Waipapa Taumata Rau University of Auckland

ResBaz RESEARCH BAZAAR

Working With Personally Identifiable Research Data

June 2026

Jean Love

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

..and were you in Research Data Management this morning?

Working With Personally Identifiable Research Data

Learning outcomes:

- Identify institutional, regulatory, and ethical requirements for working with personal information
- Recognise common types of personal information and other identifiable data collected in research
- Apply methods for managing identifiable research data at key points in the project lifecycle (collection, storage, analysis, sharing, and publication)
- Select and apply appropriate techniques for de-identification of research data

Requirements for managing identifiable data

Why is it important to protect personally identifiable data?

Privacy Act

Privacy principles covered by the Privacy Act 2020

Governs how organisations collect and use **personal information.**National Ethics Advisory Committee

Principles for the safe and effective use of data and analytics, 2018

Stats NZ & the Privacy Commission

1. Collect only the information you need	5. Store it safely and securely	9. Store only a long as required
2. Collect it from the person directly	6. Provide people access to their own data	10. Use it **only** for the purposes it was collected for
3. Inform the person about collection and use	7. Allow corrections	11. Only disclose for consented and lawful reasons
4. Be fair, lawful, and reasonable in context	8. Ensure it is accurate and up to date	12. Only share outside NZ with adequate protections

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

Privacy Act

Privacy principles covered by the Privacy Act 2020

Governs how organisations collect and use **personal information.**National Ethics Advisory Committee

Principles for the safe and effective use of data and analytics, 2018

Stats NZ & the Privacy Commission

1. Collect only the information you need

2. Collect it from the perso… directly

3. Inform the perso… collection and u…

4. Be fair, lawful, a… reasonable in conte…

5. Store it…

…ctions

…. Ensure it is accurate and up to date

9. Store only a long as required

10. Use it **only** for the purposes it was collected for

11. Only disclose for consented and lawful reasons

12. Only share outside NZ with adequate protections

13. Be careful with unique identifiers. (You are creating extra identifiable information!)

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

Privacy Act – Deidentification clauses

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."***

https://www.legislation.govt.nz/act/public/2020/31/en/latest/#LMS23342

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)

Biometric Processing Privacy Code 2025:

Covers the collection and use of Biometric Characteristics including:

Body (part or whole), face, fingerprints, palmprints, iris, retina, voice, vein patterns, gestures, gait, voice, heartbeat, eye movements, keystroke pattern, signature or handwriting...

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
- 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

Data and privacy policies

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

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/

Why is it important to protect personally identifiable data?

Why is it important to protect personally identifiable data?

Your own moral compass and duty to your participants.

Identifiable information throughout the research lifecycle

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

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
- Exclude all identifiable data
- Use only the bare minimum
- Aggregate and perturbate
- Take care linking multiple datasets

Assume that all public data is potentially re-identifiable at some point in future.

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

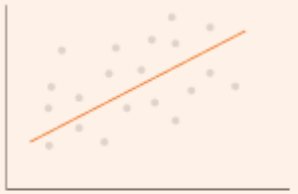
Data which has had statistical methods applied to it to protect against disclosing unauthorised information.

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.

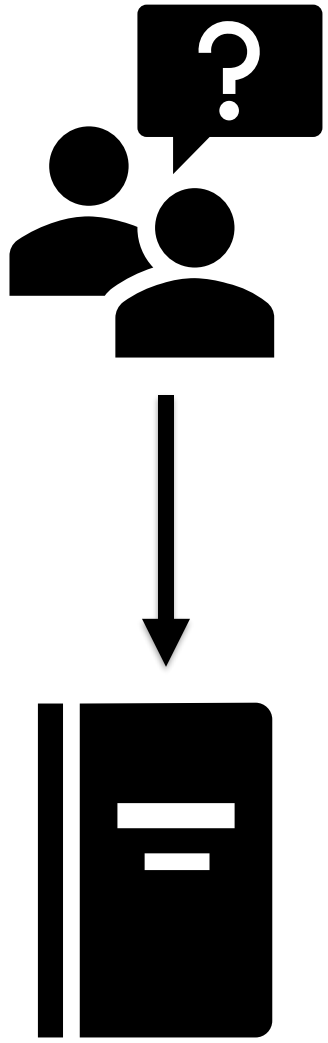
Key Moments In the research lifecycle

- Be aware of when Identifiable information is collected, Stored and shared
- What is the risk at these moments?
- Manage identifiable information to manage these risks against the utility of your data

Data Collection

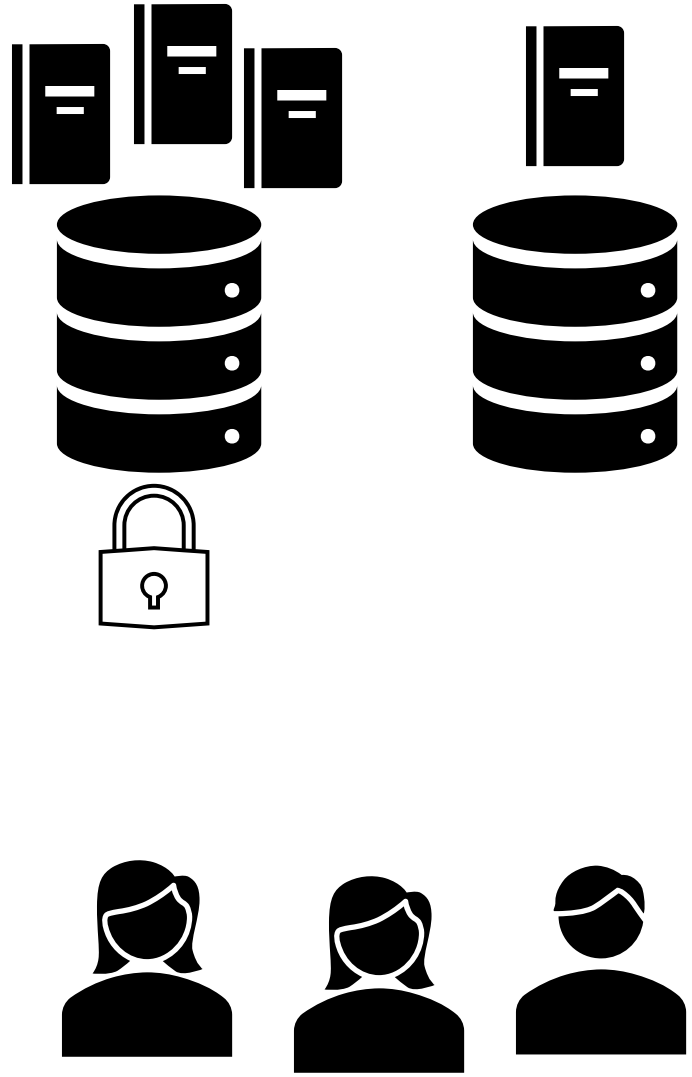
Minimize the identifiable data you collect

- **Collect only what you need** and only in the form you need it
  - e.g., if Names is not required do not collect or store names (use sample IDs instead)
- When transcribing audio use pseudonyms and remove revealing information from the recording
- Use manual and automatic validation to ensure accuracy
- Have clear and traceable collection metadata



Management and Storage

- Manage access to identifiable data
  - Control who can see or modify what based on requirements and risks
- Store multiple copies of the data with different levels of identifiability and access rights
- Use secure, resilient, trusted, and approved storage and systems
  - Encryption
  - Clear delineation of access roles
  - Audit trails of access and modifications



Access Controls

- 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
  - **Clear metadata and audit trails**

Sharing data with a collaborator

- When sharing data to an outside party mitigate risks by sharing only the information required for their use
- Provide a **de-identified** copy of the data for their use
- Assure how data will be handled in transit and in their care with appropriate agreements
- Consider other data the collaborator holds that could be used for re-identification when combined with yours
- Only share data for consented and approved purposes

Publishing your data

- When providing a public copy of your data re-identification risk and requirements to protect participants are much higher
- Data should be **confidential** (not able to be linked back to individuals) Remove **any and all identifying information**
- Use **statistical disclosure controls** to mitigate re-identification from quantitative data
- Consider public data sources (e.g., social media, databanks, other research) that could be linked with your data
- Perform and output check to ensure that re-identification risk is minimized
- If the full data cannot be shared share a metadata only record and access instruction

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

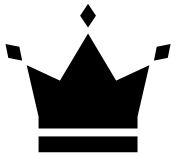
Identify stakeholders
Plan:
- how you will communicate
- what happens next once data is shared or released
- what to do if things go wrong

The De-Identification Decision-Making Framework

Methods and techniques for de-identification & confidentialisation

Example Datasets

First Name	Last Name	Date of Birth	Blood Type	RBC	Home Address
Abraham	Van Helsing	Feb 07 1833	O	6.1M cells/μL	145 Bramzeil, Amsterdam
Johnathan	Harker	Oct 15 1869	B-	4.5M cells/μL	138 Piccadilly, Green Park, London
Mina	Murray	Sep 12 1867	AB+	3.6M cells/μL	138 Piccadilly, Green Park, London
Lucy	Westenra	May 09 1871	AB	4.5M cells/μL	Castletown House, Celbridge, County Kildare
Arthur	Holmwood	Nov 23 1865	B-	5.1M cells/μL	73 The Green London
John	Seward	Apr 14 1861	B	4.9M cells/μL	551 Victoria Road, Carfax, Oxford
Quncey	Morris	Feb 25 1863	A+	4.7M cells/μL	440 Jackson Street, Dallas



Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
43	10
50	4
310	1

What methods could you use to protect the identity of individuals in this dataset?

