# Working with sensitive data and the use of secure data environments

**Dr. Rajeev Samarage**

Melbourne Institute: Applied Economic & Social Research,
The Faculty of Business and Economics
The University of Melbourne

HASS RDC and IRC Computational Skill Summer School, 7-8 February 2023, Rydges World Square, Sydney

THE UNIVERSITY OF MELBOURNE

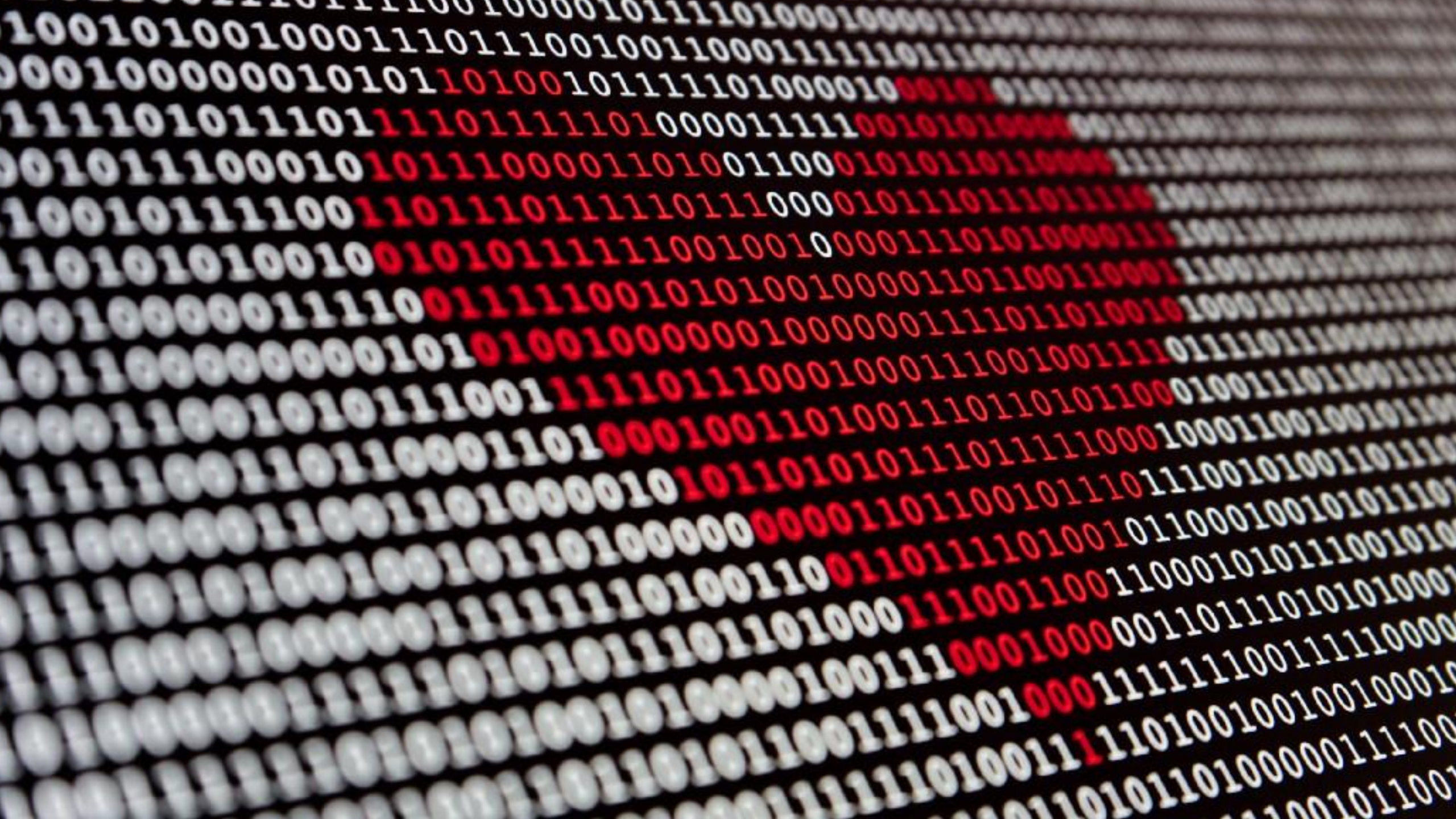
MELBOURNE INSTITUTE
APPLIED ECONOMIC & SOCIAL RESEARCH

# Acknowledgement of country

I acknowledge the Traditional Custodians of the land on which we gather, the Gadigal people of the Eora Nation.

I recognise their continued connection to the land and waters of this beautiful place, and acknowledge that they never ceded sovereignty.

I respect all Wurundjeri Woi Worrung Elders and Ancestors, and any First Nations people here today.



# Privacy Act 1988

**Let's start with some definitions.**

***personal information*** means information or an opinion about an identified individual, or an individual who is reasonably identifiable:

(a) whether the information or opinion is true or not; and

(b) whether the information or opinion is recorded in a material form or not.

- an individual's name, signature, address, phone number or date of birth
- employee record information
- photographs
- internet protocol (IP) addresses
- voice print and facial recognition biometrics (because they collect characteristics that make an individual's voice or face unique)
- location information from a mobile device (because it can reveal user activity patterns and habits)

***sensitive information*** means:

(a) information or an opinion about an individual's:

(i) racial or ethnic origin; or

(ii) political opinions; or

(iii) membership of a political association; or

(iv) religious beliefs or affiliations; or

(v) philosophical beliefs; or

(vi) membership of a professional or trade association; or

(vii) membership of a trade union; or

(viii) sexual orientation or practices; or

(ix) criminal record;

