Waipapa Taumata Rau
University of Auckland

Working With Personally Identifiable Research Data

James Love

July 2025

Introductions – In the zoom chat)
Who are you (name, role, institution)
… What are you looking to get out of this session?

Working With Personally Identifiable Research Data

Summary:

- **Understand the requirements for working with personally identifiable data**
- **Identify types of identifiable, de-identified and confidentialisation data and how to move between these**
- **Apply workflows and frameworks for identifiable data management and de-identification**
- **Select and apply techniques and technologies for de-identification of quantitative and qualitative data**

Why is managing identifiable data important?

Why is it important to protect personally identifiable data?

Privacy Act

Privacy principles covered by the Privacy Act 2020

National Ethics Advisory Committee

Principles for the safe and effective use of data and analytics, 2018
Stats NZ & the Privacy Commission

Information privacy principle 10.1.b.ii
*"...will be used for statistical or research purposes and **will not be published in a form that could reasonably be expected to identify the individual concerned."***

Further details >> https://www.privacy.org.nz/assets/New-order/Privacy-Act-2020/Privacy-Act-2020/Privacy-Act-2020-information-sheets-full-final-set-A711970.pdf

Health information Privacy Code

- Information privacy principles for the health sector
- Covers health information about identifiable individuals
- Applies to all health agencies (including departments of tertiary institutions that train healthcare professionals)
- Similar to, but **more specific** than Privacy Act 2020
  - Collection and consent requirements
  - Disposal, publication and retention rules
  - Ethics committee approval for research (when required)

https://www.privacy.org.nz/privacy-principles/codes-of-practice/hipc2020/hipc-factsheet-1-overview/

International Regulations

- EU: General Data Protection Regulation (GDPR) - guidelines
- USA: HIPAA – Health Data
  - HIPAA guidance for de-identification
- USA: Executive Orders
- AUS: Privacy Act

Privacy Commissioner
Te Mana Mātāpono Matatapu

Do I need to comply with the GDPR?

The European Union General Data Protection Regulation (GDPR) may apply to your agency if it handles the personal information of anyone living in the European Union (EU).

The GDPR will almost certainly apply to New Zealand agencies that have offices in EU countries, but it can also apply to agencies that don't.

You are likely to be covered by the GDPR if your agency is operating within the EU.

The GDPR will also apply to an agency outside the EU that targets individuals in the EU by offering goods and services, or that monitor the behaviour of individuals in the EU.

Some of the factors to consider are whether your agency:

- has websites in European languages, with the possibility of ordering goods and services in that other language
- accepts European currency
- frequently sells goods or services to EU citizens
- provides data processing services to EU-based companies.

Find out more about the GDPR here.

If you're based solely in New Zealand and you only occasionally sell something to a

Indigenous data sovereignty

Indigenous Peoples have inherent rights and responsibilities to **Indigenous data**.

- CARE principles for Indigenous data sovereignty
  Collective Benefit, Authority to Control, Responsibility, and Ethics
- Māori Data Sovereignty principles
  Rangatiratanga (Authority), Whakapapa (Relationships), Whanaungatanga (Obligations), Kotahitanga (Collective benefit), Manaakitanga (Reciprocity) & Kaitiakitanga (Guardianship)
- Pacific Data Sovereignty

Consider early as these impact the funding application, planning ethics application, consent, storage, metadata, sharing, and publishing of research findings and data throughout the research data lifecycle.

https://www.temanararaunga.maori.nz/

Data and privacy policies

- Institutional
- Data Provider
- Publisher
- Funder
- Government
- Professional Codes of Conduct
- Infrastructure – e.g. storage
- Collaborator

What is Identifiable Data?

Identifiable data

- Identifiable Data - from which the identity of a specific individual/entity can possibly be deciphered. E.g. Microdata (in surveys and census) is information at the level of individual respondents.
- In the New Zealand context, data is seen as taonga (something sacred, precious, or significant). A taonga should be actively cared for in a manner that preserves its integrity and value.

https://neac.health.govt.nz/national-ethical-standards/part-two/12-health-data

What are some examples of data values that could be used to personally identify someone?

(bonus points for less unique identifiers e.g. "Full Name" or "NHI number")

Examples of Identifiable Data

Identifiable - Direct

- Names and ID numbers
- Locations
- Phone number
- Online identities

Identifiable - Indirect

- Date of birth
- Relatives or employees
- Identification of employers
- And more (context specific)

https://neac.health.govt.nz/national-ethical-standards/part-two/12-health-data/

Examples of Identifiable Data

Identifiable - Direct
- Names and ID numbers
- Locations
- Phone number
- Online identities

Identifiable - Indirect
- Date of birth
- Relatives or employees
- Identification of employers
- And more (context specific)

De-identified data
- Encrypted unique IDs or codes
- Event dates
- Ethnicity & gender
- Rough location
- Aggregated status (e.g., deprivation index)

Confidential data

Assume that all data is potentially re-identifiable at some point in future.
- Exclude all identifiable data
- Use only the bare minimum
- Aggregate and perturbate
- Take care linking multiple datasets

https://neac.health.govt.nz/national-ethical-standards/part-two/12-health-data/

Degree of identification in data

Identifiable

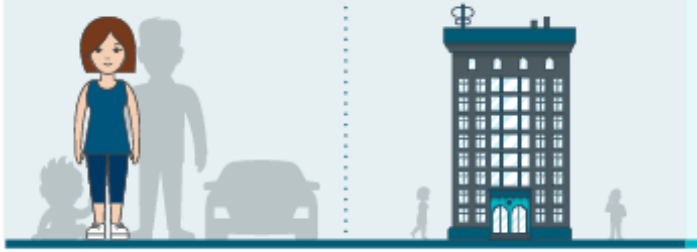
Data that directly or indirectly identifies an individual or business.

Individual

Name	Hēnī
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

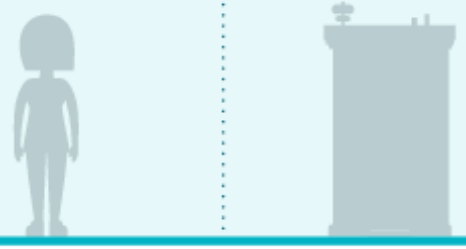
Data which has had information removed from it to reduce risk of spontaneous recognition.

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

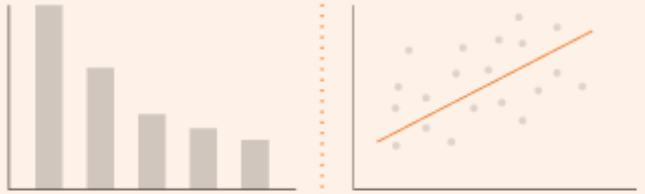
Data which has had statistical methods applied to it to protect against disclosing unauthorised information.

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.

Managing identifiable data

- Understand risks associated with data any release.
- Minimise risk within the data.
- Establishing a well-defined data management plan **before starting your research** is the best way to meet ethical and privacy requirements through access control and data security.
- Use appropriate secure storage options.
- Data should not be stored longer than is required.
- Robust policies, processes, and procedures must be in place to manage data throughout its life cycle (collection, analysis, output).

Five Safes Framework

Framework to understand Sensitive data access and governance

Used by Stats NZ IDI

Used internationally from repositories to regulations

The **Five Safes are** dimensions for evaluating sensitive data.

Not a checklist.

Each should be considered and weighted based on context.

Safe Output: are published results from the data non-disclosive?

Safe Data: Has risk been minimised within the data itself?