Methods

- **Tokenisation –** Replacing identifiers with non-identifying IDs
- **Encryption –** Algorithmically control reading of values
- **Generalization/Aggregation -** Grouping values
- **Suppression –** Hiding or removing values
- **Perturbation –** Randomly altering values
- **Synthetic data –** statistically synthesizing values

Replacing values with Identifiers

- 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

Replacing values with Identifiers

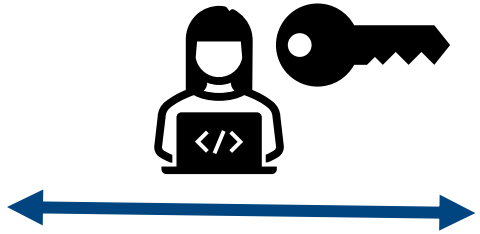
First Name	Last Name	Date of Birth	Blood Type	RBC	Home Address
Abraham	**Van Helsing**	Feb 07 1833	O	6.1M cells/µL	145 Bramzeil, Amsterdam
Johnathan	**Harker**	Oct 15 1869	B-	4.5M cells/µL	138 Piccadilly, Green Park, London
Mina	**Murray**	Sep 12 1867	AB+	3.6M cells/µL	138 Piccadilly, Green Park, London
Lucy	**Westenra**	May 09 1871	AB	4.5M cells/µL	Castletown House, Celbridge, County Kildare
Arthur	**Holmwood**	Nov 23 1865	B-	5.1M cells/µL	73 The Green London
John	**Seward**	Apr 14 1861	B	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
Quncey	**Morris**	Feb 25 1863	A+	4.7M cells/µL	440 Jackson Street, Dallas

SampleID	Date of Birth	Blood Type	RBC	Home Address
PID0001	Feb 07 1833	O	6.1M cells/µL	145 Bramzeil, Amsterdam
PID0002	Oct 15 1869	B-	4.5M cells/µL	138 Piccadilly, Green Park, London
PID0003	Sep 12 1867	AB+	3.6M cells/µL	138 Piccadilly, Green Park, London
PID0004	May 09 1871	AB	4.5M cells/µL	Castletown House, Celbridge, County Kildare
PID0005	Nov 23 1865	B-	5.1M cells/µL	73 The Green London
PID0006	Apr 14 1861	B	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
PID0007	Feb 25 1863	A+	4.7M cells/µL	440 Jackson Street, Dallas

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

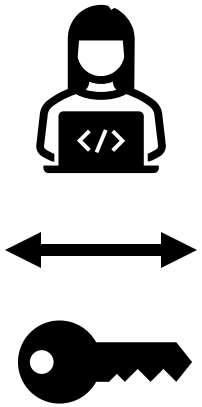
Name
Jane Doe
John Smith
Alan Doe
Mary Doe



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

Encryption

First Name	Last Name	Date of Birth	Blood Type	RBC	Home Address
Abraham	**Van Helsing**	Feb 07 1833	O	6.1M cells/µL	145 Bramzeil, Amsterdam
Johnathan	**Harker**	Oct 15 1869	B-	4.5M cells/µL	138 Piccadilly, Green Park, London
Mina	**Murray**	Sep 12 1867	AB+	3.6M cells/µL	138 Piccadilly, Green Park, London
Lucy	**Westenra**	May 09 1871	AB	4.5M cells/µL	Castletown House, Celbridge, County Kildare
Arthur	**Holmwood**	Nov 23 1865	B-	5.1M cells/µL	73 The Green London
John	**Seward**	Apr 14 1861	B	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
Quncey	**Morris**	Feb 25 1863	A+	4.7M cells/µL	440 Jackson Street, Dallas



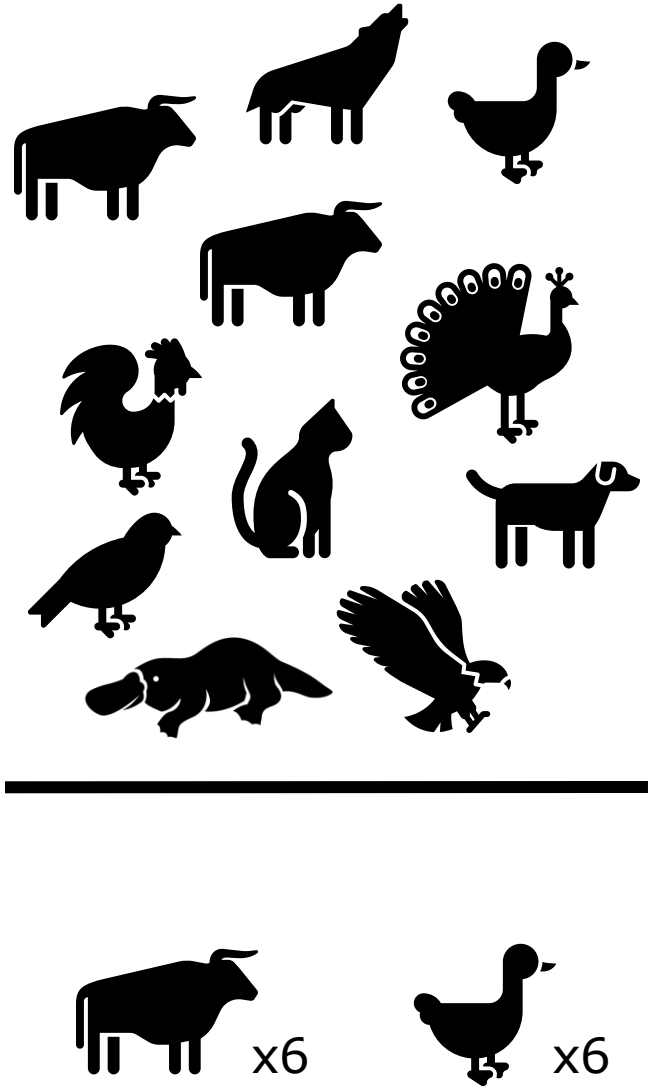
Name Encrypted	Date of Birth	Blood Type	RBC	Home Address
!^Kc7M %dM6tx	Feb 07 1833	O	6.1M cells/µL	145 Bramzeil, Amsterdam
6*Aqih 2mh9F!	Oct 15 1869	B-	4.5M cells/µL	138 Piccadilly, Green Park, London
S319qKE!C rteyytvl	Sep 12 1867	AB+	3.6M cells/µL	138 Piccadilly, Green Park, London
%H2ECyfQ GYGiX8m4 R	May 09 1871	AB	4.5M cells/µL	Castletown House, Celbridge, County Kildare
zwBuf8zU# Zxa#Ro&6	Nov 23 1865	B-	5.1M cells/µL	73 The Green London
^rab%@w TFOE^91H xI	Apr 14 1861	B	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
@j#Cf&5S #U@G&b% Rm	Feb 25 1863	A+	4.7M cells/µL	440 Jackson Street, Dallas

Aggregation

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

- A type of **Recoding**
- 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**~~

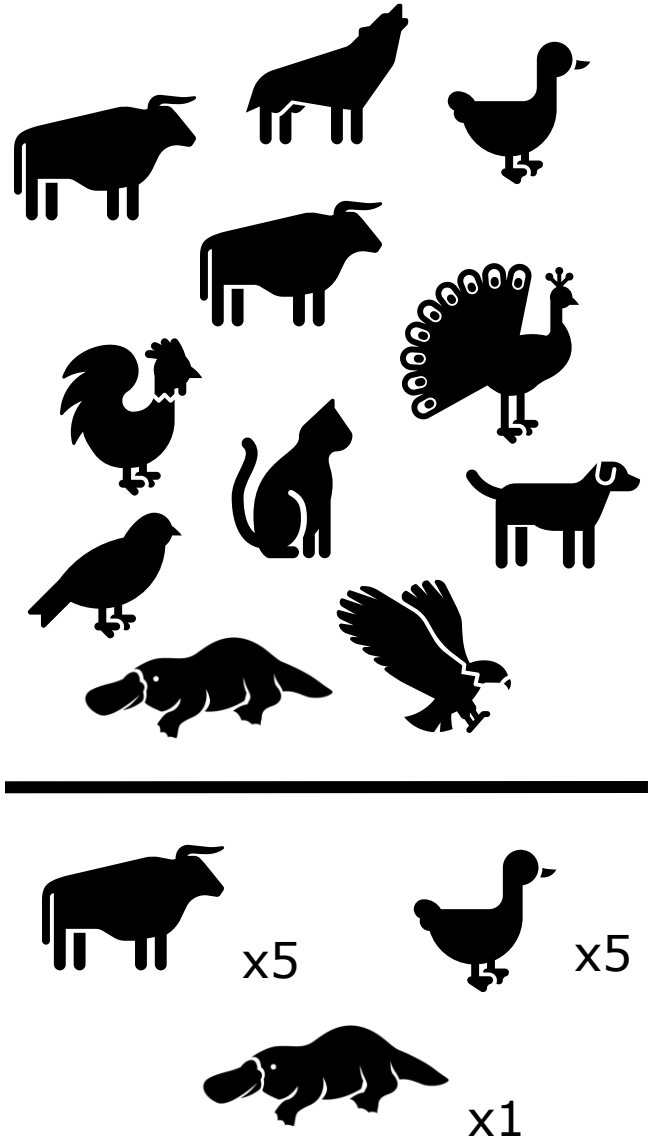
https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse



Aggregation

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

- A type of **Recoding**
- 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**

