The Seven States of Data

Khaled El Emam
Eloise Gratton
Jules Polonetsky
Luk Arbuckle

8th November 2016

Key Observations

- Pseudonymous data is still personal information
- There are degrees of pseudonymous data
- Pseudonymous data with a well defined set of conditions on it has less risk of re-identification than generic pseudonymous data
- This type of pseudonymous data can be treated more flexibly even though it is still PII

2

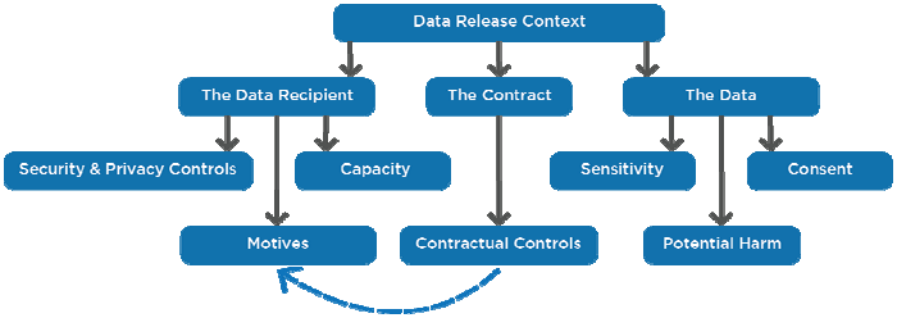
Typical Direct and Quasi-identifiers

Examples of direct identifiers: Name, address, telephone number, fax number, MRN, health card number, health plan beneficiary number, VID, license plate number, email address, photograph, biometrics, SSN, SIN, device number, clinical trial record number

Examples of quasi-identifiers: sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high level diagnoses and procedures