Safe People: Can people be trusted to use the data appropriately

Safe Projects: Is the data used for appropriate and ethical ends?

Safe Settings: Does the environment limit inappropriate use of data?

https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/

Access Controls

Safe People

Safe Settings

- Control who can see what and under what circumstances
- **Via:**
  - Trusted hardware (controlled machines, air gaped facilities)
  - Authentication (identities, multifactor)
  - Secure systems and software
  - Encryption
  - Background checks and due diligence
  - Ensuring systems, domain, and data knowledge
  - Frequent review of controls

F

Findable

A

Accessible
(others know how to access)

I

Interoperable

R

Reusable

Metadata (descriptive information), DOI and process for access are external facing, human and machine readable E.g., internet search results, bibliographic databases.

Metadata is with data and disciplinarily specific. Combining and using data are enabled by format and file type(s).
E.g., Data Management Plan, Protocol, README.txt

F A I R

Findable Accessible Interoperable Reusable

C

Collective benefit

A

Authority to control their data

R

Responsibility to engage respectfully with those communities

E

Indigenous Peoples' **ethics** should inform the use of data across time

Everyone

Specific people and purpose

De-identifying data

- De-identified data removes the details to reduce the risk a particular participant/entity can be identified from the data.
- It is important to note that keeping both the original and de-identified data is considered best practice to allow for data verification and correction/deletion required by regulation
- De-identification is not a fixed or end-state
- Assume your data may be able to be re-identified at some point in the future
- Zero risk is usually not compatible with useful data

De-identification Decision-Making Framework

Data situation audit

Describe data situation
Know your data
Understand:
- legal responsibilities
- use case

Meet ethical obligations

Risk analysis and control

Identify:
- processes needed to assess disclosure risk
- relevant **disclosure control processes**

Impact management

Identify stakeholders
Plan:
- how you will communicate
- what happens next once data is shared or released
- what to do if things go wrong

The De-Identification Decision-Making Framework

Degree of identification in data

Identifiable

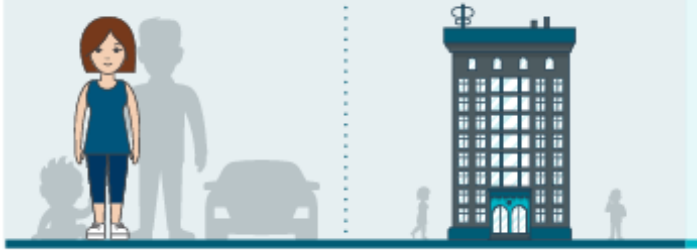
Data that directly or indirectly identifies an individual or business.

Individual

Name	Hēnī
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

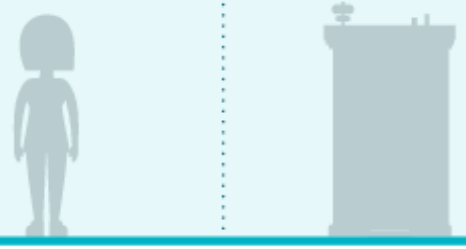
Data which has had information removed from it to reduce risk of spontaneous recognition.

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

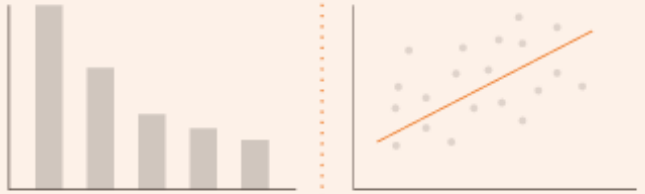
Data which has had statistical methods applied to it to protect against disclosing unauthorised information.

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.

Steps for data de-identification

1. Remove direct identifiers

2. Remove alter or obscure other information that could potentially be used to re-identify an individual/entity (secondary identifiers)

and/or

3. Use controls and safeguards

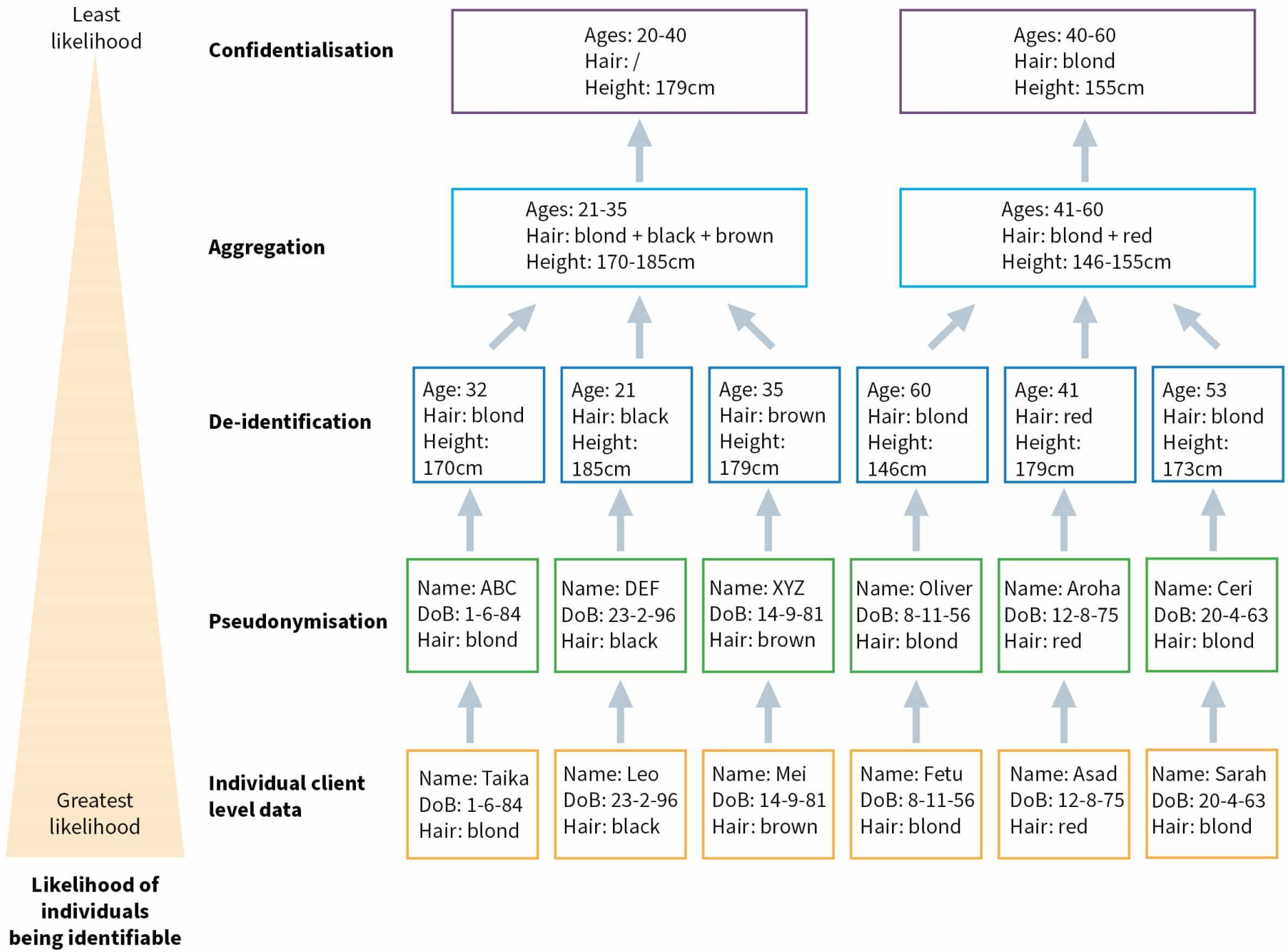
in the data access environment to prevent re-identification