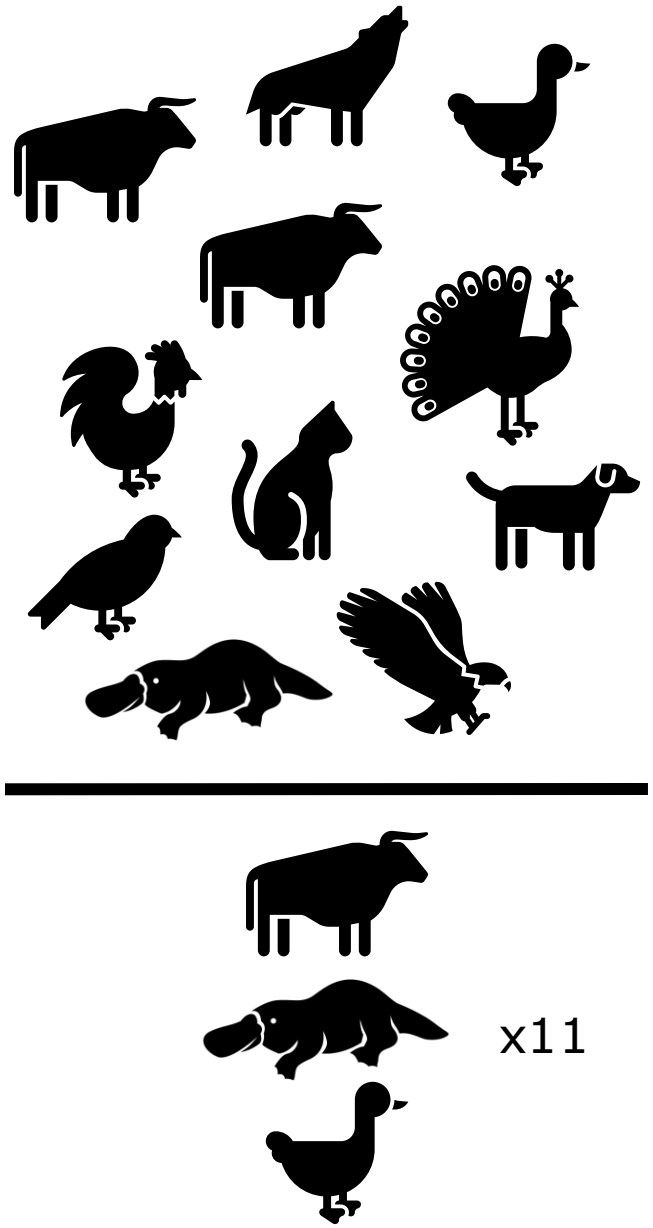


https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Aggregation

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

- A type of **Recoding**
- 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

Global recoding

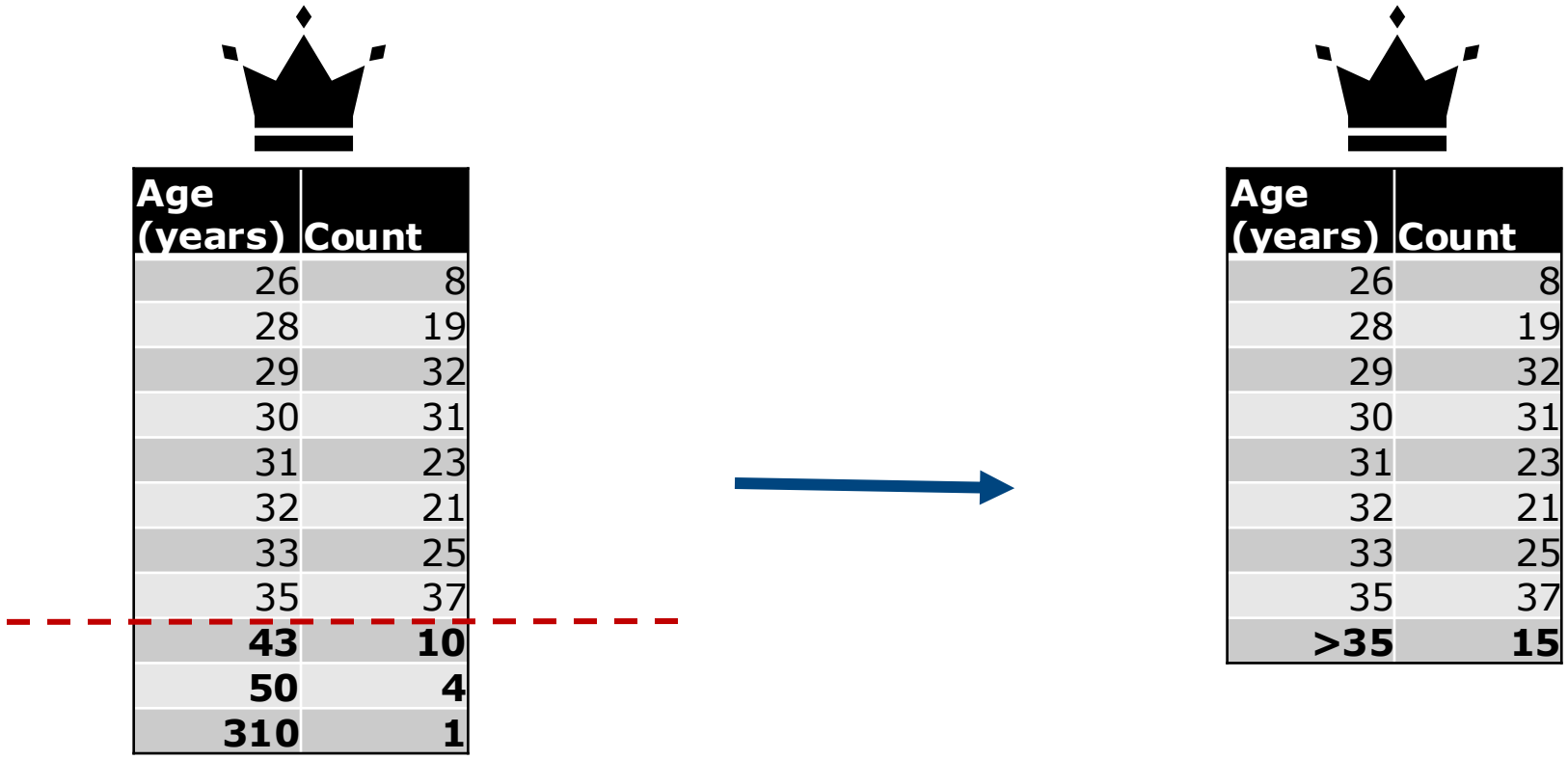
Name Encrypted	Date of Birth	Blood Type	RBC	Home Address
!^Kc7M %dM6tx	Feb 07 1833	**O**	6.1M cells/µL	145 Bramzeil, Amsterdam
6*Aqih 2mh9F!	Oct 15 1869	**B-**	4.5M cells/µL	138 Piccadilly, Green Park, London
S319qKE!Cr teyytvl	Sep 12 1867	**AB+**	3.6M cells/µL	138 Piccadilly, Green Park, London
%H2ECyfQG YGiX8m4R	May 09 1871	**AB**	4.5M cells/µL	Castletown House, Celbridge, County Kildare
zwBuf8zU#Z xa#Ro&6	Nov 23 1865	**B-**	5.1M cells/µL	73 The Green London
^rab%@wT FOE^91HxI	Apr 14 1861	**B**	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
@j#Cf&5S# U@G&b%R m	Feb 25 1863	**A+**	4.7M cells/µL	440 Jackson Street, Dallas

Blood Type	Count
O	1
B	5
A	3

Blood Type	Count
O	1
B	3
AB	2
A	1

Categorization/Aggregation

Top and bottom coding

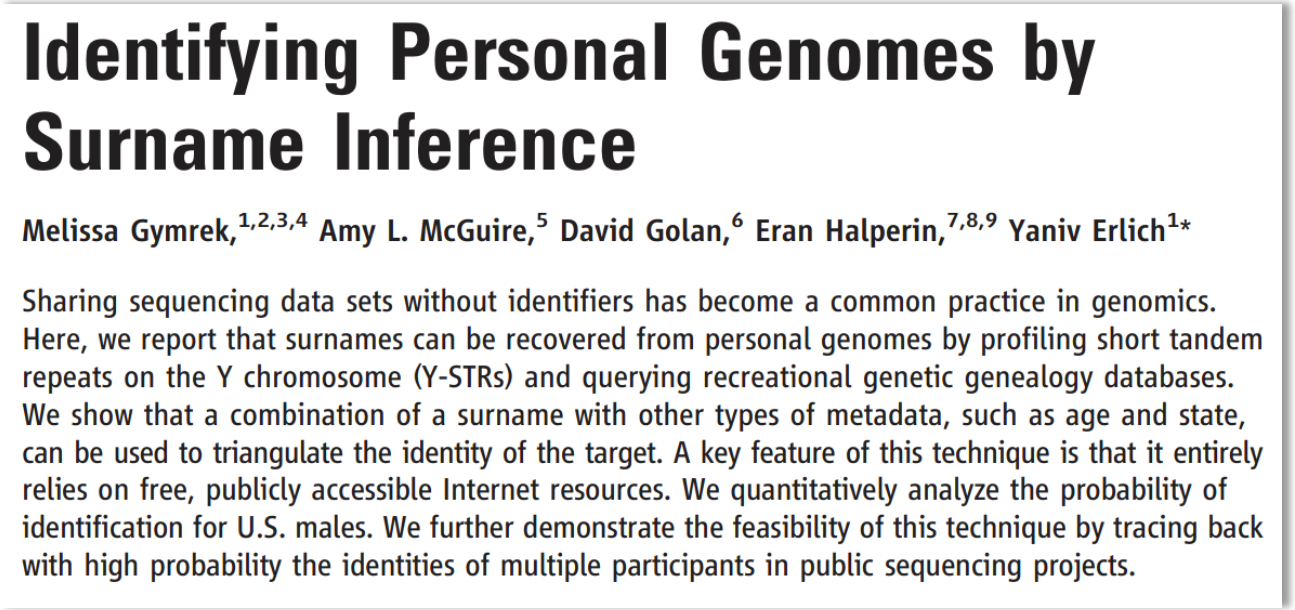


Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
43	**10**
50	**4**
310	**1**

Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
>35	**15**

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

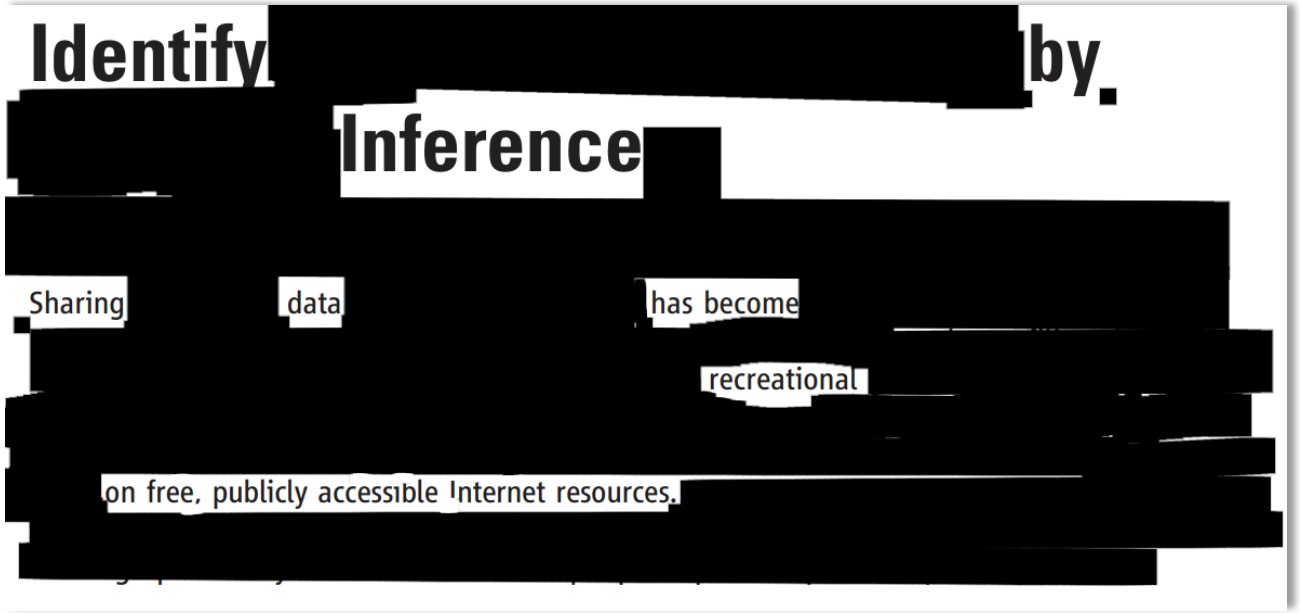
Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

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

Suppression

Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
43	**10**
50	**4**
310	**1**

Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
>35	**Suppressed**

Suppression

Name Encrypted	Date of Birth	Blood Type	RBC	Home Address
!^Kc7M %dM6tx	Feb 07 1833	**O**	6.1M cells/µL	145 Bramzeil, Amsterdam
6*Aqih 2mh9F!	Oct 15 1869	**B-**	4.5M cells/µL	138 Piccadilly, Green Park, London
S319qKE!Crt eyytvl	Sep 12 1867	**AB+**	3.6M cells/µL	138 Piccadilly, Green Park, London
%H2ECyfQG YGiX8m4R	May 09 1871	**AB**	4.5M cells/µL	Castletown House, Celbridge, County Kildare
zwBuf8zU#Z xa#Ro&6	Nov 23 1865	**B-**	5.1M cells/µL	73 The Green London
^rab%@wTF OE^91HxI	Apr 14 1861	**B**	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
@j#Cf&5S# U@G&b%Rm	Feb 25 1863	**A+**	4.7M cells/µL	440 Jackson Street, Dallas

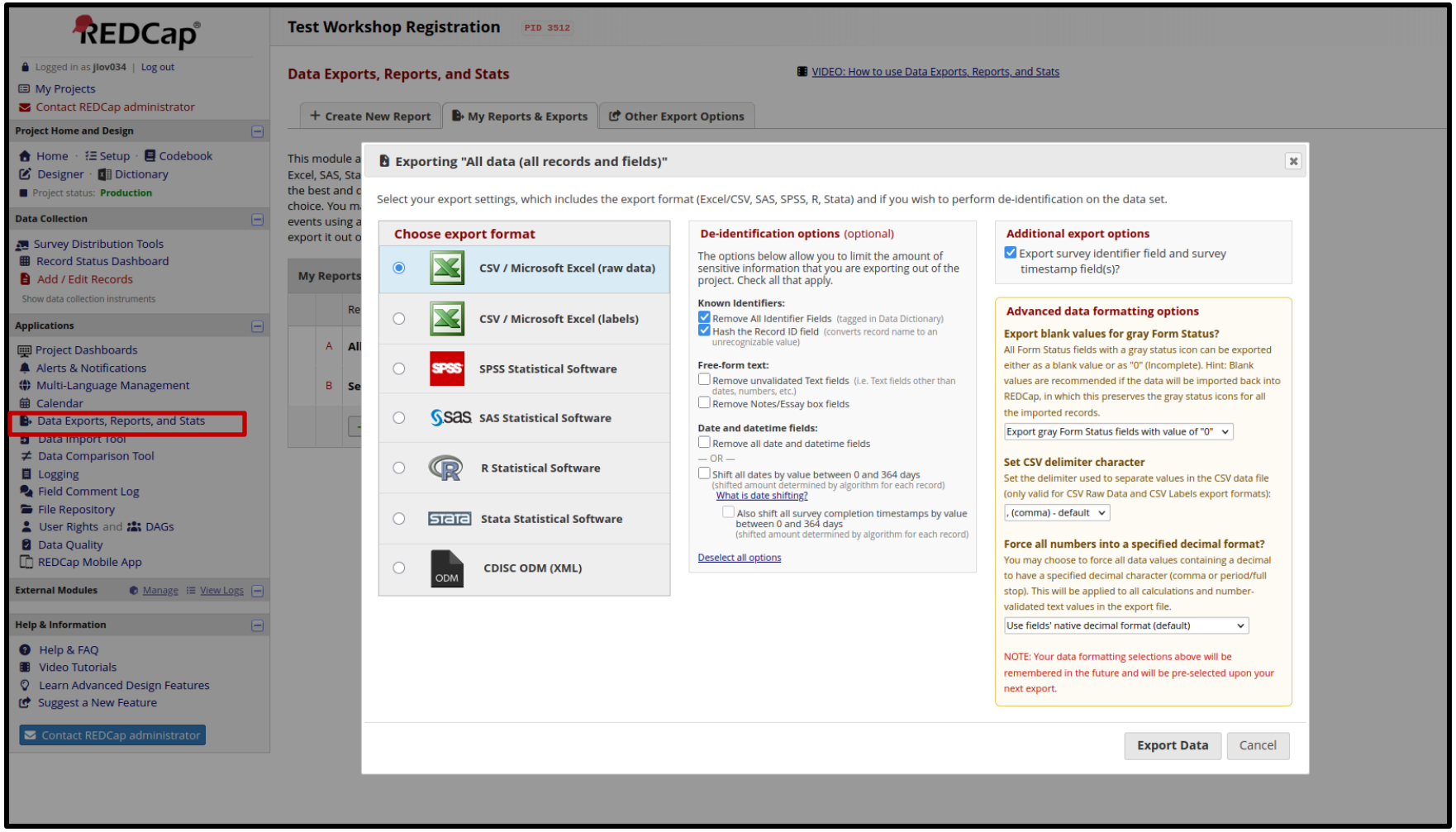
Name Encrypted	Date of Birth	Blood Type	RBC	Home Address
!^Kc7M %dM6tx	Feb 07 1833	**Suppressed**	6.1M cells/µL	145 Bramzeil, Amsterdam
6*Aqih 2mh9F!	Oct 15 1869	**Suppressed**	4.5M cells/µL	138 Piccadilly, Green Park, London
S319qKE!Cr teyytvl	Sep 12 1867	**Suppressed**	3.6M cells/µL	138 Piccadilly, Green Park, London
%H2ECyfQG YGiX8m4R	May 09 1871	**Suppressed**	4.5M cells/µL	Castletown House, Celbridge, County Kildare
zwBuf8zU#Z xa#Ro&6	Nov 23 1865	**Suppressed**	5.1M cells/µL	73 The Green London
^rab%@wT FOE^91HxI	Apr 14 1861	**Suppressed**	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
@j#Cf&5S# U@G&b%R m	Feb 25 1863	**Suppressed**	4.7M cells/µL	440 Jackson Street, Dallas

REDCap

- De-identification of marked Identifier Fields upon export
- Removal of free-text

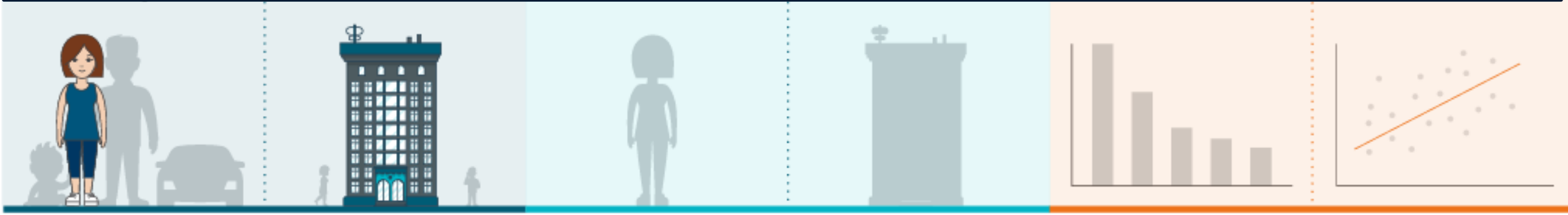
Useful for sharing with collaborators.

