The Spectrum of Human Subjects' Privacy

There are different levels of data protection and different strategies for stripping out identifiers that could directly or indirectly re-identify subjects and produce inadvertent harms to them.

The main challenge is to use and share research data while protecting human subjects' privacy.

Types of **Identifiable Data**

Direct identifiers Unique to individuals Examples:

- Name
- Email
- SSN
- IP address
- Phone number
- Full-face images
- Medical record number

Quasi-identifiers Attributes that combined can disclose one's identity Examples:

- Race or ethnicity
- Age
- Gender
- Zipcode
- Political opinion
- Religious orientation
- Affiliation/Profession

Risk of Re-identification

Identifiable data

One or more direct identifiers are present in the dataset.

Pseudonymized data

Direct identifiers are removed or transformed, but quasi-identifiers remain intact.

De-identified data

Direct and known quasi-identifiers are removed or transformed.

Anonymized data

Direct, quasi and indirect identifiers are removed or manipulated using computational techniques.

Very High

Moderate

Residual

Very Low

Some Techniques to Mitigate Re-identification:

- Scrambling: mixes or obfuscates letters
- Encryption: makes the original data unintelligible and the process is only reversed with a decryption key
- Masking: important/unique parts of the data are hidden with random characters or other data
- Tokenization: keeps specific data fully or partially visible for processing and analytics while sensitive information is kept hidden
- Data blurring: creates an approximation of data values to render their meaning obsolete and/or make it impossible to identify individuals

Want to learn more? Contact us: rds@library.ucsb.edu

UC SANTA BARBARA Library



