Identity-Reduction: The Technical Perspective

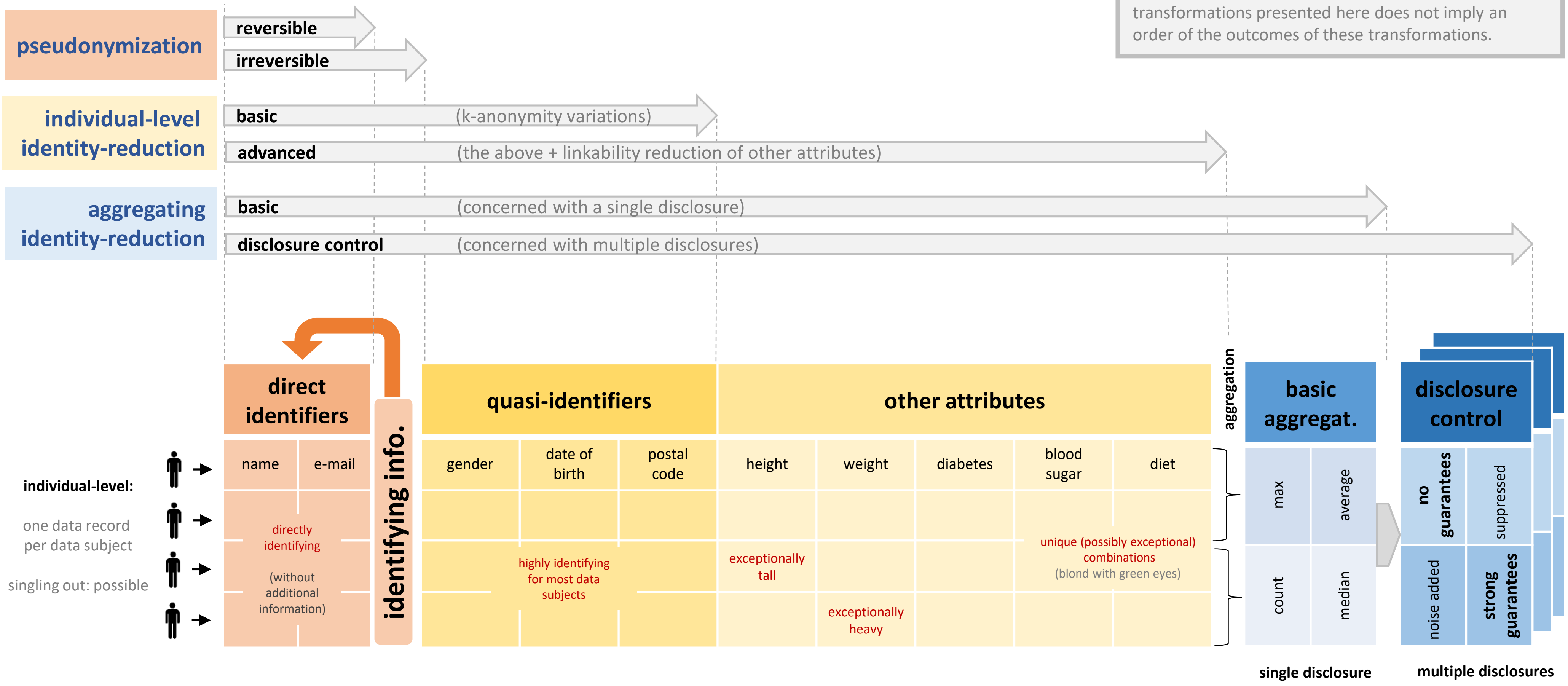
1 The Scope of Identity-Reduction Transformations

Disclaimer:
This taxonomy cannot attempt to answer the question of when data can be considered to be anonymous.

This depends on the data, on the parameters of the transformations, on the available additional information, the state of the art of re-identification, the motivation and resources of possible attackers, …

The outcome of applying a given transformation type can therefore vary widely. Thus, the order of transformations presented here does not imply an order of the outcomes of these transformations.

Transformation



pseudonymization
- **reversible**
- **irreversible**

individual-level identity-reduction
- **basic** (k-anonymity variations)
- **advanced** (the above + linkability reduction of other attributes)

aggregating identity-reduction
- **basic** (concerned with a single disclosure)
- **disclosure control** (concerned with multiple disclosures)

individual-level:

one data record per data subject

singling out: possible

ULD for AnoMed Funded by Funded by the European Union NextGenerationEU

February 2024, Version 0.9.2

Feedback to research@datenschutzzentrum.de

Identity-Reduction: The Technical Perspective

2 A Taxonomy of Identity-Reduction Transformations

Identity Reduction Type		Transformation of Data Elements	Re-Identification Attacks	Possible Outcomes
data pseudonymization	**reversible**	**Direct identifiers** are **eliminated** or **transformed** (but identifying information is kept)	• Spontaneous recognition • Linkage on: • Inversion secret • quasi-identifiers • Unique combinations of other attributes (indiv.-level: singling out is trivial)	*Pseudonymous Data*
	irreversible	In addition: **Identifying information is eliminated**	Same as above, minus: linkage on inversion secret (indiv.-level: singling out is trivial)	*Pseudonymous Data*
individual-level identity-reduction (aka. ***record-level***, ***micro data***)	**basic**	In addition: **Quasi-identifiers** are transformed such that for each possible tuple of quasi-identifiers, there are **at least K-1 tuples with undistinguishable values** • Distinction is based on equality or similarity (depending on variance of the quasi-identifiers) • Transformations include generalization and suppression	Same as above, minus: Linkage on quasi-identifiers (indiv.-level: singling out is trivial)	*Advanced Pseudonymous Data* *Supposedly Anonymous Data*
	advanced	In addition: **Other attributes** are transformed **to protect against linkage** • Transformations include generalization, suppression, top- and bottom-coding, slicing, data swapping, and noise injection	Same as above, but: Spontaneous Recognition and linkage on other attributes is rendered more difficult or impossible (indiv.-level: singling out is trivial)	*Advanced Pseudonymous Data* *Supposedly Anonymous Data* *Successfully Anonymous Data*
aggregating identity-reduction	**basic**	**For a single disclosure, all individual-level data** is transformed such that **the resulting values relate to groups of at least C persons**	Singling out (followed by linking) possible by inference over multiple disclosures. (reconstruction attacks [↪])	*Advanced Pseudonymous Data* *Supposedly Anonymous Data* *Successfully Anonymous Data*
	disclosure control see Art. 2(4) Commission Regulation 557/2013	In addition: The aggregate values **are further protected** against known or even arbitrary singling out attacks **across multiple disclosures**.	Singling out over multiple disclosures is rendered difficult or impossible.	*Supposedly Anonymous Data* *Successfully Anonymous Data*

