CYBER 759 – Assignment 1
Gunnar Yonker

1. Pseudonymization is a privacy technique that involves replacing real names and data with pseudonyms to protect sensitive information while retaining the data utility. Some of the techniques are:

i) Masking – Masking is hiding parts of sensitive data with symbols or characters. An example of this would be when using a Social Security Number, it could be masked like this: XXX-XX-9999.

ii) Tokenization – Tokenization replaces identifiable information with unique tokens using a lookup table, for example, replacing an email like email@uww.edu with a token 872df10c2da64.

iii) Hashing – Hashing converts data into a fixed-length string of characters (hash) using an algorithm, usually MD5 or SHA-256. For example, an email like email@uww.edu might be transformed using an MD5 hash algorithm into 7b4b35d39d89844e19b62dd03f145e2f.

iv) Format-preserving encryption – This technique retains the format of the original data while encrypting it, for example, preserving the format of an email address while encrypting its content.

2. Personally Identifiable Information (PII) vs. Person-related Data

PII refers to data directly identifying an individual, such as full name, SSN, or email address. Person-related data, on the other hand, includes information about an individual that may not directly identify them but can still reveal sensitive details like religious beliefs or political affiliation. It is important to consider a wide range of person-related sensitive data because individuals have varying privacy preferences, and seemingly non-sensitive data can be used to infer private information. For example, someone's location data over time can reveal their daily routines and, indirectly, sensitive information.

3. Anonymization is the process of removing or altering identifying information in a dataset to prevent tracing data back to specific individuals or entities. Some anonymization techniques are:

i) Suppression – Removing specific columns from datasets that contain identifying information like names and ages.

ii) Aggregation – Combining or summarizing rows of data to reduce granularity, such as aggregating income data by country.

iii) Transformation – Scrambling or altering data values while preserving the overall statistical distribution, which could be done by replacing ages with age groups.

4. Differential Privacy is a mathematical framework for quantifying and controlling privacy loss when releasing data. It seeks to address the limitations of basic privacy techniques like pseudonymization and anonymization. Some specific limitations of these techniques are inadequate privacy protection, difficulty in measuring privacy loss and levels, and static protection. The intuition behind differential privacy is to make it difficult for an attacker to determine if a specific individual's data is in a dataset, regardless of what they know about others. It provides formal guarantees of privacy by ensuring that the risk of identifying an individual remains low, even when an adversary has access to auxiliary information.

5. The information that the median salary increased from $72,000 to $77,000 after Adam joined does give us some clues about Adam's salary. It doesn't reveal Adam's exact salary, but it does indicate that Adam's salary is likely higher than the initial median of $72,000. This inference is based on the fact that the median increase, which implies that at least one new salary (Adam's) must be higher than the previous median for it to increase. Adam's exact salary is not provided, but it gives a general idea that his salary is higher than $72,000.