that is also personal information; or

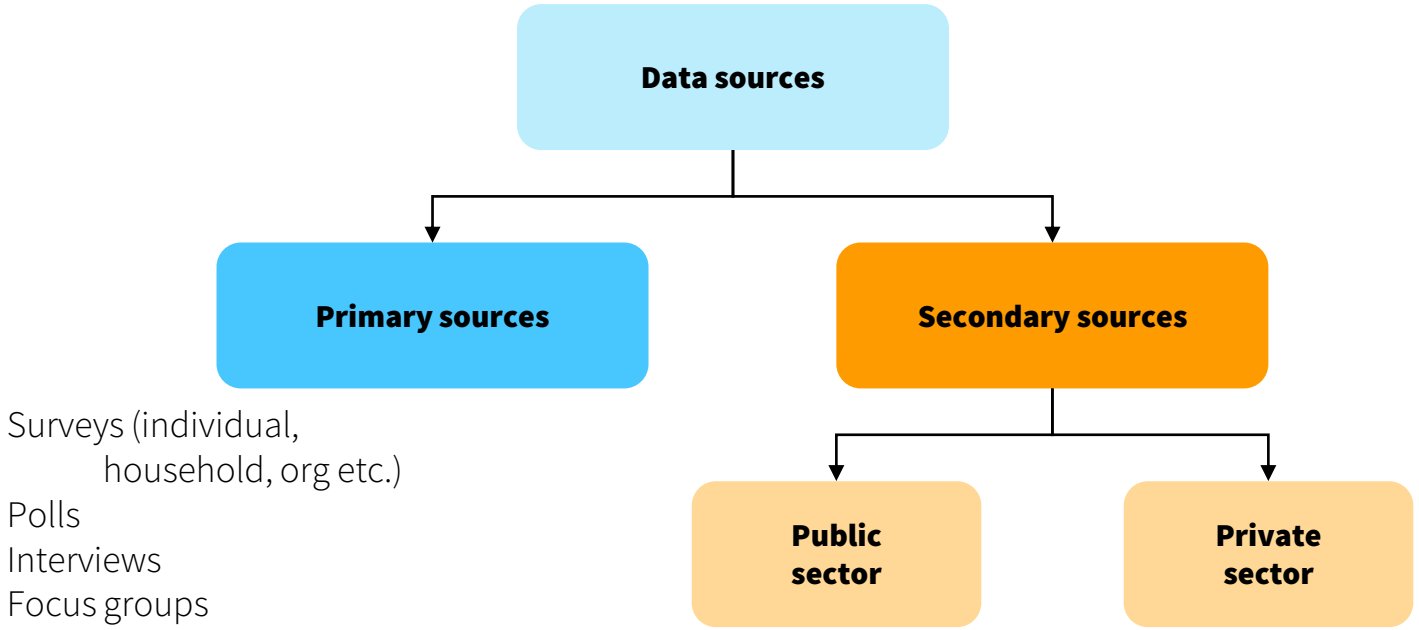
(b) health information about an individual; or

(c) genetic information about an individual that is not otherwise health information; or

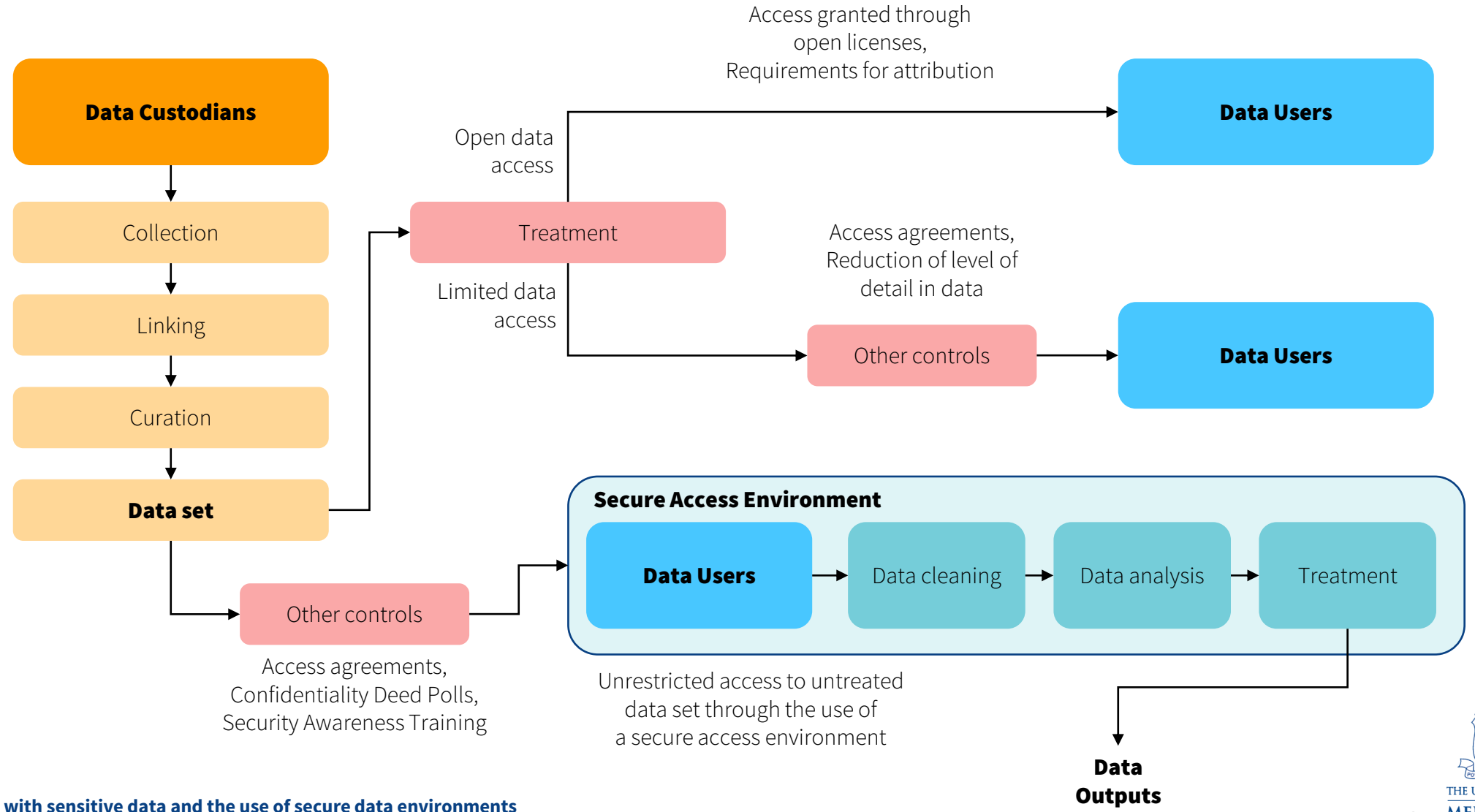
(d) biometric information that is to be used for the purpose of automated biometric verification or biometric identification; or

(e) biometric templates.