Remember to use Access controls.



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** **Expenditure**	34 $398,000	**Address**	Postcode 6012 Wellington	**Employees** **Expenditure**	30 - 40 $398,000	**Address**	Wellington	**Employees** **Expenditure**	10 - 100 Under $500,000



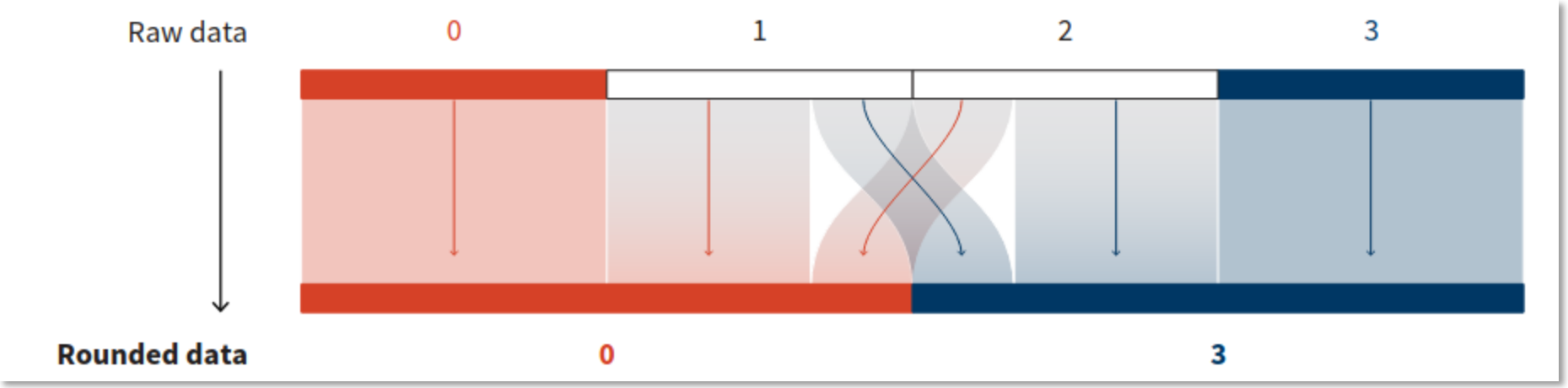
Generalisation

Name Encrypted	Date of Birth	Blood Type	RBC	Home Address
!^Kc7M %dM6tx	**Feb 07 1833**	O	6.1M cells/µL	**145 Bramzeil, Amsterdam**
6*Aqih 2mh9F!	**Oct 15 1869**	B-	4.5M cells/µL	**138 Piccadilly, Green Park, London**
S319qKE!Crt eyytvl	**Sep 12 1867**	AB+	3.6M cells/µL	**138 Piccadilly, Green Park, London**
%H2ECyfQG YGiX8m4R	**May 09 1871**	AB	4.5M cells/µL	**Castletown House, Celbridge, County Kildare**
zwBuf8zU#Z xa#Ro&6	**Nov 23 1865**	B-	5.1M cells/µL	**73 The Green London**
^rab%@wTF OE^91HxI	**Apr 14 1861**	B	4.9M cells/µL	**551 Victoria Road, Carfax, Oxford**
@j#Cf&5S# U@G&b%Rm	**Feb 25 1863**	A+	4.7M cells/µL	**440 Jackson Street, Dallas**

Name Encrypted	Age	Blood Type	RBC	Home City
!^Kc7M %dM6tx	**57**	Suppressed	6.1M cells/µL	**Amsterdam**
6*Aqih 2mh9F!	**21**	Suppressed	4.5M cells/µL	**London**
S319qKE!Cr teyytvl	**23**	Suppressed	3.6M cells/µL	**London**
%H2ECyfQG YGiX8m4R	**19**	Suppressed	4.5M cells/µL	**Kildare**
zwBuf8zU#Z xa#Ro&6	**25**	Suppressed	5.1M cells/µL	**London**
^rab%@wT FOE^91HxI	**29**	Suppressed	4.9M cells/µL	**Oxford**
@j#Cf&5S# U@G&b%R m	**27**	Suppressed	4.7M cells/µL	**Dallas**

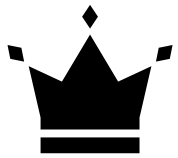
Perturbation

- 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.

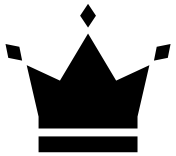


https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/manage-a-privacy-programme/making-personal-information-safe-for-reuse

Perturbation



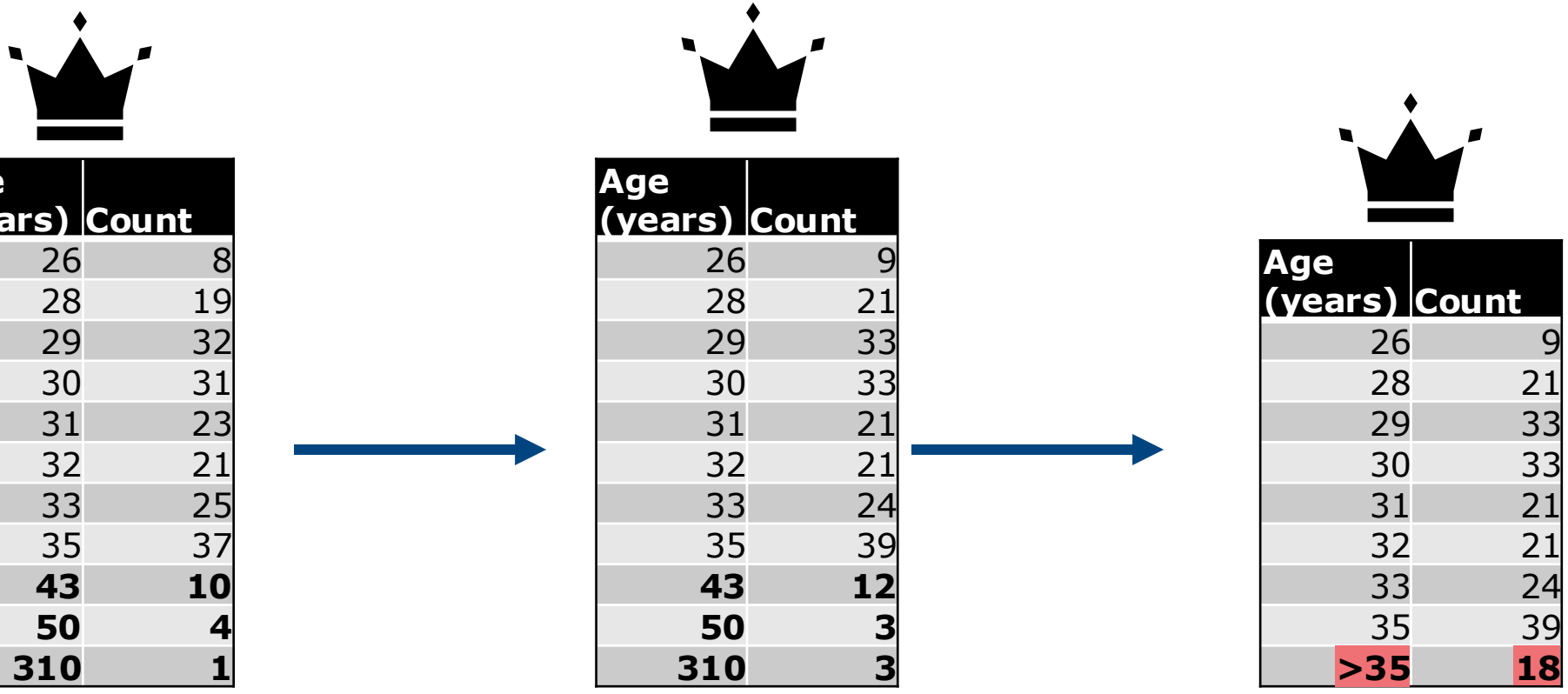
Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
43	**10**
50	**4**
310	**1**



Age (years)	Count
26	9
28	21
29	33
30	33
31	21
32	21
33	24
35	39
43	**12**
50	**3**
310	**3**

