Information Systems

Dr. Armin Gerl

WS 2021/2022

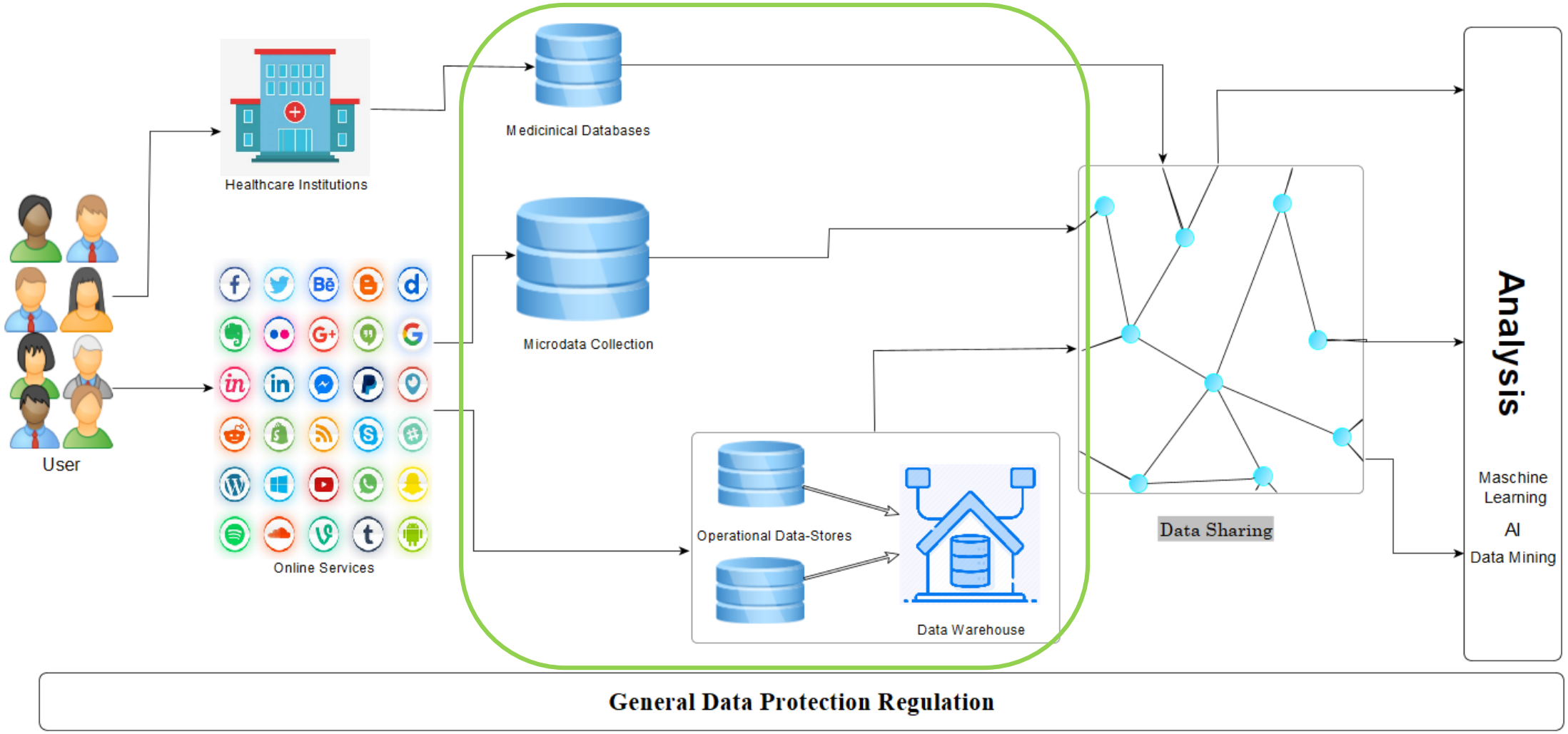
Chapter 3.1:
Information Systems

Privacy-Preservation Technologies in Information Systems

Dr. Armin Gerl

WS 2021/2022

Information Systems



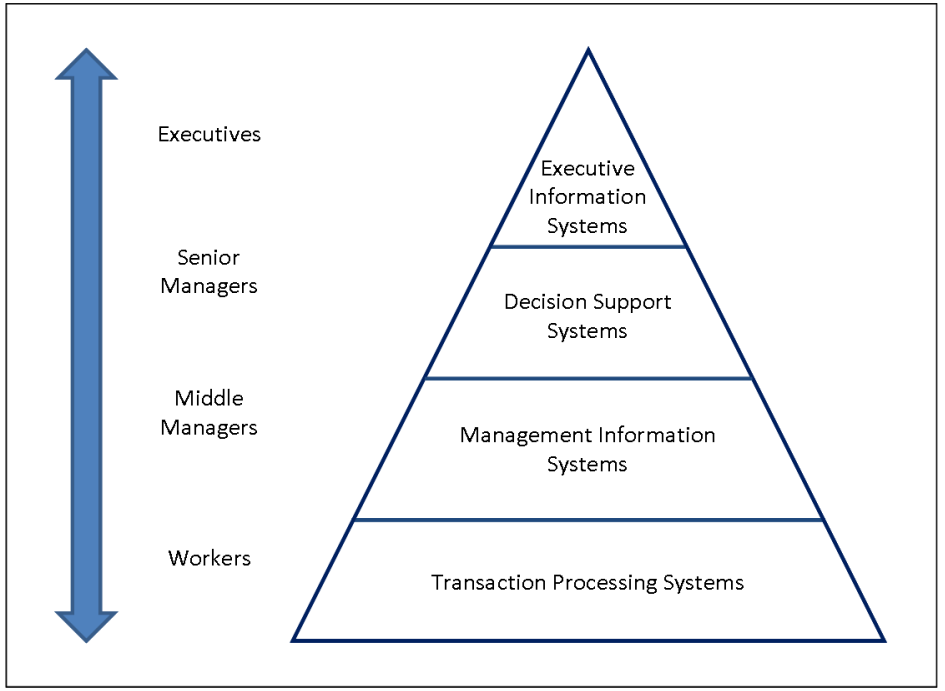
Information Systems

Information Systems (IS) is an academic study of systems with a specific reference to information and the complementary networks of hardware and software that people and organizations use **to collect, filter, process, create and also distribute data**. An emphasis is placed on an information system **having a definitive boundary, users, processors, storage, inputs, outputs and the aforementioned communication networks**.

The basic components of a Information System (IS) are:

- Hardware
- Software
- Databases
- Networks
- Procedures

- Example IS: Data Warehouse, Search Engines, Decision Support Systems, Expert Systems, etc.