# What about data?



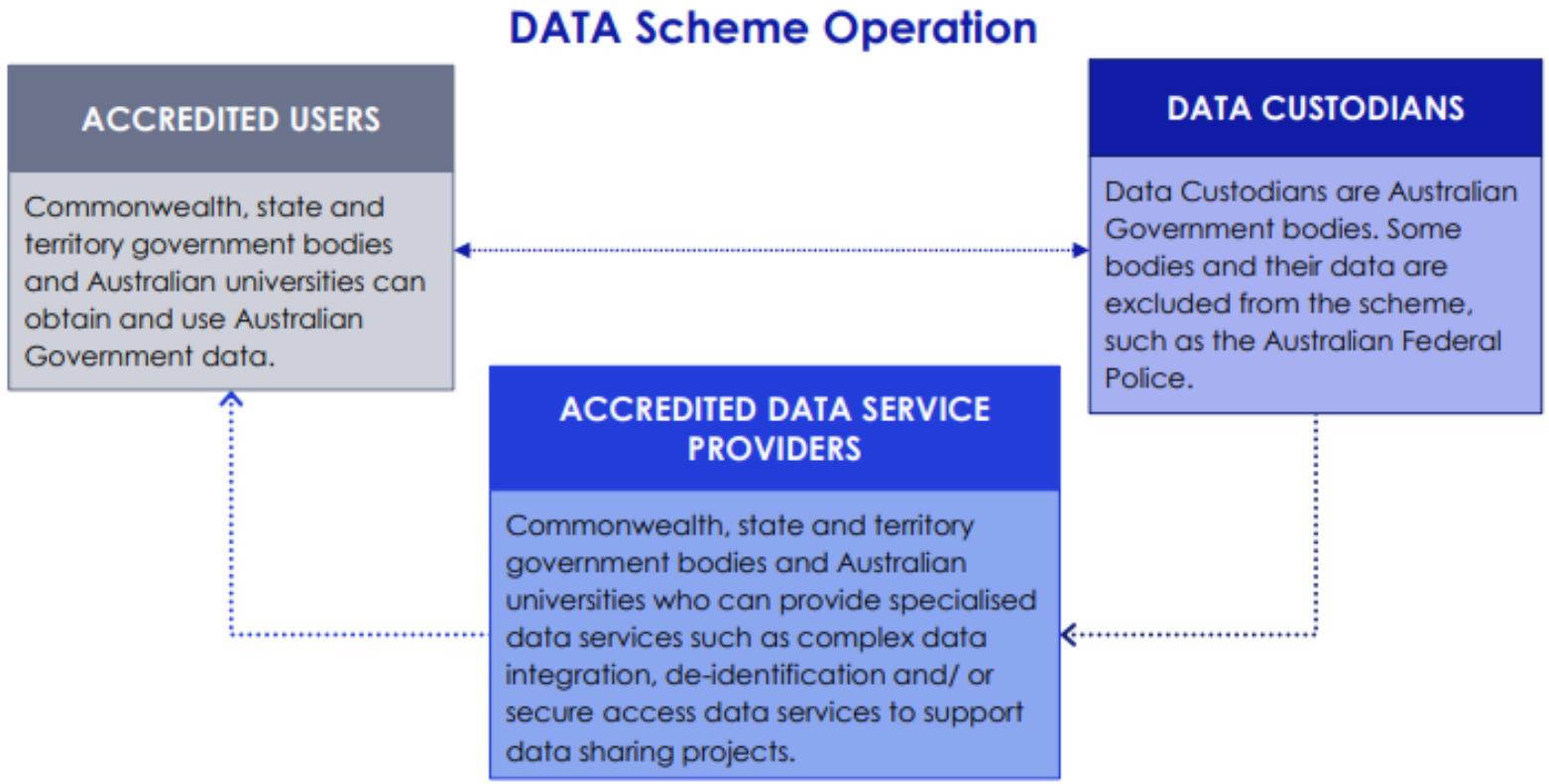
# How do you access these data?



# Recent changes in legislation

**How we access data from Australian Government is changing…**

- In 2022, the Data Availability and Transparency Act (2022) was passed to establish a new, best practice scheme for sharing Australian Government data using secure, safe and efficient practices.
- The Office of the National Data Commissioner (ONDC) supports the National Data Commissioner who is responsible for overseeing the DATA Scheme.
- **Dataplace**: the digital platform for scheme applicants and others to manage data requests/administration of the DATA Scheme



# Data classification systems

## University of Melbourne

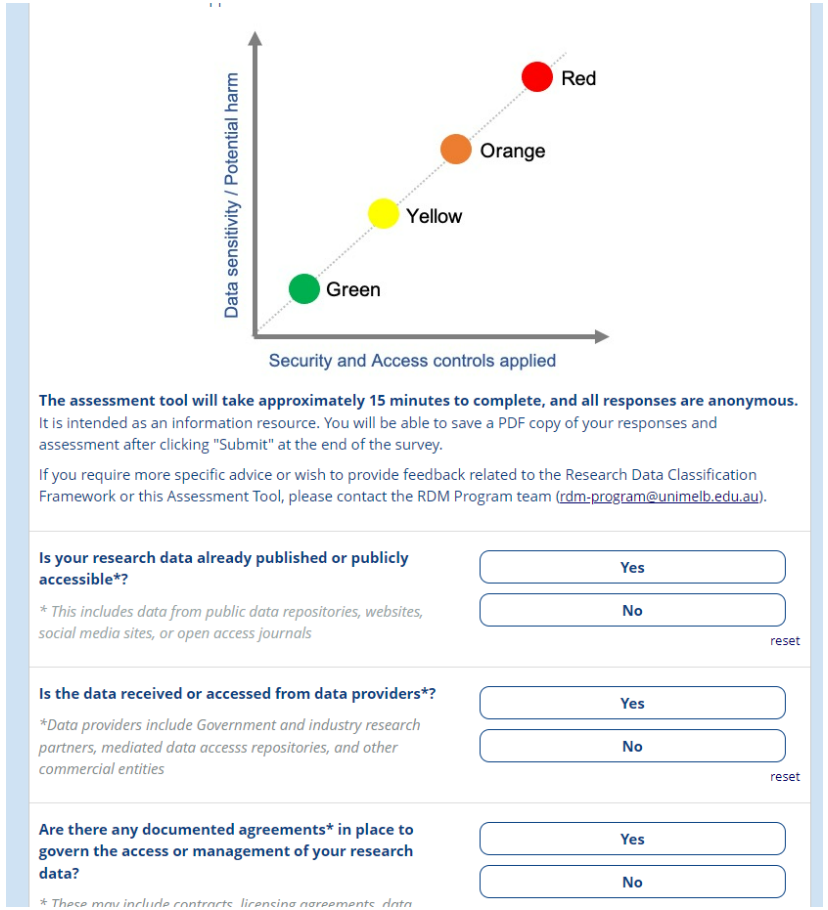
| | Green | Yellow | Orange | Red |
|---|---|---|---|---|
| Risk* | **Negligible** | **Limited** | **Significant** | **Severe** |
| Examples | • Information intended for public disclosure / consumption<br>• Published manuscripts or datasets<br>• Data from public websites or social media that does not relate to an identified or identifiable individual | • De-identified or aggregated data that does not relate to an identifiable individual or present any risk of significant harm to a community or group<br>• Unpublished research data and outputs that do not fall into all other categories, e.g., drafts of research publications, novel creative works etc.<br>• Novel analyses or transformations of publicly available data or information<br>• Data generated by instruments, imaging systems or computational models that are not linked to a specific identifiable entity | • Personally identifiable data including name, contact details, financial details, individual medical records, etc.<br>• Genetic or biometric information<br>• Re-identifiable data i.e., when the identity of a specific individual or other sensitive entity can be reasonably ascertained by data linkage or other activities<br>• Culturally and ecologically sensitive data | • Personally identifiable data containing "sensitive" information as defined by Victorian Privacy legislation<br>• Assets and information for defence research, or that have the potential to be adapted for military or 'dual use' applications<br>• Data involving Sensitive Security Biological Agents (SSBAs) |

**University of Melbourne Data Classification Framework**

*risk of material, social, reputational, legal, or other harm to individuals, communities, or organisations in the event of accidental or unauthorised access, disclosure, alteration, misuse, loss, or destruction.

