Table of De-Identification Techniques

Name of Technique	Description / Examples	Pros	Cons
Redaction	**Erasing or expunging** sensitive data from a record.	Reduces risk if data are disclosed inadvertently or through unauthorized access; useful when the erased data elements are not needed for analysis (typical with direct identifiers).	Not effective if done improperly (e.g., if the erasure can be reversed or if enough indirect identifiers remain).
Suppression	**Removing** data (e.g., from a cell or row in a table, or data element(s) in a record) prior to dissemination to prevent the identification of individuals in small groups or those with unique characteristics. Examples: • Suppressing the value of a single field, such as a field in a patient record containing a very rare disease. • Not reporting observations for those patients where the number of patients for any combination of zip code, age, and diagnosis is below a given threshold (e.g., 5 people).	Useful when multiple indirect identifiers pose a risk for re-identification. More easily done with tabular data. Helpful when presenting analysis of findings to the institution that provided the data. Helpful in public health reporting.	May result in minimal data being produced for small populations, and it usually requires additional suppression of non-sensitive data to ensure adequate protection of PII (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals). Can be difficult to perform properly. Is less likely to be effective if there are additional data available elsewhere.
Blurring: • Aggregation • Generalization • Pixelation	**Reducing precision** of data **by combining** one or more data elements. **Aggregation:** combining individual subject data with a sufficient number of other subjects to disguise the attributes of a single subject (e.g., reporting a group	Minimizes risk of identification by focusing on collective data rather than individual data. Useful for "big picture" analyses.	Decreases reliability of data and increases potential for false conclusions. Aggregation: may not be possible with a small pool of subjects.

Name of Technique	Description / Examples	Pros	Cons
	average instead of an individual value). **Generalization:** collecting or reporting values in a given range (e.g., using age or age-range instead of date of birth); including individual data as a member of a set (e.g., creating categories that incorporate unique cases); or reporting rounded values instead of exact amounts. **Pixelation:** modifying or obscuring visual information (e.g., blurring out faces in a photograph).		Generalization: unhelpful for case studies or in situations where details or specificity enhance findings. Pixelation: technology exists to reverse such modifications; other factors in a photo can lead to re-identification, such as setting and clothing.
Masking: • Pseudonymization Coding • Perturbation Randomization • Swapping Shuffling • Scrambling Encryption • Noise Differential Privacy	**Replacing** one data element with either a random or made-up value, or with another value in the data set; can be done manually or by using an algorithm. **Pseudonymization/Coding:** replacing a real name with a made up name or a real value with a made-up value. **Perturbation:** replacing sensitive info with realistic but inauthentic data or modifying original data based on predetermined masking rules (which may include randomization). Example: an algorithm which replaces the date of birth of subjects. **Swapping/Shuffling:** data for one or more variables are switched with another record, so that the data user	Attempts to retain the functional usability of the data while concealing information that could lead to identification. Pseudonymization/Coding: allows for a unique descriptor to trace data across multiple records; useful for multiple data instruments. Perturbation: reduces the likelihood of reverse identification. Swapping/Shuffling: useful for creating data sets for software testing where fields must be	Can decrease accuracy of computations in some cases, affecting validity of data. Techniques may be ineffective for small data sets. Algorithms used for masking can be reverse engineered.

Name of Technique	Description / Examples	Pros	Cons
	does not know whether the real data values correspond to certain records (i.e., all the values in the data set are real, but are assigned to the wrong people). **Scrambling/Encryption:** data are algorithmically scrambled and only those with access to the appropriate key can view the encrypted data. **Noise/Differential Privacy:** statistical technique that introduces errors by randomly misclassifying values of categorical variable(s).	present and have realistic looking values. Noise/Differential Privacy: allows for quantification of potential privacy loss, enabling a more accurate risk assessment; useful for large data sets.	
Subsampling	**Releasing either a representative or random subsample** of data instead of an entire data set.	Minimizes risk of identification by reducing the amount of data reported.	May not yield representative and generalizable estimates of a study's overall subject population.