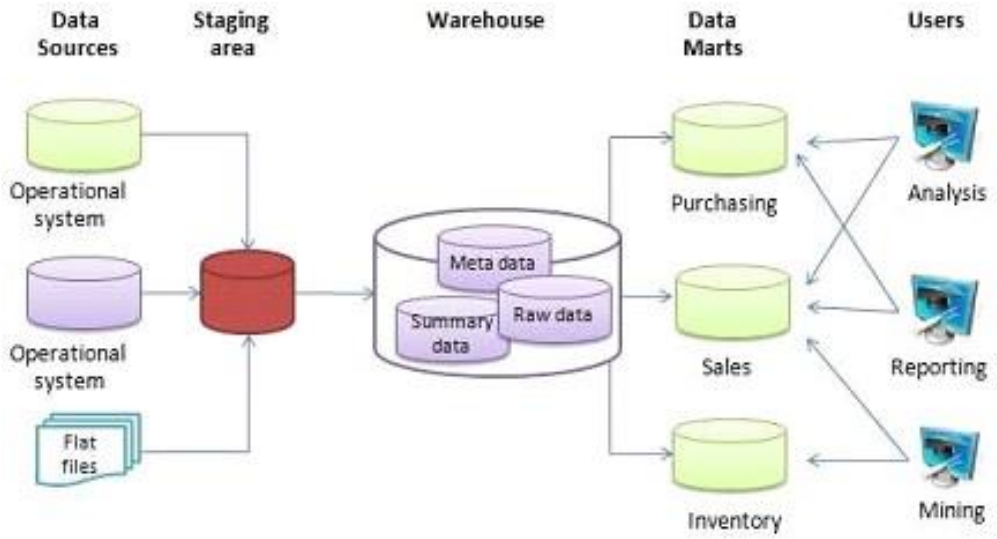


Data Warehouse as an example IS

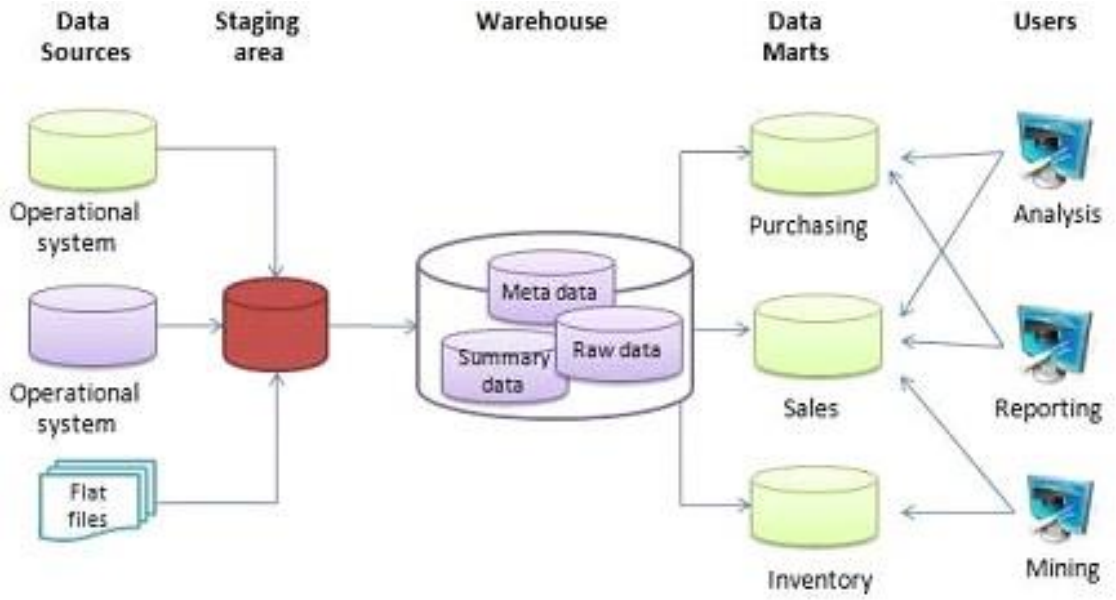
Data Warehouse (DW) is a system used for reporting and data analysis. It is the considered as a core component for business intelligence. DW are central repositories of integrated data from disparated sources.

Properties of a DW:

- **Subject-oriented:** Unlike the operational systems, e.g. databases, the data in the data warehouse revolves around specific subjects/topics. No normalization of schema.
- **Integrated:** The data in DW comes from several operational systems. Data is processed to remove inconsistences.
- **Time-variant:** DW data represents current and historical data (a long time horizon up to 10 years).
- **Nonvolatile:** Data in DW is optimized read-only, which means it should not be updated, created, or deleted.



Data Warehouse as an example IS



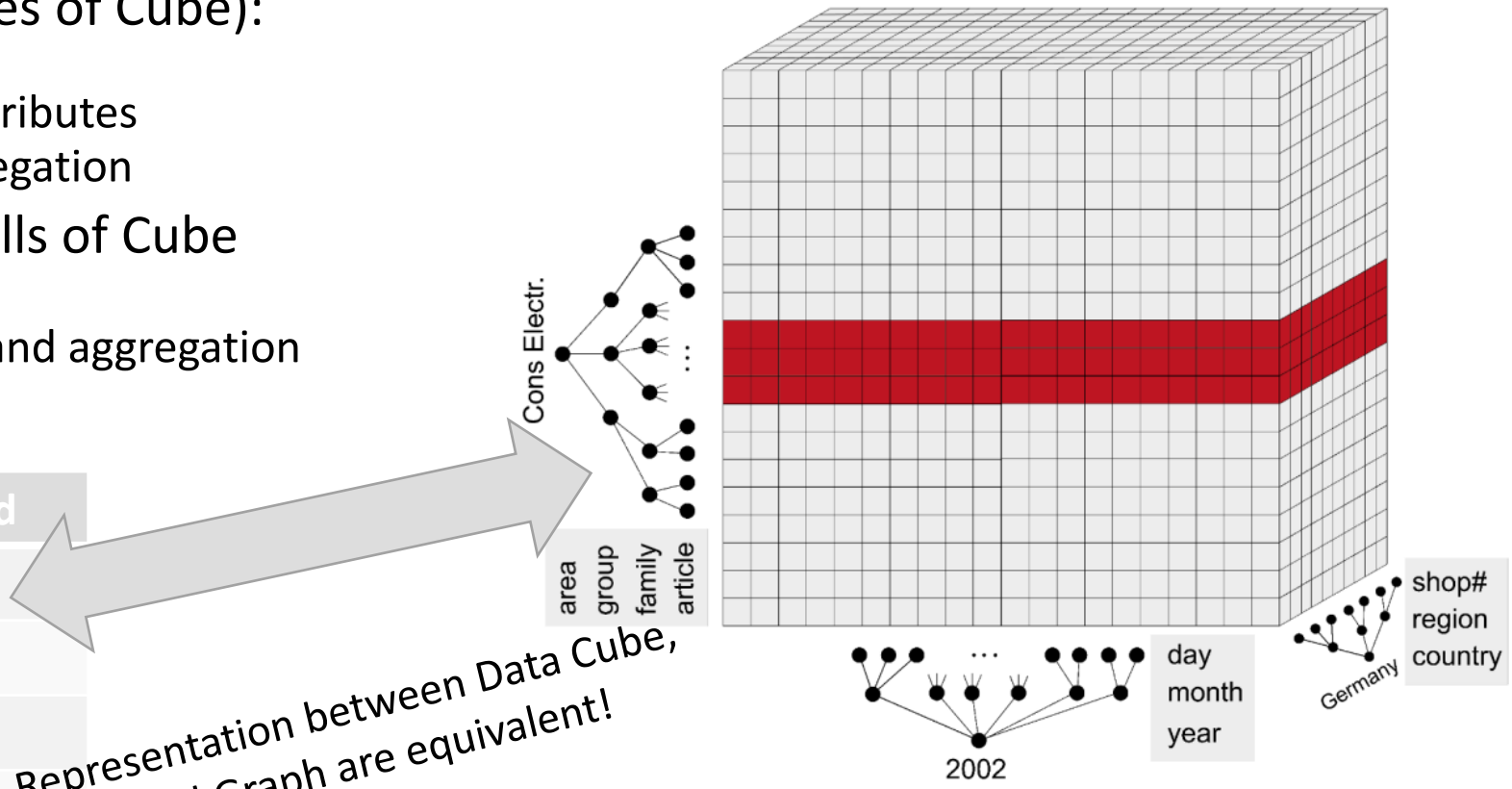
Question when to anonymize in DW Life-Cycle will be answered in later chapter!

