A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

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

<table>
<thead>
<tr>
<th>Explicitly Personal</th>
<th>Potentially Identifiable</th>
<th>Not Readily Identifiable</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DIRECT IDENTIFIERS</strong></td>
<td><strong>INDIRECT IDENTIFIERS</strong></td>
<td><strong>SAFEGUARDS and CONTROLS</strong></td>
</tr>
<tr>
<td>Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)</td>
<td>Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)</td>
<td>Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals</td>
</tr>
<tr>
<td>Intact</td>
<td>Intact</td>
<td>Not relevant due to nature of data</td>
</tr>
<tr>
<td>Partially Masked</td>
<td>Partially Masked</td>
<td>Limited or None in Place</td>
</tr>
<tr>
<td>Key Coded</td>
<td>Pseudonymous</td>
<td>Controls in Place</td>
</tr>
<tr>
<td>Protected Pseudonymous</td>
<td>De-Identified</td>
<td>Limited or None in Place</td>
</tr>
<tr>
<td>Protected De-Identified</td>
<td>Anonymous</td>
<td>Controls in Place</td>
</tr>
<tr>
<td>Aggregate Anonymous</td>
<td></td>
<td>Not relevant due to high degree of data aggregation</td>
</tr>
</tbody>
</table>

### 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 68:41:6D:35:65:03)
- **Same as Potentially Identifiable except data are also protected by safeguards and controls** (e.g., hashed MAC addresses & legal representations)
- **Clinical or research datasets where only curator retains key** (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)
- **Clinical or research datasets where only curator retains key** (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)
- **Clinical or research datasets where only curator retains key** (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)
- **Data are suppressed, generalized, perturbed, swapped, etc.** (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)
- **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)

### Data Identifiability

- **Direct Identifiers**
  - Name, address, phone number, SSN, government-issued ID
  - Unique device ID, license plate, medical record number, cookie, IP address
  - Same as Potentially Identifiable except data are also protected by safeguards and controls

- **Indirect Identifiers**
  - Clinical or research datasets where only curator retains key

- **Safeguards and Controls**
  - Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals

### Data Types

- **Pseudonymous Data**
  - Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.
  - Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z)

- **De-Identified Data**
  - Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.
  - Clinical or research datasets where only curator retains key

- **Anonymous Data**
  - Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.
  - Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)