4. Retain a key

that matches your de-identified data with the identifying information

De-identification vs confidentialisation

	De-identification	Confidentialisation
Definition	removing or masking identifiable elements e.g. names, addresses, or social security numbers	process of irreversibly removing identifiable elements from data, ensuring that individuals cannot be re-identified
Re-identification risk	potentially re-identifiable if combined with other datasets	cannot be re-identified, even when combined with other datasets
Common techniques	masking, data suppression, tokenization	generalization, randomization, differential privacy
Use Cases	Often used in healthcare and finance to comply with regulations like The Health Information Privacy Code, while still allowing data analysis and sharing	Used when data needs to be shared or published without any risk of re-identification, such as in public datasets or research


Least likelihood
Greatest likelihood
Likelihood of individuals being identifiable
Confidentialisation
Aggregation
De-identification
Pseudonymisation
Individual client level data
Ages: 20-40
Hair: /
Height: 179cm
Ages: 40-60
Hair: blond
Height: 155cm
Ages: 21-35
Hair: blond + black + brown
Height: 170-185cm
Ages: 41-60
Hair: blond + red
Height: 146-155cm
Age: 32
Hair: blond
Height: 170cm
Age: 21
Hair: black
Height: 185cm
Age: 35
Hair: brown
Height: 179cm
Age: 60
Hair: blond
Height: 146cm
Age: 41
Hair: red
Height: 179cm
Age: 53
Hair: blond
Height: 173cm
Name: ABC
DoB: 1-6-84
Hair: blond
Name: DEF
DoB: 23-2-96
Hair: black
Name: XYZ
DoB: 14-9-81
Hair: brown
Name: Oliver
DoB: 8-11-56
Hair: blond
Name: Aroha
DoB: 12-8-75
Hair: red
Name: Ceri
DoB: 20-4-63
Hair: blond
Name: Taika
DoB: 1-6-84
Hair: blond
Name: Leo
DoB: 23-2-96
Hair: black
Name: Mei
DoB: 14-9-81
Hair: brown
Name: Fetu
DoB: 8-11-56
Hair: blond
Name: Asad
DoB: 12-8-75
Hair: red
Name: Sarah
DoB: 20-4-63
Hair: blond

Re-Identification

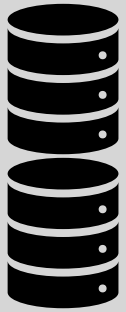
Re-identification

- When the identity of an entity in the data is determined, even in de-identified data
- The risk we are trying to minimize with de-identification techniques
- May be deliberate and malicious, or accidental and spontaneous
  - e.g., identifying someone via context clues or linking previously unlinked data such as public social media information
- Can be mitigated by controlling context and access
- Risk increases as **attributes are disclosed**

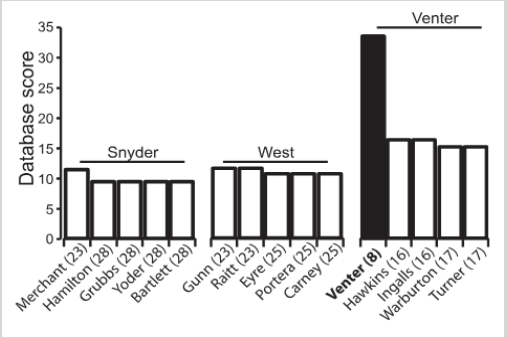
https://www.abs.gov.au/about/data-services/data-confidentiality-guide/understanding-re-identification

Example: Genomic Surname Inference

Y Chromosome mutation data



Public genealogy & population databases



Surname ID

Table 1. Comparison of CEU identification cases.

Feature	Pedigree 1		Pedigree 2		Pedigree 3
Genome for surname recovery	Paternal grandfather	Maternal grandfather	Paternal grandfather	Maternal grandfather	Father
Surname freq. in U.S.*	Rare	Rare	Common	Rare	Rare
Meioses between target and source	3	5	5	7	2
Relationship between target and source	Nephew	First cousin once removed	Great-great nephew	Second cousin once removed	Grandchild
Supporting evidence	State of residency, pedigree structure, age, and maiden name are the same		State of residency, pedigree structure, age, and maiden name are the same		State of residency, pedigree structure are the same (ages are not given)
P (random match)†	$<5 \times 10^{-9}$		$<5 \times 10^{-6}$		$<10^{-5}$

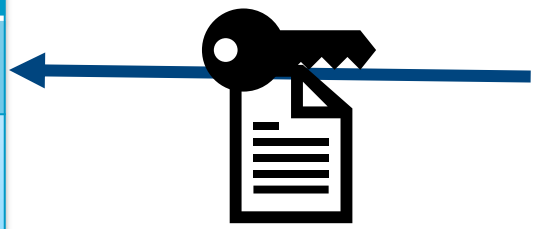
*Common: surnames with a prevalence of $>10^{-4}$; Rare: surnames with a prevalence of $\leq 10^{-4}$. †The estimated probability of finding at least one family with the same characteristics after scanning all Utah households.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., and Erlich, Y., "Identifying Personal Genomes by Surname Inference", science, vol. 339, no. 6117, p. 321, 2013. doi:10.1126/science.1229566.

Reversing De-identification

- Data may need to be re-identified for regulatory compliance validation or integrity checks.
  - Sometimes this is part of the intended use of the data (e.g., health data also used for diagnostics)
- This is often achieved by a key, function, or file that can re-map de-identified data back to identifiable values.
- This Key/File should be treated as highly identifiable information and **kept securely**

Survey Response	ID
"I often avoid checking my bank account"	1c272047233576d77a9b9a1acfdf741c
"I secretly rely on credit cards to maintain a lifestyle I can't actually afford."	6117323d2cabbc17d44c2b44587f682c



Name
Jane Doe
John Smith

Methods and techniques for de-identification & confidentialisation

What methods could you use to protect the identity of individuals in a dataset when collecting home address, height in mm, and name?

Methods

- **Tokenisation –** Replacing values with non-identifying IDs
- **Encryption –** Algorithmically access protecting values
- **Generalization -** Grouping categorical values
- **Perturbation –** Randomly altering values
- **Suppression –** Hiding or removing values
- **Synthetic data –** statistically synthesizing values

Tokenisation

- Replace values with a secure non-identifying value that is generated from/maps back to the original value
- May be reversible only by specific parties
- Shared Tokenized IDs can be used to link datasets without revealing identity

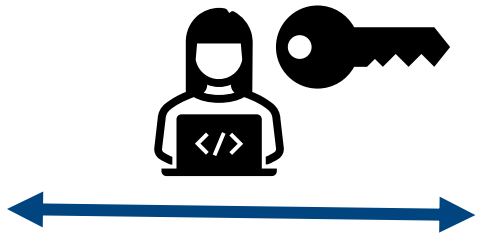
Survey Response	ID
"I often avoid checking my bank account"	1c272047233576d77a9b9a1acfdf741c
"I secretly rely on credit cards to maintain a lifestyle I can't actually afford."	6117323d2cabbc17d44c2b44587f682c

Name
Jane Doe
John Smith

Encryption

- Protect data (values, files, entire storage) using algorithms that convert them into unreadable code that cannot be read except by holders of a matching key
- Can be Asymmetric (One key encrypts, another key decrypts)
- Both **de-identifies** the data and **provides access control**
- Both tokenisation and encryption methods must be evaluated for if they are easily broken and do not leak information in context of the data