Order of operations matters

Perturbation then aggregation

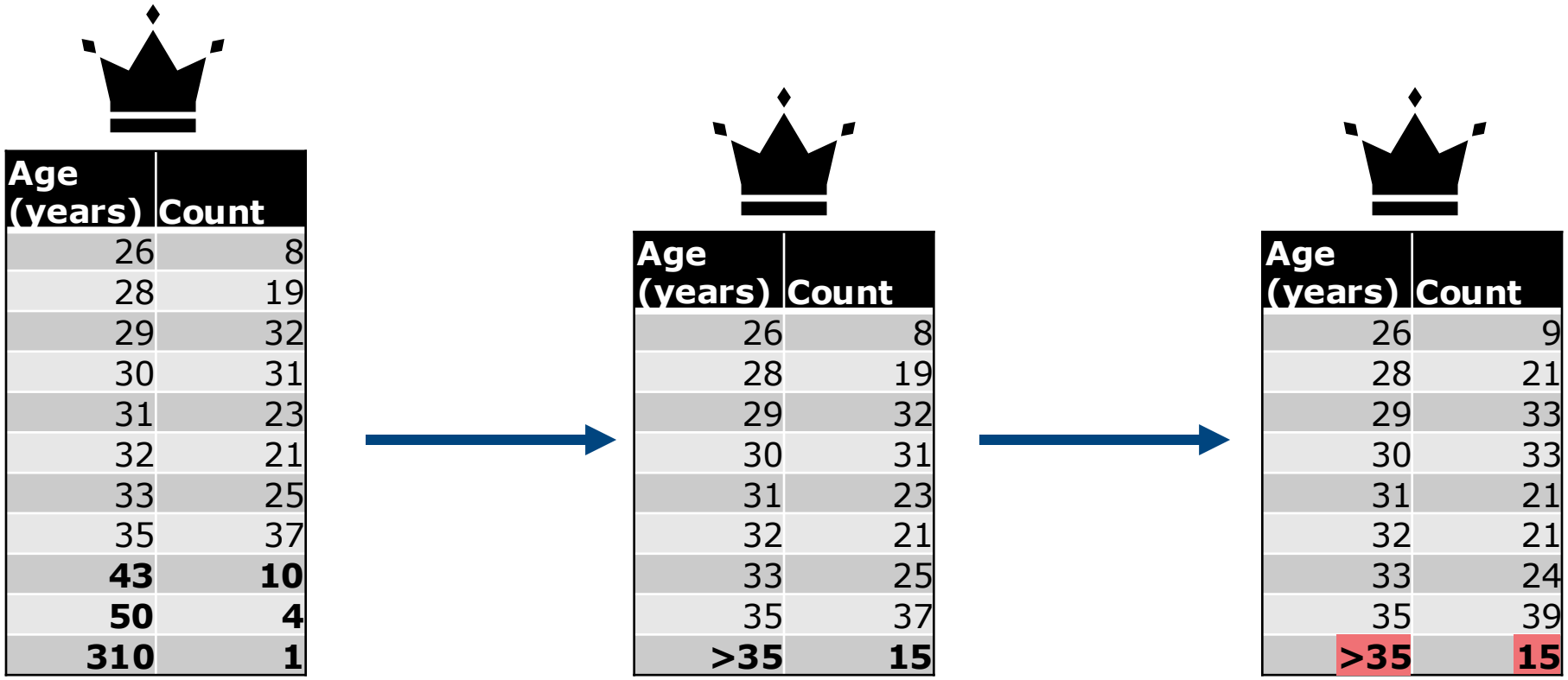


Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
43	**10**
50	**4**
310	**1**

Age (years)	Count
26	9
28	21
29	33
30	33
31	21
32	21
33	24
35	39
43	**12**
50	**3**
310	**3**

Age (years)	Count
26	9
28	21
29	33
30	33
31	21
32	21
33	24
35	39
>35	**18**

Aggregation then perturbation



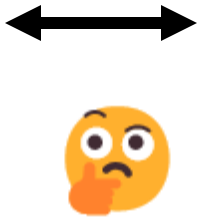
Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
43	**10**
50	**4**
310	**1**

Age (years)	Count
26	8
28	19
29	32
30	31
31	23
32	21
33	25
35	37
>35	**15**

Age (years)	Count
26	9
28	21
29	33
30	33
31	21
32	21
33	24
35	39
>35	**15**

Perturbation loses information

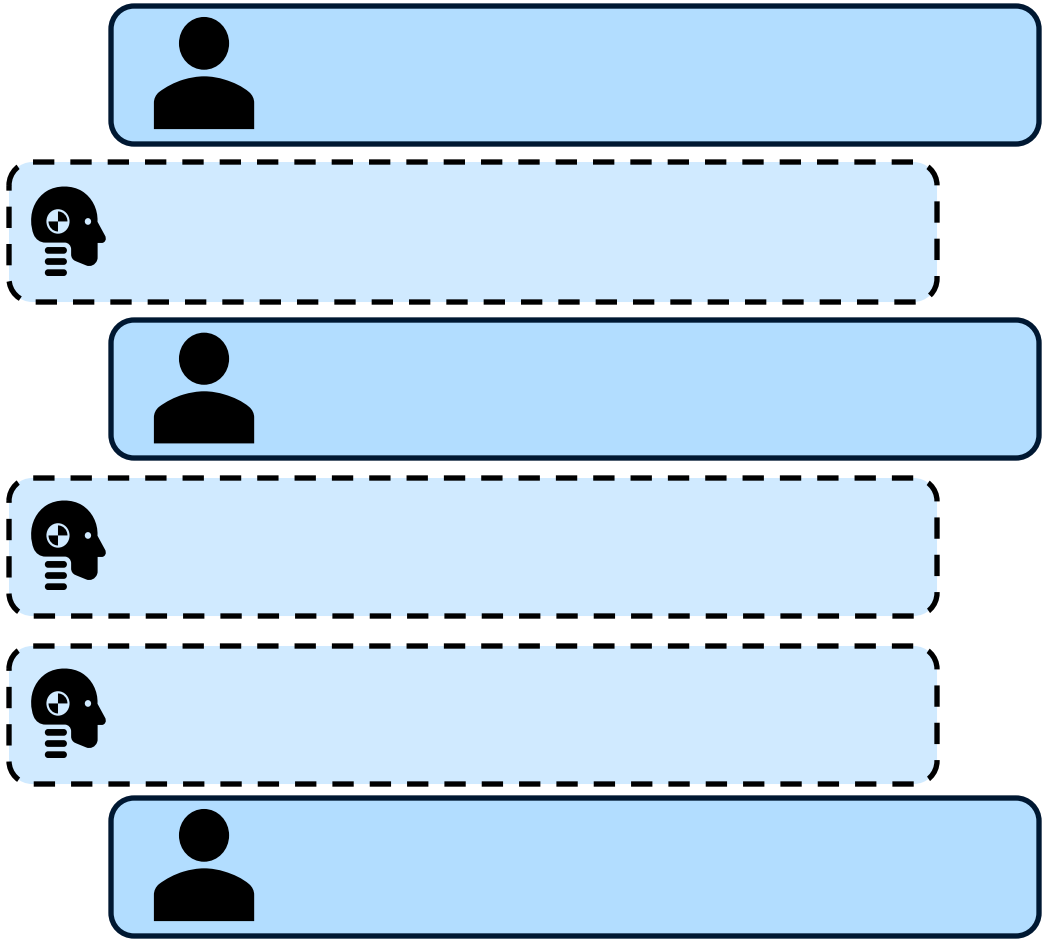
Name Encrypted	Age	Blood Type	RBC	Home city
!^Kc7M %dM6tx	57	Suppressed	**6.1M cells/µL**	Amsterdam
6*Aqih 2mh9F!	21	Suppressed	**4.5M cells/µL**	London
S319qKE!Crt eyytvl	23	Suppressed	**3.6M cells/µL**	London
%H2ECyfQG YGiX8m4R	19	Suppressed	**4.5M cells/µL**	Kildare
zwBuf8zU#Z xa#Ro&6	25	Suppressed	**5.1M cells/µL**	London
^rab%@wTF OE^91HxI	29	Suppressed	**4.9M cells/µL**	Oxford
@j#Cf&5S# U@G&b%Rm	27	Suppressed	**4.7M cells/µL**	Dallas



Name Encrypted	Age	Blood Type	RBC	Home city
!^Kc7M %dM6tx	57	Suppressed	**6M cells/µL**	Amsterdam
6*Aqih 2mh9F!	21	Suppressed	**5M cells/µL**	London
S319qKE!Cr teyytvl	23	Suppressed	**3M cells/µL**	London
%H2ECyfQG YGiX8m4R	19	Suppressed	**4M cells/µL**	Kildare
zwBuf8zU#Z xa#Ro&6	25	Suppressed	**5M cells/µL**	London
^rab%@wT FOE^91HxI	29	Suppressed	**4M cells/µL**	Oxford
@j#Cf&5S# U@G&b%R m	27	Suppressed	**5M cells/µL**	Dallas

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

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)
- Measure **utility** lost as well as risk
  - e.g., change in Summary Statistics or record counts
  - Consider user needs

This data is still unique!

First Name	Last Name	Date of Birth	Blood Type	RBC	Home Address
Van	**Helsing**	Feb 07 1833	O	6.1M cells/µL	145 Bramzeil, Amsterdam
Johnathan	**Harker**	Oct 15 1869	B-	4.5M cells/µL	138 Piccadilly, Green Park, London
Mina	**Murray**	Sep 12 1867	AB+	3.6M cells/µL	138 Piccadilly, Green Park, London
Lucy	**Westenra**	May 09 1871	AB	4.5M cells/µL	Castletown House, Celbridge, County Kildare
Arthur	**Holmwood**	Nov 23 1865	B-	5.1M cells/µL	73 The Green London
John	**Seward**	Apr 14 1861	B	4.9M cells/µL	551 Victoria Road, Carfax, Oxford
Quncey	**Morris**	Feb 25 1863	A+	4.7M cells/µL	440 Jackson Street, Dallas

Name Encrypted	Age	Blood Type	RBC	Home city
!^Kc7M %dM6tx	57	Suppressed	**6M cells/µL**	Amsterdam
6*Aqih 2mh9F!	21	Suppressed	**5M cells/µL**	London
S319qKE!Cr teyytvl	23	Suppressed	**3M cells/µL**	London
%H2ECyfQG YGiX8m4R	19	Suppressed	**4M cells/µL**	Kildare
zwBuf8zU#Z xa#Ro&6	25	Suppressed	**5M cells/µL**	London
^rab%@wT FOE^91HxI	29	Suppressed	**4M cells/µL**	Oxford
@j#Cf&5S# U@G&b%R m	27	Suppressed	**5M cells/µL**	Dallas

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

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 (names) **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

Initial Data

	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

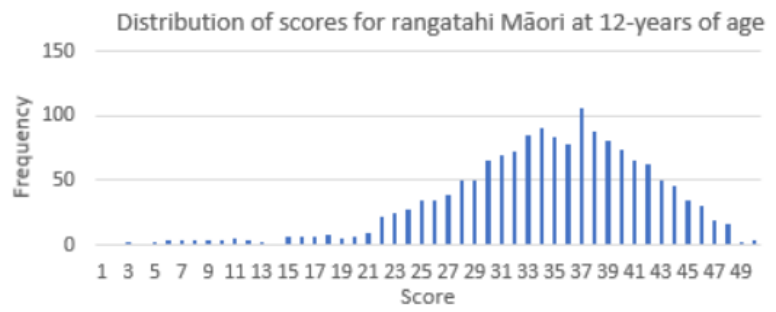
Aggregate

	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

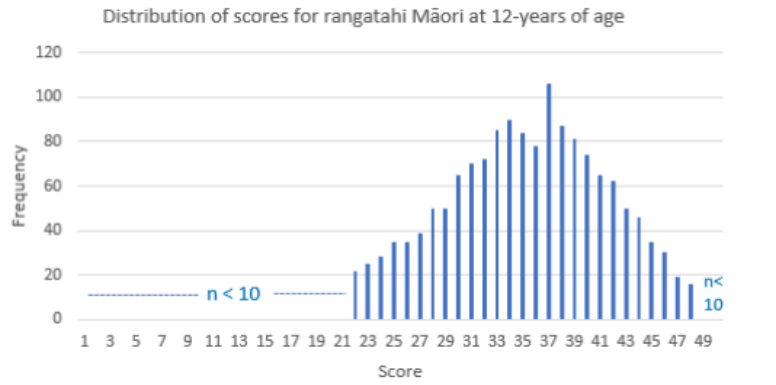
Perturbate

	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



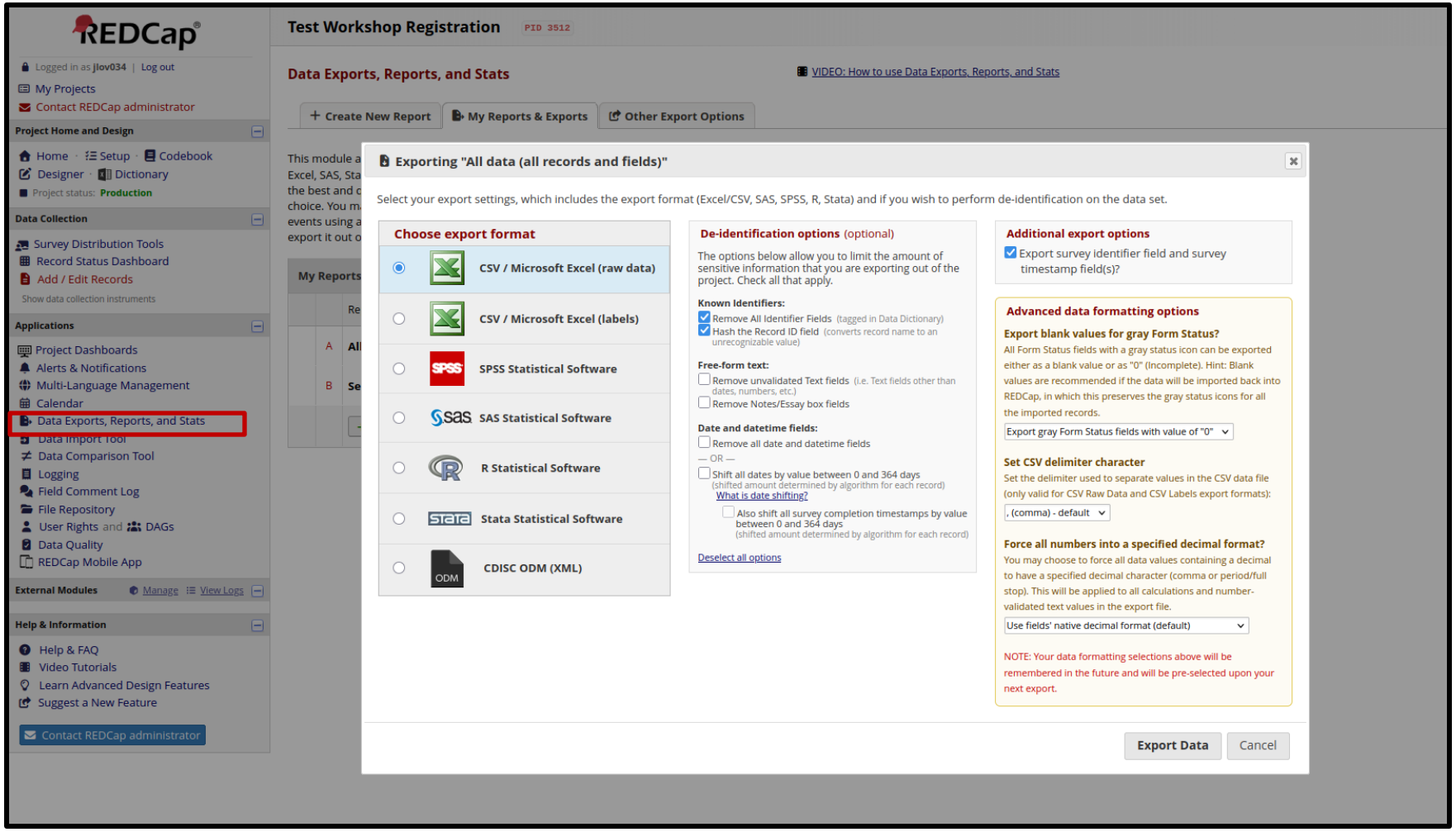
Tools for de-identification and evaluation

REDCap

- De-identification of marked Identifier Fields upon export
- Removal of free-text

Useful for sharing with collaborators.

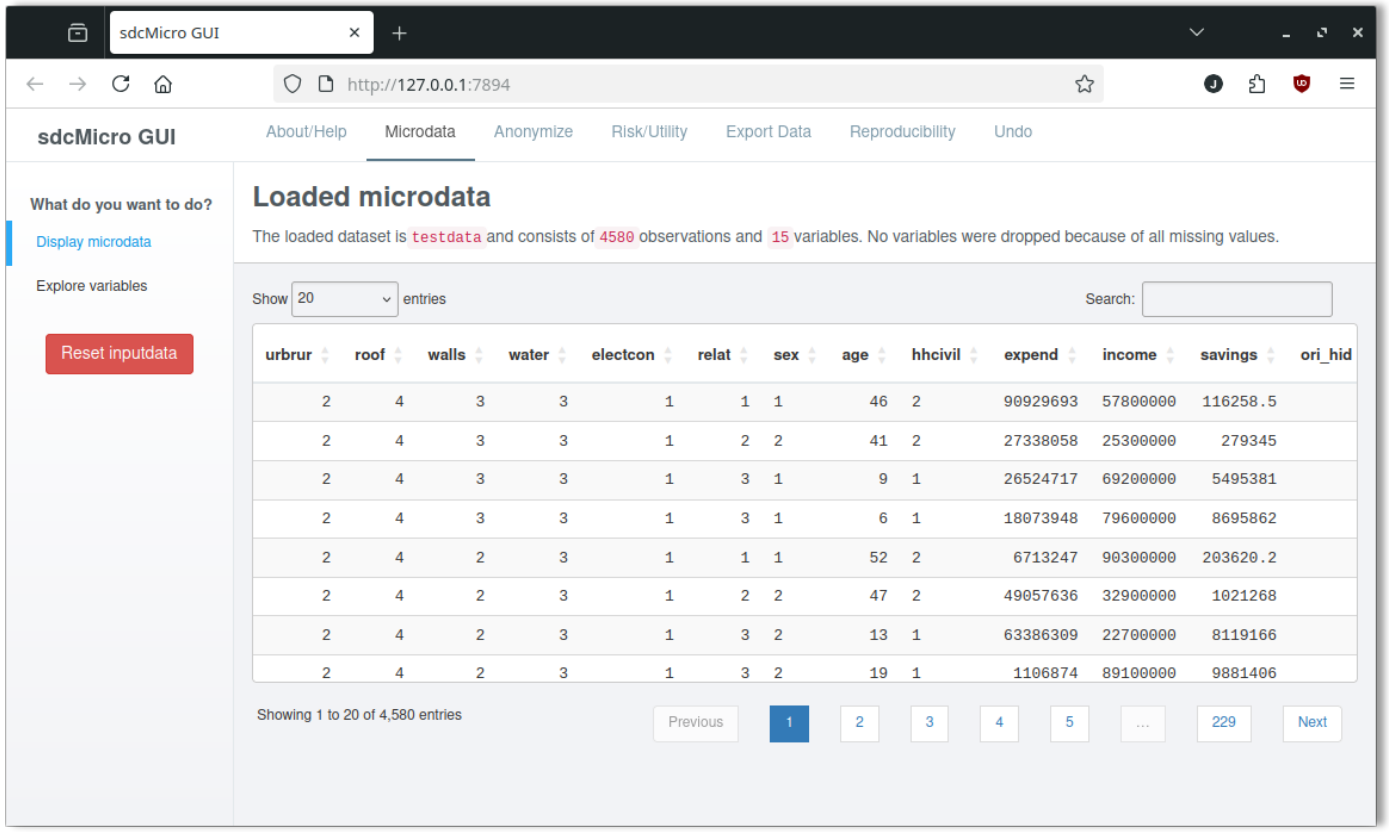
Remember to use Access controls.



sdcTable/sdcApp (R)

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

```
install.packages('sdcMicro')
library(sdcMicro)
sdcApp()
```

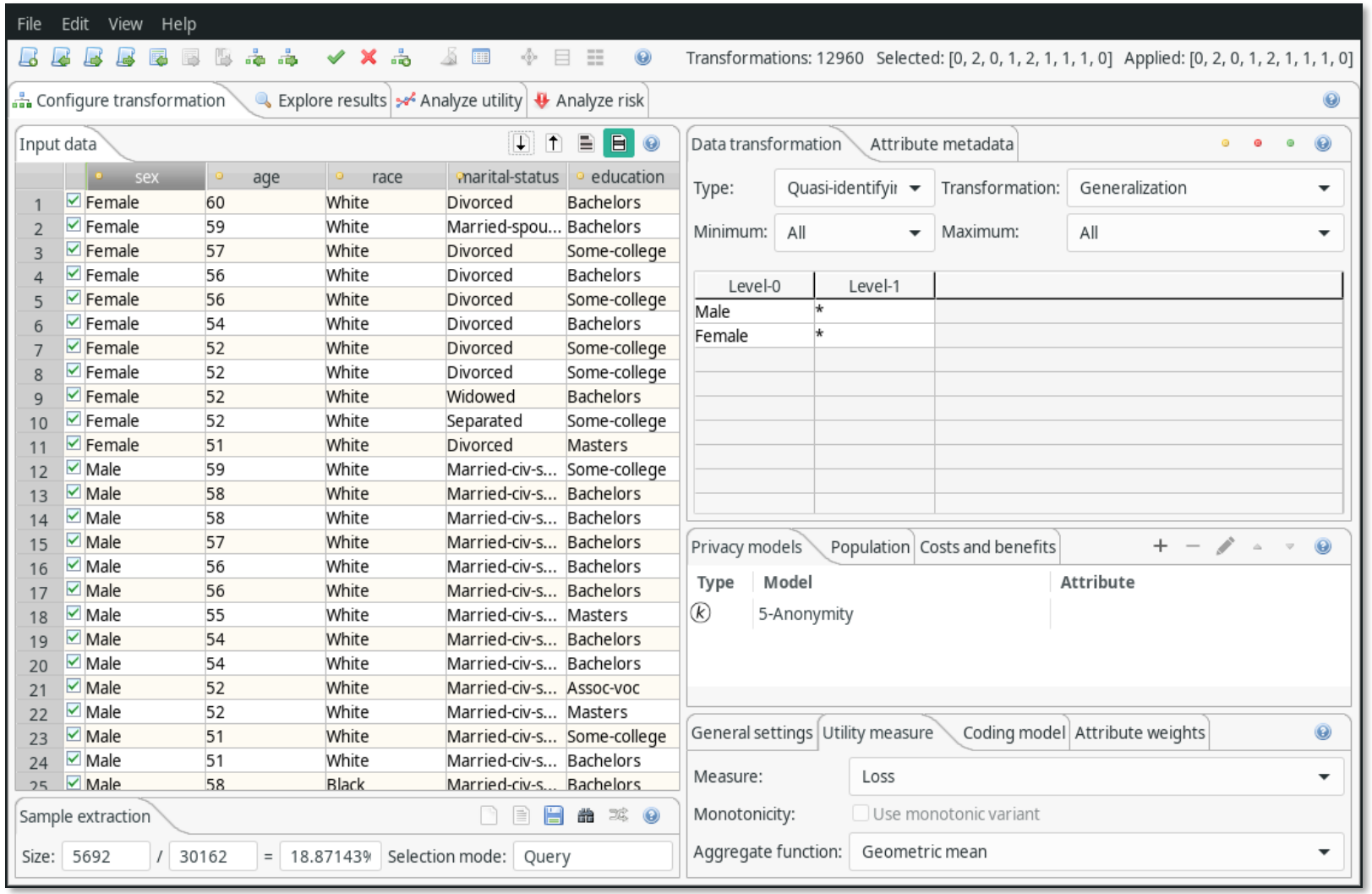


- **sdcTable** Vignette
- SDC Practice guide

ARX Data Anonymization Tool

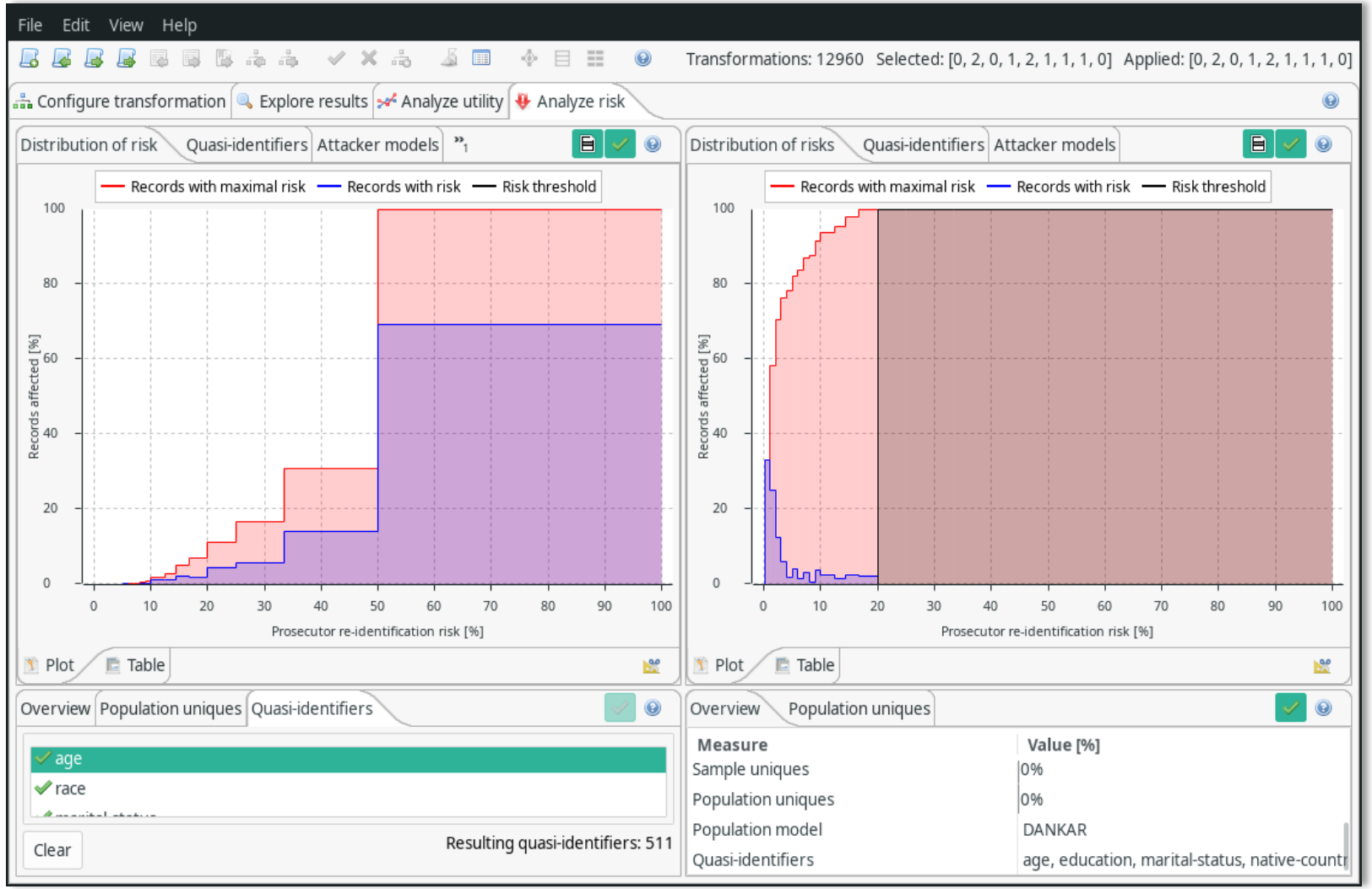
- Gui interface for de-identifying datasets
- Feature rich (perhaps a bit too much so)
- Customize hierarchies, privacy models, sensitivity

[ARX Download]



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



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**
- **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.

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 at each stage of the research lifecycle

Apply de-identification

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

Waipapa
Taumata Rau
University
of Auckland

Thank you!

Jean Love

June 2026

ResBaz
RESEARCH BAZAAR