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

Information containing direct and indirect identifiers.



PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.



DE-IDENTIFIED DATA

DE-IDENTIFIED

-11-1.

×...

ELIMINATED or

TRANSFORMED

×...

ELIMINATED or

TRANSFORMED

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.



Produced by

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.



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)



SAFEGUARDS and CONTROLS

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

> **SELECTED EXAMPLES**

Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)

NOT RELEVANT

EXPLICITLY

PERSONAL

INTACT

IDENTIFIABLE

POTENTIALLY



INTACT

LIMITED or

NONE IN PLACE

Unique device ID,

cookie, IP address

(e.g., MAC address

68:A8:6D:35:65:03)

record number.

license plate, medical

PARTIALLY MASKED PARTIALLY MASKED



NOT READILY

IDENTIFIABLE



INTACT

KEY

CODED



-X

ELIMINATED or

TRANSFORMED

INTACT



PSEUDONYMOUS

-12|.

×

ELIMINATED or

TRANSFORMED

LIMITED or **NONE IN PLACE**



INTACT

PROTECTED

PSEUDONYMOUS

-12.

-×...

ELIMINATED or

TRANSFORMED



female = gender: male)



PROTECTED

DE-IDENTIFIED

TRANSFORMED



and controls

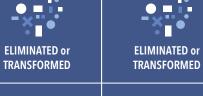
ELIMINATED or **TRANSFORMED**



NOT RELEVANT CONTROLS IN PLACE due to nature of data

Same as De-Identified, except data are also protected by safeguards

AGGREGATED ANONYMOUS ANONYMOUS









NOT RELEVANT due to high degree of data aggregation

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)



representations)

CONTROLS IN PLACE

Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123

CONTROLS IN PLACE

Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)

except data are also and controls

Same as Pseudonymous, Data are suppressed, protected by safeguards

generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: