I. State of the art overview

A) Anonymization technics

Data anonymization plays a crucial role in balancing data utility with individual privacy in our information-driven age. Various advanced techniques are currently employed, reflecting the evolving landscape of data security:

1. Data Masking

This prominent method involves concealing or altering values within a dataset to prevent re-engineering of original values (Ohm, 2010). It encompasses static masking, applied to a copied dataset, and dynamic masking, which alters data in real-time during queries or transfers. Common techniques include k-anonymity (Sweeney, 2002), encryption (Kamara & Lauter, 2011), and differential privacy (Dwork & Roth, 2014).

2. K-Anonymity

This technique strives to anonymize data by ensuring indistinguishability within a crowd. Every record, representing an individual, must be indistinguishable from at least k-1 others based on specified identifying attributes (Sweeney, 2002). Imagine a dataset where zip codes are replaced with broader city names. This grants "anonymity" within a city, but re-identification becomes easier with smaller k values or additional information. While simple and interpretable, its vulnerability to attackers with partial knowledge remains a concern (Machanavajjhala et al., 2007).

3. Differential Privacy

This method takes a different approach, focusing on the outcome of queries rather than raw data. It injects carefully calculated noise into statistical queries about the data, adding a cloak of uncertainty (Dwork et al., 2006). Even if an observer sees the anonymized results, they cannot pinpoint the exact contribution of any single individual. Think of it like analyzing the movement of a crowd in a thick fog; you gain insights about the group's behavior without identifying any specific person. The strength of differential privacy lies in its provable guarantees, quantifying the risk of re-identification. However, this cloak comes at a cost: the added noise can introduce inaccuracies in the results, demanding a delicate balance between privacy and data utility.

4. Data Pseudonymization

This involves masking direct identifiers by replacing them with artificial identifiers or pseudonyms, like replacing email addresses with numerical codes. While encouraged by GDPR for risk reduction (Ohm, 2010), it is reversible and does not address indirect identifiers, limiting its anonymization efficacy compared to other techniques. Pseudonymized data remains within the jurisdiction of the original data, as it does not address indirect identifiers as anonymization techniques do.

5. Data Generalization

This broadens the view of a dataset, making individual characteristics less distinguishable (Domingo-Ferrer & Torra, 2005). This involves mapping various values to a single value or range, such as grouping specific ages into ranges. Two primary approaches include automated generalization, which algorithmically balances privacy and accuracy, and declarative generalization, which involves manual determination of distortion levels.

6. Data Perturbation

This technique introduces deliberate randomness to data elements, adding vagueness predictably without compromising analytical accuracy (Lipton et al., 2002). It randomizes sensitive numerical values or alters categorical variables randomly. This technique is often utilized to protect electronic health records (EHR) or ensure privacy in surveys while estimating responses accurately.

7. Data Swapping

This technique, also known as data shuffling or permutation, involves rearranging attribute values in a dataset, disrupting correspondence with original data (Drineas et al., 2006). This is beneficial in machine learning to reduce biases and enhance model performance by providing representative testing batches.

8. Synthetic Data

Generated by algorithms, this data closely mimics real sensitive data. It is extensively used for training and validating machine learning and artificial intelligence models, eliminating the need for large volumes of sensitive personal information (King et al., 2020). Predictions indicate a significant rise, with Gartner estimating that 60% of data used in AI development and analytics projects will be synthetic within the next two years.

B) The Evolving Landscape of Linking and De-linking (Formerly "Profiling") in Data Anonymization

While the term "profiling" previously encompassed a broader sense of identifying individuals in anonymized data, the intricacies of this issue now lie in the distinct processes of linking and de-linking. Let's delve into the state of the art, navigating through impactful research and crucial ethical considerations.

Linking: Unveiling the Hidden Identity

Imagine tracing someone through digital breadcrumbs. Linking methods utilize external information, often from separate datasets, to connect anonymized records to specific individuals. Location data intertwined with social media profiles paints a revealing picture, bridging the gap between anonymity and identity. Research unveils the tools fueling this practice:

- Machine learning: Studies like Fredriksen et al. (2014) highlight the power of supervised and unsupervised learning models in uncovering hidden patterns and correlations within anonymized data, potentially revealing identifiers.
- Network analysis: Narayanan and Shmatikov (2009) explore how graph-based methods analyze relationships between data points, uncovering social connections and behavior patterns that can lead to individual identification.
- Homogeneity attacks: Backes et al. (2008) delve into exploiting the lack of diversity within specific groups in an anonymized dataset. By isolating unique characteristics, attackers can potentially identify individuals.

De-linking: Severing the Ties

Think of masking faces in a crowd photo. De-linking strives to disconnect an individual's identity from potentially identifying data while preserving valuable information. Research explores various techniques:

- Differential privacy: Dwork et al. (2006) propose introducing controlled noise into data analysis, blurring individual details while preserving usability for statistical purposes.
- Federated learning: Kairouz et al. (2019) discuss distributing computations across multiple data holders, making it difficult for any single entity to access complete datasets and hinder linkage attempts.
- Homomorphic encryption: Gentry (2009) explores enabling computations on encrypted data, protecting identifiers while allowing analysis.

Balancing data analysis needs with individual privacy demands responsible use of these powerful tools. Ohm (2010) emphasizes the importance of transparency and

accountability throughout the data lifecycle, ensuring individuals understand how their information is used and protected.

Technology and attacker strategies constantly evolve. Staying informed about advancements in linking and de-linking methods, embracing ethical considerations, and continuously refining countermeasures are crucial for navigating this dynamic landscape. As Meidan et al. (2014) highlight, ongoing research is vital to mitigate potential risks and uphold individual privacy in a data-driven world.

Method	*Impact on Mobility*	*Impact on Profiling*
Data Masking	Low impact: data anonymized but location data may still be identifiable.	Moderate impact: noise may disrupt profiling, but additional information can often reveal individuals.
K-Anonymity	Moderate impact: location may be generalized (e.g., city instead of street), reducing precision.	Low impact: profiles can still be built based on other attributes with k-anonymity.
Differential Privacy	High impact: location data obfuscated with noise, reducing accuracy for both mobility analysis and profiling.	High impact: noise makes profiling highly inaccurate, but information from other sources could still be used.
Data Pseudonymization	Limited impact: location data remains identifiable by the pseudonymization key.	No impact: pseudonyms don't anonymize, profiles can be built using the linked data.
Data Generalization	Moderate impact: location generalized (e.g., regional area), reducing precision for mobility analysis.	Moderate impact: profiles can still be built based on other attributes with generalization.

Data Perturbation	Moderate impact: location data slightly randomized, affecting mobility analysis accuracy.	Moderate impact: noise disrupts profiling but additional information can be used to overcome it.
Data Swapping	Low impact: location data remains largely unchanged.	No impact: data swapping doesn't anonymize, profiles can be built using the original data.
Synthetic Data	High impact: simulated location data used, preserving privacy but potentially lacking real-world accuracy.	High impact: profiles built with synthetic data are inaccurate and reveal no real information.