# Data classification systems

## Australian Universities




**Data Classification Framework, University of Melbourne**

MONASH University

Understanding research data classifications

| Data Types | Very Sensitive | Sensitive | Restricted | Public |
|---|---|---|---|---|
| **Data** classified by Human and Animal Ethics Committees | Classified as **Very Sensitive** | | | |
| **Data** involving information on children or young persons | Classified as **Very Sensitive** | | | |
| **Identifiable data** containing direct identifiers e.g. Name, Medical Record Number (MRN), DOB and contact details | Classified as **Very Sensitive** | | | |
| **Re-identifiable data** where direct identifiers have been removed but indirect identifiers may be present e.g. Postcode | Classified as **Sensitive or Higher** | | | |
| **Data** in datasets which is not combined with personally identifiable information | Classified as **Sensitive or Higher** | | | |
| **De-identified data** with no identifying information in aggregate data | Classified as **Restricted or Higher** | | | |
| **Data** from instruments and imaging systems. sensors, detectors, cameras, recorders that does not contain identifiable or re-identifiable information | Classified as **Restricted or Higher** | | | |

**Security Classifications for Research Data, Monash University**

Monash University Library | July 2020

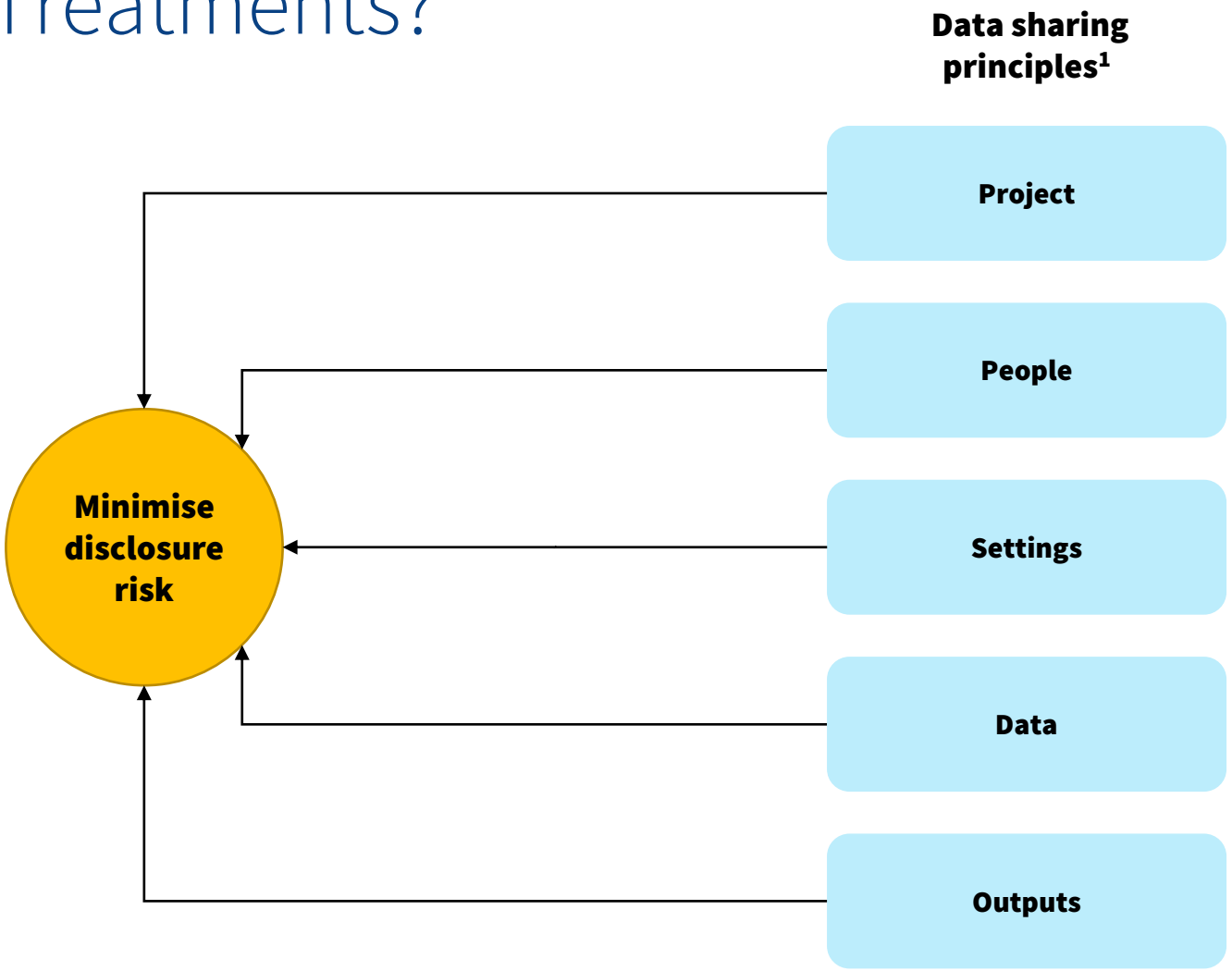
# Data classification systems

**Australian Government**

| | | | Sensitive information | Security classified information | | |
|---|---|---|---|---|---|---|
| | UNOFFICIAL | OFFICIAL | OFFICIAL: Sensitive | PROTECTED | SECRET | TOP SECRET |
| | No business impact | 1 Low business impact | 2 Low to medium business impact | 3 High business impact | 4 Extreme business impact | 5 Catastrophic business impact |
| **Compromise of information confidentiality would be expected to cause ➔** | **No damage.** This information does not form part of official duty. | **No or insignificant damage.** This is the majority of routine information. | **Limited damage** to an individual, organisation or government generally if compromised. | **Damage** to the national interest, organisations or individuals. | **Serious damage** to the national interest, organisations or individuals. | **Exceptionally grave damage** to the national interest, organisations or individuals. |

**Protective Security Policy Framework (PSPF), Attorney General's Department**

# Treatments?




**Why** is the data being used?
Is the project an appropriate project or program of work?

**Who** is using the data?
Is the data made available for appropriate persons?

**Where** is the data being used/accessed?
Is the data shared, collected and used in an appropriately controlled environment?

**What** data is being shared/accessed?
Are appropriate protections applied to the data?

**How** are the results of this activity being used?
Are appropriate protections applied to outputs using this data before they are released?