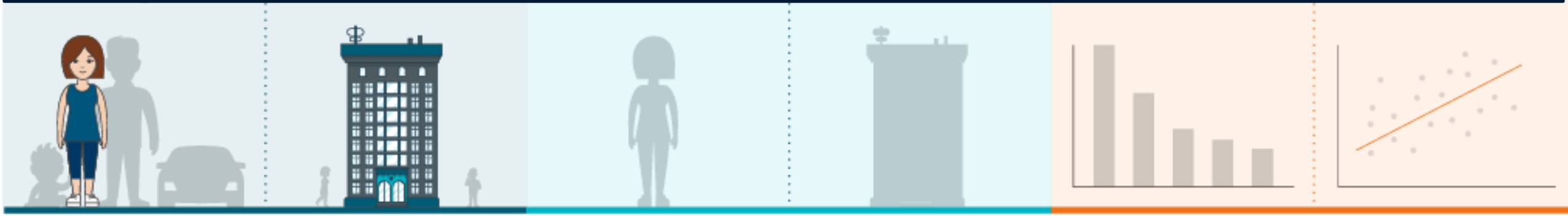
Encrypted Name
!^Kc7M %dM6tx
6*Aqih 2mh9F!
!^Kc7M p72Po#
!^Kc7M *R7Rne



Name
Jane Doe
John Smith
Alan Doe
Mary Doe

Generalisation

Gender	Female	**Type**	Paper Stationery Manufacturing	**Gender**	Female	**Type**	Manufacturing	**Gender**	Female	**Type**	Manufacturing
DOB	31/01/1985			**DOB**	1985			**Age**	30 - 40 years		
Address	28 My Road Postcode 6012 Wellington	**Employees**	34	**Address**	Postcode 6012 Wellington	**Employees**	30 - 40	**Address**	Wellington	**Employees**	10 - 100
		Expenditure	$398,000			**Expenditure**	$398,000			**Expenditure**	Under $500,000

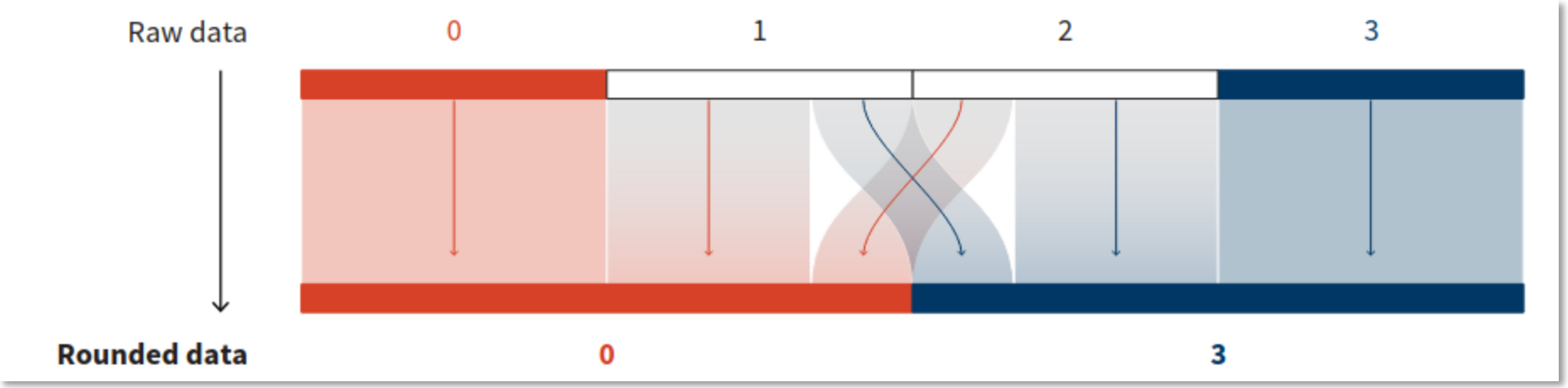


Statistical Disclosure Control

- A technique to assess and lower of re-identification within the data when it is published or released.
- SDC methods suppress or alter values in the data results in information loss compared to the original data.
- The goal of a well-implemented SDC process is to find the optimal point where utility for end users is maximized at an acceptable level of risk.
- SDC cannot achieve total risk elimination but can reduce the risk to an acceptable level- based on legal and ethical concerns.
- The **lower the disclosure risk**, the higher the information loss and the lower the **data utility for end-users**.

Perturbation

- Adding random noise to data outputs.
- Use **consistently** random methods (identical seeds, same random value per cell)
- Example **Randomised Rounding:**
  - By uniformly rounding to base 3 smaller values (where re-identification risk is higher) are proportionally changed more relative to higher values.

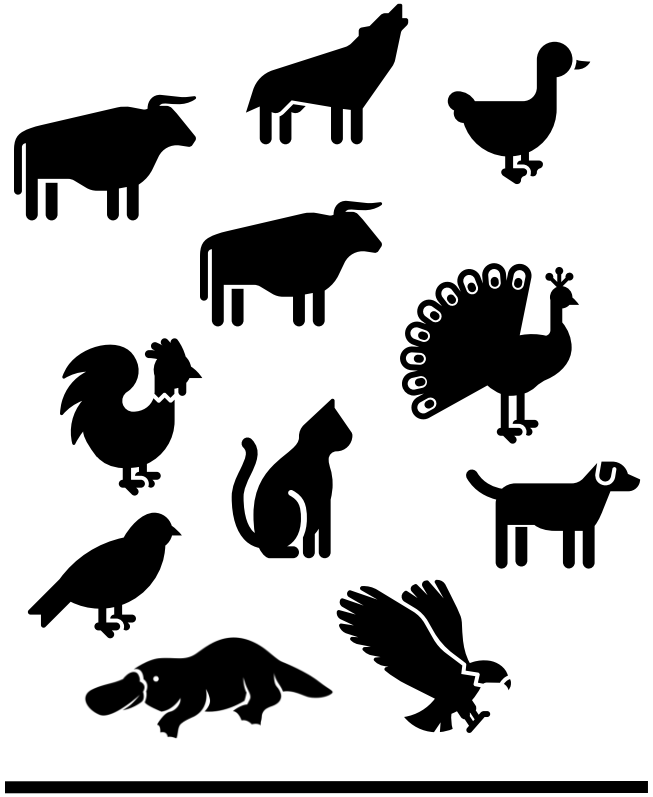


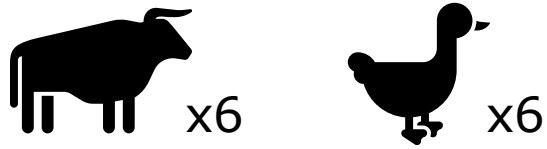
https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Aggregation

Similar to **categorization/Generalisation** (binning quantitative values)

- Combining and/or simplifying data outputs.
- Aggregation will reduce the specificity of the data and possibly introduce bias
- Requires contextual and subject matter knowledge of how categories are defined and combined
- Requires an understanding of what SDC methods have already been used, including other aggregations
- **Data Classifications** and **standards** for aggregation must be **unambiguous**, **exhaustive**, and ~~**mutually exclusive**~~