ULD for AnoMed Funded by Funded by the European Union NextGenerationEU

February 2024, Version 0.9.2

Feedback to research@datenschutzzentrum.de

Identity-Reduction: The Technical Perspective

3 Categories of Data

Possible Outcomes of Identity-Reduction Transformations

Disclaimer:
The data category cannot be determined from the data alone.

While there are indicators for data being personal, no technical test exists that guarantees anonymity. Data categories are therefore the result of a risk assessment which takes factors beyond just the data into account.

Data Category	Possibilities of (Re-)Identification
Fully Identified Personal Data	• **direct identification** is possible (since data is unchanged)
(Basic) ***Pseudonymous Data*** *personal data* (Recital 26 GDPR)	• direct identification is no longer possible • **only indirect identification** using **additional information** is possible
Advanced Pseudonymous Data *likely still personal data*	• direct identification is no longer possible • **even indirect identification** is rendered **difficult** or **prevented** (but with unknown success)
Supposedly Anonymous Data *likely anonymous* *but future re-identification cannot be excluded*	• **all relevant known re-identification attacks are excluded** • **thorough assessment of re-identification risk** results in low risk
Successfully Anonymous Data certainly anonymous *future practical re-identification can be excluded*	• **re-identification can be practically**[1] **excluded** • strong guarantees or thorough assessment of re-identification risk

[1] *practically* here means considering any party who can reasonably likely gain access to the data, its reasonably likely means, and taking into account technological developments.

Identity-Reduction: The Technical Perspective

4 Glossary

Direct Identifier: A direct identifier is a value or value combination that is commonly known to be related to a given natural person or where a known procedure of limited effort can be used to establish such a relation. Direct Identifiers are often unique in a given context. Examples include a person's name, address, phone number, coordinates of residence, etc.

Relation to a natural person: A value is related to a natural person if, with a significant likelihood, the person has (positive relation) or has not (negative relation) a certain property described by that value.

Quasi-Identifier: A quasi-identifier is a value that is expected to be known about a natural person or easy to find out. Combinations of quasi-identifiers are often unique for a majority of persons. Examples include age, gender, and place of birth.

Singling Out: Singling out is a processing step executed on a data set that, for at least one data subject, results in some data value that is related to a (possibly unknown) person. Such processing can be a trivial lookup in the data set or require sophisticated inference that possibly uses additional information. Singling out through inference can also require the combination of multiple data sets as for example used in reconstruction attacks of statistical data[↪].

Inference: Inference is the process of deriving information from a data set that is not evident. Inference typically applies knowledge of functional dependencies between values, known correlations, known probability distributions, or other dependencies of values that can be expressed with models (including machine learning models). Types of inference include *attribute inference* where the result of the inference are new values that are related to the same data subject, and *membership inference* where, based on some known values of a person, it can be established that this person is indeed a data subject.

Linkage: Linkage is the process of establishing a relation between a singled-out value and an actual natural person. Simple forms of linkage *match* combinations of values of the data set with an external data set that contains direct identifiers. More sophisticated forms of linkage match on values derived by inference or use inference without matching. Linkage is only possible if at least one value relating to the data subject can be singled out.

Matching: Matching is a kind of Linkage based on comparison. The comparison can be based on equality of invariant values or the similarity or closeness of values that change.

Spontaneous Recognition: Spontaneous recognition is a kind of Linkage in which a human observer of a data set matches a singled out combination of values to the known values of a familiar person (relative, colleague, acquaintance, etc.). It uses additional information about the data subject that is knowledge much rather than materialized as data.

Aggregation: Aggregation is a mapping from values relating to multiple persons to a value that relates to a group of persons. Examples include statistics, machine learning models, and decision trees.

Genaralization: Generalization maps values to a coarser scale of measurement such that the number of possible values is reduced. Examples include re-classification of nominal values and the definition of intervals of ordinal, ratio or interval values. Genealization can involve multiple values as in mapping weight and height into a body mass index or mapping possible coordinates to districts or zones.

Suppression: Suppression eliminates values from the data set. This can be a single (for example exceptional) value, all values (i.e., a record) of a given data subject, or an attribute for all data subjects.

Top- and Bottom-Coding: Top- and Bottom-Coding is a transformation in which all values above or below a certain threshold are mapped to the same output value that represents (e.g., "above 220 cm")

Noise Injection: Noise injection is a transformation that adds random noise to data values.

Slicing: Slicing is a transformation that splits a high-dimensional data set into multiple lower-dimensional ones.

Data swapping: Data swapping is a transformation in which values belonging to different data subjects (typically belonging to some group) are swapped.

ULD for AnoMed Funded by Funded by the European Union NextGenerationEU

Feedback to research@datenschutzzentrum.de