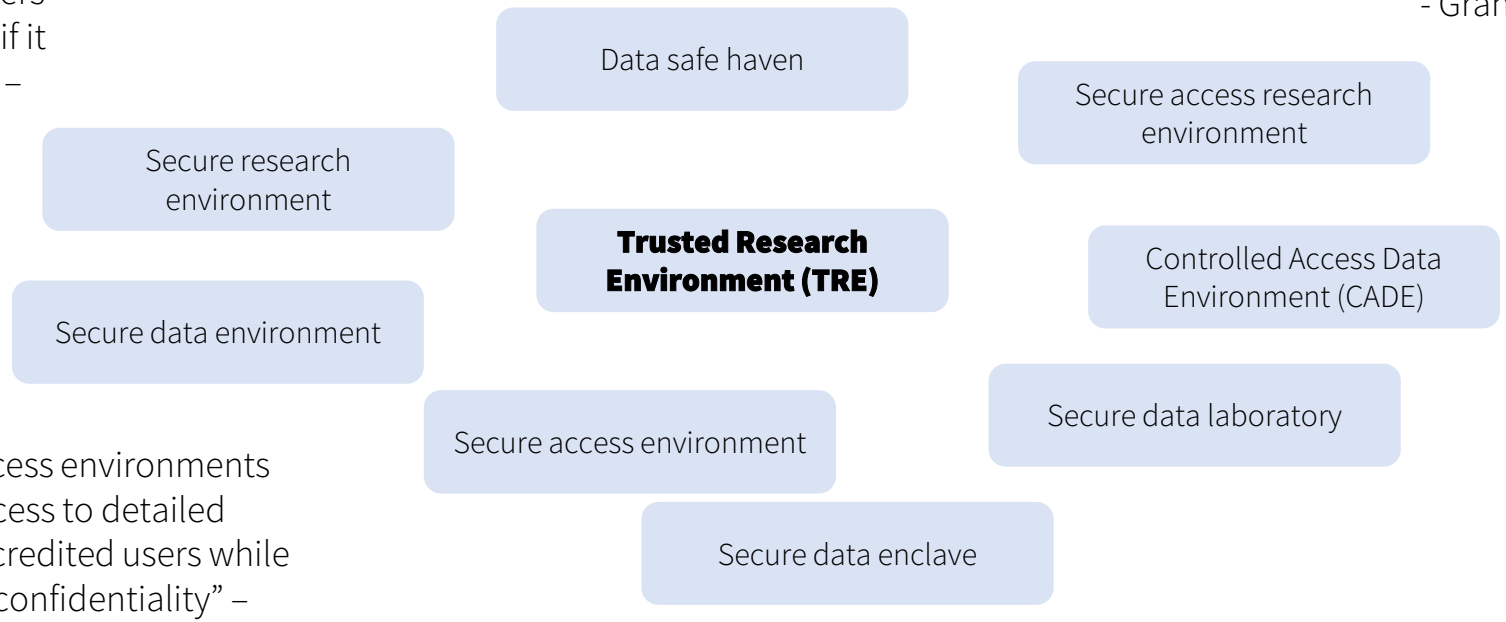
[1]Derived from the Five Safes Framework by Desai et al. (2016)

# Safe settings – what is a secure environment?

**Ongoing debate about terminology**

Remote RDCs "use thin-client technology to allow researchers to have full access to data as if it held on their own machines" – Desai et al. (2016)

"A TRE is a controlled computing environment that provides remote access to health data for approved researchers via a virtual desktop." - Graham et al. *J Med Ethics* (2022)




"Secure access environments support access to detailed data for accredited users while upholding confidentiality" – Australian Data Strategy, *DPMC* (2022)

# Safe settings – certification

**Regulatory compliance / certification / security assessments**

- ONDC Accredited Data Service Provider
- IRAP (Infosec Registered Assessors Program)
- ISO 27001 Certification
- FEDRAMP Assessment (US)

| | |
|---|---|
| **Security Governance** | Planning<br>Assurance and review processes<br>Investigation and response<br>Governance of contracted service providers |
| **Information Security** | Assessment of information holdings<br>Access control methods and practice<br>Safeguarding from cyber threats |
| **Personnel Security** | Screening and vetting processes for personnel<br>Ongoing assessment and training<br>Risk assessments, need to know |
| **Physical Security** | Measures of physical protection<br>Disposal mechanisms<br>Facilities and site controls |

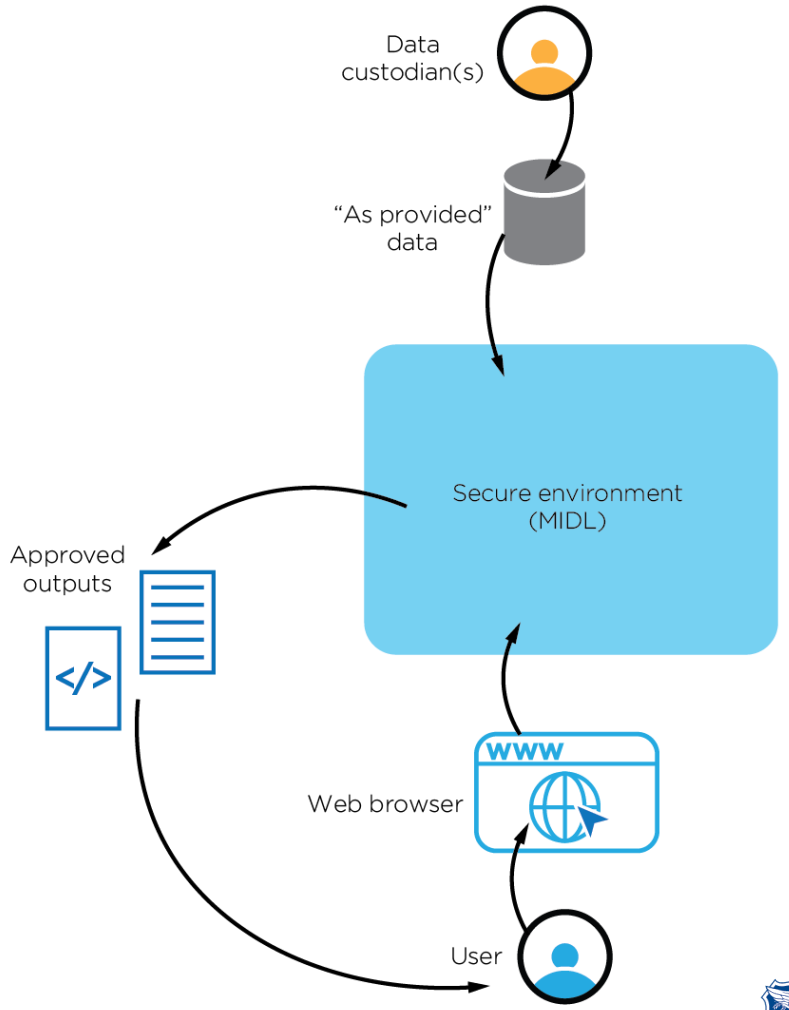
Protective Security Policy Framework (PSPF), Attorney General's Department

# Safe settings – an example

**The Melbourne Institute Data Lab (MIDL)**

The MIDL's features include:

- Citrix remote access desktop environment with additional security controls to ensure data confidentiality is maintained
- Strong security posture (PROTECTED) which includes strong authentication and identity management controls including the enforcement of multi-factor authentication protocols and 24x7 ongoing security monitoring of facilities and environment (SIEM/SOC)
- Citrix-based solution for file ingress/egress vetting and approvals before files/outputs can be transferred in to or out of the MIDL environment
- Access to a range of additional resources that speed up research through the setup of Shared Data Environments
    - "Research-ready" data assets that have been carefully curated by the MI's Foundation Fellow Program with oversight by MI researchers.
    - An in-environment wiki solution (MIDL Wiki) for sharing, documenting project activities.