- "Raw" Data from various Data Sources is integrated in DW using ETL process on a regular basis, e.g., daily or weekly
- DW contains Data in non-normalized form, typically a Data Cube
- (Optional) DW derivate data is stored in specialized Data Marts (DM)
- Users use DW or DM to perform strategic tasks (Analysis, Reporting, Mining, etc.)

Data Model in Data Warehouse

- Qualitative Information (Edges of Cube): Dimensions
  - Hierarchies, Dimensional Attributes
  - Used for Selection and Aggregation
- Quantitative Information: Cells of Cube
  - Facts (Measures)
  - Information to be analysed and aggregation

Sales		Electronics	Food
Germany	2019		
	2020		
	2021		
France	2019		
	2020		



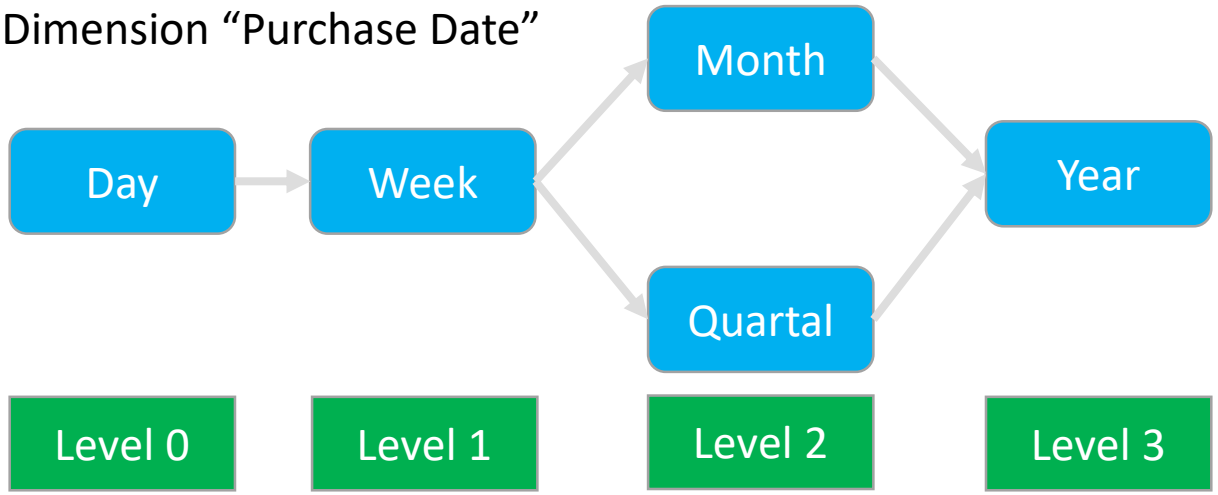
Representation between Data Cube, Table and Graph are equivalent!