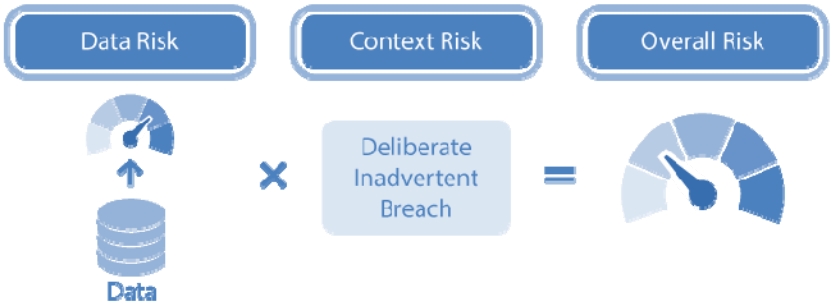
3

Data Sharing Context



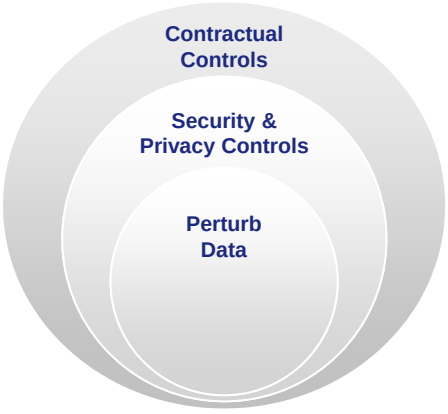
4

Measuring the Risk of Re-identification



5

Multiple Layers of Protection



6

Resources for Risk-based De-identification

7

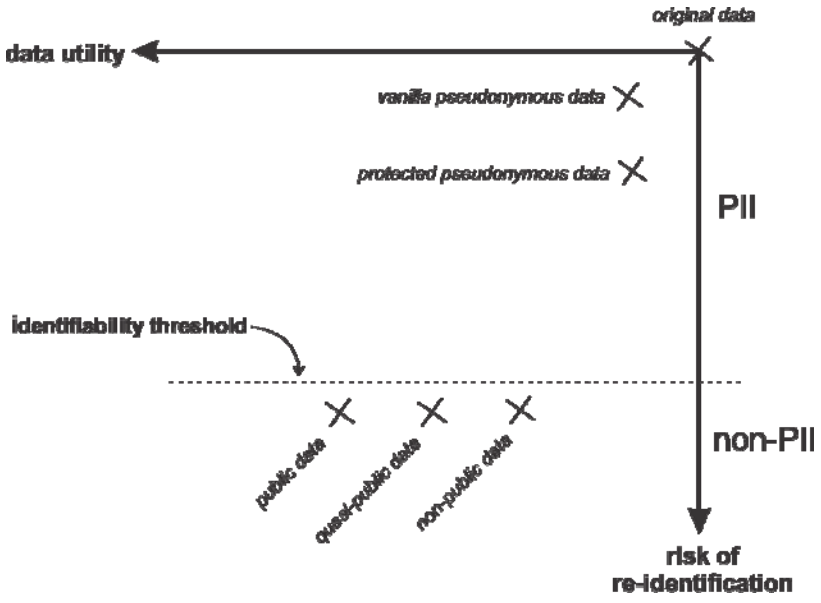
Pseudonymous Data

Examples of direct identifiers: name, address, telephone number, [illegible], health card number, health plan beneficiary number, VID, license plate number, email address, photograph, biometrics, SSN, SIN, device identifiers, medical record number

Examples of quasi-identifiers: sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high level diagnoses and procedures

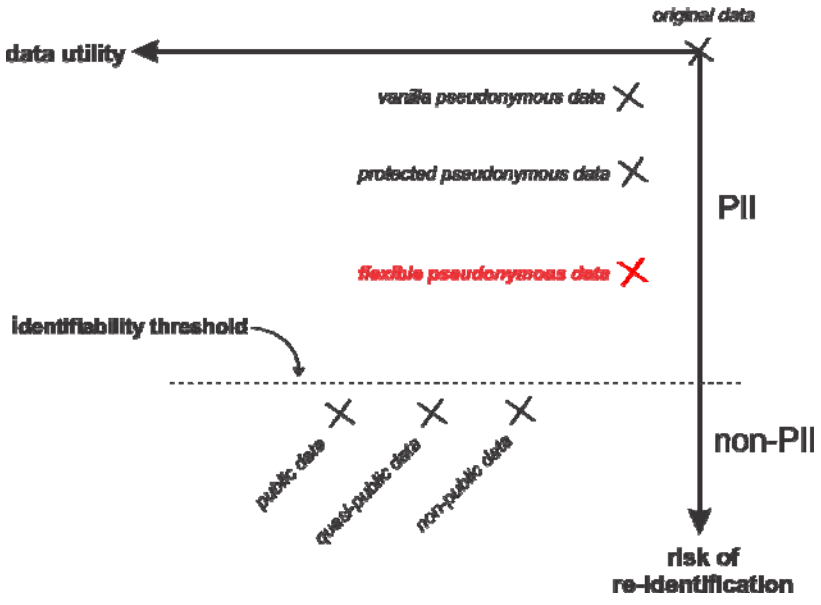
8

The Six States of Data



9

Adding Flexible Pseudonymous Data



10

Conditions on Flexible Pseudonymous Data

1. Strong privacy, security, and contractual controls
   - Risk from deliberate attempt of re-identification is very small

2. All processing is automated and data is transient
   - Risk from inadvertent re-identification is very small

3. No sensitive data is processed
   - Potential harm from a successful re-identification is minimized

4. No inference of identity information from analysis results

11

Afforded Flexibility

1. No need for opt-in consent (opt-out consent would be considered acceptable, or maybe just notice)
   - The risk of re-identification would be close to the threshold and main risk is that of a breach (the same risk exists for PII)
   - If there is a data breach then the same notification requirements would apply

2. Reduced restrictions on secondary purposes (reduced reliance on legitimate interests)

3. Because this applies only to sensitive information, then data such as health information would be outside this (as it would be considered sensitive)
   - See extensive discussion of sensitive information in the paper

4. This is consistent with existing risk management frameworks (ie, we are not introducing a new way of looking at identifiability)

12


Questions