# SDE Demo

**Deep dive in to a secure environment...**

# Visualisation Demo

**A look at outputs hosted from a secure environment**

# For more information on MIDL

**Website:**

- https://melbourneinstitute.unimelb.edu.au/data/midl

**Email:**

- MIDL-services@unimelb.edu.au (for new users/projects)

# Safe outputs

**Statistical Disclosure Control**

- Statistical Disclosure Control (SDC) aims to:
  - prevent the identity of a data subject from being revealed;
  - and/or releasing associated confidential information belonging to that data subject.
- Traditionally, SDC has been applied to:
  - statistical tables (often produced by National Statistical Institutes) prior to their release;
  - microdata, for the purposes of creating anonymised versions of original data.

# Safe outputs

## Critical concepts

- **Primary disclosure:** occurs when a data subject is identified from statistical result.

  For example, from descriptive statistics in the previous table identifying person with maximum income.

- **Secondary disclosure:** occurs when one released result is combined with other information to produce new statistics that are disclosive.

  For example, two tables of descriptive statistics produced from one data source could show various breakdowns in different ways. A single observation could be isolated, for example, if one table is deducted from another.

## Possible rules to apply

- **Base de-identification:**
  - No identifiable information
  - Top coding/bottom coding
- **Thresholds (Rule of N):**
  - All non-modelled output (descriptive) should have at least N unweighted counts.
  - N is typically 10
- **Dominance:**
  - This is the idea that one observation could account for most of the value in a statistical measure, and therefore be identifiable.
    - The largest contributor of any number should not exceed 50% of the total for that cell/statistic.
    - In any descriptive table, a single item of a category should not contain more than 90% of the observations of the sum of the categories.

# Safe outputs

**What do we mean by disclosive outputs?**

Here are some examples…

- Frequency tables

| ETHNIC BACKGROUND/AGE | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|
| African_Asian | 0 | 0 | 0 | 0 | 0 |
| Bangladeshi | 3 | 4 | 6 | 5 | 4 |
| Black African | 3 | 5 | 7 | 6 | 8 |
| Caribbean_West Indian | 0 | 4 | 2 | 4 | 3 |
| Chinese | 0 | 2 | 1 | 1 | 0 |
| Far Eastern | 1 | 0 | 1 | 2 | 0 |
| Indian | 5 | 9 | 4 | 4 | 4 |
| Middle Eastern | 1 | 1 | 0 | 0 | 1 |
| Mixed Caribbean_West Indian | 0 | 0 | 1 | 0 | 2 |
| Mixed Indian | 0 | 0 | 0 | 0 | 0 |
| North African | 1 | 0 | 1 | 0 | 1 |
| Pakistani | 6 | 10 | 5 | 4 | 3 |
| Sri Lankan | 0 | 0 | 2 | 0 | 0 |
| Turkish | 1 | 1 | 1 | 0 | 1 |
| White | 54 | 135 | 141 | 146 | 130 |

Consider grouping columns or rows?

Understandable labels

Small cell frequencies

IMPORTANT: Please note that all data displayed are completely fake and used for illustrative purposes.

# Safe outputs

**What do we mean by disclosive outputs?**

Here are some examples…

- Frequency tables
- Descriptive statistics

| | MIN | 1ST QUARTILE | MEDIAN | MEAN | 3RD QUARTILE | MAX |
|---|---|---|---|---|---|---|
| Income | 0 | 2894 | 6046 | 7178 | 10350 | 41113 |
| Age | 16 | 30 | 41 | 43 | 54 | 111 |
| Turnover | 140 | 3345993 | 7100730 | 8327156 | 12034046 | 38333022 |

Consider rounding or average values for e.g. 10 observations?

Precise information, probably about one observation

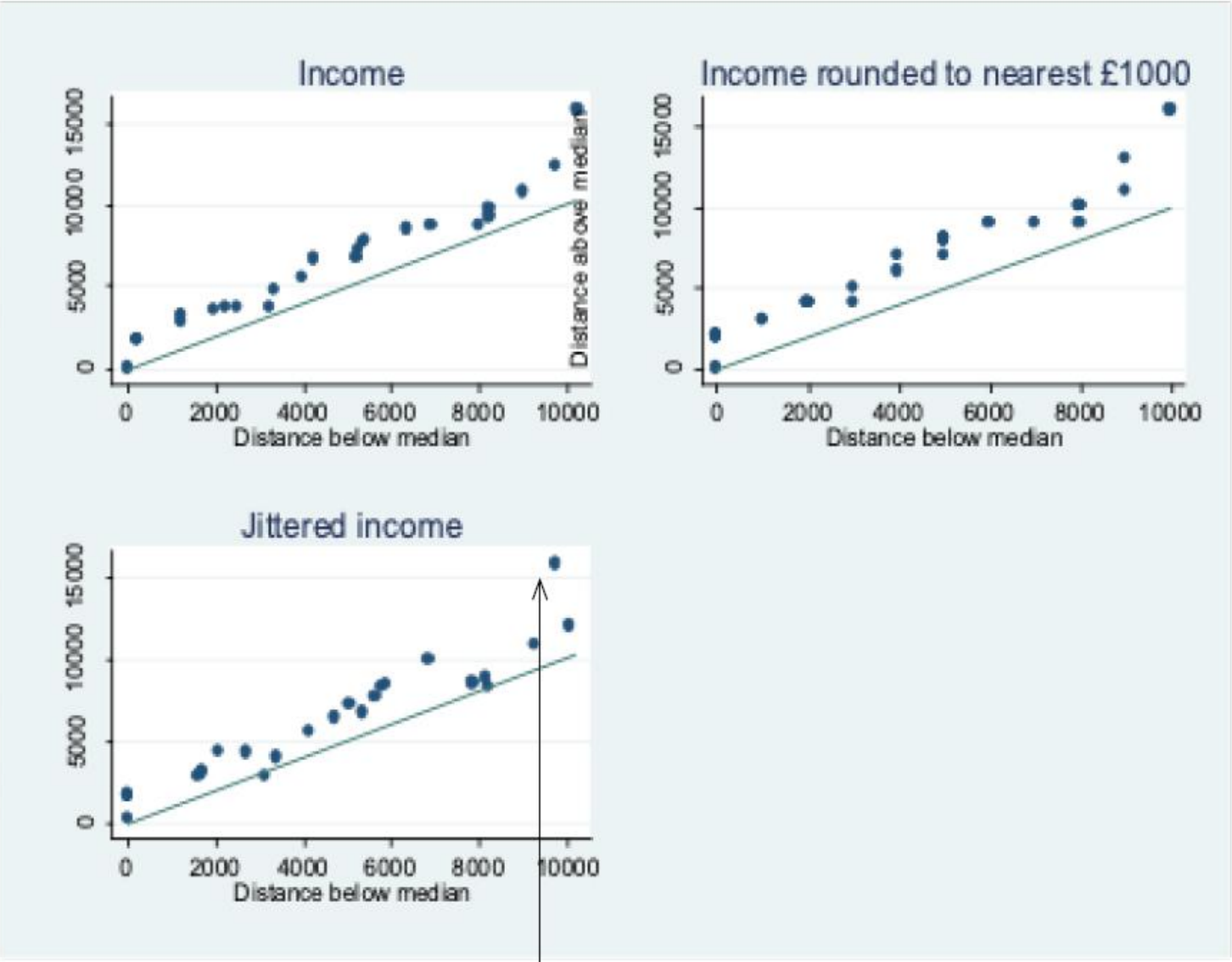
IMPORTANT: Please note that all data displayed are completely fake and used for illustrative purposes.