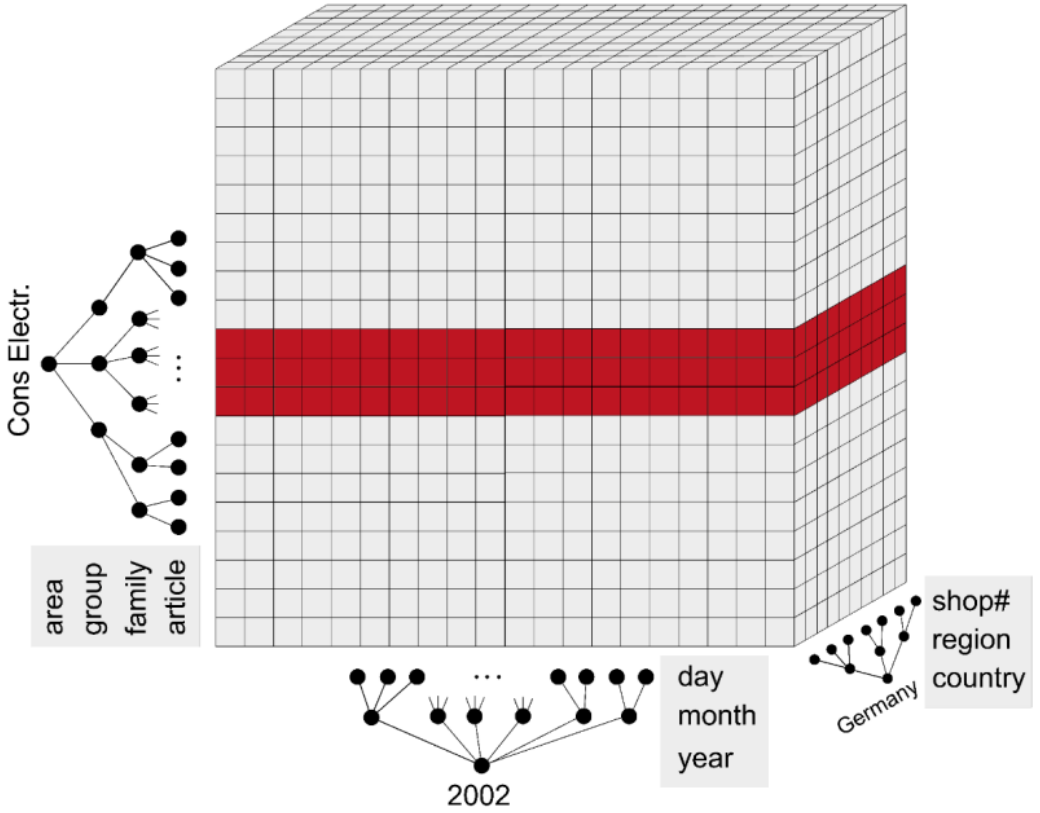
Data Hierarchies

- Dimensions (Qualitative Information) have typically various Levels of Hierarchy
- Required in DW for different levels of Analysis
- The lower the Hierachy Level the more detailed the information (High Information Value)
- The higher the Hierarchy Level the less detailed the information (Low Information Value)
- Hierarchies are not always unique

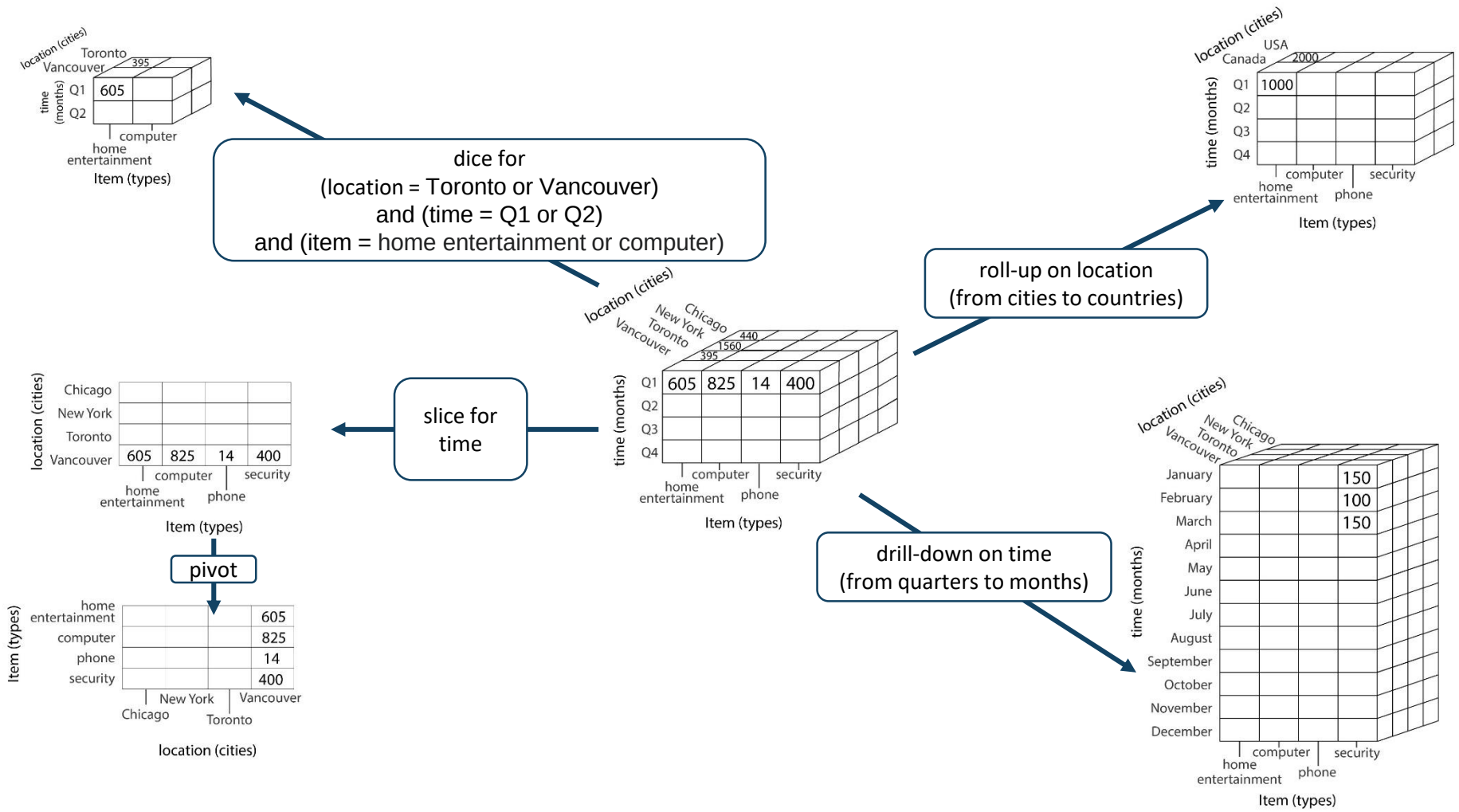
➢ Data Hierarchies are also used in Anonymization, i.e. Generalization

Dimension “Purchase Date”





Operations in Data Warehouse



Chapter 3.2:
Basics on Personal Data

Privacy-Preservation Technologies in Information Systems

Dr. Armin Gerl

WS 2021/2022

Personal Data (GDPR)

- **Data Subject = Natural Person**:
  - Identifiable natural person is one who can be identified **directly or indirectly**
  - Identification via a reference to an identifier, e.g. name, id number, location data, online identifier, identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person

- Personal Data:
  - **Any information** relating to an identified or identifiable natural person
  - Very broad and general definition, but also more specific definitions:
    - **Genetic Data**: personal data relating to the inherited or acquired genetic characteristics
    - **Biometric Data**: personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics
    - **Data concerning Health**: personal data related to the physical or mental health

Attribute Classifications

Explicit identifier (EI)

Directly lead to re-identification

Social security number

Quasi-identifier (QI)

A combination of them can lead to re-identification

Address or birth date

Sensitive attributes (SA)

Don‘t directly lead to re-identification but need to be protected from being matched

Illnesses or income

Non-sensitive attributes (NSA)

Every attribute that doesn‘t fit into the other categories

Weather data

Explicit Identifier (EI-)Attributes

- Attributes that clearly identify the entity to which the record refers to, because they are used as a primary key for the entity in an official context (SSN) or are sufficiently rare to assume a connection with 99% certainty
- Dependent on the specific data set but easy to notice, examples are:
  - Social security number
  - Tax identification number
  - Passport number
  - ISBN (books)
- If present in a data set, all other attributes are immediately linked to a specific entity
- Therefor it is common practice to remove all identifier attributes from a set before performing further anonymisation
- The act of linking an (unsuccessfully) anonymized record to the entity it refers to is called “re-identification”

