What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

**DEGREES OF IDENTIFIABILITY**

- **Explicitly Personal**
  - Direct Identifiers: Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)
  - Indirect Identifiers: Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)
- **Potentially Identifiable**
- **Not Readily Identifiable**
  - Safeguards and Controls: Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals
- **Key Coded**
- **Pseudonymous**
  - Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.
- **Protected Pseudonymous**
- **De-Identified**
  - Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.
- **Protected De-Identified**
  - Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

**SELECTED EXAMPLES**

- Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)
- Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:48:6D:35:65:03)
- Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses or legal representations)
- Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HbA1c 5.1 g/dL = Csrk123)
- Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z)
- Same as Pseudonymous, except data are also protected by safeguards and controls
- Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)
- Same as De-Identified, except data are also protected by safeguards and controls
- For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)
- Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)