# Safe outputs

**What do we mean by disclosive outputs?**

Here are some examples…

- Frequency tables
- Descriptive statistics
- Scatter plots




IMPORTANT: Please note that all data displayed are completely fake and used for illustrative purposes.

# Safe outputs

**What do we mean by disclosive outputs?**

Here are some examples…

- Frequency tables
- Descriptive statistics
- Scatter plots
- Spatial data

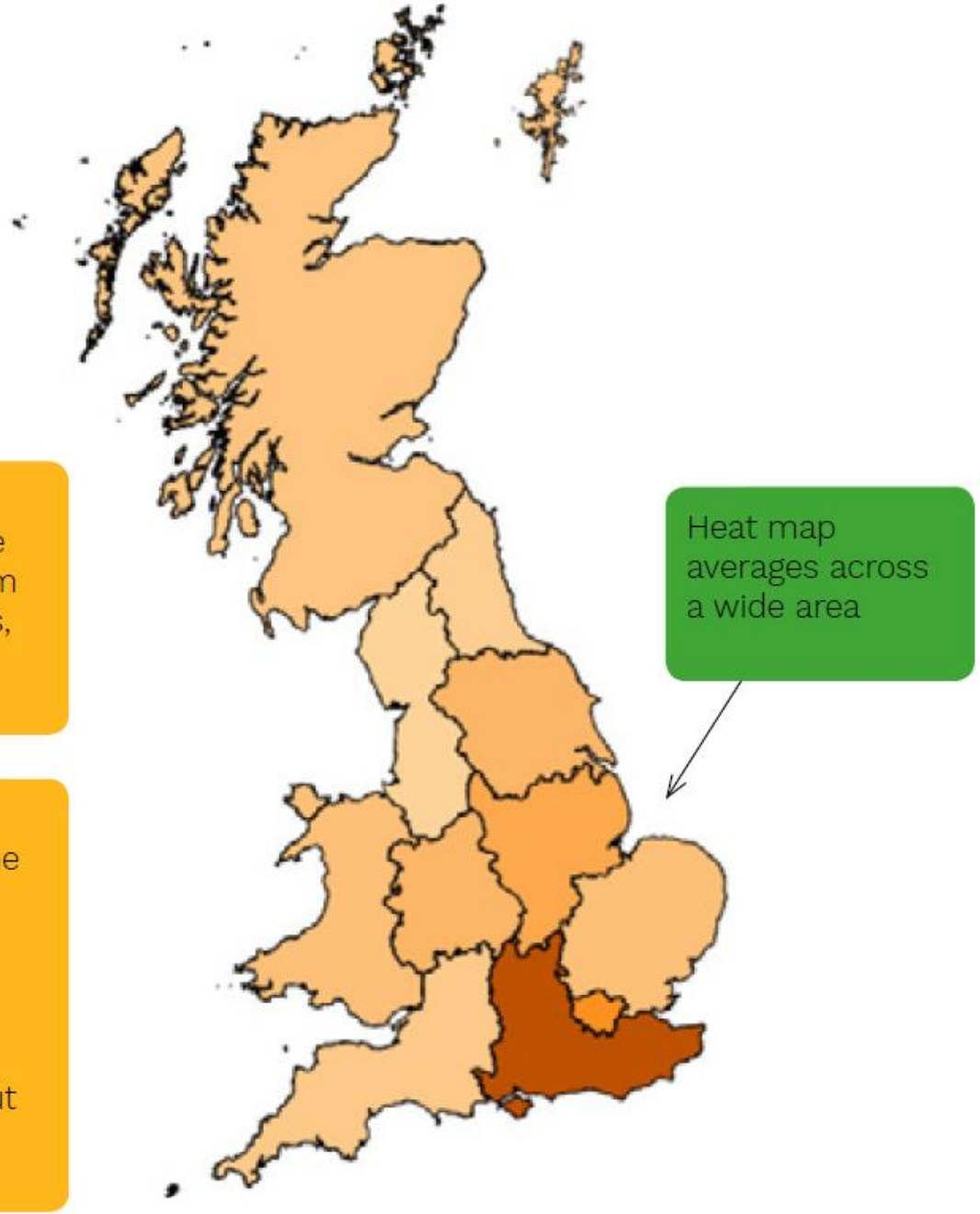

IMPORTANT: Please note that all data displayed are completely fake and used for illustrative purposes.

# Safe outputs

**What do we mean by disclosive outputs?**

Here are some examples…

- Frequency tables
- Descriptive statistics
- Scatter plots
- Spatial data




IMPORTANT: Please note that all data displayed are completely fake and used for illustrative purposes.

# Safe outputs

**Example from the MIDL Output Vetting Policy**

## 7.2 Treatment of specific output types

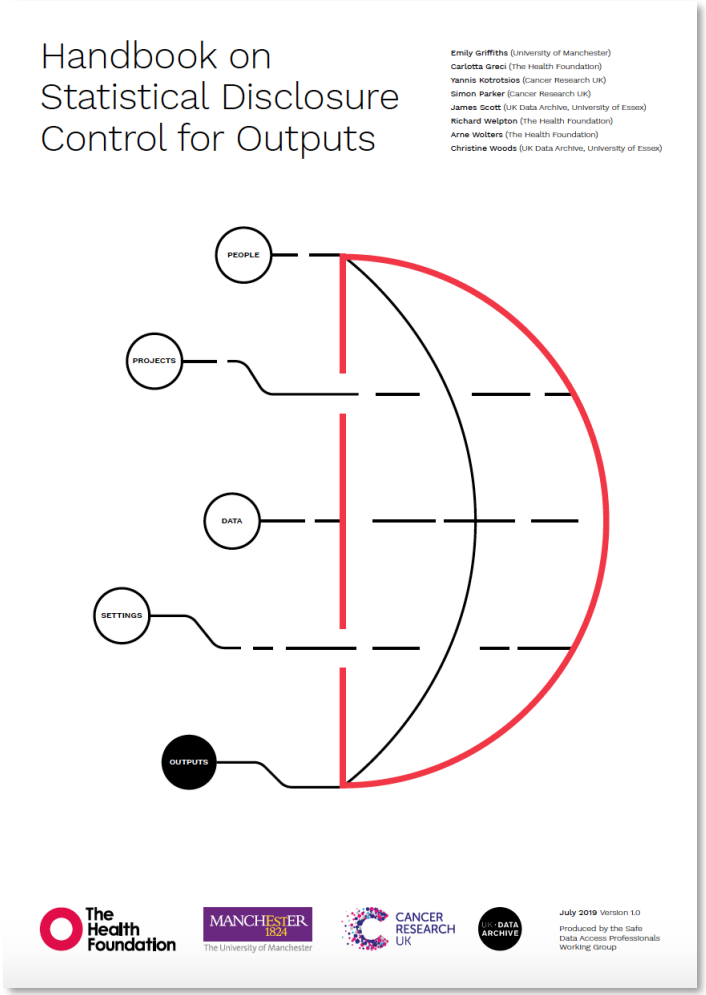
Outputs from research activity can take multiple forms. The table below specifies how specific output types are to be treated generally. The general classifications used here are derived from Bond, Brandt and de Wolf (2009) "Guidelines for the checking of output based on microdata research", Data Without Boundaries.

| Type of statistic | Type of output | General classification | Base rules to be applied | Other considerations |
|---|---|---|---|---|
| **Descriptive statistics** | Frequency tables | Unsafe | • Rule of N<br>• Group disclosure rule<br>• Dominance rule | • Is the data in the frequency table confidential? If not, then counts of 1 could be acceptable.<br>• If the data in the table are direct responses from a survey, census or administrative data, then they are confidential and we must ensure that units of observations could not be reidentified.<br>• Counts less than 10, but greater than 1, can be released if the likelihood of identification is low enough to be considered negligible. Considerations here include: likelihood of other relevant information being available to assist identification; age of the data; geographical aggregation; sample design; whether the rank ordering of contributors are known. Also to be included is the question: is this output useful, including reliability of the estimate? If not, then the rule of thumb should be applied.<br>• Counts of 2 should be treated with caution as one member of the pair could potentially identify the other. |
| | Magnitude tables | Unsafe | • Rule of N<br>• Group disclosure rule<br>• Dominance rule | • There may be circumstances in which the cell count rule of ten is inappropriate. For example, a Data User may believe that an output is non-disclosive even though cell counts do not meet the required threshold of N units. The onus is then on the Data User to explain why this is the case, and the individual responsible for checking output will be required to make a decision as to whether this output can be released. |
| | Maxima, minima and percentiles (incl. median) | Unsafe | • Maxima and minima are not released since they usually refer to only one unit.<br>• A percentile (or centile) is the value of a variable below which a certain percentage of observations fall. Percentile data, therefore, usually represent the value of the variable for an individual respondent and are not generally released. | • Data users need to provide counts to help make some assessments.<br>• For maxima, in particular, care needs to be taken to ensure disclosure is not possible, because usually the largest/oldest/wealthiest etc. respondent is known. Issues of dominance and group disclosure need to be considered in addition to the Group of N rule before providing clearance for a maximum value.<br>• For percentiles, the question to be considered by the Output Vetting Specialist is whether the respondent could be identified from the value published.<br>• For minima, maxima and percentiles, the Data User will always have to provide additional information to assure that the risk of disclosure is negligible |
| | Mode | Safe | • Rule of N<br>• Group disclosure rule | • Same as "Magnitude" principles |
| | Means, indices, | Unsafe | • Rule of N | • Same as "Magnitude" principles |

# Safe outputs

**Resources on SDC for safe outputs**

- There is an excellent series on confidentiality by the ABS:
  - 1160.0 – ABS Confidentiality Series, Aug 2017 (link)
- Which has the following sub-sections that are of particular importance:
  - Managing the risk of disclosure: treating aggregate data (link)
  - Managing the risk of disclosure: treating microdata (link)
- This covers topics like:
  - Assessing disclosure risk
  - Treatment methods such as:
    - Limit the number of variables
    - Modification of cell values through rounding or perturbation
    - Combining categories (i.e. age ranges instead of single years for age)
    - Top/bottom coding
    - Suppression of values that have less than 10 counts




*Handbook on Statistical Disclosure Control for Outputs*, Safe Data Access Professionals Working Group, 2019

# Key takeaways

**Think about the data sharing principles as you look for data for your project.**

- **Data:** Put effort in to knowing more and understanding more about the data set itself.

  This includes:

  - What is captured in the data? How are they measured? What are the privacy implications of accessing this data? Think about what level of detail you are willing to compromise for access?

- **Project:** Does the data access/availability affect your project and its intended purposes?
- **People:** What training do I/people in my team when it comes to working with sensitive data?
- **Settings:** Not much say here but would be a conversation with the data custodian that feeds off "data" above i.e. reduction of detail ► Reduction of controls required for settings.
- **Outputs:** How do I work with the data to better capture what I need for my project? What are my intended outputs? How does Statistical Disclosure Control affect me?

# Other things to remember…

You will still need to follow best practices to ensure data is safe outside of analysis. This would mean you need to remember:

- **Store paper forms securely:** use your filing cabinet (in the office) when storing documents relating to sensitive data. This includes deeds of confidentiality.
- **Handle detachable media securely:** confidential data stored on transportable media such as a CD, hard drive or USB keep need to be encrypted with a secure password and/or stored securely in a safe or locked filing cabinet
- **Better passwords:** use better passwords. Use a password generator if possible (i.e. LastPass) to create long, complex passwords. Combinations of actual random words are powerful and easier to remember!
    - QuickFox#JumpsLAZYdog?
- **Account control:** never share your account details with anyone. Do not let anyone access the MFA apps on your phone either! Regularly check your backup phone / contact options.
- **Never email data:** data should never ever be emailed around. Regardless of security classification. Even if it is public data and your doing a research project. Send your co-author a link to the location on OneDrive. Removes unnecessary duplication of data and leads to less errors – good research practice
- **If something goes wrong, tell someone:** if you do something wrong, don't stay quiet. Just tell someone. There may be processes that need to be followed to mitigate risks of disclosure. The sooner we act, the better.
- **If using an external system:** adhere to their data access and use policies. We do not want to lose access to the data due to a breach!

# Thank you

THE UNIVERSITY OF MELBOURNE

MELBOURNE INSTITUTE APPLIED ECONOMIC & SOCIAL RESEARCH