Country	Name	Surname	Sex
USA	John	Smith	m
USA	William	Johnson	m
Germany	Peter	Müller	m
USA	James	Davis	m
Germany	Anna	Jäger	f

Neither Surname nor Name are EIs, even both columns together wouldn‘t justify the classification since the names are very common and a match just from a single value is impossible

Quasi-identifiers

- When an attribute doesn‘t immediately lead to re-identification but needs to be combined with others to do so, these attributes are classified as quasi-identifying
- No easy way to determine which combination of attributes might be compromising and usually no way to preventively remove them because of the information loss that results
- Classification depends strongly on assumed knowledge of attackers and publicly available data that could be used

Combined quasi-identifiers lead to re-identification

ZIP Code	Sex	Age
85535	m	75
85535	f	42
60629	f	69
60629	m	19
85535	f	33

QIs are context dependent e.g.:
Eden, Arizona (85535): approx. 20 people
Metropolitan area, Chicago (60629): >100000 people

- While “quasi-identifier” or QI classifies an entire column X^j, the term “quasi-identifier groups” or QI* is used for attribute value combinations that yield re-identification risk for single or few records
- According to a 2000 paper by L. Sweeney [1], 87% of the US population could be identified by combining birth date, ZIP code and sex
- Good classification requires knowledge about the subject and experience

[1] “Simple Demographics Often Identify People Uniquely” Latanya Sweeney: https://dataprivacylab.org/projects/identifiability/paper1.pdf

Sensitive Attributes (SA)

- Also refered to as confidential attributes in some sources
- These attributes hold private information about the respondent (e.g. sexual orientation, salary, medical conditions and illnesses etc.)
- Preventing attackers from linking these attributes with the respondent is the primary goal of anonymisation techniques for microdata
- Secondary goals also include averting the possibility of setting bounds or inferences to such attributes
- Sensitive attributes are often part of the subject of the data release, therefore removal is not always feasible

Name	Complaint	Appointment	Diagnose
H. Dampf	Cough	22.03.2021	Common Cold
J. Doe	Cramp	22.03.2021	Leg Spasm
P. Müller	Weakness	23.03.2021	HIV+
Pumuckel	Fear of drawers	24.03.2021	Klabautermann

An exemplary log of doctors appointments containing sensitive information

Non-sensitive Attributes (NSA)

- All information that doesn‘t directly or indirectly imply a connection to entity the record belongs to
- Attributes which are neither EI, QI or don‘t hold any sensitive information are called non-sensitive attributes
- Also referred to as „non-confidential“ in some sources
- Air pressure would be an example but in combination with additional weather data and timestamps, it could still lead to a diclosure of location
- The context is important, even harmless things such as room temperature could allow inference of information

Date	Day	Temp	Weather
19-Oct-21	Sat	12°C	Sunny
20-Oct-21	Sun	6°C	Sun/cloud
21-Oct-21	Mon	3°C	Sleet
22-Oct-21	Tue	10°C	Rainy
23-Oct-21	Thu	16°C	Sun/rain

A weather report containing no sensitive information

Attribute Classifications and the GDPR

- These classifications form the basis for anonymization and privacy models
- In the context of the GDPR, information held by attributes from all of the afore mentioned categories pose a problem, as long as the information can be connected with an individual person
- For our classifications, while not explicitly mentioned, the GDPR states, that the above connection must neither occur:
  - “directly” -> explicit identifiers
  - or “indirectly” -> quasi-identifiers
- The regulations imposed by the GDPR on data processing loosen, when direct or indirect re-identification is prevented

Attribute/Data Types

Categorical

Blood Type	State
A	Iowa
B	Kansas
AB	Ohio
O	Texas
A	Utah

Ordinal

Grade	Rank
A	Private
B	Corporal
C	Sergeant
D	Colonel
E	General

Discrete/Numerical

Inhabitants	Votes
1520123	520417
15120	5125
125600	3333
12769	9214
420	420

Continuous

Weight	Time
58,859kg	16:43.16,50
72,727kg	23:59.59,98
95,678kg	00:00.00,01
80,001kg	16:20.00,00
125,286kg	12:34.56,78

- These classifications are not exclusive e.g.:
  - The military rank example is also categorical
  - Any numerical value is orderable by size and therefore also ordinal
  - Continuous attribute values are almost always represented by real numbers and consequently also numerical
- Continuous data that is (digitally) stored is always truncated or rounded but an infinite range of values is possible between any 2 data points
- Finer subdivisions for statistical analysis (binary, binomial, count, real-valued additive/multiplicative)

Further Attribute Classifications

- Categorical vs. continuous attributes:
  - If a sufficiently large or infinite range of values is possible, the attribute's domain is continuous

    (e.g. Temperature, height, speed,…)
  - Otherwise, a limited domain of values belongs to a categorical (or discrete) attribute

    (e.g. Gender, State, Date,…)
- Diverging Terminology: In some sources, Explicit identifiers are called "direct identifiers", while QIs along with SAs are summarized as "indirect identifiers"

Definition of Microdata

- Microdata refers to information about a specific entity, besides a private citizen, this can also be a company (Macrodata in contrast refers to aggregated or grouped data)
- To serve data regulations in place, microdata related to private citizens (personal data) needs to be anonymized before being published
- A microdata set X is a tabular collection of information about n respondents and m attributes represented as a matrix
- The value of the j-th attribute of the i-th record is referred to as x_i^j (e.g. “Pumuckel is not vaccinated” $\equiv$ "$x_5^7 = n$“)

Name	Age	Blood Type	Weight	Body Fat	Antibodies	Vaccinated
J. Doe	72	A	84kg	10%	y	y
H. Dampf	23	O	100kg	23%	n	y
A. Smith	27	B	64kg	5%	y	y
P. Müller	58	AB	140kg	31%	y	n
Pumuckel	17	O	56kg	8%	n	n

An (imagined) microdata example *X* containing personal data collected to study the effects of some vaccine

