Degrees of identification in data

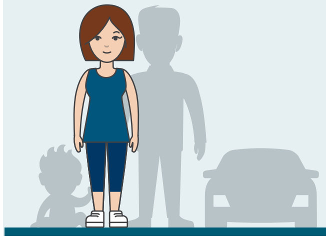
What do statisticians, data scientists and data analysts mean when they talk about confidentiality? How does identifiable data differ from de-identified or confidentialised information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability. This is a primer on how to distinguish different categories of data in the NZ context.

Identifiable

Data that directly or indirectly identifies an individual or business.

Individual

Name	Hēni
Gender	Female
DOB	31/01/1985
Address	28 My Road Postcode 6012 Wellington



Business

Name	Puzzles
Type	Paper Stationery Manufacturing
Employees	34
Expenditure	$398,000

Data that identifies a person without additional information or by linking to information in the public domain. Where an individual can be identified through connecting up information.

Personal, identifiable data like this are protected, and should only be released to the public providing we have explicit permission to do so.

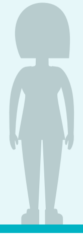
For example: Name, Date of birth, Gender.

De-identified

Data which has had information removed from it to reduce risk of spontaneous recognition.

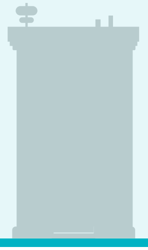
Individual

Name	*Unknown*
Gender	Female
DOB	1985
Address	Postcode 6012 Wellington



Business

Name	*Unknown*
Type	Manufacturing
Employees	30 - 40
Expenditure	$398,000



De-identified: Data which has had information removed from it to reduce risk of spontaneous recognition (likelihood of identifying a person, place or organisation without any effort).

For example: Data held within Stats NZ's Integrated Data Infrastructure and Longitudinal Business Database is de-identified before approved researchers can access in a secure data lab environment.

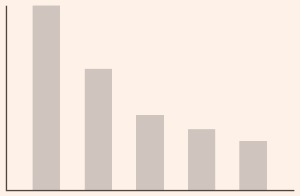
Partially confidentialised: Data which has been modified to protect the confidentiality of respondents while also maintaining the integrity of data.Modification involves applying methods such as top-coding, data swapping, and collapsing categorical variables to the unit records.

Confidentialised

Data which has had statistical methods applied to it to protect against disclosing unauthorised information.

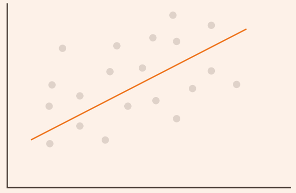
Individual

Name	*Unknown*
Gender	Female
Age	30 - 40 years
Address	Wellington



Business

Name	*Unknown*
Type	Manufacturing
Employees	10 - 100
Expenditure	Under $500,000



Statistical methods include suppression, aggregation, perturbation, data swapping, top and bottom coding, etc. These prevent the unauthorised identification of individuals, households, or organisations. This data is publicly available.

For example: Stats NZ nz.stat datasets.