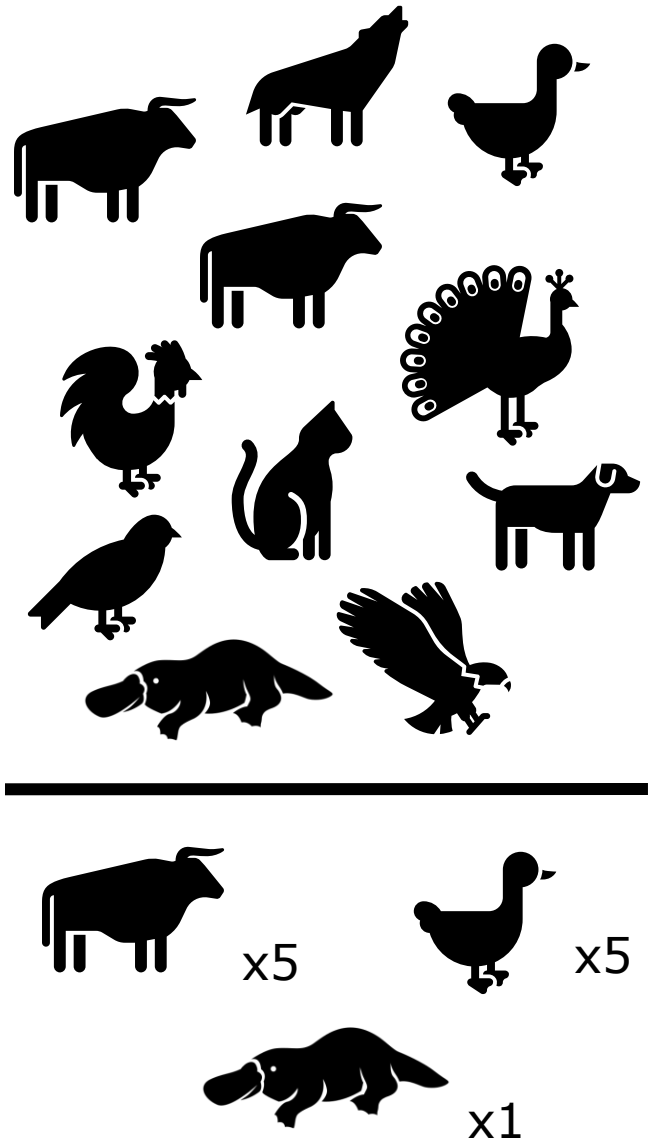


https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Aggregation

Similar to **categorization/Generalisation** (binning quantitative values)

- Combining and/or simplifying data outputs.
- Aggregation will reduce the specificity of the data and possibly introduce bias
- Requires contextual and subject matter knowledge of how categories are defined and combined
- Requires an understanding of what SDC methods have already been used, including other aggregations
- **Data Classifications** and **standards** for aggregation must be **unambiguous**, **exhaustive**, and **mutually exclusive**

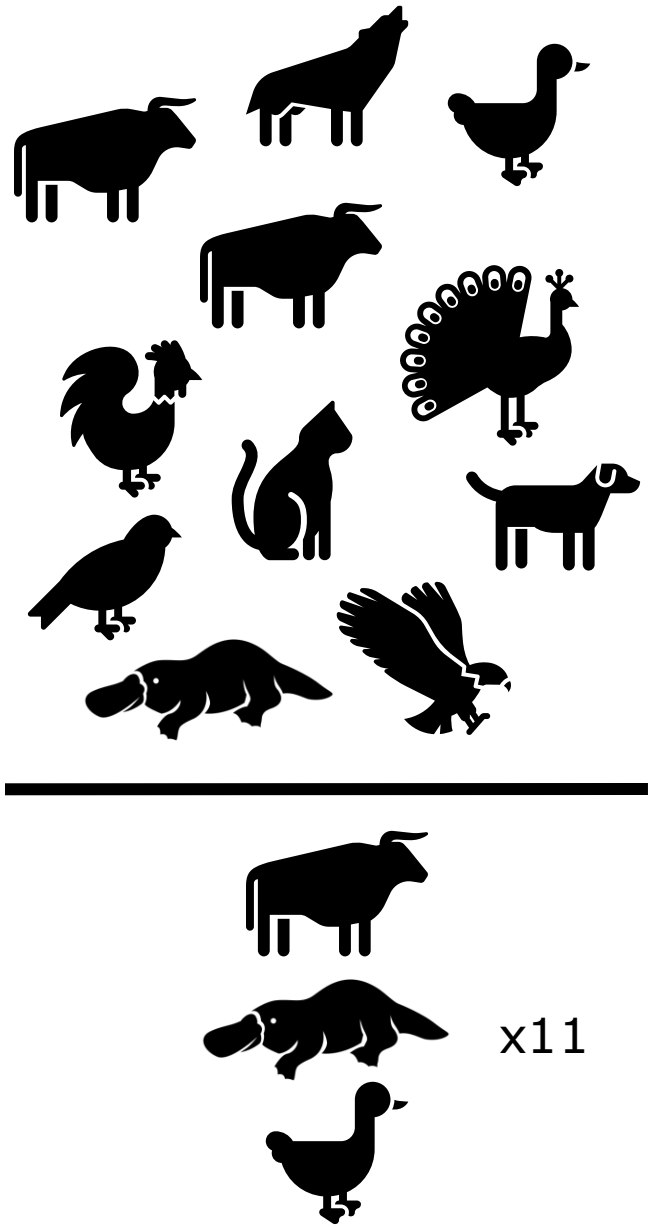


https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Aggregation

Similar to **categorization/Generalisation** (binning quantitative values)

- Combining and/or simplifying data outputs.
- Aggregation will reduce the specificity of the data and possibly introduce bias
- Requires contextual and subject matter knowledge of how categories are defined and combined
- Requires an understanding of what SDC methods have already been used, including other aggregations
- **Data Classifications** and **standards** for aggregation must be **unambiguous**, **exhaustive**, and **mutually exclusive**



https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Suppression

- Not Reporting or removing values
- Similar to **Masking** or **Redaction**
- You may need to hide secondary values to prevent inference of primary values
- Often performed by **automated tools**
- Be clear and consistent why and how data was removed/replaced

Identifying Personal Genomes by Surname Inference

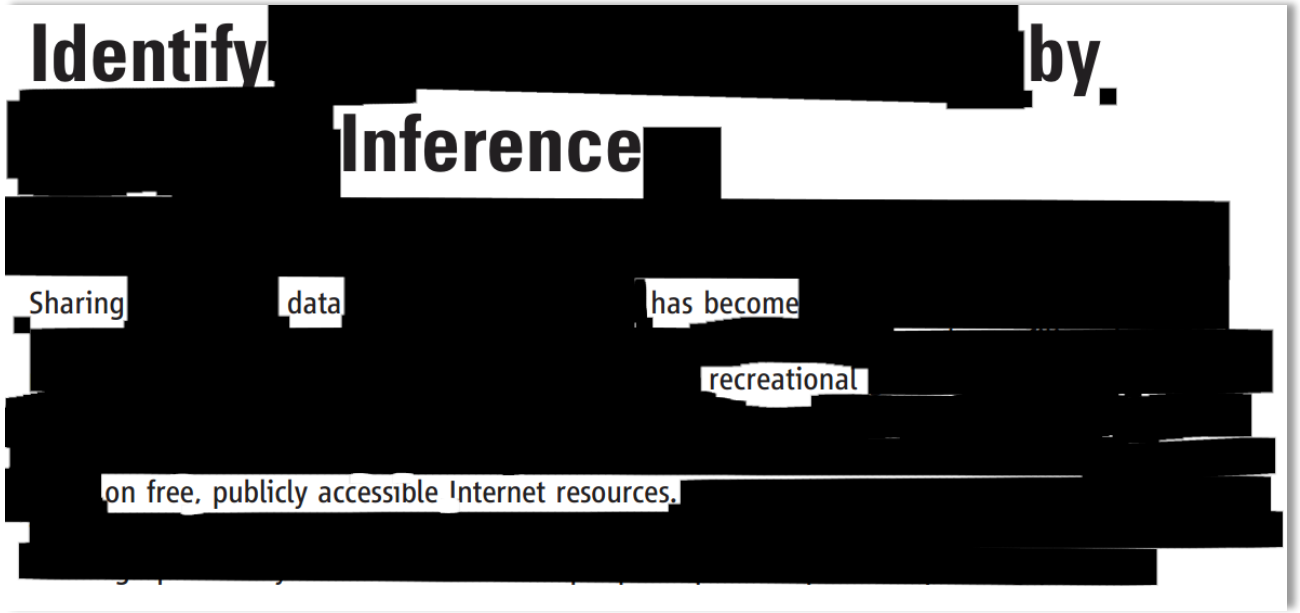
Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Suppression

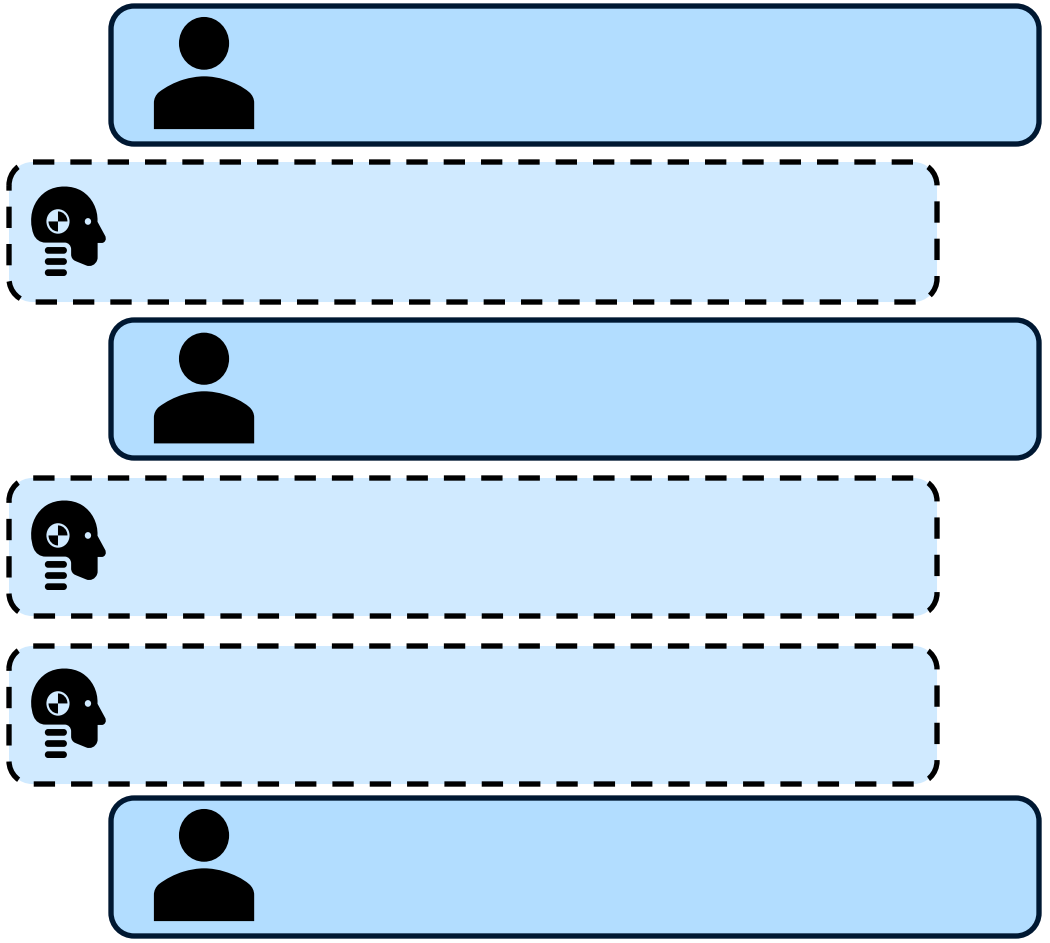
- Not Reporting or removing values
- Similar to **Masking** or **Redaction**
- You may need to hide secondary values to prevent inference of primary values
- Often performed by **automated tools**
- Be clear and consistent why and how data was removed/replaced



https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Synthetic Data

- Effective for protection against re-identifications from statistical results or machine learning weights
- Enhances and enhanced by other privacy techniques
- Trade-offs between similarity to real data and privacy
- Creating a synthetic dataset is not a simple task



https://ieeexplore.ieee.org/abstract/document/10568134

Review outputs and reassess disclosure

Before any data is output or shared assess possible re-identification risks

- Have de-identification methods been consistently applied to all data?
- Has an expert with domain or statistical knowledge reviewed the data?
- Evaluate in light of any novel developments in terms of data and technology

Measuring risk in outputs

- ***k*-anonymity –** A measure for if an individual in the dataset cannot be distinguished in a group of at least *k*-1 other individuals
  - Not a perfectly mathematical measure, relies on assumptions of what values are identifier
  - Cannot account for when attributes are disclosed that information outside the dataset may allow for identification
- **Special Unique Detection Algorithm (SUDA)**
  - Identifies and measures uniqueness of records in the data for combinations of variables
  - **Python and R libraries** (available as part of SdcTable and sdcMicro)

Example - data practices at

- Longitudinal study of child health
- Running since 2009 (before some children were born)
- Collect data on 6,000 New Zealand children and their families
- Collect a huge amount of quantitative (health, key dates) and qualitative data (questionaries, wellbeing, culture) on children and their parents (these are linked)

Example - data practices at Growing Up in New Zealand

Data Access:

Only the data needed is shared

- **Data application** Reviewed by Data Access Committee
- Clear aims & methodology
- Detail exactly the datasets and values needed and who will access them
- Specified datasets made available via secure platform

The Data:

Data is de-identified during use

- Direct Identifiers **tokenised** as IDs
- Tables can be merged on IDs (e.g. mother to child)
- Extreme, rare and specific (e.g. dates) data are **categorized** or **suppressed**
- Free Text **suppressed**

Data Output:

Only safe data is released

- Sensitive cells (count <10) are suppressed or aggregated
- Random rounding is applied to all counts
- Rules applied to graphs and models
- Derived variables are contributed back to GUINZ
- Must **apply to publish**

Example - data practices at Growing Up in New Zealand

Data Output:

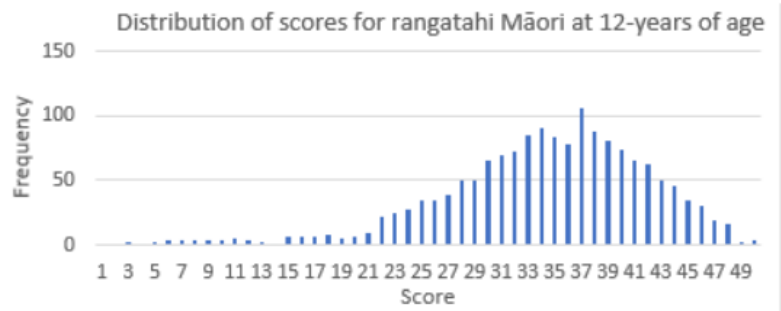
Only safe data is released

	Group 1	Group 2	Group 3
Group 1	11	47	58
Group 2	27	32	33
Group 3	4*	31	20
Total	42	110	111

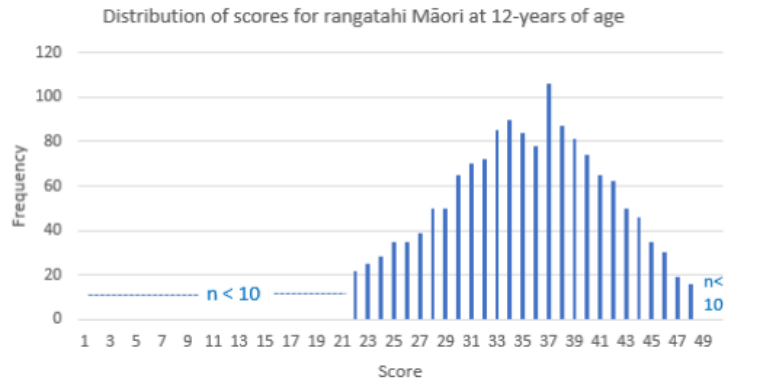
	Group 1	Group 2	Group 3
Group 1	<20	47	58
Group 2	27	33	33
Group 3	<10	33	21
Total	42	110	111

	Group 1	Group 2	Group 3
Group 1	<20	48	57
Group 2	27	32	33
Group 3	<10	31	18
Total	42	108	111

Before suppression is applied



After suppression has been applied



Tools for de-identification and evaluation

Tools for de-identification and evaluation

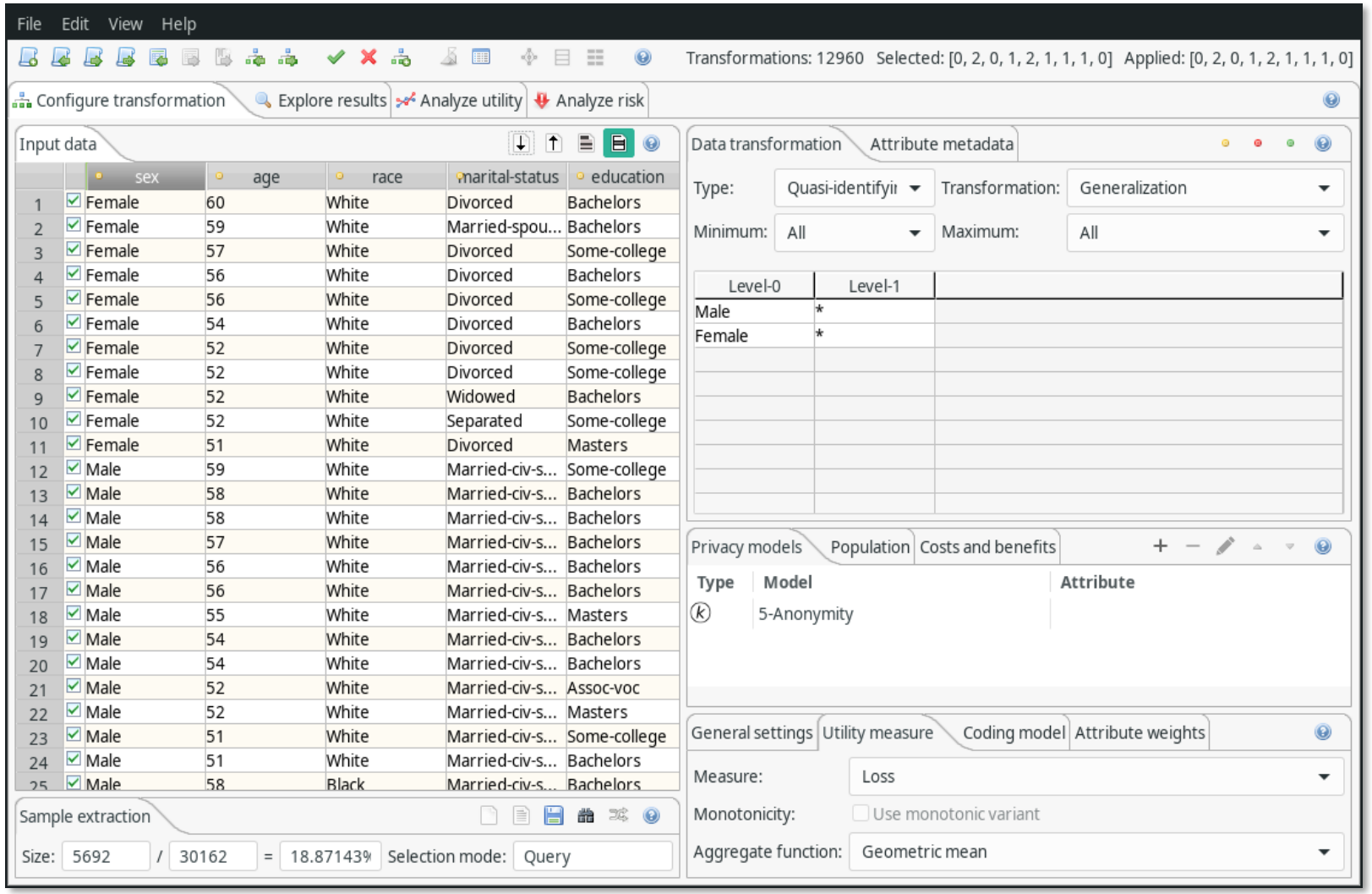
- Tau-ARGUS - Gui based, Open Source, controlled rounding and suppression in tabular data
- Amnesia - Designed to meet GDPR standards
- Privacy Analytics Eclipse - Enterprise software for
- **REDCAP – a secure survey platform with built in packages for suppression of sensitive data when exporting**
- **Microsoft Presidio – Open-source Microsoft tool for automatic identifier suppression**
- deepdefacer - Machine learning tool for confidentialising facial images
- And many more... https://guides.library.jhu.edu/protecting_identifiers/software

Software Packages and Libraries

- SdcTable - R package – performs statistical disclosure control and suppression in tabular data. (also sdcMicro - only microdata and sdcApp - Graphical Shiny interface)
- DicomAnonymizer - Python package for de-identifying DICOM images.
- Pydeface - Python package for removing facial structures from MRI images
- **Python Deidentification – Python package for de-identifying text documents**
- Anonymizer MySQL - This simple tool will allow you to make anonymized clone of your database.