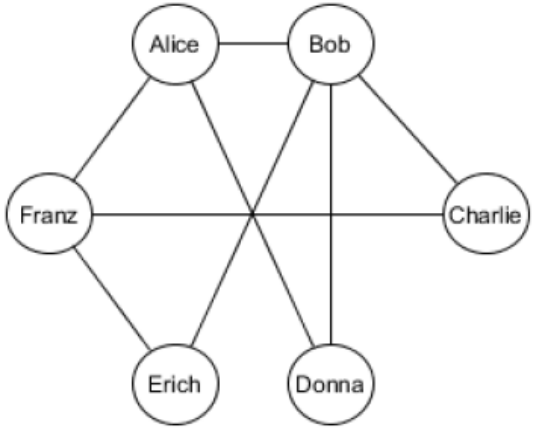
Additional Data(-set) Types

Multidimensional Data
(or Rational Data)
Each Record = Row in Data Table

Longitudinal Data
Correlated; Clustered; Temporal Order
e.g. repeated measurements of patient

Name	*Date*	*Time*	*Diabetes*	*Blood Glucose mg/dL*
Bob	01.03.2019	10:00	Type 1	110
Bob	01.03.2019	14:00	Type 1	95
Bob	01.03.2019	19:00	Type 1	130

Graph Data
Set of vertices and a set of edges
Denoting relationships



Transaction Data
DB holding Transactions of Users
High dimensionality and sparsity of data

Name	*Baguette*	*Croissant*	*Cheese*	Coffee	*Wine*
Alice	1	0	1	0	0
Bob	0	0	0	0	1
Charlie	0	1	0	1	0

Time Series Data
Temporal Order; High Dimensionality;
Semantic Relationship; Single Variable

ID	*Name*	*Postal-Code*	*Salary 2016*	*Salary 2017*	*Salary 2018*	*Salary 2019*
1	Alice	94032	25.000	27.000	29.000	30.000
2	Bob	94036	30.000	30.000	35.000	35.000
3	Charlie	94405	26.000	31.000	31.000	28.000

Chapter 3.3:
Basics on Anonymization

Privacy-Preservation Technologies in Information Systems

Dr. Armin Gerl

WS 2021/2022

Anonymization

- The goal of anonymization is the prevention of the leakage of the identity of a user based upon personal data
  - In other words: personal data of a user should not reveal the identity of the user
- Processing of personal data is required in many domains, therefore multiple anonymization methods have been developed that intend to preserve the anonymity of the user while the processing of personal data remains possible
- The anonymisation of data is performed with a wide variety of different statistical tools, each with their own strength and weaknesses
- Our running Example:

EI		QI			SD	NSD
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
2	Bob	M	33	94036	35.000	1337
3	Charlie	M	29	94405	28.000	404

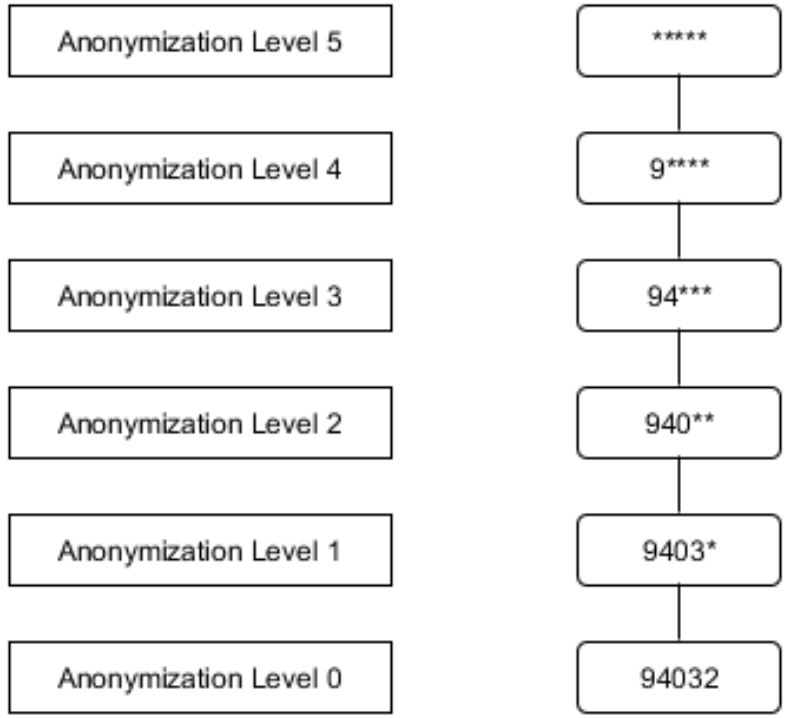
Anonymizing the Data

- We will discuss the technical aspects in more detail later, but most anonymisation techniques make use of the following 3:

Suppression	Generalization	Deletion
• Values are being replaced by an replacement character, e.g., asterix (*)	• Using Hierarchies to cluster attribute values in similar groups	• Delete singular entries, that carry high re-identification risk or would require inappropriate effort to anonymize

Suppression

- Suppression is replacement of parts of the original value with replacement characters
- **Replacement Character** and **Replacement Strategy** have to be carefully chosen to preserve the semantics of the original value in its specific domain
- **Replacement Character**
  - A character that does not interfere with the original **semantic** of the attribute
  - E.g., german postal-codes -> no numerical replacement character
- **Replacement Strategy**
  - Consider **syntax** (structure and format) of the original value
  - E.g., for a date in the german date format dd.mm.yyyy
  - first replace the days dd, month mm, last year beginning from the last character yyy**y**



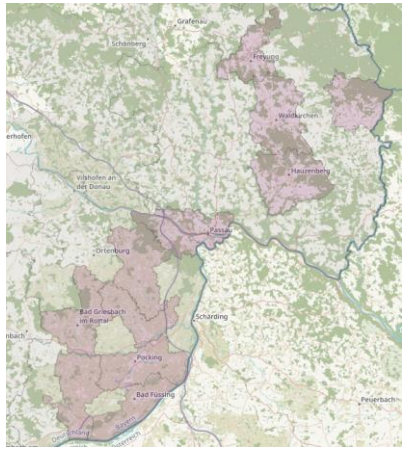
Suppression

- Use Running Example
- Suppression on QI Attribute Postal-Code for “Bob”
  - Replacement Character “*”
  - Replacement Strategy: Last to First
    - Reason: Addressed Region of Postal-Core increases
      - ➢ sequential reduction of information
- Result: Suppression of Value “94036” to “940**”
  - Anonymization Level: 2 (arbitrary chosen)

EI		QI			SD	NSD
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
2	Bob	M	33	94036	35.000	1337
3	Charlie	M	29	94405	28.000	404

EI		**QI**			**SD**	**NSD**
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
2	Bob	M	33	**940****	35.000	1337
3	Charlie	M	29	94405	28.000	404

FYI: Region addressed by Postal Code 940**

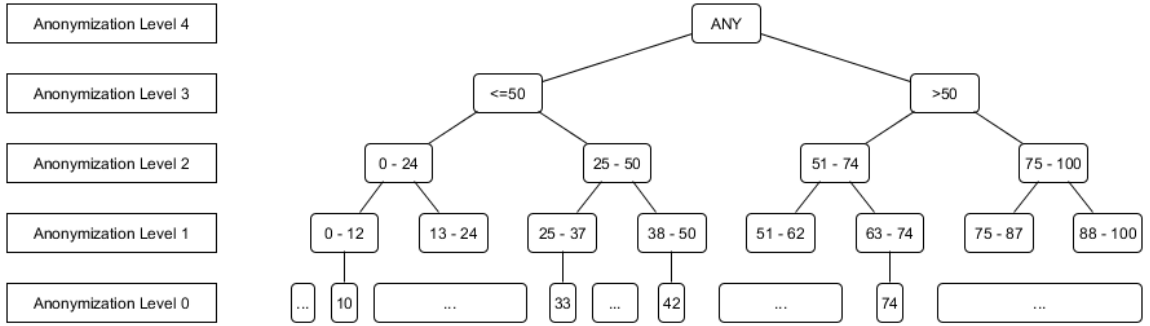


Generalization

- Generalization is the replacement of the original value with another value of a taxonomy denoting a more general description of the value
- Creation of the taxonomy (hierarchy) is essential
  - ➢ Preserve the **semantic** of the value!

➢ Hierarchy Creation:

- In general: Choose hierarchy in a way, such that the resulting values are still useful for intended usage, e.g., analysis, reports, etc.
- "Utility": Quality or usefulness of the data
- Example for **nominal** attribute "sex": define group terms, e.g., simplified "any" value
- Example for **cardinal** attribute "age": use intervals of ages

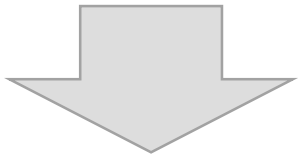


Generalization

- Use Running Example
- Use Generalization for the QI Attributes Sex and Age of Bob
- Use Taxonomies/Hierarchies as detailed before
- Generalize “sex” to Level 0: M -> ANY
- Generalize “age” to Level 1: 33 -> 25-37

- Comparison :
  - **Generalization**: Taxonomies/Hierarchies Values have to be defined beforehand
  - **Suppression**: Hierarchy Values can be calculated on-the-fly (given the Replacement Character and Strategy are known)
- Suppression and Generalization hide information in Attributes (commonly QI) that influence the **Utility**

EI		**QI**			**SD**	**NSD**
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
2	Bob	M	33	**940****	35.000	1337
3	Charlie	M	29	94405	28.000	404



EI		**QI**			**SD**	**NSD**
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
2	Bob	**ANY**	**25 - 37**	940**	35.000	1337
3	Charlie	M	29	94405	28.000	404

Generalization

- Depending on the data type of the attribute, hierarchies are formed differently during the generalization process
- Examples would be:
  - Clustering postal codes into regions (e.g. 60000-99999 into Southern Germany)
  - Clustering ages into intervals (e.g. 0-6, 7-14, 15-18, 19-25,…)
- It is worth mentioning, that an attributes data type might change during the generalization process, as was the case in the postal code example
- Within a data warehouse, similarities between the qualifying data (e.g. dimensional tables), that structure the quantifiable data, and the clustering into hierarchies during generalization exist
- This allows for easier anonymisation inside a data warehouse, by creating the hierarchies based on already clustered data
- This is easier to perform on simple rather than complex data types like multimedia

Deletion

- Deletion is the extreme version of both Suppression and Generalization, leaving no trace of the original value.
- Easy and straight-forward way to anonymize personal data is to delete it
  - Remove any information from the data value that can identify an individual person
  - No semantic information is preserved (no remaining utility)

$$\text{Deletion}(\text{Value}) \longrightarrow \varnothing$$

- Deletion is important for EI
  - Generalization or Suppression on EI might not be sufficient, but for QI or SD
  - Use Deletion for Attributes explicitly identifies a user

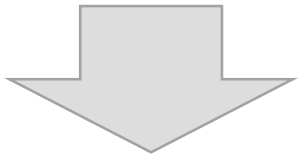
Note: Deletion can go wrong!
The “Deletion Value” should not be distinguishable from the default or empty value, otherwise the information can be derived that some value has been there before
Imagine Health Data of a VIP: You don’t want to disclose that this person had Cancer or other sensitive Health Issue

Deletion

- Use Running Example
- Use Deletion for the anonymization of the ID and Name of Bob
- Replace the Value of "ID" with ""
- Replace the Value of Name with ""

EI		QI			SD	NSD
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
2	Bob	**ANY**	**25 - 37**	940**	35.000	1337
3	Charlie	M	29	94405	28.000	404

- Note: An attacker could derive information about the data-set composition
  - E.g., High Probability of Sequential ID
- Note: attacker could know that Bob is in data-set and is over 30 years old
  - ➢ Salary can be derived with external knowledge



EI		QI			SD	NSD
ID	*Name*	*Sex*	*Age*	*Postal-Code*	*Salary*	*Lucky#*
1	Alice	F	27	94032	30.000	1234
		ANY	25 - 37	940**	35.000	1337
3	Charlie	M	29	94405	28.000	404

- Anonymization should be applied to whole data-set and not only localized record/attributes

Basic Anonymization Methods

Anonymization Method	Target	Strategy
Suppression	EI, QI, SD	Replacement Strategy/Character
Generalization	QI, SD	Generalization Hierarchy
Deletion	EI	Value Removal

Microdata Anonymisation

- To avoid disclosure, a modified data set *Y* is released instead of the original set *X*
- *Y* is refered to as the sanitized, anonymized or protected version of *X*
- The first step in almost any anonymisation effort is the removal of Identifier Attributes
- There is a trade-off between privacy and informational value of the released data => publishers of data sets and statistics cannot retain maximum informational value of their release and simultaneously ensure minimum disclosure risk when applying anonymisation techniques
- Anonymisation of microdata is done by masking the original data using a wide range of methods, some of which are introduced next

Defining the State of anonymised Data

- We have now seen various methods and transformations on microdata that can be used to anonymise microdata
- Each method has its own strengths and weaknesses depending on the types of data they are used on
- These anonymisation techniques are the tools used to achieve certain privacy goals that an anonymised data set has to meet
- In order to messure the effectiveness of the performed transformations on a data set with respect to privacy, researchers developed a wide variety of models to solidify these goals
- The most common ones are based on the attribute classifications (EI, QI, SA, NSA) that were mentioned throughout the lecture, even though other approaches exist

External Matching

- Problem: Even if a given data set has been anonymized, so that no singular individual can be re-identified, matching with background information might still lead to re-identification
- Publically available census data or publications from other sources are good candidates for conducting an external matching attempt
- For illustration purposes, lets transform the following secret data set:

Surname	Name	1st Treatment	2nd Treatment	Diagnose
Thomas	Meier	01.10.2020	13.10.2020	Lung cancer
Beate	Wimmer	01.10.2020	09.10.2020	Pelvic fractur
Maximilian	Huber	05.10.2020	07.10.2020	HIV

External Matching (cont.)