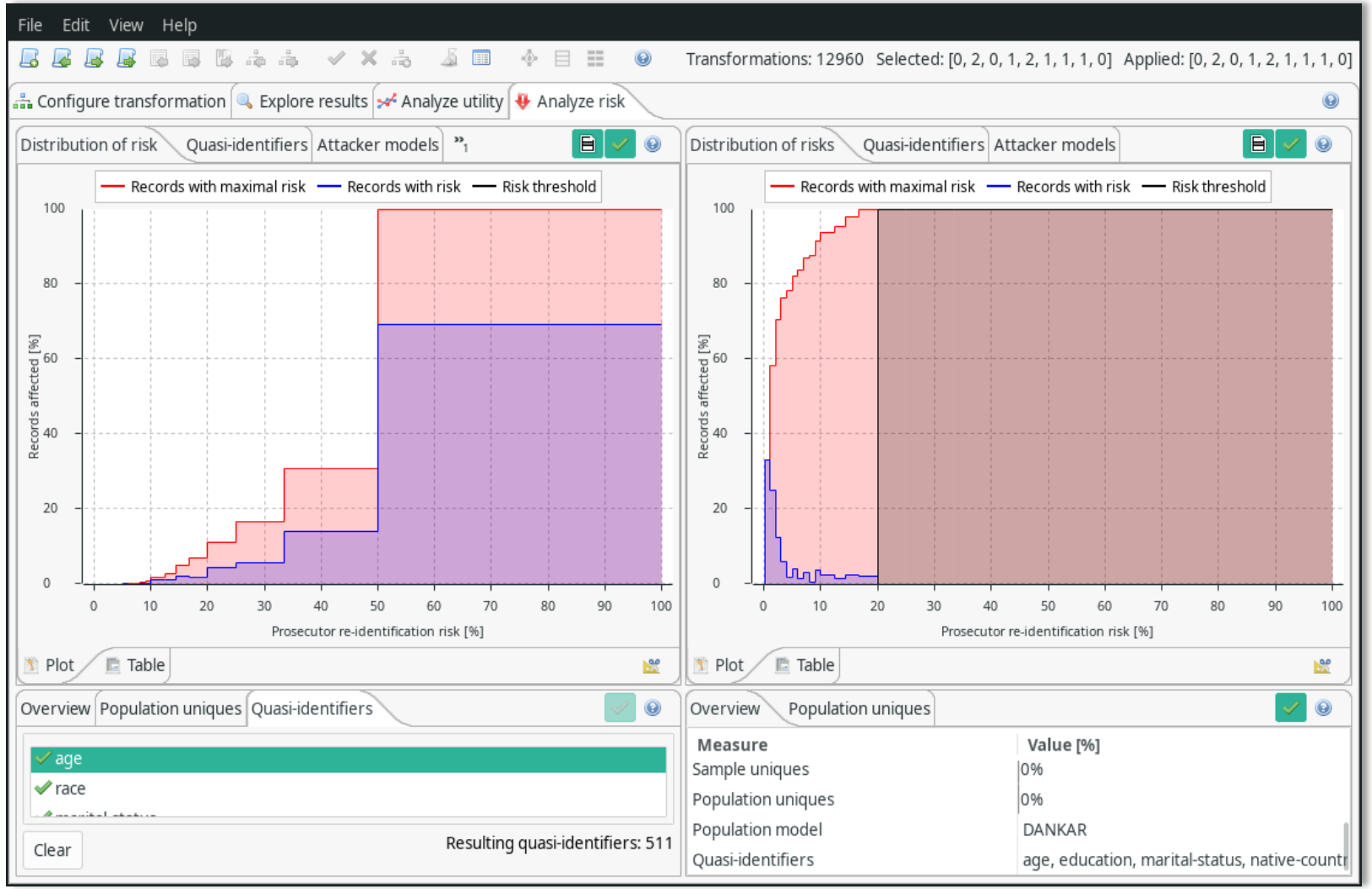
Tool Highlight – ARX Data Anonymization Tool

- Gui interface for de-identifying datasets
- Feature rich (perhaps a bit too much so)
- Customize hierarchies, privacy models, sensitivity



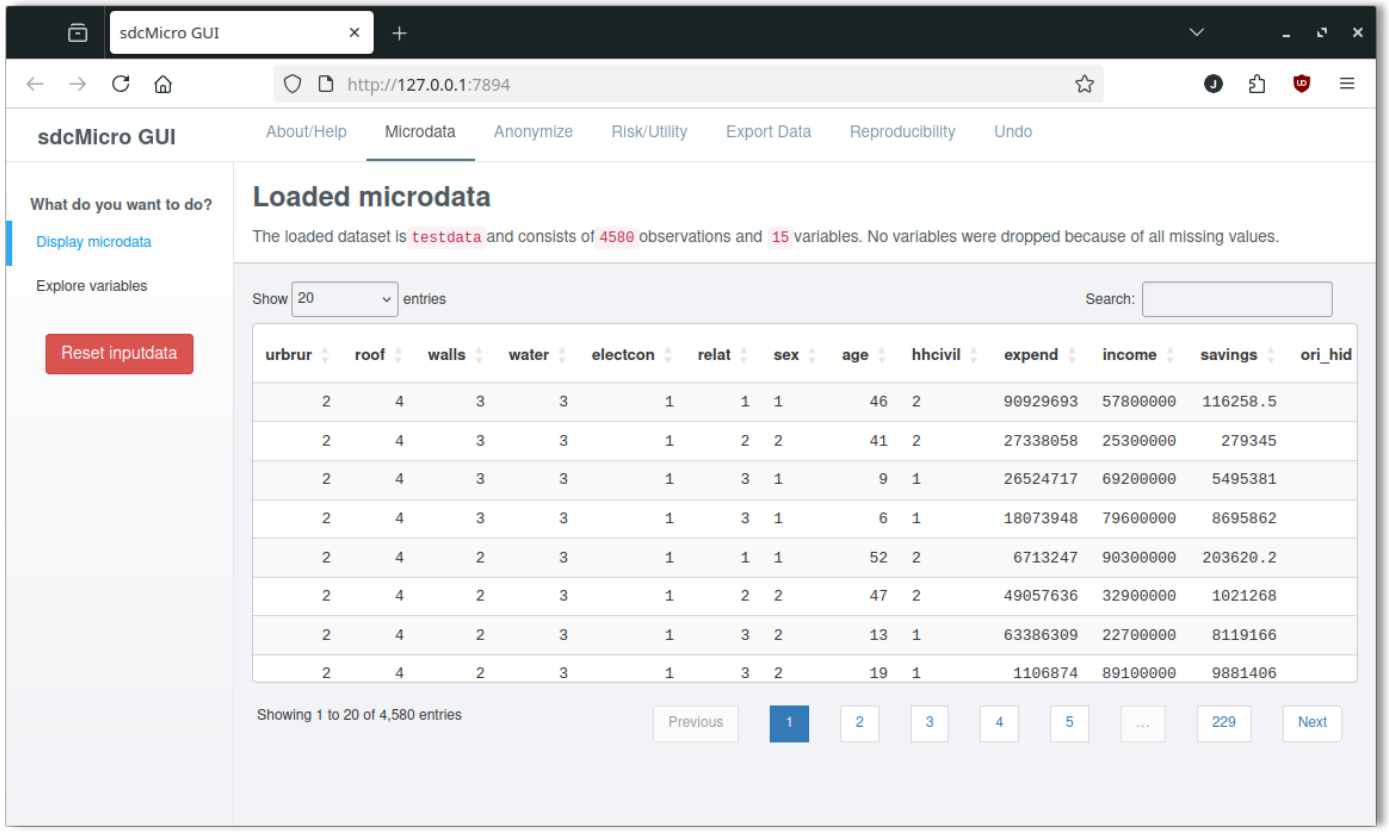
Tool Highlight – ARX Data Anonymization Tool

- Gui interface for de-identifying datasets
- Feature rich (perhaps a bit too much so)
- Customize hierarchies, privacy models, sensitivity
- Visual risk and utility assessments



Tool Highlight – sdcTable/sdcApp (R)

- **R libraries – Can be built into existing data workflows**
- **Shiny App for gui interface**



To Run Shiny App:

install.packages('sdcMicro')

library(sdcMicro)

sdcApp()

sdcTable Vignette

Show and Tell: Do you use any of these tools, or any I have not mentioned?

Summary – Working With Identifiable Data

Audit and assess identifiers

- Identify what parts of your data are direct and indirect identifiers
- Determine risk and **requirement**s of **re-identification**

Apply de-identification

- Remove, generalise and suppress identifiers
- Tokenise or encrypt data used to link tables
- Apply Statistical disclosure controls

- Apply data specific expertise and automated methods

Evaluate data risk

- Review data and techniques
- Evaluate privacy metrics of de-identified data
- Release balancing utility and privacy based on context

UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

Questions? Get in touch…

researchdata@auckland.ac.nz