- Lets say that the hypothetical anonymisation process replaces the Names with codes and censors the Diagnoses, resulting in the sanitized version of the set, that is then published

Surname	Name	1st Treatment	2nd Treatment	Diagnose
Akdmclkm	Indncdc	01.10.2020	13.10.2020	*
Bkjdnckjn	Dgdhcd	02.10.2020	09.10.2020	*
Kodcndcd	Ldcdoo	05.10.2020	07.10.2020	*

- In addition to the sanitized set, an attacker also got the following information on the opening hours of treatment centres from public sources:

Monday	Tuesday	Wednesday	Thursday	Friday
HIV-Consulting	Irradiation	HIV-Consulting	Irradiation	Surgery

- For a potential employer, it would be easy to infer Mr. Hubers HIV-diagnose, if he had to cancel his invitation to a job interview on the 5th and 7th

Analysing the Data Set

- This shows that it isn‘t always sufficient to only look at the data that is to be protected
- **Recombination attacks**, like the one we just illustrated, also need to be taken into account by privacy models
- In order to better distinguish the risks the attributes of tabular data pose to the individuals they represent, many models are based on a classification into 4 subject-dependent categories
- These classifications aren‘t absolute, the same column in different data sets aren‘t always sorted into the same categories -> context is important

Attack Models and Disclosure Risks

- Trade-off between re-identification risk and informational value of the released data set when classifying attributes => Modelling of outside knowledge/ressources available to attacker
- Publicly available or easily obtainable information often used as basis for model, since complete statistical evaluation either impossible or disproportionally complex
  => Attribute classification mostly based on assumptions of background information
- When releasing a data set, 2 types of privacy disclosure can occur:

Identity disclosure:
Def.: If a record in the protected data set can be linked with the identity of the respondent it belongs to, identity disclosure takes places

Attribute disclosure:
Def.: Attribute disclosure takes place when an attribute of an individual or entity can be determined more accurately with access to the data set then otherwise possible (sometimes refered to as „inferential disclosure“)

- Identity disclosure violates the anonymity, while someones privacy is violated if attribute disclosure occurs

Independency of Disclosure Risks

- Attribute and identity disclosure may occur independant from one another
- An attacker may be able to learn important information about his target without disclosing his identity
- On the other hand, even if a record in the data set is successfully linked to an individual, attribute disclosure does not automatically arise if the confidential attributes have been properly masked
- The example shows what information an attacker obtains about his target without identifying it (attribute disclosure without identity disclosure)

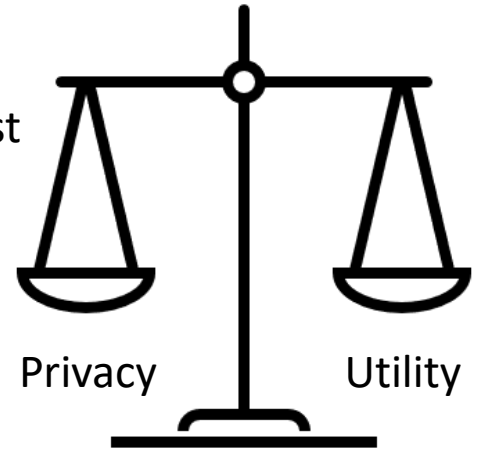
Department	Position	Salary	Sex
Sales	Accountant	80.000	m
Inventory	Accountant	70.000	f
Costumer Service	Manager	120.000	f
Costumer Service	Team Leader	70.000	m
IT	Manager	160.000	m
Sales	Manager	200.000	f
IT	Team Leader	100.000	m

Prior knowledge: Job position = manager, target is part of set
=> Target doesn‘t work in Inventory & Salary ϵ [120.000, 200.000]

- With identity disclosure being the more serious of the two, measures to determine this risk before releasing will be discussed at the end of the chapter

Utility Requirements

- There is an inherent trade-off between anonymity and utility of a data set
- A hypothetical data set with the highest achievable anonymity would then contain almost no information
- Everyone has the right to privacy, but:
- Since there are no reasons to store a data set without information, maximum anonymisation isn't a good thing either
- For data collectors, a balance between costs and risks is the name of the game
- Practical implementations of the privacy models we've seen so far need to:
  - Produce a data set that satisfies the model
  - Limit the usefulness of the data as little as possible at the same time
- Measurements for data privacy are as important as measurements for its utility



Measuring Utility

- Data utility is strongly connected to the subject of the data set
- Our approach to measuring it will be a mathematical and objective one
- The following illustrations show principles, that are applicable to singular attributes (e.g. postal code, illness or sex)
- It is important to differentiate between categorical and numerical attributes
- When measuring the whole data set, numerical attributes need to be aggregated and sometimes attached with weights
- At first, we will take a look at generally applicable methods, as many measurements are taylored to specific use-cases

Data Protection Challenges

- Technical progress provides the means to look deeper into the private lives of citizens than ever before
- Even anonymized data sets can be mined for useful information with advanced analysis tools in data warehouses
- Actors in the data business are primarily objected to obey the legal restriction imposed upon them
- Moral concerns are only starting to be acknowledged in recent years due to public backlash after data protection scandals and leaks
- A framework suitable for examining large and complex data sets on privacy issues would solve many problems for data processors and sources alike
- The academic field that tries to tackle this problem is still expanding
- A variety of privacy models for different areas of applicability already exist, each with their own benefits and limitations

3.3
Summary

Recap of Chapter

- Introduction of Information Systems, with Data Warehouse as a example
- Notion of Personal Data (Computer Scientist vs. GDPR)
- Anonymization Methods on Attributes/Values
- Attack Models and Disclosure Risks on Data-Sets
- Data Protection Challenges

- Next Chapter: Deeper Look on Privacy Models

Overview of Lecture Topics

We are here

Chapter	Est. Extent
Chapter 1: Introduction	~1 Lecture
Chapter 2: From GDPR to Privacy Languages	~3 Lecture
Chapter 3: Basics on Data Anonymization in IS	~1 Lecture
Chapter 4: Privacy Risks and Anonymization Techniques	~4 Lectures
Chapter 5: Privacy in Health-Care	~2 Lectures
Chapter 6: Privacy in Data Warehouses	~2 Lectures
Chapter 7: Privacy in Social Networks	~2 Lectures
Exam Preparation Lecture	1 Lecture